Communications in Computer and Information Science 273

Editorial Board

Simone Diniz Junqueira Barbosa
Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil

Phoebe Chen
La Trobe University, Melbourne, Australia

Alfredo Cuzzocrea
ICAR-CNR and University of Calabria, Italy

Xiaoyong Du
Renmin University of China, Beijing, China

Joaquim Filipe
Polytechnic Institute of Setúbal, Portugal

Orhun Kara
TÜBİTAK BİLGEM and Middle East Technical University, Turkey

Tai-hoon Kim
Konkuk University, Chung-ju, Chungbuk, Korea

Igor Kotenko
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Russia

Dominik Ślęzak
University of Warsaw and Infobright, Poland

Xiaokang Yang
Shanghai Jiao Tong University, China

Ana Fred Joaquim Filipe Hugo Gamboa (Eds.)

Biomedical Engineering Systems and Technologies

4th International Joint Conference, BIOSTEC 2011
Rome, Italy, January 26-29, 2011
Revised Selected Papers

Springer

Volume Editors

Ana Fred
IST - Technical University of Lisbon
Av. Rovisco Pais, 1
1049-001 Lisbon, Portugal
afred@lx.it.pt

Joaquim Filipe
INSTICC and IPS
Estefanilha, 2910-761 Setúbal, Portugal
E-mail: joaquim.filipe@estsetubal.ips.pt

Hugo Gamboa
Institute of Telecommunications
Av. Rovisco Pais, 1
1049-001 Lisboa, Portugal
hgamboa@gmail.com

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-29751-9 e-ISBN 978-3-642-29752-6
DOI 10.1007/978-3-642-29752-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012955350

CR Subject Classification (1998): J.3, H.2.8, F.1, F.2.2, G.1, K.4.1, H.3-5

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.
The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The present book includes extended and revised versions of a set of selected papers from the Fourth International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2011), sponsored by the Institute for Systems and Technologies of Information Control and Communication (INSTICC), technically co-sponsored by the IEEE Engineering in Medicine and Biology Society (EMB), the IEEE EMBS Portugal, COST Action NeuroMath, the European Society for Engineering and Medicine (ESEM), the Biomedical Engineering Society (BMES) and in cooperation with Workflow Management Coalition (WfMC), Association for the Advancement of Artificial Intelligence (AAAI) and Euromicro.

The purpose of the International Joint Conference on Biomedical Engineering Systems and Technologies is to bring together researchers and practitioners, including engineers, biologists, health professionals and informatics/computer scientists, interested in both theoretical advances and applications of information systems, artificial intelligence, signal processing, electronics and other engineering tools in knowledge areas related to biology and medicine.

BIOSTEC is composed of four co-located conferences; each specializes in one of the afore-mentioned main knowledge areas, namely:

- BIODEVICES (International Conference on Biomedical Electronics and Devices) focuses on aspects related to electronics and mechanical engineering, especially equipment and materials inspired from biological systems and/or addressing biological requirements. Monitoring devices, instrumentation sensors and systems, biorobotics, micro-nanotechnologies and biomaterials are some of the technologies addressed at this conference.

- BIOINFORMATICS (International Conference on Bioinformatics Models, Methods and Algorithms) focuses on the application of computational systems and information technologies to the field of molecular biology, including, for example, the use of statistics and algorithms to understanding biological processes and systems, with a focus on new developments in genome bioinformatics and computational biology.

- BIOSIGNALS (International Conference on Bio-inspired Systems and Signal Processing) is a forum for those studying and using models and techniques inspired from or applied to biological systems. A diversity of signal types can be found in this area, including image, audio and other biological sources of information. The analysis and use of these signals is a multidisciplinary area including signal processing, pattern recognition and computational intelligence techniques, among others.

- HEALTHINF (International Conference on Health Informatics) promotes research and development in the application of information and communication technologies (ICT) to healthcare and medicine in general and to the specialized

support to persons with special needs in particular. Databases, networking, graphical interfaces, intelligent decision support systems and specialized programming languages are just a few of the technologies currently used in medical informatics. Mobility and ubiquity in healthcare systems, standardization of technologies and procedures, certification, privacy are some of the issues that medical informatics professionals and the ICT industry in general need to address in order to further promote ICT in healthcare.

The joint conference, BIOSTEC, received 538 paper submissions from 63 countries in all continents. In all, 48 papers were published and presented as full papers (i.e., completed work, 10 pages/30-minute oral presentation), 158 papers reflecting work-in-progress or position papers were accepted for short presentation, and another 121 contributions were accepted for poster presentation. These numbers, leading to a "full-paper" acceptance ratio of about 9% and a total oral paper presentation acceptance ratio close to 38%, show the intention of preserving a high-quality forum for the next editions of this conference.

The conference included a panel and six invited talks delivered by internationally distinguished speakers, namely: Sergio Cerutti, Alberto Cliquet Jr., Mário Forjaz Secca, Tanja Schultz, Bruno Sobral and Bradley Nelson.

We must thank the authors, whose research and development efforts are recorded here. We also thank the keynote speakers for their invaluable contribution and for taking the time to synthesize and prepare their talks. Finally, special thanks to all the members of the INSTICC team, whose collaboration was fundamental for the success of this conference.

September 2011

Ana Fred
Joaquim Filipe
Hugo Gamboa

Organization

Conference Co-chairs

Joaquim Filipe	Polytechnic Institute of Setúbal / INSTICC, Portugal
Ana Fred	Technical University of Lisbon / IT, Portugal
Hugo Gamboa	CEFITEC / FCT - New University of Lisbon, Portugal

Program Co-chairs

BIODEVICES

Pedro Vieira	Universidade Nova de Lisboa, Portugal

BIOINFORMATICS

Marco Pellegrini	IIT-CNR, Italy

BIOSIGNALS

Fabio Babiloni	Sapienza University of Rome, Italy

HEALTHINF

Vicente Traver	ITACA, Universidad Politécnica de Valencia, Spain

Organizing Committee

Sérgio Brissos	INSTICC, Portugal
Helder Coelhas	INSTICC, Portugal
Vera Coelho	INSTICC, Portugal
Andreia Costa	INSTICC, Portugal
Patrícia Duarte	INSTICC, Portugal
Bruno Encarnação	INSTICC, Portugal
Frederico Fernandes	INSTICC, Portugal
Liliana Medina	INSTICC, Portugal
Carla Mota	INSTICC, Portugal
Raquel Pedrosa	INSTICC, Portugal
Vitor Pedrosa	INSTICC, Portugal
Daniel Pereira	INSTICC, Portugal
José Varela	INSTICC, Portugal
Pedro Varela	INSTICC, Portugal

BIODEVICES Program Committee

W. Andrew Berger, USA
Luciano Boquete, Spain
Bernard Courtois, France
Fernando Cruz, Portugal
Pedro Pablo Escobar, Argentina
Michele Folgheraiter, Germany
Marcos Formica, Argentina
Juan Carlos Garcia Garcia, Spain
Maki Habib, Egypt
Leonore Heiland, Germany
Leonid Hrebien, USA
Takuji Ishikawa, Japan
Bozena Kaminska, Canada
Ondrej Krejcar, Czech Republic
Hongen Liao, Japan
Rui Lima, Portugal
Jordi Madrenas, Spain
Dan Mandru, Romania
Eric McAdams, France
Joseph Mizrahi, Israel
Raimes Moraes, Brazil
Umberto Morbiducci, Italy
Toshiro Ohashi, Japan
Kazuhiro Oiwa, Japan
Mónica Oliveira, Portugal
Nathalia Peixoto, USA
Wim L.C. Rutten, The Netherlands
Mario Sarcinelli-Filho, Brazil
Fernando Schlindwein, UK
Rahamim Seliktar, USA
Milutin Stanacevic, USA
Federico Vicentini, Italy
Pedro Vieira, Portugal
Peter Yves-Alain, Canada
Qing Zhu, USA

BIODEVICES Auxiliary Reviewers

Thomas Chmielewski, USA
Getúlio Igrejas, Portugal
Yingkan Lin, USA
Zuyan Shen, USA

BIOINFORMATICS Program Committee

Tatsuya Akutsu, Japan
Kiyoshi Asai, Japan
Tim Beissbarth, Germany
Erik Bongcam, Sweden
Carlos Brizuela, Mexico
Egon L. van den Broek, The Netherlands
Chris Bystroff, USA
Alessandro Cestaro, Italy
Antoine Danchin, France
Thomas Dandekar, Germany
Charlotte Deane, UK
Eytan Domany, Israel
Francisco Domingues, Germany
Richard Edwards, UK
George Eleftherakis, Greece
Nadia El-Mabrouk, Canada
Jose-Jesus Fernandez, Spain
Fabrizio Ferre, Italy
Liliana Florea, USA
Gianluigi Folino, Italy
Andrew French, UK
Andreas Gisel, Italy
Julian Gough, UK
Weisen Guo, Japan
Michael Habeck, Germany
Bo Jin, USA
Yuki Kato, Japan
Sunduz Keles, USA
Inyoung Kim, USA
Sophia Kossida, Greece
Malgorzata Kotulska, Poland

Lukasz Kurgan, Canada
Yinglei Lai, USA
Guohui Li, China
Shuangge Ma, USA
Veli Mäkinen, Finland
Xizeng Mao, USA
Elena Marchiori, The Netherlands
Carlo Mastroianni, Italy
Francesco Masulli, Italy
Tommaso Mazza, Italy
Imtraud Meyer, Canada
Radhakrishnan Nagarajan, USA
Giri Narasimhan, USA
Christine Nardini, China
Jean-Christophe Nebel, UK
Kay Nieselt, Germany
John Novembre, USA
José Luis Oliveira, Portugal
Oscar Pastor, Spain
Florencio Pazos, Spain
Marco Pellegrini, Italy
Matteo Pellegrini, USA
Esa Pitkänen, Finland
Xiaoning Qian, USA
Huzefa Rangwala, USA
Jaques Reifman, USA
Fabio Rinaldi, Switzerland
Eric Rivals, France
Paolo Romano, Italy
Simona E. Rombo, Italy
Juho Rousu, Finland
Jianhua Ruan, USA
Yvan Saeys, Belgium
David Sankoff, Canada
Joao C. Setubal, USA
Hamid Shahbazkia, Portugal
Thomas Simonson, France
Victor Solovyev, UK
Haixu Tang, USA
Tatiana Tatusova, USA
Silvio C.E. Tosatto, Italy
Ricardo Vencio, Brazil
Thomas Wilhelm, UK
Yufeng Wu, USA
Yanbin Yin, USA
Jingkai Yu, China
Erliang Zeng, USA
Louxin Zhang, Singapore
Leming Zhou, USA
Qing Zhou, USA

BIOINFORMATICS Auxiliary Reviewers

Joel Arrais, Portugal
Huiling Chen, USA
Wen-Chi Chou, USA
Qin Ma, USA
Fenglou Mao, USA
José Mombach, Brazil
Wenxin Niu, China
Leena Salmela, Finland
Xiaojia Tang, USA
Jiayin Wang, USA
Beisi Xu, China
Han Zhang, China
Fengfeng Zhou, USA

BIOSIGNALS Program Committee

Fabio Aloise, Italy
Fernando Alonso-Fernandez, Spain
Fabio Babiloni, Italy
Egon L. van den Broek, The Netherlands
Tolga Can, Turkey
Francis Castanie, France
M. Emre Celebi, USA
Joselito Chua, Australia
Albert C.S. Chung, China
Jan Cornelis, Belgium
Fernando Cruz-Roldán, Spain

Jose-Jesus Fernandez, Spain
Eugene Fink, USA
Arfan Ghani, UK
Juan I. Godino-Llorente, Spain
Christian Jutten, France
Kaustubh Kale, USA
Georgios Kontaxakis, Spain
Lenka Lhotska, Czech Republic
Marco Loog, The Netherlands
Mahdi Mahfouf, UK
Mohammad Mahfuz, Canada
Corinne Mailhes, France
Mihaela Morega, Romania
Tadashi Nakano, Japan
Kazuhiro Oiwa, Japan
George Panoutsos, UK
Sever Pasca, Romania
Hanchuan Peng, USA
Gennaro Percannella, Italy
Tuan Pham, Australia
Vitor Pires, Portugal
Lucia Rita Quitadamo, Italy
Dick de Ridder, The Netherlands
Marcos Rodrigues, UK
Jianhua Ruan, USA
Wim L.C. Rutten, The Netherlands
Saeid Sanei, UK
Carlo Sansone, Italy
Andres Santos, Spain
Gerald Schaefer, UK
Peter Schelkens, Belgium
Emanuele Schiavi, Spain
Tapio Seppänen, Finland
Iryna Skrypnyk, Finland
Daby Sow, USA
Ana Susac, Croatia
Asser Tantawi, USA
Wallapak Tavanapong, USA
Gianluca Tempesti, UK
Carlos M. Travieso, Spain
Bart Vanrumste, Belgium
Giovanni Vecchiato, Italy
Dimitri Van De Ville, Switzerland
Eric Wade, USA
Yuanyuan Wang, China
Didier Wolf, France
Huiyu Zhou, UK

BIOSIGNALS Auxiliary Reviewers

Rahil Garnavi, Australia
Sae Hwang, USA
Henning Müller, Switzerland
Julie Wall, UK

HEALTHINF Program Committee

Filippos Azariadis, Greece
Adrian Barb, USA
Isabel Segura Bedmar, Spain
Bert-Jan van Beijnum, The Netherlands
Riccardo Bellazzi, Italy
Elizabeth Borycki, Canada
Egon L. van den Broek, The Netherlands
Federico Cabitza, Italy
Cristina De Castro, Italy
James Cimino, USA
Donald Craig, Canada
Anna Divoli, USA
Stephan Dreiseitl, Austria
José Fonseca, Portugal
Christoph M. Friedrich, Germany
Ioannis Fudos, Greece
Jonathan Garibaldi, UK
Enrique J. Gómez, Spain
Alfredo Goñi, Spain
David Greenhalgh, UK
Tiago Guerreiro, Portugal
Cigdem Gunduz-Demir, Turkey

Alexander Hörbst, Austria
Chun-Hsi Huang, USA
Ana Iglesias, Spain
Andreas Kerren, Sweden
Georgios Kontaxakis, Spain
Giuseppe Liotta, Italy
Martin Lopez-Nores, Spain
Emilio Luque, Spain
Paloma Martínez, Spain
Sally Mcclean, UK
Gianluigi Me, Italy
Gerrit Meixner, Germany
Sai Moturu, USA
Hammadi Nait-Charif, UK
Goran Nenadic, UK
Chaoyi Pang, Australia
Danilo Pani, Italy
José J. Pazos-arias, Spain
John Puentes, France
Rosario Pugliese, Italy
Juha Puustjärvi, Finland
Arkalgud Ramaprasad, USA
Marcos Rodrigues, UK
Valter Roesler, Brazil
George Sakellaropoulos, Greece
Nickolas S. Sapidis, Greece
Bettina Schnor, Germany
Heiko Schuldt, Switzerland
Arash Shaban-Nejad, Canada
Carla Simone, Italy
Zoran Stevic, Serbia
Francesco Tiezzi, Italy
Adrian Tkacz, Poland
Vicente Traver, Spain
Alexey Tsymbal, Germany
Aristides Vagelatos, Greece
André Zúquete, Portugal
Jana Zvarova, Czech Republic

HEALTHINF Auxiliary Reviewers

Dimitris Alexandrou, Greece
Eduardo C. Cabrera, Spain
Cinzia Cappiello, Italy
Manel Taboada, Spain

Invited Speakers

Sergio Cerutti	Polytechnic University of Milan, Italy
Alberto Cliquet Jr.	State University of Campinas, Brazil
Mário Forjaz Secca	Universidade Nova de Lisboa, Portugal
Tanja Schultz	Karlsruhe Institute of Technology, Germany
Bruno Sobral	Virginia Bioinformatics Institute, USA
Bradley Nelson	ETH Zurich, Switzerland

Table of Contents

Invited Speaker

Informatics-Driven Infectious Disease Research 3
Bruno Sobral, Chunhong Mao, Maulik Shukla, Dan Sullivan, and Chengdong Zhang

Part I: Biomedical Electronics and Devices

A Simple and Low Cost Device for Automatically Supervising Urine Output of Critical Patients 15
Abraham Otero, Francisco Palacios, Andrey Apalkov, and Roemi Fernández

Photoplethysmogram Processing Using an Adaptive Single Frequency Phase Vocoder Algorithm 31
Walter Karlen, Chris Petersen, Jennifer Gow, J. Mark Ansermino, and Guy Dumont

A Configurable Integrated Circuit for Biomedical Signal Acquisition 43
Jakob M. Tomasik, Wjatscheslaw Galjan, Kristian M. Hafkemeyer, Dietmar Schroeder, and Wolfgang H. Krautschneider

Movement Disorder Assessment and Attenuation Techniques for Removal of Tremor 57
Wesley Teskey, Mohamed Elhabiby, and Naser El-Sheimy

A New Approach for Diagnostic Investigation of Total Hip Replacement Loosening 74
Catherine Ruther, Ulrich Timm, Andreas Fritsche, Hartmut Ewald, Wolfram Mittelmeier, Rainer Bader, and Daniel Kluess

Wearable Monitoring Unit for Swimming Performance Analysis 80
Ana S. Silva, Antonio J. Salazar, Carla M. Borges, and Miguel V. Correia

Part II: Bioinformatics Models, Methods and Algorithms

3D Protein Surface Segmentation through Mathematical Morphology . . . 97
Virginio Cantoni, Riccardo Gatti, and Luca Lombardi

Simple Tool for Semi-automated Evaluation of Yeast Colony Images 110
Jan Schier and Bohumil Kovář

Reconstructing Gapless Ancestral Metabolic Networks 126
Esa Pitkänen, Mikko Arvas, and Juho Rousu

Discretized Kinetic Models for Abductive Reasoning in Systems Biology 141
Gabriel Synnaeve, Katsumi Inoue, Andrei Doncescu, Hidetomo Nabeshima, Yoshitaka Kameya, Masakazu Ishihata, and Taisuke Sato

Quasi-Steady State Approximations and Multistability in the Double Phosphorylation-Dephosphorylation Cycle 155
Guido Dell'Acqua and Alberto Maria Bersani

Improving Latent Semantic Analysis of Biomedical Literature Integrating UMLS Metathesaurus and Biomedical Pathways Databases 173
Francesco Abate, Elisa Ficarra, Andrea Acquaviva, and Enrico Macii

Part III: Bio-inspired Systems and Signal Processing

Towards a Patient-Specific Model of Lung Volume Using Absolute Electrical Impedance Tomography (aEIT) 191
Suzani Mohamad Samuri, George Panoutsos, Mahdi Mahfouf, G.H. Mills, M. Denaï, and B.H. Brown

Auditory Processing Inspired Robust Feature Enhancement for Speech Recognition 205
Hari Krishna Maganti and Marco Matassoni

A Robust and Efficient Spatio-Temporal Feature Selection for Interpretation of EEG Single Trials 219
Yehudit Meir-Hasson, Andrey Zhdanov, Talma Hendler, and Nathan Intrator

Time Series Clustering Algorithm for Two-Modes Cyclic Biosignals 233
Neuza Nunes, Tiago Araújo, and Hugo Gamboa

Non-contact Pulse Wave Velocity Assessment Using Optical Methods . . . 246
T. Pereira, M. Cabeleira, P. Matos, E. Borges, V. Almeida, H.C. Pereira, J. Cardoso, and C. Correia

Fractal-Based Brain State Recognition from EEG in Human Computer Interaction 258
Olga Sourina, Qiang Wang, Yisi Liu, and Minh Khoa Nguyen

Understanding Cerebral Activations during the Observation of Marketing Stimuli: A Neuroelectrical Perspective 273
Giovanni Vecchiato, Laura Astolfi, Fabrizio De Vico Fallani, Jlenia Toppi, Fabio Aloise, Anton Giulio Maglione, Febo Cincotti, Donatella Mattia, and Fabio Babiloni

Part IV: Health Informatics

Semantic-Based Conformance Checking of Computer Interpretable Medical Guidelines 285
M.A. Grando, M.H. Schonenberg, and W. van der Aalst

REST-Style Architecture and the Development of Mobile Health Care Solutions 301
François Andry, Lin Wan, and Daren Nicholson

Biomedical Literature Retrieval Based on Patient Information 312
Ana Jimenez-Castellanos, Izaskun Fernandez, David Perez-Rey, Elisa Viejo, Francisco Javier Díez, Xabier García de Kortazar, Miguel Garcia-Remesal, Victor Maojo, Antonio Cobo, and Francisco del Pozo

Evaluating Information Security Effectiveness with Health Professionals 324
Robin Krens, Marco Spruit, and Nathalie Urbanus

Using Social Network Analysis to Study the Knowledge Sharing Patterns of Health Professionals Using Web 2.0 Tools 335
Samuel Alan Stewart and Syed Sibte Raza Abidi

A Semantic-Based Platform for Medical Image Storage and Sharing Using the Grid 353
Daniela Giordano, Carmelo Pino, Concetto Spampinato, Marco Fargetta, and Angela Di Stefano

Securing the Access to Electronic Health Records on Mobile Phones 365
Alexandra Dmitrienko, Zecir Hadzic, Hans Löhr, Ahmad-Reza Sadeghi, and Marcel Winandy

Similarity Grouping of Human Sleep Recordings Using EEG and ECG 380
Amro Khasawneh, Sergio A. Alvarez, Carolina Ruiz, Shivin Misra, and Majaz Moonis

Author Index 395

Invited Speaker

Informatics-Driven Infectious Disease Research

Bruno Sobral, Chunhong Mao, Maulik Shukla, Dan Sullivan, and Chengdong Zhang

Virginia Bioinformatics Institute at Virginia Tech, Blacksburg, Virginia, U.S.A.

Abstract. Informatics-driven approaches change how research and development are conducted, who participates, and enables systems-oriented views of science and research. Most life sciences researchers have a very strong desire for the full integration of data and analysis tools delivered through a single interface. Infectious disease (ID) research and development provides a uniquely challenging and high impact opportunity. The biological complexity of infectious disease systems, which are composed of multiple scales of interactions between potential pathogens, hosts, vectors, and the environment, challenges information resources because of the breadth of organism-organism and organism-environment interactions. Applications of integrated data for ID serves a variety of constituencies, such as clinicians, diagnostician, drug and vaccine developers, and epidemiologists. Thus there is a complexity that makes ID an opportune area in which to develop, deploy and use CyberInfrastructure.

1 Introduction

A key challenge in infectious disease research is integrating information about microbes, their response to environmental factors, and their interactions with hosts and vectors. This challenge is driven, in large part, by two factors: a variety of technologies used in the infectious disease studies and a variety of uses cases that employ the data generated by these technologies.

Sequencing technologies have enabled the elucidation of over three thousand bacteria genomes (at various stages of completion) and the generation of comprehensive phylogenetic trees for classes of bacteria important to infectious disease research, such as the alphaproteobacteria [1] and gammaproteobacteria [2]. Bioinformatics analysis of microbial genomes illuminate processes relevant to infectious diseases, such as the role of horizontal gene transfer even within bacteria with a preferred intracellular lifecycle [3] and the evolution of virulence factors, such as secretion systems [4].

Technologies for studying the transcriptome have provided insights into host-pathogen interactions during infections [5-7] and patterns of co-expression in pathogens [8]. RNA sequencing technology (RNA-Seq) is well adapted to identifying mechanisms of pathogenesis [9], generating genome-wide maps of transcriptional start sites and operons of pathogens [10], and facilitating discovery of regulatory mechanisms, such as small RNA [11]. Mass spectrometry is used across infectious disease research domains. It has been used to identify extracellular proteins that enhance the virulence of particular pathogens, such as those in community-associated

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 3–11, 2013.
© Springer-Verlag Berlin Heidelberg 2013

methicillin resistant *Staphylococcus aureus* [12]; it is also a promising technology for clinical diagnostics [13]. Flow cytometry provides insights into immune response and has been particularly valuable to tuberculosis researchers and clinicians [14] as well as the study of host response to infection by *S. aureus* [15] and *Clostridium difficile* [16]. It is not surprising that different technologies generate different types of data about related biological phenomenon; what may be surprising is how difficult it is to integrate these data to provide a comprehensive view of biological processes. The Pathosystems Rescource Integration Center (PATRIC, patricbrc.org) has addressed this problem with a range of techniques, including identifying and linking to experiment data associated with particular organisms. (See Figure 1).

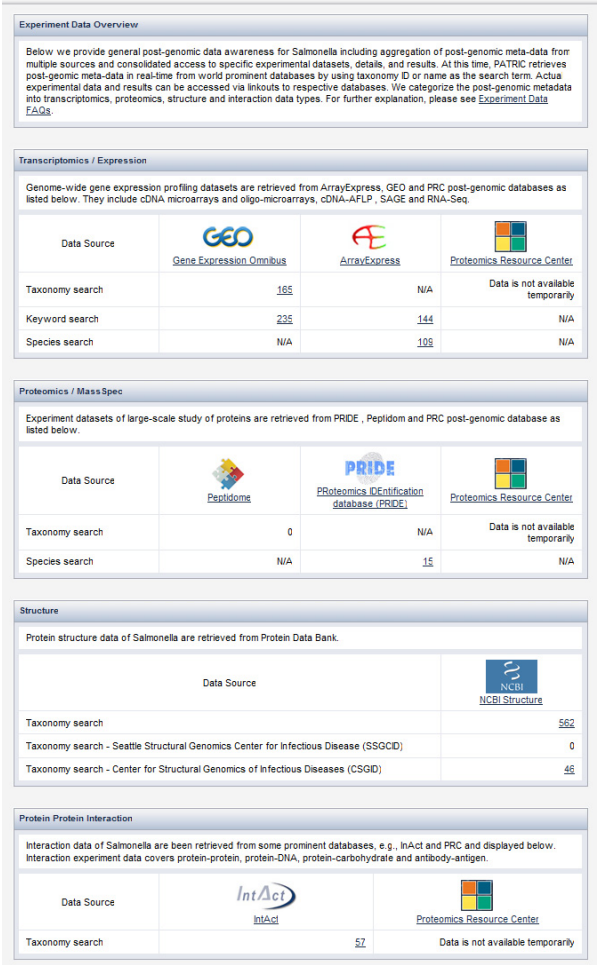


Fig. 1. PATRIC supports infectious disease research by providing context sensitive access to a wide variety of experiment data. In this example, transcriptome, proteome, structure and protein-protein interaction experiments on *Salmonella* are displayed (Source: Pathosystems Resource Integration Center (PATRIC) http://.patricbrc.org [17]).

The second factor contributing to the data integration challenge is that there is no single way to use the data. An organism specialist studying a single genome may be well served by existing genome browsers (see Figure 1) but the same cannot be said for a comparative genomics researcher studying an entire class. Vaccine researchers may be particularly interested in signaling factors at specific steps of a virus lifecycle and need to search genomic data for promising targets [18]. In the areas of emerging disease research and biodefense, genomics data is also proving valuable for identifying, predicting and preventing biological threats [19]. Each of these research areas can benefit from available bioinformatics data but only if it is accessible and in a format that allows for integrating multiple types of data according to the specific needs of each research area. PATRIC provides multiple tools, from sequence alignment to comparative pathway and text mining resources.

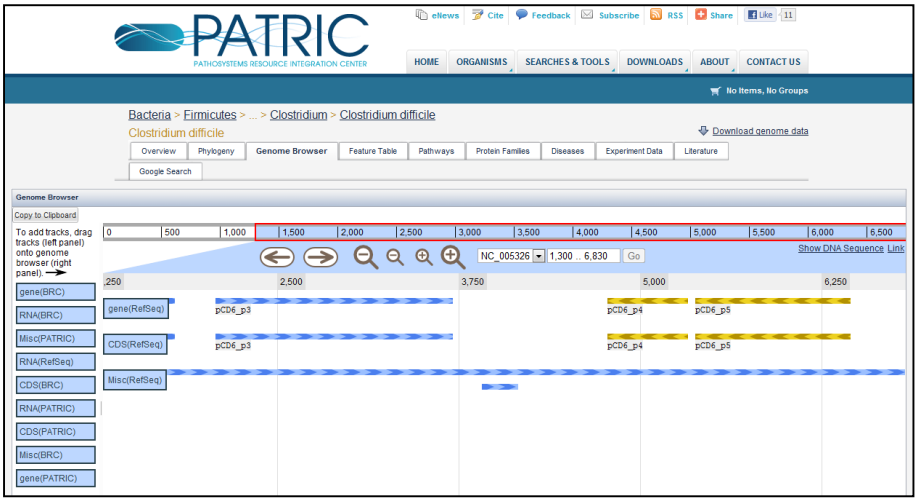


Fig. 2. Genome browsers are well suited to visualizing small numbers of genes and genomes. Advances in visualization are required to support large scale comparative genomics (Source: Pathosystems Resource Integration Center (PATRIC) http://patricbrc.org [17]).

2 Bioinformatics Needs of Infectious Disease Researchers

The range of bioinformatics needs in infectious disease research reflect the diversity of the field. At one end of the research spectrum, we must understand how genes, proteins, signaling mechanisms, and metabolic pathways function and interact to produce virulent phenotypes [20]. At the other end of the infectious disease spectrum we face epidemiological scale issues, such as understanding rates of transmission and identifying especially vulnerable populations. These are not independent problems, though. For example, genetic variations can change virulence characteristics that in turn help to determine infection rates and transmission through a population.

We envision using bioinformatics to support at least four classes of users (though some may fit in more than one class) in infectious disease research:

- Microbiologists and immunologists who will tend to focus on genomics, metagenomcis, metabolomics, signaling pathways, host-pathogen interactions and related areas.
- Biomedical and pharmaceutical researchers interested in vaccines, diagnostics and therapeutics in general as well as more specific research areas, such as antibiotic resistance and the role of the human mircrobiome in human health.
- Clinical and medical practitioners, such as infectious disease control specialists working in hospitals who need to control the potential spread of hospital acquired infections, such as *S. aureus* and *C. difficile*.
- Public health officials focused on surveillance and responses to outbreaks (epidemiology).

The ideal cyberinfrastructure for infectious disease research will integrate data related to genetics and phenotypes of microbes (both pathogenic and non-pathogenic), experiments conducted under a broad array of conditions, clinical data sufficient to support medical practitioners, and surveillance data to sufficient to guide public health response to potential outbreaks. These data are generated from a combination of laboratory and computational technologies.

3 Bioinformatics-Related Technologies in Infectious Disease Research

The life sciences are currently dominated by "wet lab" disciplines founded on the study of organisms and the systems they create. They are also information rich and therefore amenable to computational approaches as well, though such approaches have not permeated the health sciences as much. The bioinformatics research pipeline typically starts with organisms or their constituent parts and ends with structured data populating community resources such as the databases at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) and the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/). This pipeline requires both biology-centric and computing-centric technologies.

3.1 Biology-Centric Technologies

The more biology-centric work in bioinformatics consists of gene sequencing, transcriptome analysis, proteome analysis, metabolomics and phenotype testing. Genomic sequencing generates data about the sequence of nucleic acids in chromosomes and plasmids which is then analyzed to annotate the location of genes and other structures, such as regulons and operons. Sequencing technology, like computing, is advancing in ways that lowers the cost of sequencing while improving throughput. Older, semi-automated sequencing techniques that required preparing DNA, cloning into a plasmid, and inserting it into *E. coli* are being replaced by high-throughput methods commonly referred to as next-generation sequencing [21]. The newer technology utilizes massively parallel sequencing machines that are accelerating the pace with which we can sequence bacteria, virus and eukaryotic pathogens' genomes.

Transcriptome experiments, such as microarrays and RNA-Seq are used to study genomes under varying conditions, thus revealing the dynamics of these systems. These generate data about the existence of transcribed RNA under various conditions. For example, the pre-invasion growth conditions of *Salmonella* [22], genetic diversity of *Mycobacterium tuberculosis* in clinical isolates [23], and host responses to *Helicobacter pylori* [24] have all been studied using microarray technologies. RNA-Seq experiments are increasingly preferred for studying transcriptomes because they yield more precise measurement of transcripts than other techniques [25] and have added benefits such as detecting unanticipated sequences. RNA-Seq experiments of interest to infectious disease researchers includes the study of the transition from saprophytism to virulence in *Listeria monocytogenes*, a cause of food-borne infections [26] and ethanol-induced pathogenesis mechanisms [9].

Commonly used proteome analysis technologies include mass spectrometry and yeast two-hybrid experiments. The former generates data about the existences of peptides in a biological sample which is then used to predict the presence of specific proteins. Mass spectrometry data sets are relatively large and it is not unusual to generate tens of gigabytes of data in a single experiment. Yeast two-hybrid experiments demonstrate the potential ability of two proteins to interact. These data sets are much smaller than mass spectrometry data sets and yield binary type of data indicating the presence or absence of an interaction.

The aim of metabolomics is to identify the set of metabolites in an organism under a particular set of conditions. Data about metabolites complements genomics and proteomics data to provide a more complete picture of the biological state of an organism. Metabolomic studies have been used in the study of biomarkers and metabolic pathway regulation [27]. Metabolite data may be in the form of binned spectra, peak lists, or chemical concentration.

Phenotype microarrays [28] generate data about the growth of organisms under a wide range of conditions. This type of assay is useful, for example, to determine the function of each gene in a genome wide knock-out study. Phenotype microarray data is similarly structured to cDNA microarrays and typically reflects the ability to the organism to grow under a particular condition.

No single data type is sufficient to elucidate the structure and function of biological entities relevant to infectious disease research. Unfortunately, the experiments that are used often generate data that is not easily linked to other data, such as positional data about the location of a gene and its regulator, the abundance measure of that gene's expression under experimental conditions, the presence and phosphorylation form of the protein encoded by the gene, and the gene's effect on the organism phenotype as measured by a phenotype microarray. Before these data can be integrated, they must be subject to a variety of bioinformatics analysis processes and annotated with additional data to support linking in both general and specialized bioinformatics resources.

3.2 Computational Technologies

Bioinformatics is a relatively young discipline but it has produced a substantial ensemble of techniques for meeting the computational and information management challenges of infectious disease research. Open access databases have been essential

to advancing knowledge across the field of genomics. GenBank [29], UniProt [30], KEGG [31], GEO [32], ArrayExpress [33], Peptidome [34], PubChem [35] are just several of the many databases available to the bioinformatics community. Those resources provide resources for a broad spectrum of infectious disease research but more specialized databases such as PATRIC [17], EuPathDB [36], Influenza Research Database [37], VectorBase [38] and ViPR (http://www.viprbrc.org) will be of special interest to infectious disease researchers. These specialized resources are funded by the National Institute of Allergy and Infectious Diseases to focus on NIH Category A-C pathogens and to respond to the needs for the infectious disease research community.

In addition to data about genes, proteins and pathways, infectious disease researchers now have access to genome-scale metabolic models. These are computationally derived models for predicting phenotype and metabolic characteristics of an organism under a particular set of condition. Metabolic models are constructed using data about genomes, reactions, metabolic pathways, phenotype array data as well as gene essentiality data [39]. Prediction and modeling techniques are used in many aspects of infectious disease research, ranging from predictions of host-pathogen protein-protein interactions [40] to modeling mucosal immunity [41].

Text mining and ontologies are increasingly important to making formerly unstructured data in the scientific literature and experiment metadata more amenable to automated analysis. Some of the more important text mining techniques for infectious disease research are named entity recognition [42] and information extraction [43]. The National Center for Biomedical Ontology's BioPortal [44] is a repository of over 200 ontologies and 1.4 million terms (as of December 2010); one of the most relevant to this area is the Infectious Disease Ontology (http:// http://infectiousdiseaseontology.org).

Like experiment data, computational tools that are kept in silos and do not integrate well with other tools are a barrier to advancing infectious disease research. PATRIC address this problem by allowing researchers to identify a subset of PATRIC data and analyze that data with a variety of tools, such as comparative pathway and protein family sorting tools.

4 Breaking Down Barriers and Integrating Full Spectrum of Analysis and Data

A driving principal in the CyberInfrastructure Division of the Virginia Bioinformatics Institute is that the research objective, not the data or computational tools, should provide the organizing rationale for bioinformatics applications. The PATRIC resource (http://patricbrc.org), for example, integrates data from multiple sources and organizes them in a context sensitive way. A *Mycobacterium* researcher has access to taxonomy, phylogeny, gene, protein, pathway, experiment data, disease data and literature from a single location within the resource. Furthermore, visualizations are used to show the relations between hosts and pathogens, an obvious area of importance for infectious disease researchers. Projects such as these have built on lessons learned creating earlier bioinformatics resources [45].

In spite of the progress made toward integrating bioinformatics data sources and computational resources, much work remains. Both microbiologists and clinical practitioners are working toward more efficient sharing of knowledge so that what is learned on the wet lab bench is rapidly translated into benefits at the bedside. The potential risks of emerging diseases and bioterrorism are important factors supporting research into the development of new vaccines, diagnostics and therapeutics. Public health officials should be supported in their response to outbreaks as well. The infectious disease research community has taken on the formidable task of understanding and controlling the risk of pathogen borne diseases and must serve many constituencies. Bioinformatics can make significant contributions to supporting all of these facets of infectious disease research.

References

1. Williams, K.P., Sobral, B.W., Dickerman, A.W.: A Robust Species Tree for the Alphaproteobacteria. J. Bacteriol. 189(13), 4578–4586 (2007)
2. Williams, K.P., et al.: Phylogeny of Gammaproteobacteria. The Journal of Bacteriology 192(9), 2305–2314 (2010)
3. Wattam, A.R., et al.: Analysis of ten Brucella genomes reveals evidence for horizontal gene transfer despite a preferred intracellular lifestyle. J. Bacteriol. 191(11), 3569–3579 (2009)
4. Gillespie, J.J., et al.: An Anomalous Type IV Secretion System in Rickettsia is Evolutionarily Conserved. BMC Microbiology (2009)
5. Xu, Q., Dziejman, M., Mekalanos, J.J.: Determination of the transcriptome of Vibrio cholerae during intraintestinal growth and midexponential phase in vitro. Proc. Natl. Acad. Sci. U.S.A 100(3), 1286–1291 (2003)
6. Gilchrist, C.A., et al.: Impact of intestinal colonization and invasion on the Entamoeba histolytica transcriptome. Molecular and Biochemical Parasitology 147(2), 163–176 (2006)
7. van Erp, K., et al.: Role of strain differences on host resistance and the transcriptional response of macrophages to infection with Yersinia enterocolitica. Physiol Genomics 25(1), 75–84 (2006)
8. Mazandu, G.K., Opap, K., Mulder, N.J.: Contribution of microarray data to the advancement of knowledge on the Mycobacterium tuberculosis interactome: Use of the random partial least squares approach. Infection, Genetics and Evolution 11(1), 181–189 (2011)
9. Camarena, L., et al.: Molecular Mechanisms of Ethanol-Induced Pathogenesis Revealed by RNA-Sequencing. PLoS Pathog 6(4), e1000834 (2010)
10. Sharma, C.M., et al.: The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 464(7286), 250–255 (2010)
11. Sharma, C.M., Vogel, J.: Experimental approaches for the discovery and characterization of regulatory small RNA. Current Opinion in Microbiology 12(5), 536–546 (2009)
12. Ventura, C.L., et al.: Identification of a Novel <italic>Staphylococcus aureus</italic> Two-Component Leukotoxin Using Cell Surface Proteomics. PLoS ONE 5(7), e11634 (2010)
13. Dowling, P., et al.: Recent advances in clinical proteomics using mass spectrometry. Bioanalysis 2(9), 1609–1615 (2010)

14. Fuhrmann, S., Streitz, M., Kern, F.: How flow cytometry is changing the study of TB immunology and clinical diagnosis. Cytometry Part A 73A(11), 1100–1106 (2008)
15. Bernstein, J.M., et al.: Further observations on the role of Staphylococcus aureus exotoxins and IgE in the pathogenesis of nasal polyposis. The Laryngoscope, n/a–n/a (2010)
16. Kim, H., et al.: Inflammation and Apoptosis in Clostridium difficile Enteritis Is Mediated by PGE2 Up-Regulation of Fas Ligand. Gastroenterology 133(3), 875–886 (2007)
17. Snyder, E.E., et al.: PATRIC: the VBI PathoSystems Resource Integration Center. Nucleic Acids Res. 35(Database issue), D401–D406 (2007)
18. de la Torre, J.C.: Reverse genetics approaches to combat pathogenic arenaviruses. Antiviral Research 80(3), 239–250 (2008)
19. Fricke, W.F., Rasko, D.A., Ravel, J.: The Role of Genomics in the Identification, Prediction, and Prevention of Biological Threats. PLoS Biol. 7(10), e1000217 (2009)
20. Crasta, O.R., et al.: Genome sequence of Brucella abortus vaccine strain S19 compared to virulent strains yields candidate virulence genes. PLoS ONE 3(5), e2193–e2193 (2008)
21. Shendure, J., Ji, H.: Next-generation DNA sequencing. Nat. Biotech. 26(10), 1135–1145 (2008)
22. Ibarra, J.A., et al.: Induction of Salmonella pathogenicity island 1 under different growth conditions can affect Salmonella-host cell interactions in vitro. Microbiology 156(4), 1120–1133 (2010)
23. Gao, Q., et al.: Gene expression diversity among Mycobacterium tuberculosis clinical isolates. Microbiology 151(1), 5–14 (2005)
24. Wen, S., et al.: Inflammatory Gene Profiles in Gastric Mucosa during Helicobacter pylori Infection in Humans. The Journal of Immunology 172(4), 2595–2606 (2004)
25. Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10(1), 57–63 (2009)
26. Cossart, P., Archambaud, C.: The bacterial pathogen Listeria monocytogenes: an emerging model in prokaryotic transcriptomics. Journal of Biology 8(12), 107 (2009)
27. Vinayavekhin, N., Homan, E.A., Saghatelian, A.: Exploring Disease through Metabolomics. ACS Chem. Biol. 5(1), 91–103 (2010)
28. Bochner, B.R.: Global phenotypic characterization of bacteria. FEMS Microbiology Reviews 33(1), 191–205 (2009)
29. Benson, D.A., et al.: GenBank. Nucleic Acids Research 39(suppl. 1), D32–D37 (2011)
30. Boutet, E., et al.: UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. Methods Mol. Biol. 406, 89–112 (2007)
31. Kanehisa, M., et al.: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research 38(suppl. 1), D355–D360 (2010)
32. Barrett, T., et al.: NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Research 37(suppl. 1), D885–D890 (2009)
33. Parkinson, H., et al.: ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Research 33(suppl. 1), D553–D555 (2005)
34. Ji, L., et al.: NCBI Peptidome: a new repository for mass spectrometry proteomics data. Nucleic Acids Research 38(suppl. 1), D731–D735 (2010)
35. Wang, Y., et al.: An overview of the PubChem BioAssay resource. Nucleic Acids Research 38(suppl. 1), D255–D266 (2010)
36. Aurrecoechea, C., et al.: ApiDB: integrated resources for the apicomplexan bioinformatics resource center. Nucleic Acids Research 35(suppl. 1), D427–D430
37. Squires, B., et al.: BioHealthBase: informatics support in the elucidation of influenza virus host–pathogen interactions and virulence. Nucleic Acids Research 36(suppl. 1), D497–D503 (2008)

38. Lawson, D., et al.: VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Research 37(suppl. 1), D583–D587 (2009)
39. Henry, C.S., et al.: High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology 28(9), 969–974 (2010)
40. Dyer, M.D., Murali, T.M., Sobral, B.W.: Computational prediction of host-pathogen protein–protein interactions. Bioinformatics 23(13), i159–i166 (2007)
41. Bassaganya-Riera, J., Hontecillas, R.: CLA and n-3 PUFA differentially modulate clinical activity and colonic PPAR-responsive gene expression in a pig model of experimental IBD. Clinical Nutrition 25(3), 454–465 (2006)
42. Ananiadou, S., Sullivan, D., Black, B., Levow, G.-A., Gillespie, J.J., Mao, C., Pyysalo, S., Kolluru, B., Tsujii, J., Sobral, B.: Named Entity Recognition for Bacterial Type IV Secretion Systems (2010) (submitted)
43. Pyysalo, S., Tomoko, O., Cho, H.-C., Sullivan, D., Mao, C., Sobral, B., Tsujii, J., Ananiadou, S.: Towards Event Extraction from Full Texts on Infectious Diseases. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, Uppsala (2010)
44. Noy, N.F., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37(suppl. 2), W170–W173 (2009)
45. Sullivan, D.E., Gabbard Jr., J.L., Shukla, M., Sobral, B.: Data integration for dynamic and sustainable systems biology resources: challenges and lessons learned. Chemistry & Biodiversity 7(5), 1124–1141 (2010)

PART I

Biomedical Electronics and Devices

A Simple and Low Cost Device for Automatically Supervising Urine Output of Critical Patients

Abraham Otero[1,⋆], Francisco Palacios[2],
Andrey Apalkov[3], and Roemi Fernández[3]

[1] Department of Information Systems Engineering, University San Pablo CEU,
28668 Madrid, Spain
abraham.otero@gmail.com
[2] Critical Care Unit, University Hospital of Getafe, Getafe
Carretera Toledo KM 12.500, 28901 Madrid, Spain
franciscodepaula@gmail.com
[3] Centre for Automation and Robotics CSIC-UPM
Ctra. Campo Real Km. 0,2 La Poveda, Arganda del Rey 28500 Madrid, Spain
{andrey.apalkov,roemi}@iai.csic.es
http://biolab.uspceu.com/

Abstract. Nowadays, patients admitted to critical care units have most of their physiological parameters sensed by sophisticated commercial monitoring devices. These devices also supervise whether the values of the parameters lie within a preestablished range of normality set by the clinician. If any of the parameters leaves its normality range, an alarm will be triggered. The automation of the sensing and supervision of physiological parameters discharges the healthcare staff of a considerable workload. It also avoids human errors, which are common in repetitive and monotonous tasks.

Urine output is a physiological parameter that, despite being of great relevance in the treatment of critical care patients, is still measured and supervised manually. This paper presents a device capable of sensing and supervising urine output automatically. The device uses reed switches that are activated by a magnet that is attached to a float in order to measure the amount of urine collected in two containers. An electronic unit sends the state of the reed switches to a PC, which supervises the achievement of therapeutic goals.

Keywords: Biosensors, Urine output, Critical care, Patient monitoring, Fuzzy logic.

1 Introduction

Current critical care units are equipped with sophisticated commercial monitoring devices that allow clinicians to sense nearly any physiological parameter that may provide

⋆ This work was supported by the Spanish MEC and the European FEDER under the grant TIN2009-14372-C03-03 and by the University San Pablo-CEU under the grant USP BS PPC05/2010. A. Apalkov acknowledges the financial support from Ministry of Science and Innovation of Spain under Juan de la Cierva Program. R. Fernández acknowledges the support received from CSIC under JAE-DOC Programme.

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 15–30, 2013.
© Springer-Verlag Berlin Heidelberg 2013

information relevant for interpreting the patient's state. In most cases, these devices can also supervise that the values of the physiological parameters they sense remain within a preestablished range set by the clinician. This range represents the values considered as normal for each parameter. If a parameter does not fall within its acceptable range, the corresponding sensing device alerts the healthcare staff by means of an audible warning [18], [5].

These devices discharge the healthcare staff of a considerable workload, since they need not continuously supervise if the physiological parameters of every patient lie within the acceptable range. They also avoid human errors, which are common in any repetitive and monotonous task such as the one at hand [9], [15].

Arguably, the most relevant physiological parameter which still is measured and supervised manually by healthcare staff is the patient's urine output. Urine output is the best indicator of the state of the patient's kidneys. If the kidneys are producing an adequate amount of urine it means they are well perfused and oxygenated. Otherwise, it is a sign that the patient is suffering from some complication. Urine output is required for calculating the patient's water balance, which is essential in the treatment of burn patients [14]. Finally, it is also used in multiple therapy protocols to check whether the patient reacts properly to treatment [22]. When urine output is too low the patient is said to have oliguria. If the patient does not produce urine at all, then he/she is said to have anuria. Sometimes, urine output can be too high; in these cases the patient is said to have polyuria.

Currently, in critical care units, urine is collected in a graduated container that is connected to the patient's bladder through a Foley catheter (see Figure 1). Periodically the nursing staff manually records the reading of the container of every patient, and operates a valve which releases the urine into a larger container. In critical care units of first world countries, this procedure is usually performed every hour, 24 times a day, 365 days a year. In the case of emerging countries, often only the burn patients –for whom urine output monitoring is of paramount importance– have this parameter registered every hour, while the remaining critical patients have it recorded every 2 or 3 hours. This disparity in criteria is due to the more reduced availability of healthcare staff in the critical care units of emerging countries. It can even be argued that the monitoring interval currently employed in first world countries also is a compromise between avoiding risk states for the patient and relieving the nursing staff of an excessive burden. Hence the interest in developing a device capable of automatically monitoring urine output. Such a device not only would decrease the workload associated with this task, but it would permit the supervision of the therapeutic goals to take place on a more continuous basis.

This paper presents a patent-pending device capable of sensing and supervising urine output [1]. Section 2 reviews related work. Our device is described in Section 3. It uses reed switches that are activated by a magnet that is attached to a float in order to measure the amount of urine collected by two containers. An electronic unit checks the state of the reed switches and sends it to a PC, which supervises the achievement of the therapeutic goals that have been established for urine output. Section 4 discusses the results of this work. Finally, a series of conclusions and lines of future extension are given.

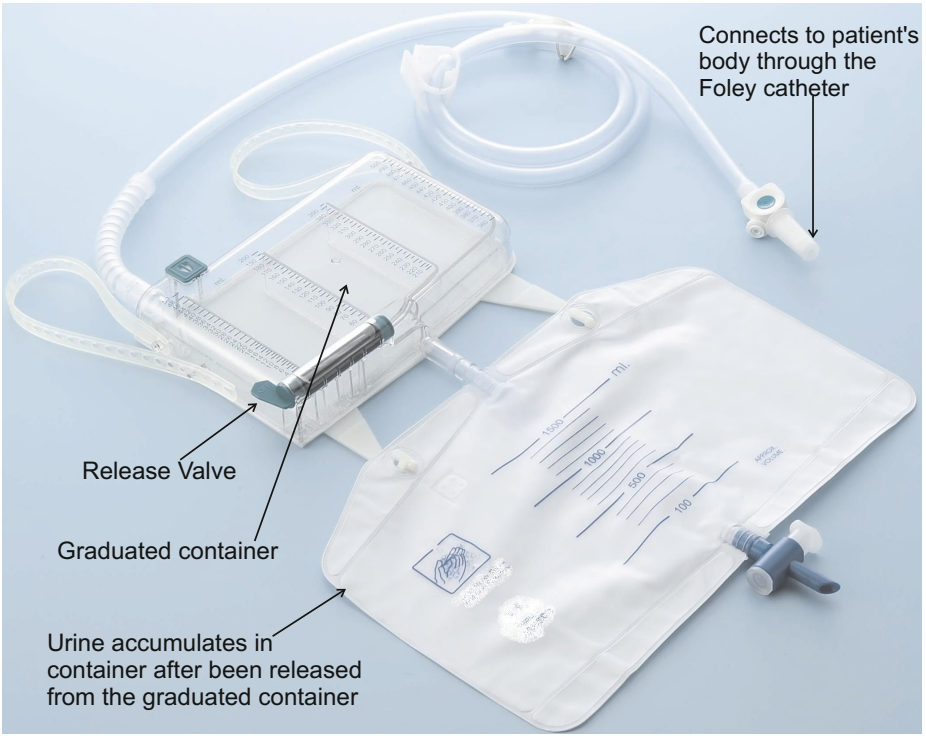


Fig. 1. Typical commercial urine meter currently used in critical care units to measure urine output

2 Related Work

Automating the supervision of urine output is a problem that has only begun to be addressed by the biomedical engineering community very recently. There are several problems that have contributed to this delay. On the one hand, it requires devices capable of measuring very small amounts of flowing liquid (up to 3-5 milliliters per hour). This precludes the use of common industrial solutions such as ultrasound sensors [8] or commercial flowmeters. On the other hand, any component of the device that is in contact, or may be in contact, with the urine cannot be reused on other patients and must be replaced approximately every 4-7 days for hygienic reasons. Therefore, any component of the device that is or may be in contact with the urine must be easy to dispose of, and should have a low price. This precludes the use of expensive high precision laser based solutions [7]. Furthermore, contact with the urine also means that the component is in indirect contact with the patient's bladder through a Foley catheter. Therefore, the components that come into contact with the urine have to be sterilized before use. Finally, urine contains Uric acid, Sodium, Potassium, Chlorine and other components which make it corrosive, especially for metals.

The first device proposed to automatically measure urine output was Urinfo 2000, developed by the Israeli company Medynamix [6]. Urinfo 2000 was designed to automate the hourly urine output measurement, but not to take more frequent measures. Its operation is based on counting the number of drops of urine produced by the patient, and from this count it estimates urine output. The average error of the device when used to take hourly measurements was 8% ($\pm$25 ml). Urinfo 2000 cannot supervise therapeutic goals and its readings are not transmitted to a central station/PC. Thus it requires the nursing staff to take the urine output measures from the device's display, placed next to the patient's bedside.

The authors of this paper proposed a device to automate the supervision of urine output in [19]. Several technical and legal problems precluded us from moving this device beyond the laboratory validation phase. Using the knowledge we gained building it, we developed a second device with the main objective of conducting a series of clinical studies based on more continuous and accurate monitoring of urine output throughout the stay of patients in a critical care unit [17] [16]. This device uses a high precision scale to measure the weight of a commercial urine meter. On the scale's pan there is a support frame made up of Bosch profiles that isolates the scale from force transmission from the patient's bed, and guarantees that the urine flows properly through the urine meter's input tube. The maximum measurement error of this device is under 1.5%.

Currently this device is in use at a research unit associated with the University Hospital of Getafe, in Spain. In this research unit a series of experiments aimed at the study of sepsis in an animal model –pigs– are being conducted. Sepsis is a serious medical condition characterized by a systemic inflammatory state that has developed in response to an infection and affects the patient´s entire body. Patients affected by this pathology are treated in an intensive care unit, and often require dialysis to support the functioning of the kidneys [11]. In these experiments, sepsis is induced by the administration of Escherichia coli bacteria. Animals are anesthetized and mechanically ventilated via tracheostomy. Blood pressure, pulmonary arterial pressure, renal artery flow and central flow are monitored invasively during the experiment. Urine output is continuously monitored by our device [17].

The aim of these experiments is twofold. On one hand, we would like to gain a better understanding of the pathophysiological mechanisms underlying systemic and renal hemodynamics during sepsis –hence the interest in a continuous and accurate monitoring of urine output. On the other hand, we would like to define the optimum monitoring interval for urine output, and to determine the level of accuracy required for this task.

The high accuracy and acquisition rate of this device make it ideal for carrying out clinical studies. However, this device was not intended to be used in the clinical routine. Its size and operation make it somewhat tedious to be used in a critical care unit. Conversely, in using this device we have learned that its accuracy –1.5%– and measurement interval –up to 10 sec.– are superior to what is required for properly supervising urine output. Our experience shows that an accuracy of 5% and a monitoring interval of 5-10 minutes is enough. Therefore, these parameters can be relaxed to obtain a cheaper and simpler device. That is the goal of the device presented this paper.

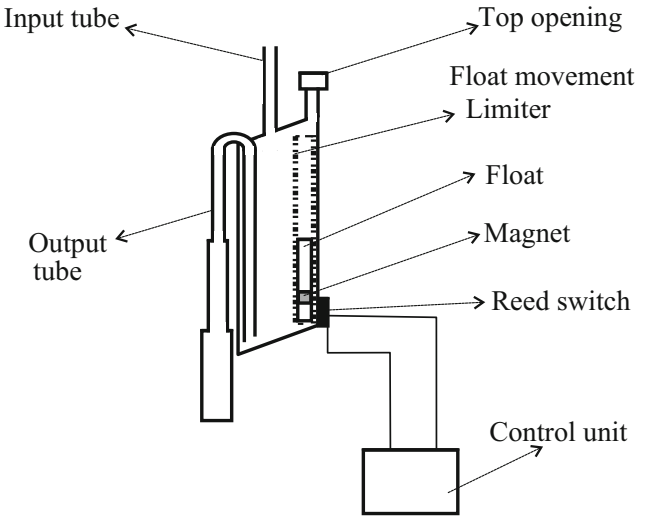


Fig. 2. Diagram of the small container of our prototype

3 The Device

In this section we first describe the general operating principles on which our device is based. Although we will use the device to measure urine output, the principles on which it is based are general and could be used to measure the volume of any liquid flowing through a tube. Then we describe the prototype we have built, and the calibration procedure we have carried out to determine its operating parameters. Finally, we describe how the supervision of therapeutic goals is carried out.

3.1 General Operating Principles

In our device the urine arising from the patient's bladder is collected through an input tube in a container. This container is equipped with an output tube that empties its contents when it gets filled by taking advantage of the siphon principle (see Figure 2). One end of the output tube will be placed in the bottom wall of the container. Ideally the bottom wall of the container will be inclined or conical and the tube will be in the lowest part. This favors a more complete evacuation of the container. Both ends of the output tube are always open. When the container begins to receive liquid, the liquid goes up inside of the output tube until it reaches the elbow located in its top part. When it reaches the elbow, liquid begins to fall down the portion of the tube located on the exterior of the container. By the principle of the siphon, the liquid will flow down the tube and the container will be emptied.

For the system to function properly, the output tube should not be too wide. If the tube is too wide, when the water starts flowing, air could rise through the tube towards the container, breaking the siphon principle. On the other hand, to drain the container as quickly as possible, it is desirable that the tube be as wide as possible. A compromise

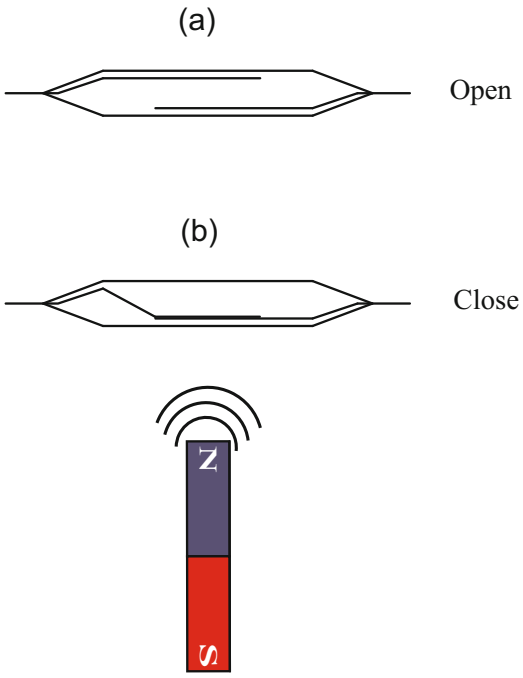


Fig. 3. Operation of a reed switch. In the presence of a magnetic field, the switch is closed. In the absence of magnetic field, it is open.

between these two goals can be achieved using a variable diameter tube, so that its diameter grows (staggered or progressively) from the elbow of the tube to the end which is outside the container (see Figure 2). In this way, the principle of the siphon will work properly because the tube is narrow enough in its upper part, but when the water begins to fall it finds a wider tube which provides less resistance, thereby increasing flow rate.

The container must also be equipped with a filtered top opening that is used to equalize the internal and external pressures without risk of bacterial contamination. This opening should be located at a height H_1 above the elbow of the output tube. The liquid will not start flowing immediately when it reaches the elbow height; but it will continue to accumulate in the container until the height of the column of liquid which is above the elbow height overcomes the surface tension of the liquid against the walls of the output tube. The surface tension increases as the diameter of the output tube decreases. The height H_1 and the effective volume of the container V^S (the volume of water contained just before the draining starts) must be determined experimentally.

Ideally, the upper wall of the container will also be inclined or conical in shape. The top opening to equalize the internal and external pressures should be placed in the uppermost part of the wall (see Figure 2). This prevents bubbles from forming in the upper wall of the container, bubbles that would occupy volume and distort the measures.

One or more reed switches are placed on the outside of the container that will receive the urine. A reed switch contains two or more magnetizable, flexible, metal reeds hermetically sealed in a tubular glass envelope whose end portions are separated by a

small gap. Under these conditions, the switch is open (see Figure 3a). A magnetic field properly applied will cause the reeds to bend, and the contacts to pull together, thus closing the switch (see Figure 3b).

A float within a structure designed to limit its movement so that the float can only move vertically is located inside of the container, near the container wall where the reed switches are placed (see Figure 2). The float has a magnet attached which interacts with the reed switches closing them when the magnet is approximately at the same height as each of the reed switches.

An electronic unit is connected to the reed switches, continuously checking their state. For the general case in which there are N reed switches on the outside wall of the container, the procedure for measuring the volume of liquid flowing into the container is as follows. At least one reed switch should be located in a position such that when the container is empty, the magnet located on the float closes the reed switch. As the liquid begins to flow into the container, the float, and therefore the magnet, begins to raise. At some point, the magnet will stop interacting with the first reed switch and, therefore, it opens. At that point, a volume V_1 of liquid has flowed into the container. When enough liquid has been accumulated in the container, the magnet will close the second reed switch. At this point an additional volume V_2 of liquid has flowed into the container, being the total amount of liquid accumulated $V_1 + V_2$. When the magnet moves higher, the second reed switch opens again and an additional volume V_3 of liquid has flowed –being the total volume of liquid $V_1 + V_2 + V_3$.

In general, when the reed switch N is closed, we will add the volume V_{2N-2} to the the volume of liquid that has flowed, and when the reed switch N is opened again, we will add the volume V_{2N-1}. When the container is emptied, the float with the magnet will go back to the bottom of the container, and therefore it will close the first reed switch. At this point, the effective volume of the container $-V^C-$ has flowed, and the measurement procedure is resumed.

The volumes V_{2N-2} and V_{2N-1} for each reed switch, and the effective volume V^C must be determined experimentally.

3.2 Our Prototype

The general operating principle described in the previous section suffers from a problem: the volume of liquid that flows into the container from the time the container begins to empty through the siphon mechanism, until it is completely empty, will not be measured. Depending on the specific application and characteristics of the container, this may or may not be tolerable. The problem that concern us, the measurement of the amount of urine produced by a patient, requires an early warning about deviations from the therapeutic goals. Thus, a small volume container must be used –approximately 5 ml in the prototype we have built. A container of such a small size needs to be emptied a large number of times, which may make the inability to measure the liquid that flows during the discharge of the container intolerable if the patient has polyuria, i.e., the patient is producing a large amount of urine. If the patient is producing normal amounts of urine, or if he/she has oliguria, the amount of urine that will flow during discharge of the container will be nil or negligible.

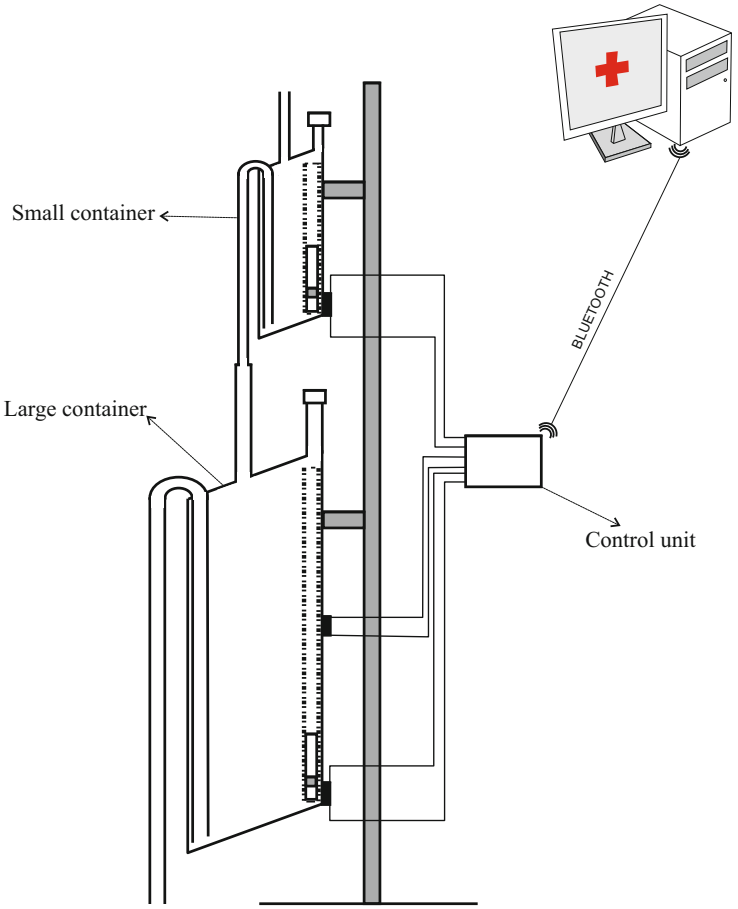


Fig. 4. Diagram of our prototype device. Each of the containers operates according to the principle shown in Figure 2.

It could be argued that in the case of patients that have polyuria it is not required to have a high degree of accuracy in the measurement of the urine output. The more urine is produced, the less important it is to have an accurate measure; only when the patient is producing small amounts of urine it is important to measure accurately [17]. Therefore, the operation of the device as is described in the previous section could be acceptable. In any case, this problem can easily be solved using two different size containers, each of them working according to the principle presented in the previous section. The output of the smaller container is connected to the input of the larger container. Thus, although the volume of liquid that flows during the discharge of the first container is not measured in the small container, it will be measured in the larger one.

Chaining two containers as described here causes another problem. It can happen that when the small container releases its content, the larger one is nearly full. In the worst

scenario, the first drop that falls from the small container would trigger the draining of the large container. Thus, the content of the small container would not be measured in the large one. The maximum error this may cause is $(V^S/V^L)\%$, being V^S and V^L the effective volumes of the small and large container, respectively. On average, the large container will start to drain when the small container is emptied halfway. Therefore, the average error caused by this effect will be $((V^S/2)/V^L)\%$. As long as $V^L >> V^S$, this error will be small.

The prototype we have built uses two containers. The first one was built to have a volume of approximately 5 ml, and the second a volume of approximately 375 ml (see Figure 4). The first container is equipped with a single reed switch that is activated when the float is on its lowest position. The staggered diameter of its output tube increases in three steps. The output tube is connected to the input tube of the larger container, which is equipped with two reed switches. The tube of the large container is connected to a 2.5 liters plastic bag that collects the liquid once it has been measured.

An electronic unit continuously checks the status of the reed switches, and reports any change via Bluetooth to a Java application that runs on a PC. From these changes the application calculates the volume of urine that has flowed and displays a chart with this information.

3.3 Device Calibration

The calculation of the effective volumes of the containers –V^S and V^L– must be determined experimentally. By effective volume we mean the volume of liquid that triggers the emptying of the container through the siphon mechanism. This volume will be slightly higher than the volume corresponding to the height of the top of the siphon mechanism because the liquid does not begin to flow until the pressure overcomes the surface tension force of the liquid against the walls of the output tube. The smaller the diameter of the output tube, the greater the height in the column of liquid required to overcome the surface tension will be.

The effective volume of each container was determined separately. A saline solution with properties similar to those of urine and an dropper were used to simulate the flow of urine (see Figure 5). Each of the containers were placed so that they would release their content on a bowl placed on the plate of a high-precision industrial scale –a PGW 4502e, built by Adam Equipment Inc. This scale has an accuracy guaranteed by the manufacturer of 0.01 g. Given that the density of the saline solution was known, we can determine the volume of liquid that the container releases into the bowl from the weight.

The PGW 4502e is equipped with a serial port that permits querying for readings. We built a program that periodically takes measurements from the scale. The program together with the dropper allow us to automate the process of carrying out multiple measures of the volume of liquid released by the containers.

Using this set up we took 200 measurements of the volume of liquid released by the small container, and 50 measures of the volume of liquid released by the large container. From these measures we calculated the effective volumes of the containers: 5.87 ± 0.32 ml and 376.72 ± 1.11 ml (mean $\pm$ standard deviation). At the time of writing, the authors are working to determine the volumes V_1^L, V_2^L and V_3^L, i.e., the

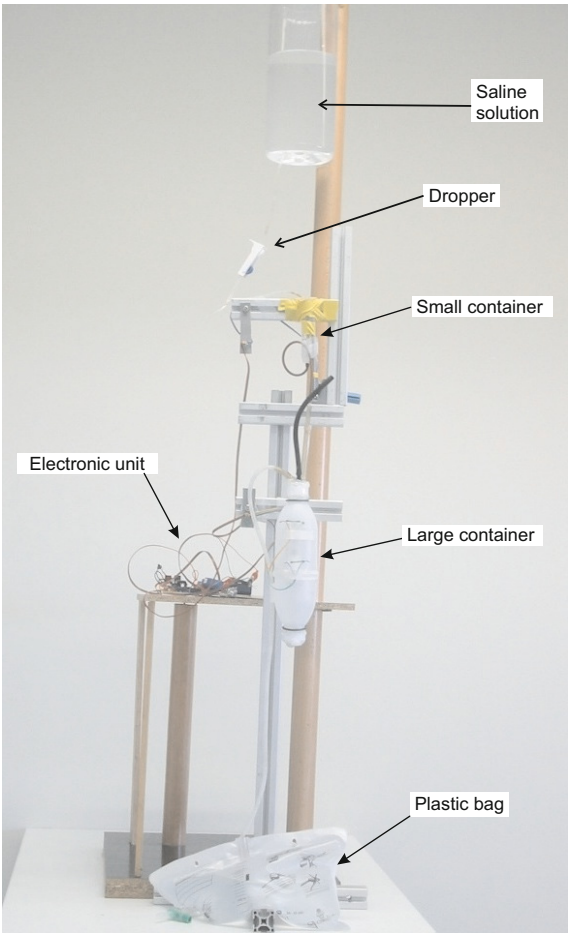


Fig. 5. Picture of the prototype device with the saline solution and eye dropper used in its validation

volumes corresponding with the opening of the first reed switch of the large container, with the closing of the second reed switch, and whit the opening of second reed switch, respectively. This is challenging because it requires measuring with high accuracy the volume of liquid in the container before the liquid is released, without interfering with the normal operation of the device. These volumes are not necessary for the system to work correctly, but they would allow us to correct the measurements obtained from the small container –which are less precise–, not just one time but four times for each release cycle of the large container.

3.4 Therapeutic Goals Supervision

The state of the reed switches placed in the containers is sent from the electronic unit to a Java application installed on a PC. From these states and from the values of V^S

and V^L, the amount of liquid flow can be calculated. The Java application allows the healthcare staff to inspect a graph showing urine output, and to set the therapeutic goals for the urine output. These therapeutic goals are represented with the aid of the Fuzzy Set Theory, a tool which has proved its value for handling and representing medical knowledge [3].

We shall introduce some basic concepts of the Fuzzy Set Theory. Given a discourse universe $\mathbb{R}$ we define a *fuzzy value C* by means of a possibility distribution π^C defined over $\mathbb{R}$ [23]. Given a precise value $x \in \mathbb{R}$, $\pi^C(x) \in [0,1]$ represents the possibility of C being precisely x. A *fuzzy number* [10] is a normal ($\exists x \in \mathbb{R},\ \pi^C(x) = 1$) and convex ($\forall\, x,\, x',\, x'' \in \mathbb{R}, x' \in [x,x''], \pi^C(x') \geq min\{\pi^C(x), \pi^C(x'')\}$) fuzzy value. Normality and convexity properties are satisfied by representing π^C, for example, by means of a trapezoidal representation. In this way, $C = (\alpha, \beta, \gamma, \delta)$, $\alpha \leq \beta \leq \gamma \leq \delta$, where $[\beta, \gamma]$ represents the core, $core(C) = \{x \in \mathbb{R} |\ \pi^C(x) = 1\}$, and $]\alpha, \delta[$ represents the support, $supp(C) = \{x \in \mathbb{R} | \pi^C(x) > 0\}$.

A fuzzy number C can be obtained from a flexible constraint given by a possibility distribution π^C, which defines a mapping from $\mathbb{R}$ to the real interval $[0,1]$. A fuzzy constraint can be induced by an item of information such as *"x has a high value"*, where *"high value"* will be represented by $\pi^{C=high}$. Given a precise number $x \in \mathbb{R}$, $\pi^{C=high}(x) \in [0,1]$ represents the possibility of C being precisely x; i.e., the degree with which x fulfills the constraint induced by *"high value"*.

Physicians are accustomed to expressing the therapeutic goals for urine output in milliliters of urine produced per kilogram of patient body mass per hour $-ml/kg \cdot h$. Our tool allows them to indicate the weight of the patient -*P*- and the therapeutic goals represented by the trapezoidal possibility distribution π^U. π^U can be interpreted as a computational projection of the piece of clinical knowledge "adequate UO". The minimum and maximum values acceptable for urine output are the beginning and the end of the support of the distribution, respectively. If the patient produces less urine than the amount corresponding with the beginning of the support, the patient is clearly in oliguria. If he/she produces more urine than the amount corresponding with the end of the support, the patient is clearly in polyuria. The beginning and end of the core are the limits of the interval within which ideal values of urine output lie (see Figure 6).

If u_i is the urine output in $ml/kg \cdot h$, the degree to which the therapeutic goals established by the clinician are being met is $\pi^U(u_i)$. If $\pi^U(u_i) = 1$, the urine output is within the range of ideal values. If $\pi^U(u_i) = 0$, either the urine output is less than the minimum acceptable value -the patient has oliguria or anuria-, or greater than the maximum acceptable –the patient has polyuria. In both cases, the program produces an audible warning. The closer $\pi^U(u_i)$ is to 1, the closer the amount of urine produced by the patient is to the ideal value, and the closer $\pi^U(u_i)$ is to 0, the closer the patient is to oliguria or polyuria.

In the tool $\pi^U(u_i)$ is represented by a color code used when drawing the graph of urine output. Red represents the null compatibility –the patient is clearly in oliguria or in polyuria–, followed by orange, yellow, green and black, which represents the total compatibility -the urine output lies within the range of ideal values. Therefore, the urine output graph provides instant visual feedback on the patient's state.

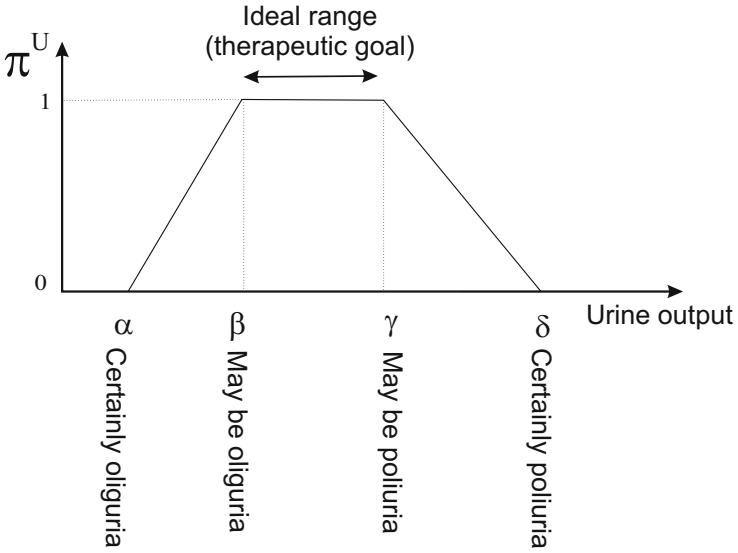


Fig. 6. The therapeutic goals for urine output are represented by a trapezoidal possibility distribution

The tool generates an audible warning when the small container is not filled within the maximum time allowed by π^U; i.e. when the patient clearly has oliguria. The time at which the alarm is triggered is given by:

$$t_{alarm\ oliguria} = \alpha \cdot P/(V^S + 2 \cdot \sigma_{V^S}) \tag{1}$$

where α is the beginning of the support of π^U, P is the patient's weight, V^S is the effective volume of the small container, and σ_{V^S} is the standard deviation of the effective volume of the small container. The correction $V^S + 2 \cdot \sigma_{V^S}$ is applied to avoid false positives caused when the small container does not release its contents when urine output reaches its average effective volume, but some more urine accumulates before releasing.

Similarly, the tool also produces an audible warning when the small container fills faster than the minimum time allowed by π^U; i.e. when the patient clearly has polyuria. This alarm is triggered when the small container fills before the time given by:

$$t_{alarm\ polyuria} = \delta \cdot P/(V^S - 2 \cdot, \sigma_{V^S}) \tag{2}$$

where δ is the end of the support of π^U. The correction $V^S - 2 \cdot \sigma_{V^S}$ is applied to avoid false positives caused when the small container releases its contents before reaching its effective volume.

4 Discussion

Given that the effective volume of the small container is 5.87 ± 0.32 ml (mean ± standard deviation), in 95% of cases the amount of urine that is really released when the

container is emptied will be in the range 5.87 ± 0.64 ml (mean $\pm$ 2·standard deviation). Therefore, 95% of the measures derived from the small container will have an error of 10.9% or less. The purpose of this small container is to provide an early warning if the patient's urine output is not within the therapeutic goals: depending on the patient's weight and state this container should be filled within 5 and 12 minutes, if the patient is within the limits of the therapeutic goals. If this does not occur, the healthcare staff is alerted by an audible warning.

To obtain more accurate measures of the total urine output, we rely on the second container. In 95% of cases, the amount of urine released by this container will be within the range 376.72 ± 2.22 ml. On the other hand, on average half of the effective volume of the small container will not be registered. Thus, the error in the measures in 95% of the cases will be 5.87/2+2.22=5.16; 1.37% of its volume. This error is low, especially when compared to errors committed by the nursing staff when they take measures visually. For example, in the commercial urine meter Unometer 500 –the one used in the ICU of the hospital we collaborate with– the separations between the coarser measurement divisions in the 500 ml graduated container where urine is collected correspond with 20 ml. Therefore, the minimum measurement error is 4%. However, this is a highly optimistic estimation of the error. It assumes that the visual measurements are taken in ideal conditions; i.e., the graduated container is not tilted in any direction and the nurse's eyes are in the same horizontal plane as the graduated container. In practice, the average visual measurements error has been reported to be as high as 26% [6].

It can be argued that the recording interval currently employed for urine output –once every hour– is a compromise between avoiding risk states for the patient and relieving the nursing staff of an excessive burden. The automation of the monitoring of urine output would decrease the workload associated with this task and, at the same time, it would permit supervision to take place on a more continuous basis. For a patient of 80 kilos, who should produce at the very least least 40 milliliters of urine per hour, our device could warn of a deviation from the therapeutic goals in less than 10 minutes – the time in which the small container should be filled if the patient produces urine within the therapeutic goals. Therefore, the device has the potential of improving patient outcome: given that it provides feedback on the patient's condition at shorter time intervals, it can allow the clinician to react more promptly to complications in the state of the patient.

On the other hand, there are abundant studies in the medical literature showing that critical patients' outcome improves when the number of nurses [13] [2] [12] and the number of clinicians [20] [4] available in the units increases. One of the factors that most influences the staffing of critical care units is the financial resources of the health system of each country [21]. Hiring highly qualified personnel, such as critical care unit staff, is very expensive. However, patient outcomes could also be improved by reducing the healthcare staff workload related to simple and repetitive tasks. This will free up time that could be used on other tasks.

From this point of view, the automation of the processes related to sensing and supervising urine output can produce significant time savings. Currently, in developed countries every hour a nurse must visit each of the patient's bed of the critical care unit to manually record urine output. The nurse must put on gloves since he/she is going to manipulate body fluids, walk to the patient's bed, take the measure visually, write it

down, open the valve that releases urine from the graduated container to the plastic bag, wait for the urine to drain, close the valve, check that the valve is properly sealed and check if the plastic bag needs to be emptied. This procedure takes at least 2 minutes. In a critical care unit with 15 patients, this means 30 minutes per hour; 12 hours a day.

Typically, nurses work in shifts of six hours. Therefore, each day two complete nursing shifts are required only for tasks related to supervising urine output. If these tasks were automated by a low cost device, we may obtain an improvement in patient outcomes equivalent to the improvement that would be obtained when two nurses per day are added to the staffing of the unit.

The containers needed to build our device are not more complex than the containers used in commercial urine meters, which often require a small container embedded within a large container, valves that communicate both containers with each other and with a plastic bag, mechanisms to prevent the containers from overflowing, etc. However, the price of our manufacturing our device is slightly higher than the price of the commercial urine meters because of the addition of four pieces: two floats, and two magnets. The rest of the pieces that are part of our solution do not have to be discarded because they are not in direct contact with the patient's urine, nor do they suffer significant degradation caused by its operation. Thus, their cost can be amortized over a long period of time and their impact on the overall cost of the solution is negligible. Given that the additional cost of our device can be offset by the reduction in the healthcare staff workload, its slightly higher price should not be a barrier to adoption.

5 Conclusions and Future Work

We have built a device capable of automatically sensing and supervising the urine output of critical care patients. The device comprises two containers of different volumes, a small one that receives the urine coming from the patient's bladder, and a greater volume container in which the first container releases its content when it gets full. Both containers release their content automatically when they are filled using a siphon mechanism.

The containers are equipped with reed switches that are activated by a magnet that is attached to a float located inside the containers. These reed switches allow us to identify the instants at which they get filled with urine. An electronic unit sends via Bluethooth the information provided by the reed switches to a PC which calculates the urine output from the switches' state, and supervises the achievement of the therapeutic goals established by the clinician. The error in measuring the patient's urine output is under 2%. The large container is the one which allows us to obtain this high accuracy, while the small one permits an early warning of deviations from the therapeutic goals.

The cost of our device is slightly higher than the cost of the commercial devices currently used in monitoring urine output. However, the device has the potential to save costs for the institutions that provide health services by freeing a considerable amount of time for the healthcare staff. Furthermore, it provides a more continuous supervision of the urine output than is currently carried out in critical care units, which may help improve patient outcomes.

As future work, we intend to take advantage of all the state changes of the reed switches of the large container to correct the urine output measures while the large

container is been filled. Currently, this correction is only performed when the large container releases the urine. We also will start to use our device in animal tests conducted in a research unit associated with Getafe University Hospital. After this phase, we intend to use it in a pilot test in the Intensive Care Unit of this hospital.

References

1. Akinfiev, T., Apalkov, A., Otero, A., Palacios, F.: Device for measuring the amount of liquid that flows and procedure for its measurement. Patent Pending, eS 201031227 (2010)
2. Amaravadi, R., Dimick, J., Pronovost, P., Lipsett, P.: ICU nurse-to-patient ratio is associated with complications and resource use after esophagectomy. Intensive Care Medicine 26(12), 1857–1862 (2000)
3. Barro, S., Marín, R., Palacios, F., Ruíz, R.: Fuzzy logic in a patient supervision systems. Artificial Intelligence in Medicine 21, 193–199 (2001)
4. Dimick, J., Pronovost, P., Heitmiller, R., Lipsett, P.: Intensive care unit physician staffing is associated with decreased length of stay, hospital cost, and complications after esophageal resection. Critical Care Medicine 29(4), 753 (2001)
5. Hande, A., Polk, T., Walker, W., Bhatia, D.: Self-powered wireless sensor networks for remote patient monitoring in hospitals. Sensors 6(9), 1102–1117 (2006)
6. Hersch, M., Einav, S., Izbicki, G.: Accuracy and ease of use of a novel electronic urine output monitoring device compared with standard manual urinometer in the intensive care unit. Journal of Critical Care 24, 629–633 (2009)
7. Ishida, S.: Liquid level indicator using laser beam. us patent 4938590 (1990)
8. Johnson, S.J.: Liquid level measurement device. United States Patent 3693445 (1978)
9. Jungk, A., Thull, B., Rau, G.: Intelligent Alarms for Anaesthesia Monitoring Based on Fuzzy Logic Approach, pp. 219–238. Physica-Verlag (2002)
10. Kaufmann, A., Gupta, M.: Introduction to Fuzzy Arithmetic. Van Nostrand Reinhold Company Inc. (1984)
11. Klenzak, J., Himmelfarb, J.: Sepsis and the kidney. Critical Care Clinics 21(2), 211–222 (2005)
12. Knauf, R., Lichtig, L., Risen-McCoy, R., Singer, A., Wozniak, L.: Implementing nursing's report card: a study of RN staffing, length of stay and patient outcomes. American Nurses Association, Washington (1997)
13. Kovner, C., Gergen, P.: Nurse staffing levels and adverse events following surgery in US hospitals. Journal of Nursing Scholarship 30(4), 315–321 (1998)
14. Mitra, B., Fitzgerald, M., Cameron, P., Cleland, H.: Fluid resuscitation in major burns. ANZ Journal of Surgery 76, 35–38 (2006)
15. Mora, F., Passariello, G., Carrault, G., Pichon, J.L.: Intelligent patient monitoring and management systems: A review. IEEE Engineering in Medicine and Biology 12, 23–33 (1993)
16. Otero, A., Akinfiev, T., Fernández, R., Palacios, F.: A device for automatic measurement of critical care patient's urine output. In: 6th IEEE International Symposium on Intelligent Signal Processing, pp. 169–174 (2009)
17. Otero, A., Akinfiev, T., Fernández, R., Palacios, F.: A device for automatically measuring and supervising the critical care patient's urine output. Sensors 10(1), 934–951 (2010)
18. Otero, A., Félix, P., Barro, S., Palacios, F.: Addressing the flaws of current critical alarms: a fuzzy constraint satisfaction approach. Artificial Intelligence in Medicine 47(3), 219–238 (2009)
19. Otero, A., Panigrahi, B., Palacios, F., Akinfiev, T., Fernández, R.: A prototype device to measure and supervise urine output of critical patients, pp. 321–324. Intech (2009)

20. Pronovost, P., Angus, D., Dorman, T., Robinson, K., Dremsizov, T., Young, T.: Physician staffing patterns and clinical outcomes in critically ill patients: a systematic review. Jama 288(17), 2151 (2002)
21. Pronovost, P., Needham, D., Waters, H., Birkmeyer, C., Calinawan, J., Birkmeyer, J., Dorman, T.: Intensive care unit physician staffing: Financial modeling of the Leapfrog standard*. Critical Care Medicine 32(6), 1247 (2004)
22. Rivers, E., Nguyen, B., Havstad, S., Ressler, J., Muzzin, A., Knoblich, B., Peterson, E., Tomlanovich, M., Group, E.G.D.T.C.: Early goal-directed therapy in the treatment of severe sepsis and septic shock. New England Journal of Medicine 345, 1368–1377 (2001)
23. Zadeh, L.: The concept of a linguistic variable and its application to approximate reasoning. Information Science 8, 199–249 (1975), part 1

Photoplethysmogram Processing Using an Adaptive Single Frequency Phase Vocoder Algorithm

Walter Karlen[1], Chris Petersen[2], Jennifer Gow[2],
J. Mark Ansermino[2], and Guy Dumont[1]

[1] Electrical and Computer Engineering in Medicine (ECEM)
The University of British Columbia, Vancouver, Canada
{walterk,guyd}@ece.ubc.ca
[2] Pediatric Anesthesia Research Team (PART)
British Columbia Children's Hospital, Vancouver, Canada
{CPetersen,JGow}@cw.bc.ca,
anserminos@yahoo.ca
http://ecem.ece.ubc.ca,
http://part.cfri.ca

Abstract. We have previously designed a pulse oximeter connected to a mobile phone, called the *Phone Oximeter*, for clinical decision making based on photoplethysmography. The limited battery and computational resources demand efficient and low-power algorithms for the *Phone Oximeter* to be effective in resource-poor and remote areas. We present two new algorithms for the fast and economical estimation of heart rate (HR) from the photoplethysmogram (PPG). One method estimates the HR frequency by adaptively modeling the PPG wave with a sine function using a modified phase vocoder. The other method uses the obtained wave as an envelope for the detection of peaks in the PPG signal. HR is computed using the vocoder center frequency or the peak intervals in a histogram, respectively. PPG data obtained from 42 subjects were processed with the vocoder algorithms and, for comparison, with two traditional methods that use filtering algorithms (Pan-Tompkins) and frequency domain transformations (Fast-Fourier Transform). We compared HR estimates obtained from these four methods to the reference HR obtained from a electrocardiogram. The two vocoder methods performed at least as well as the two traditional methods in terms of normalized root mean sqare error and robustness towards artifacts. Experiments on a mobile device prototype showed comparable speed performance of the vocoder algorithms with the Pan-Tompkins algorithm while the frequency domain approach was nearly two orders of magnitude slower. These results point to further developments using a combination of both vocoder HR estimation methods that will enable the robust implementation of adaptive phase vocoders into mobile device health applications.

Keywords: Pulse detection, Heart rate estimation, Mobile phones, Embedded systems, Photoplethysmography, Vocoder.

1 Introduction

Mobile health technology is a rapidly advancing field that holds great promise for improving medical services and changing the way that health care is delivered. A common

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 31–42, 2013.
© Springer-Verlag Berlin Heidelberg 2013

theme in this area is the use of general purpose consumer devices, in particular smart phones. An increasing number of health care applications use these mobile platforms to interface directly to physiological sensors, such as heart rate (HR) detectors. This reduces or eliminates the cost of custom embedded hardware. However, features such as increased noise level, limited battery and computational resources, and the requirement for on-line processing challenge the accurate, real-time detection of physiological parameters, such as HR peaks.

In this paper, we consider the extraction of heart rate (HR) from a photoplethysmography (PPG) signal that originates from an oximeter sensor interfaced to a mobile device. We propose a novel approach to compute HR from the PPG signal based on a dynamically adapted, single frequency phase vocoder algorithm. This algorithm is intended to form a core engine for more complex mobile signal analyzers within a fully functional low-cost, mobile phone-based pulse oximeter.

1.1 Related Work

The traditional methods used for peak detection in the electrocardiogram (ECG) signal for HR estimation, which have a long history in biomedical signal processing, can be applied to PPG signals. These HR estimation algorithms operate either in the time or the frequency domain. Common time domain algorithms include linear and non-linear filters, artificial neural networks, genetic algorithms, filter banks, and heuristic methods based on nonlinear transforms. Frequency domain algorithms include various wavelet and Fourier transforms. However, not all of these methods are suitable for on-line computation [6]. In particular, algorithms must be computationally efficient for mobile applications, where battery power and computational resources are limited. Algorithms are commonly trimmed in order to achieve computational efficiency [6]. However, this comes at the expense of accuracy and performance. Our aim was to design an accurate HR estimation algorithm for PPG that requires low processing power and so would be suitable for battery-powered mobile applications.

2 Algorithm Development

Since cardiac signals are quasi periodic, a time-frequency transformation algorithm could be appropriate to extract HR. However, Fourier or wavelet-based transformations on time series require segmentation of the signal, which is not always practical in an on-line, low-cost system. Their relatively high processing power and memory usage requirements add to this unsuitability. Instead, we propose to operate in the time domain, using a methodology inspired by the phase vocoder originally developed for the compression of voice signals in telecommunications [2]. In this method, a vocoder models the input signal with one or multiple sinusoidal waves that vary in time. The parameters that have to be estimated are the time varying amplitude and frequency of each sine wave that comprise the original signal. The phase vocoder can be seen as a filter bank consisting of a series of band-pass filters with successive center frequencies [1].

In our case, we are solely interested in the dominant frequency in the signal (the heart beat). It is, therefore, sufficient to represent the input signal by only one sinusoidal wave

that varies over time. The filter-bank with a distinguished set of frequencies is replaced with a single band-pass filter whose center frequency is adapted over time. This can be achieved by scanning the incoming signal and computing the difference in phase. The obtained frequency parameter, where the system eventually locks-in, can be seen as a filtered frequency of the incoming signal that would correspond to an averaged HR. The method can also find the location in time of each heart beat by using the output sine wave as an envelope to locate maximal peaks in the input waveform.

2.1 Algorithm Description

The raw PPG signal is high-pass filtered to remove the baseline using a second-order Butterworth filter with a cut-off frequency at 0.5 Hz (Figure 1-1). The filtered signal is then routed into two parallel streams. The signal is multiplied by a sine wave in one stream and by a cosine wave in the other (Figure 1-2). Cosine and sine waves have the same unitary amplitude and frequency w_v. The frequency w_v is set to the estimated vocoder frequency of the previous iteration (Figure 1-8). The two parallel streams are, therefore, identical with the exception of the $\pi/2$ shifted phase of the multiplying waveform. This step creates a new signal that is composed of two periodic signals with shifted frequencies of $\pm w_v$ as follows:

$$\begin{aligned} &\cos(w_v * t) * \cos(w_s * t) = \\ &\cos((w_s - w_v) * t) + \cos((w_s + w_v) * t), \end{aligned} \tag{1}$$

where w_s is the frequency of the incoming wave at timestep t. Next, each of the two streams is fed to a moving average low-pass filter (Figure 1-3). The application of this heterodyning step has two effects. First, input frequencies in proximity of the vocoder frequency w_v are shifted down close to DC and are allowed to pass the filter. All other frequency components will also be shifted but they will not go through the low-pass filter. Secondly, the heterodyning provides a way to compute the time-varying amplitude and frequency of the resulting signal in the next step. The two filtered waveforms are subsequently transformed from Cartesian to Polar coordinates to obtain a single amplitude r_v and phase θ_v (Figure 1-4). The amplitude is calculated as the square root of the sum of the squares of the two heterodyned signals as follows:

$$r_v = 4 * \sqrt{y_{sin}^2 * y_{cos}^2}, \tag{2}$$

where y_{sin} and y_{cos} are the heterodyned signals for the sine and the cosine streams, respectively. Similarly, θ_v at each point in time, t, is the angle whose tangent is the ratio of the vertical to the horizontal position as follows:

$$\theta_v = \arctan(\frac{y_{sin}}{y_{cos}}). \tag{3}$$

The phase is subsequently unwrapped. The real time-varying frequency of the original wave is then estimated by computing the difference between the actual and previous phase (Figure 1-5), and subtracting it from the current center frequency (Figure 1-6) as follows:

$$\hat{w_s} = w_v = w_v^{(t-1)} - \Delta\theta, \tag{4}$$

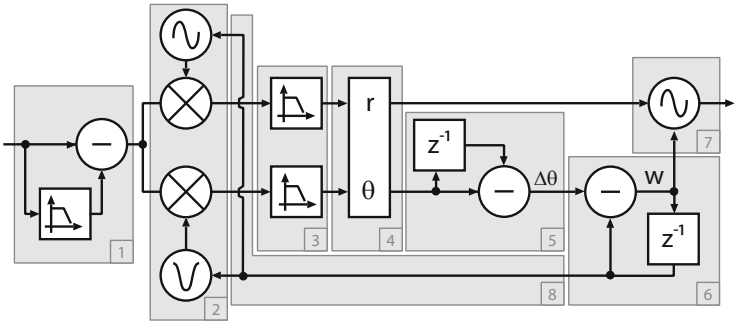


Fig. 1. Operation of the adaptive single-frequency phase vocoder: (1) high-pass filter to remove DC value, (2) heterodyning the input with both a sine and a cosine wave in parallel, (3) a low-pass filter, (4) converting the two signals from rectangular to polar coordinates and unwrapping the angular-position values, (5) subtracting successive unwrapped angular-position values, (6) subtracting the phase-difference from the previous center frequency to obtain the new center frequency, (7) generating the vocoder output wave for peak detection, and (8) feeding the center frequency back into the heterodyne function.

where $\Delta\theta = \theta_v^{(t_{-1})} - \theta_v$. The amplitude and the newly computed center frequency are used to compute the vocoder output y_{out} (Figure 1-7) as follows:

$$y_{out} = r_v * \sin(2 * \pi * w_v * t). \tag{5}$$

The output signal frequency, w_v, is also used to estimate the instantaneous HR. Algorithm 1 shows a possible implementation of the adaptive single-frequency phase vocoder in pseudo C code.

Algorithm 1. Adaptive Phase Vocoder

initialization
while TRUE **do**
  $val = input$
  $val \leftarrow highpass(val)$
  $y_{sin} \leftarrow lowpass(val * sin(freq_{-1} * t))$;
  $y_{cos} \leftarrow lowpass(val * cos(freq_{-1} * t))$;
  $phase \leftarrow unwrap(atan2(y_{sin}, y_{cos}), phase_{-1})$ ▷ unwrap the phase
  $ampl \leftarrow 4 * \sqrt{y_{sin} * y_{sin} + y_{cos} * y_{cos}}$
  $freq \leftarrow freq_{-1} - (phase_{-1} - phase)$ ▷ adjust frequency to lock onto the wave
  if $freq < 0.00001$ **then** ▷ clamp to avoid negative frequency values
    $freq \leftarrow 0.00001$
  end if
end while

To detect the location of the HR peaks, the monophone sinusoidal vocoder output, y_{out}, is used as an indicator of the approximate location of the peaks in the actual PPG waveform. The simple form of the vocoder waveform reduces the problem of detecting

Table 1. Algorithm configuration parameters

Parameter	Value
Heterodyne low-pass cut-off frequency	0.039 Hz
Initial vocoder frequency $w_v^{(0)}$	50 Hz
Initial vocoder amplitude $r_v^{(0)}$	1
Number of histogram bins	32

peaks in the noisy PPG signal to a straight-forward search for the maximum amplitude within the temporal interval of a positive half cycle of the vocoder output signal. The time elapsed from previously detected peaks is binned into a moving histogram of heart rate intervals. This serves to eliminate artifacts caused by motion and other spurious effects, with the largest histogram bin giving a direct representation of instantaneous HR. The sensitivity of this peak detection method can be adjusted by offsetting the sinusoidal vocoder signal; a positive or negative offset adjustment will result in a wider or narrower positive half cycle interval, respectively.

2.2 Algorithm Validation

In order to test the algorithm, data was gathered from 29 children (4.8 years $\pm$ 5.4, 18.5 kg $\pm$ 23.4) and 13 adults (46.3 years $\pm$ 9.0, 73.5 kg $\pm$ 24.2) who underwent general anesthesia, following institutional review board approval. The recordings obtained included ECG, capnometry, and PPG signals. All signals were recorded with Collect S/5 software (Datex-Ohmeda, Finland) using a sampling frequency of 300 Hz. An 8-min segment of reliable recording of spontaneous or controlled breathing was selected from each case. The mean HR for these segments ranged from 53 to 142 bpm. The segments are available for download from the on-line database CapnoBase.org [5].

The ECG waveform was used as the reference recording for computing HR. A technician independently validated the reference measurement using the CapnoBase Signal Evaluation Tool [5]. The mean ECG HR of a sliding window was calculated. The window size was set to 20 s; a size that corresponds to that commonly used in commercial pulse oximeters. In addition, the technician labeled the beginning and end of all artifacts in the PPG waveform.

The vocoder algorithms were written in C programming language and compiled to be embedded in a Matlab (Mathworks, Natick, USA) development framework. The adaptive vocoder parameters were tuned using the in-vivo dataset from Capnobase that was different from our test data set. The configuration parameters are shown in Table 1.

To compare the performance of the two vocoder algorithms, we implemented two other HR estimation methods; one that uses traditional filtering methods and another that uses a frequency domain-based approach. The first used the Pan-Tompkins algorithm [7]. This time domain-based method makes use of a cascade of band-pass filter, integrator, squaring, and differentiators. An adaptive threshold was applied to detect the heart beat pulses. Again, mean HR using a sliding window of 20 s duration was calculated. The second method calculated the power spectral density with a Fast Fourier Transform

Table 2. Normalized root mean square (NRMS) error between the reference HR, obtained from ECG, and the HR estimated by the vocoder frequency, vocoder peak detection, Pan-Tompkins, or FFT method, respectively

Case	**HR NRMS error** [%]				**mean HR** [bpm]
	vocoder peak	vocoder freq	Pan-Tompkins	FFT	ECG
0009l	1.97	3.22	19.99	1.39	102.46
0015l	1.28	0.55	0.07	0.45	120.07
0016l	1.87	2.19	0.94	1.59	126.33
0018l	1.62	0.79	2.14	0.28	141.15
0023l	2.58	2.28	0.99	0.84	102.17
0028l	2.49	2.33	0.98	1.68	73.27
0029l	3.90	2.48	0.62	2.49	68.53
0030l	2.44	6.39	3.57	5.18	117.00
0031l	3.21	4.00	4.89	4.02	67.76
0032l	2.58	2.25	5.39	2.72	85.63
0035l	1.59	2.87	0.88	1.55	112.62
0038l	2.37	8.31	1.88	1.96	120.39
0103l	1.32	0.86	0.06	0.55	103.36
0104l	2.90	1.94	0.21	0.45	113.61
0105l	1.32	1.14	3.10	1.19	66.27
0115l	2.65	5.67	6.69	3.34	96.01
0121l	2.23	2.07	0.26	2.29	72.37
0122l	2.13	3.29	4.17	3.55	73.03
0123l	2.31	3.92	1.19	1.30	90.16
0125l	1.82	1.14	5.12	1.04	78.27
0127l	2.34	2.39	4.08	1.79	77.11
0128l	2.26	3.13	0.34	3.34	67.99
0133l	1.63	1.87	2.61	0.87	71.32
0134l	1.92	1.98	2.29	1.49	72.38
0142l	1.69	1.89	9.61	1.44	92.03
0147l	1.33	7.62	3.07	13.67	66.00
0148l	1.52	1.32	0.16	1.55	77.92
0149l	1.42	1.63	4.97	1.60	57.13
0150l	1.64	3.34	1.72	1.16	63.55
0309l	1.23	2.00	2.17	1.30	63.92
0311l	1.36	4.34	0.29	3.06	68.97
0312l	2.91	3.97	5.00	3.38	53.91
0313l	0.92	1.40	0.08	0.88	73.53
0322l	0.83	1.99	0.15	1.11	73.72
0325l	0.86	2.69	0.76	1.11	73.20
0328l	1.14	5.69	74.38	1.72	78.61
0329l	2.99	12.76	45.11	5.81	91.08
0330l	0.68	0.77	0.09	0.60	75.06
0331l	5.11	12.53	1.67	9.25	66.68
0332l	1.22	4.90	0.14	1.54	73.86
0333l	8.73	24.09	8.51	9.61	73.84
0370l	1.85	3.78	2.53	2.46	72.77
Average	**2.15**	**3.90**	**5.54**	**2.54**	**83.69**

(FFT) algorithm [8] that was available through the C library [3]. A 50 s window size was selected, border effects were reduced using a Hanning windowing function, and the maximum power band larger than 0.4 Hz was chosen to estimate the HR. The FFT was recalculated every second by shifting the window by 75 samples.

The performance of the four algorithms was then assessed using the normalized root mean square (NRMS) error (%). The NRMS error corresponds to the square root of the sum of the squares of the differences between the test and reference HR measurements, divided by the sum of the reference measurements:

Table 3. Normalized root mean square (NRMS) error for the labeled artifact sections between the reference HR, obtained from ECG, and the HR estimated by the vocoder frequency vocoder peak detection, Pan-Tompkins, or FFT method, respectively

Case	**HR NRMS error** [%]				**mean HR** [bpm]
	vocoder peak	vocoder freq	Pan-Tompkins	FFT	ECG
0016l	2.35	1.97	1.85	1.08	128.71
0030l	0.34	5.76	0.64	1.58	107.15
0032l	3.15	1.00	4.71	3.29	85.69
0035l	3.20	5.61	2.39	3.21	118.43
0105l	1.29	1.00	13.34	0.48	65.69
0115l	2.22	3.32	7.37	2.49	97.56
0123l	3.52	3.84	4.92	0.63	89.64
0127l	4.40	0.66	5.71	0.63	75.19
0147l	2.62	2.23	13.48	0.78	64.82
0150l	1.78	5.68	6.06	0.90	61.41
0309l	2.88	2.55	2.63	1.52	62.78
0312l	5.10	3.54	8.78	4.01	54.35
0370l	2.62	0.90	7.68	1.15	76.16
Average	**2.73**	**2.93**	**6.12**	**1.67**	**83.66**

Table 4. Algorithm execution time per sample on the iPod Touch calculated over 16 measurements of one minute

	Vocoder freq	**Vocoder peak**	**Pan-Tompkins**	**FFT**
	[μs]	[μs]	[μs]	[μs]
Mean	41.25	47.33	28.84	1937
SD	0.9	0.8	0.45	4.06

$$NRMS\ error = \frac{\sqrt{\sum_{i=1}^{n}(x_i^{ref} - x_i^{alg})^2 * n}}{\sum_{i=1}^{n} x_i^{ref}}. \tag{6}$$

The NRMS error was calculated for each test measurement by comparing it with the reference measurement that was nearest to it in time. The first 50 s, which were used to initialize the high-pass filters and the vocoder sliding window, were not analyzed. The artifact labels from the technician were ignored by the algorithm in this assessment.

To evaluate the robustness of each algorithm with respect to artifacts in the PPG, the NRMS error was then calculated on the sections that were labeled as artifacts. Processing delays were taken into consideration by extending the sections by 5 s. The statistacl tests (Lilliefors,T-test, and Wilcoxon) were performed at the 5% significance level.

We built a prototype device, called the *Phone Oximeter* [4] to evaluate the computational load of the algorithm when executed on a mobile device (Figure 2). This consists



Fig. 2. The *Phone Oximeter* (mobile phone pulse oximeter) application. 1) Soft finger probe; 2) Nonin Xpod OEM pulse oximeter module; 3) iPod Touch with raw PPG waveform, SpO2 and HR trend display.

of a PureLight medium soft sensor connected to an Xpod OEM module (Nonin, Plymouth, USA), which was in turn connected to a 2^{nd} generation iPod Touch (Apple, Cupertino, USA) that features an ARM11 620 MHz microprocessor running underclocked at 533 MHz. The iPod Touch displayed the PPG waveform and recorded the continuous data stream. The PPG was recorded with a 16bit resolution at a sampling rate of 75 Hz. The algorithm was implemented in C and embedded into the iOS application software to process the raw PPG signal in real-time [4]. The computational load was measured in μs dynamic program analysis using an integrated low-overhead flat profiler.

3 Results

The two vocoder methods performed at least as well as the two traditional methods in terms of overall NRMS error (Table 2). The HR calculated from the vocoder peak detection using histograms (average NRMS error of 2.15%) was more accurate than the vocoder frequency estimations (average NRMS error of 3.9% , T-test, $p<0.01$). Both vocoder methods did not show a clear difference in their distribution from the Pan-Tompkins (average NRMS error of 2.15%) and the FFT (average NRMS error of 2.15%) method (T-test, $p>0.1$).

The vocoder methods were robust towards artefacts (Table 3). Thirteen of the 42 obtained recordings contained PPG artifacts. The total duration of artifacts in the dataset

was 113.83 s. The NRMS errors for the artifact experiment were not normally distributed (Lilliefors, $p>0.07$). The Pan-Tompkins algorithm was the algorithm most vulnerable to artifacts, with an average NRMS error of 6.12% (Wilcoxon, $p<0.031$). Both vocoder algorithms and the FFT method showed similar NRMS error distributions (Wilcoxon, $p>0.06$).

All the algorithms showed a processing speed that allows real-time HR computation. The adaptive vocoder algorithms including peak detection required an average of 47.33 μs to process a new value (Table 4). While this was slower than the Pan-Tompkins algorithm (28.84 μs), both were significantly faster than the inter-sampling distance of 13.3 ms. In contrast, the FFT HR estimation method was about 40 times slower than the vocoder algorithms. The algorithms can, therefore, be computed in real-time on the *Phone Oximeter* with sufficient processing power remaining for other computing tasks.

4 Discussion

The major design criteria of accuracy and on-line capabilities for a mobile phone HR estimation algorithm were met by the two newly developed vocoder algorithms.

4.1 Overall Accuracy

With an average accuracy of 2.15% NRMS, the HR prediction from vocoder peak detection is accurate for a mobile device. Any disagreement between it and the ECG reference HR arose from delays in the vocoder peak detection HR estimation; a 5 s to 10 s delay observed in some cases can be attributed to the size of the histogram for HR calculation. Only the cases 0331l and 0333l showed a decrease in performance greater than 4%. In these cases, the peak detection algorithm and the histogram method were not able to fully compensate for the poor performance of the vocoder frequency method which was required for the peak detection algorithm to accurately estimate HR. The vocoder peak detection was able to track fast variations and compensate for the low responsiveness of the direct vocoder frequency calculation.

The HR calculated directly from the vocoder output frequency was also accurate. However, six cases (0030l, 0038l, 0147l, 0329l, 0331l, and 0333l) exhibited higher NRMS errors than the others. Closer inspection illustrates that reduced performance of the vocoder frequency HR in these cases was mainly due to the low responsiveness of this algorithm to a rapidly changing HR of more than 10 bpm (Figure 3). The vocoder responsiveness could be improved by tuning the cut-off frequency of the low-pass filter shown in Figure 1-3. In the special case of 0333l, the vocoder locked into the double frequency of the original signal. This may have been due to a strong presence of a dicrotic notch that was amplified during the Butterworth filtering. A possible solution to circumvent this problem is to validate the vocoder frequency using the output of the peak detection algorithm.

A desired improvement to the vocoder peak HR estimation algorithm is to reduce the delay that was introduced by the histogram detection algorithm. It is, therefore, evident that the advantages and drawbacks of the two vocoder approaches are complementary. A logical step would be the combined use of these vocoder based HR extraction methods. A system that detects short-term variations with the peak detection and long-term trends with the vocoder frequency could provide increased diagnostic information.

The Pan-Tompkins algorithm showed three cases with very large NRMS errors (0009l, 0328l, and 0329l). These were the results of repeated double detection of peaks. To avoid these errors in the future, rejection of too closely located peaks will be necessary. For the FFT method, NRMS errors larger than 5% were caused by the detection of larger frequency content in the frequency band spanning over the double HR.

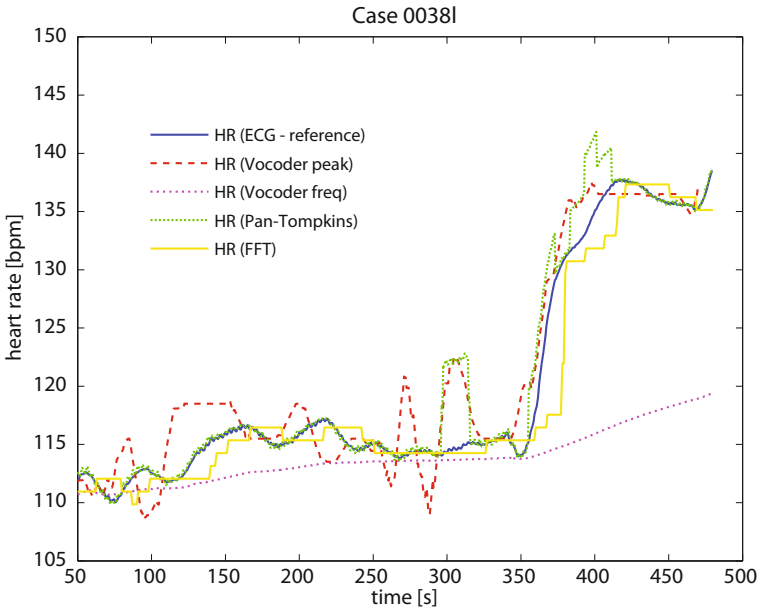


Fig. 3. HR agreement comparison for case 0038l. The HR estimated by the vocoder frequency does not respond rapidly to a large permanent change in HR after 350 s. The HR calculated by the other methods follows the ECG HR trend more closely.

4.2 Robustness to Artifacts

The adaptive vocoder frequency output filtered short variations in HR and PPG artifacts effectively. However, the algorithm was not able to detect large, rapid changes in HR and lock into the new frequency in a reasonable amount of time. On the other hand, the vocoder peak detection method was able to rapidly track changes in HR. Nevertheless, this attribute made it more vulnerable to errors when the PPG was corrupted with noise. Using a histogram to determine HR limited theses errors, but increased the detection delay. In the future, we plan to increase its robustness to poor quality PPG signals by adding an additional computation of a signal quality index to the algorithm. Active detection of poor signal will prevent the erroneous HR output and increase overall estimation performance.

The Pan-Tompkins algorithm was less robust to artifacts than the other methods. Its simple method of using a moving-average to filter the HR was not able to compensate for these artifacts, and propagated the error for the averaging window duration (Figures 4 and 5). The FFT algorithm was robust to shorter artifacts, but had a relative low time resolution, since it analyzed all frequencies inside a 50 s window. This was also

expressed by it having the largest delay of all methods, which can easily be seen in Figures 3 and 4.

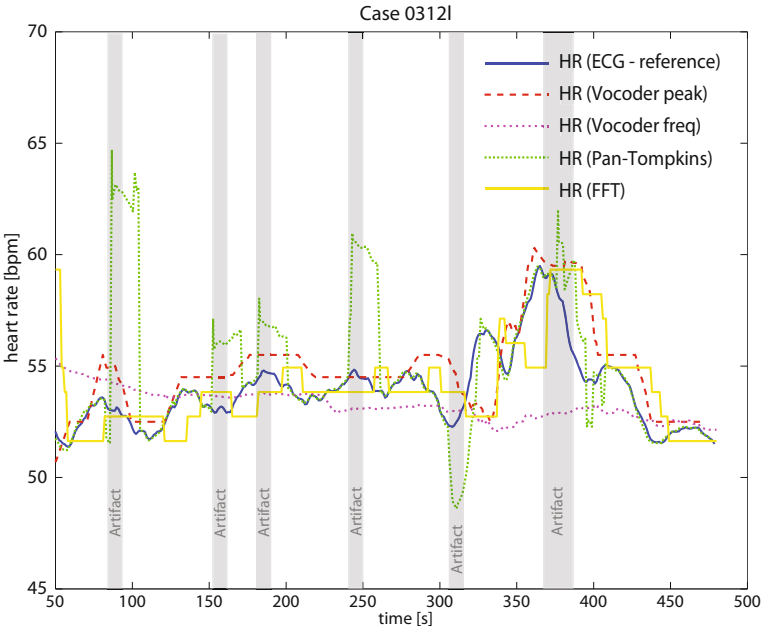


Fig. 4. HR agreement comparison for case 0312l. This case illustrates how the Pan-Tompkins algorithm is impacted by artifacts in the plethysmogram (grey bars) and the propagation of the error for the duration of the smoothing window.

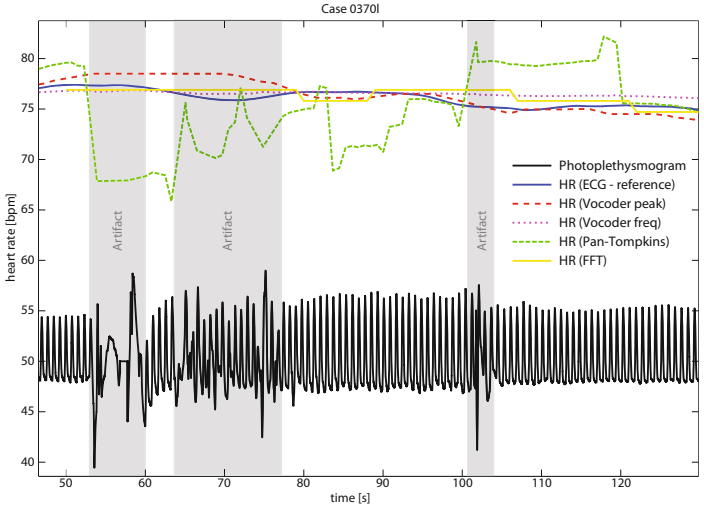


Fig. 5. HR agreement comparison for case 0370l. This case illustrates how the Pan-Tompkins algorithm is impacted by artifacts (grey bars) in the plethysmogram waveform.

4.3 Computational Costs

The speed tests on the *Phone Oximeter* showed that the proposed algorithms achieve the computation within the range of traditional filtering methods like the Pan-Tompkins algorithm and faster than the frequency domain-based approach of the FFT. Processing speed and theoverall accuracy achieved by the four tested algorithms suggested that they are all valid methods for HR estimation on mobile devices. However, the computing intensive FFT method is not optimized for low-power systems.

5 Conclusions

Two novel approaches to compute HR based on a dynamically adapted, single frequency phase vocoder algorithm were proposed. The computed HR was accurate with respect to the reference HR computed from ECG and compared favorably with other algorithms in terms of overall accuracy, robustness to artifacts and computational costs. This makes the algorithms suitable to process on-line PPG signals that are recorded from an oximeter sensor interfaced to a mobile device. We intend to use these algorithms as a core engine for more complex mobile signal analyzers (i.e. for the estimation of heart rate variability or respiratory rate) within a mobile phone based pulse oximeter called the *Phone Oximeter*. The suggested algorithms have the potential to be applied to other periodic signals whose frequency range needs to be determined in real-time.

References

1. Dolson, M.: The Phase Vocoder: A Tutorial. Computer Music Journal 10(4), 14–27 (1986)
2. Flanagan, J.L., Meinhart, D.I.S., Golden, R.M., Sondhi, M.M.: Phase Vocoder. The Journal of the Acoustical Society of America 38(5), 939–940 (1965)
3. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23), e215–e220 (2000), `http://www.physionet.org/physiotools/wfdb/psd/fft.c`
4. Karlen, W., Dumont, G., Petersen, C., Gow, J., Lim, J., Sleiman, J., Ansermino, J.M.: Human-centered Phone Oximeter Interface Design for the Operating Room. In: Traver, V., Fred, A., Filipe, J., Gamboa, H. (eds.) HEALTHINF 2011 - Proceedings of the International Conference on Health Informatics, pp. 4333–4337. SciTePress, Rome (2011)
5. Karlen, W., Turner, M., Cooke, E., Dumont, G., Ansermino, J.M.: CapnoBase: Signal database and tools to collect, share and annotate respiratory signals. In: Annual Meeting of the Society for Technology in Anesthesia (STA), p. 25. West Palm Beach (2010), `http://www.capnobase.org`
6. Kohler, B., Hennig, C., Orglmeister, R.: The principles of software QRS detection. IEEE Engineering in Medicine and Biology Magazine 21(1), 42–57 (2002)
7. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Transactions on Bio-Medical Engineering 32(3), 230–236 (1985)
8. Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical recipes in C: the art of scientific computing, 2nd edn. Cambridge University Press, New York (1992)

A Configurable Integrated Circuit for Biomedical Signal Acquisition

Jakob M. Tomasik[1], Wjatscheslaw Galjan[1], Kristian M. Hafkemeyer[2], Dietmar Schroeder[1], and Wolfgang H. Krautschneider[1]

[1] Institute of Nanoelectronics, Hamburg University of Technology
Eissendorfer Str. 38, 21073 Hamburg, Germany
[2] NXP Semiconductors Germany GmbH
Stresemannallee 101, 22529 Hamburg, Germany

Abstract. A 10 channel CMOS integrated circuit (IC) for biomedical signal acquisition is presented. Each channel of the IC includes a programmable analog front-end (AFE) and a 20 bit analog-to-digital converter (ADC). An active DC-suppression circuitry allows to tolerate DC-offsets of up to ± 1 V for a power supply voltage of 3.3 V. The AFE includes a common-mode rejection ratio (CMRR) calibration circuitry resulting in a CMRR of more than 80 dB. In low-noise mode the AFE achieves an input referred noise of less than 0.11 μV_{rms} for EEG application (0.5-70 Hz) and the power consumption of the IC is less than 30 mW in low-power mode. An experimental USB-Stick for biomedical signal acquisition has been realized using the IC.

Keywords: Biomedical signals, IC, Low-power, Low-noise, Configurable, CMRR-calibration, DC-suppression.

1 Introduction

The research in the area of integrated solutions for biomedical signal acquisition has increased substantially in the last few years. The new systems allow not only a reduction of costs for traditional medical devices, but also facilitate portable long-term applications or implantable solutions. In this context, integrated CMOS analog front-ends (AFE) [1,2] as well as complete systems including also analog-to-digital conversion (ADC) and digital IO interfacing [3,4,5] have been reported.

To cover the wide range of biomedical signals the system should be adaptable to their characteristics. The flexibility of a system means to provide an optimal setting of the overall system with respect to the signal type and system application. In this context, the AFE is the prime candidate for configurability. Two complementary application examples illustrate this: For portable long-term monitoring of electrocardiogram (ECG) signals the system's power consumption plays an essential role whereas noise constraints are rather relaxed [6,7]. In contrast to this, for extremely sensitive evoked potential (EP) recordings [8], low-noise amplification of the signal is crucial at the expense of rather high power consumption. Including these two extreme settings, a configurable biomedical signal acquisition system should be adaptable to a variety of signal types.

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 43–56, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Two additional requirements are inherently typical for biomedical signal acquisition systems: A high common-mode rejection ratio (CMRR) (in particular at 50/60 Hz) and large electrode DC-offset handling capability. The need for a high CMRR results from power line interferences inducing a common mode voltage present at the amplifier's input [9]. Large DC-offsets (i.e. hundreds of millivolts) may arise from the electrode-skin interface [9] and require counter measures to prevent saturation of the instrumentation amplifier (IA). Finally, a biomedical IC should provide a digital interface for simple configuration of the system and a high speed data output for read out of the acquired biomedical signals.

In this work we present a configurable IC for biomedical signal acquisition including CMRR-calibration and DC-offset suppression. Its applicability is demonstrated by a biomedical signal acquisition USB-stick. A first approach for the IC has been described in [10] and first simulation results of this system have been presented in [11].

2 System Overview

The block diagram of the IC is presented in Fig. 1. The main components of this IC are 10 identical channels and a digital block. Each channel includes an AFE, a $\Sigma\Delta$-modulator and a 12-bit digital-to-analog converter (DAC) for DC-offset suppression of the associated channel. The AFE's inputs are the electrode input (V_{EL}), the reference signal (V_{REF}), signal ground (V_{SGND}) and DC-suppression voltage (V_{DC}). Whereas V_{EL}, V_{REF} and V_{SGND} are familiar IA inputs, the V_{DC} input is used to suppress the input DC-offset voltage and is driven by the DAC output (V_{DAC}).

Subsequent to the AFE the $\Sigma\Delta$-modulator generates a 1 bit stream which is decimated to perform the analog-to-digital conversion. The digital block contains 10 decimation filters for this purpose. Additional functions of the digital block include the control of the DAC, timing and IC's sub-blocks. Finally, the digital block enables

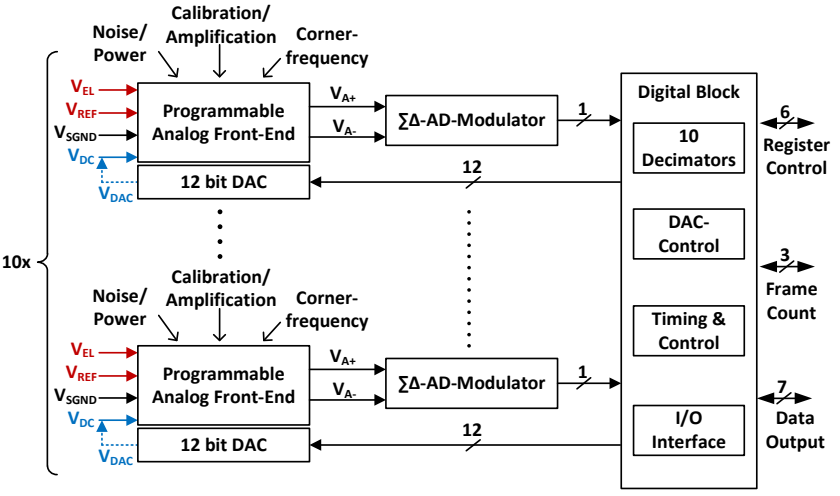


Fig. 1. Block diagram of the integrated circuit

communication to and from the chip. This comprises setting the internal registers (register control), reading out the measured data (data output) and synchronizing the measured data if several chips are used on the same bus (frame count).

3 Analog Front-End

A block diagram of the programmable analog front-end is given in Fig. 2. The IA amplifies the applied signal with a constant factor of 4. The signal is fed into the postamplifier directly or via an external high-pass filter. The postamplifier's gain is configurable with

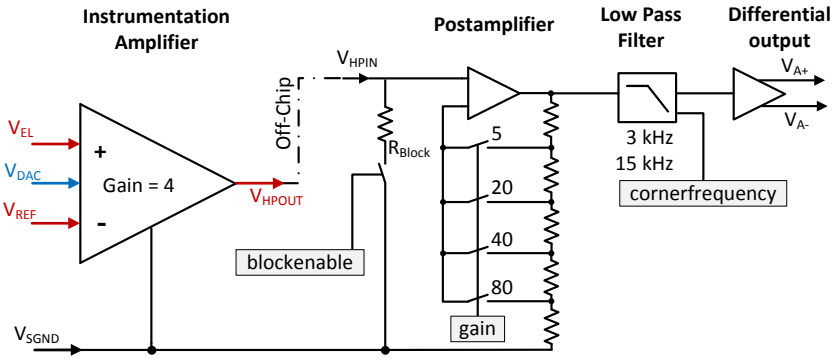


Fig. 2. Structure of the programmable analog front-end

amplification factors of 5, 20, 40 and 80 resulting in the desired channel gain factors of 20, 80, 160 and 320. Following the postamp a low-pass filter acts as an anti-aliasing filter for the ADC. The low-pass corner frequency can be set to be either 3 kHz (ECG, EEG) or 15 kHz (EMG, EP). Finally, the single-ended signal is converted to a differential one as required by the subsequent $\Sigma\Delta$-modulator.

3.1 Instrumentation Amplifier with CMRR-Calibration

The IA used as a preamplifier in the AFE is implemented by using a modified 2-opamp IA architecture and is presented in Fig. 3. The modification is composed of a third opamp for subtracting any DC-offset using the DAC generated voltage V_{DC}. The additional circuitry also incorporates a digitally controlled tunable resistor for CMRR-calibration, the resistor's step size is denoted as ΔR. For modeling parasitic PCB or bond wire resistances at the signal ground input V_{SGND} a resistor R_{PCB} has been added in Fig 3. The IA's transfer function with respect to V_{SGND} is given by

$$V_{\mathrm{HPOUT}} = V_{\mathrm{EL}}\left(1+\frac{R_2}{R_{1\mathrm{A}^*}}+\frac{R_2}{R_{1\mathrm{B}}}\right) - V_{\mathrm{REF}}\frac{R_2}{R_{1\mathrm{B}}}\left(1+\frac{R_3}{R_{4^*}}\right) - V_{\mathrm{DC}}\frac{R_2}{R_{1\mathrm{A}^*}} \quad (1)$$

with $R_{1\mathrm{A}^*} = R_{1\mathrm{A}} + n\Delta R$ and $R_{4^*} = R_4 + R_{\mathrm{PCB}}$.

A differential gain G between V_{EL} and V_{REF} is obtained from (1) by setting the values for R_1, R_2, R_3 and R_4 (neglecting R_{PCB} and assuming $R_1 = R_{1\mathrm{A}^*} = R_{1\mathrm{B}}$ and

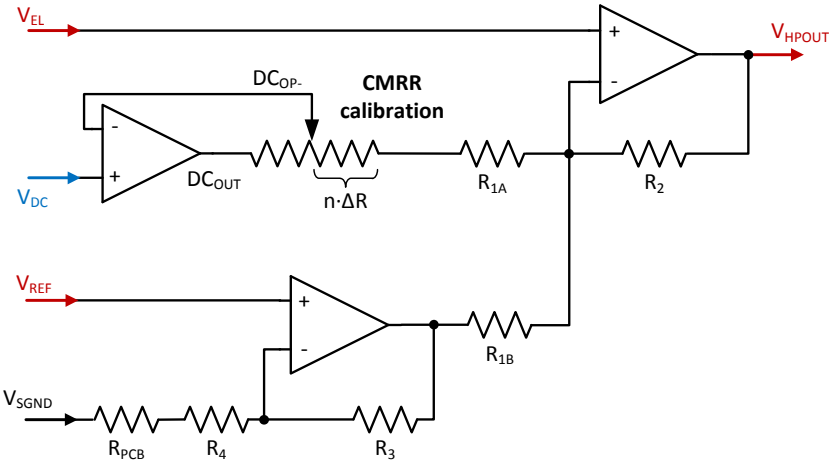


Fig. 3. Detailed view of the instrumentation amplifier

$V_{\mathrm{DC}} = 0$ V) such that the terms in front of V_{EL} and V_{REF} have both the same value G. In this case the gain G is equal to

$$G = \frac{V_{\mathrm{HPOUT}}}{V_{\mathrm{EL}} - V_{\mathrm{REF}}} = 2\frac{R_2}{R_1} + \frac{R_3}{R_4} - \frac{R_1}{R_2}. \tag{2}$$

In the presented circuit a gain value of $G = 4$ has been obtained using resistor values of $R_1 = 1\,\mathrm{k}\Omega$, $R_2 = 1.5\,\mathrm{k}\Omega$, $R_3 = 1\,\mathrm{k}\Omega$, $R_4 = 600\,\Omega$.

If any DC-offset is present at the input, the aforementioned voltage V_{DC} scaled by a factor $R_2/R_{1\mathrm{A}}$ is subtracted from the IA's output voltage V_{HPOUT}. The result of this action is a cancellation of DC-offsets up to a value of

$$V_{\mathrm{EL,dcmax}} = \frac{V_{\mathrm{DD}}}{2G}\frac{R_2}{R_1}. \tag{3}$$

Using above values for R_1, R_2 and G and a power supply voltage V_{DD} of 3.3 V, a maximal DC-offset of $V_{\mathrm{EL,dcmax}} = \pm 619$ mV is obtained. In conjunction with the external high pass filter and a postamplifer gain of 80 an offset compensation range of ± 1000 mV is reached.

The CMRR calibration compensates for any mismatch in the resistors R_1 to R_4 and wiring parasitics R_{PCB}. The calibration circuit has been located in the high-resistive feedback loop of the DC-offset compensation opamp to minimize non-idealities. By changing the tap-point in the feedback path of the DC-offset compensation opamp a value of $n\Delta R$ is added to the resistor $R_{1\mathrm{A}}$. The CMRR of the IA becomes

$$\mathrm{CMRR} = \frac{1 + \frac{R_2}{R_{1\mathrm{A}} \pm n\Delta R} + \frac{R_2}{R_{1\mathrm{B}}}}{1 + \frac{R_2}{R_{1\mathrm{A}} \pm n\Delta R} - \frac{R_2 R_3}{R_{1\mathrm{B}} R_4}} \tag{4}$$

It should be taken into account that a variation of $R_{1\mathrm{A}}$ also affects the differential gain G. Therefore, resistor values should be chosen such that the gain error between channels does not exceed 1%. A CMRR value of more than 80 dB (including R_{PCB}) is achieved

by using a step size $\Delta R < 0.26\,\Omega$ and $n = 64$ steps. Using a driven right leg system [12] for patient grounding the overall CMRR is increased by approx. 40 dB resulting in a system CMRR of 120 dB.

The configurability of the IA regarding noise and power consumption is realized using programmable operational amplifiers as described in [13,14]. These opamps use chopper modulation to reduce $1/f$-noise and can be programmed in a wide range between ultra low-noise (2 nV/$\sqrt{\text{Hz}}$) and low-power consumption (140 μW).

3.2 Postamplifier and External Highpass Filter

The postamplifier is composed of the same programmable opamp as used in the IA. It is used in a non-inverting configuration having four gain settings (5, 20, 40 and 80). These are realized by a resistor chain having four taps. Connecting one of the taps to the opamp's inverting input using a CMOS switch sets the appropriate gain factor. Care has to be taken regarding the accuracy of the resistors to fulfill the specified gain error between channels ($< 1\%$). Chopper modulation of the opamp is also needed in the postamplifier due to the low preamplifiers's gain factor of 4 for low-noise applications.

The optional external passive high pass filter can be placed between the V_{HPOUT} output pin of the IA and the postamplifier's input pin V_{HPIN} (dashed line in Fig. 2). Using external components for the high pass filter (i.e. capacitor and resistor) eases the implementation of the large time constants needed.

A drawback of the large time constant arises if events like external stimulation or movements of the electrodes induce a noticeable shift in the DC-potential of the external capacitor: the time for restoring the system back to equilibrium is determined by the time constant of the high pass filter, i.e. in range of seconds. To overcome this problem an additional switch has been placed to discharge the capacitor over the relative small resistor R_{Block}. The signal to close the switch ('block enable' in Fig. 2) can be set via the serial configuration port.

3.3 Low-Pass Filter and Symmetry Stage

The $\Sigma\Delta$-analog-to-digital conversion requires an anti-aliasing filter preceding the $\Sigma\Delta$-modulator. In contrast to Nyquist-rate ADCs the anti-aliasing filter's cut-off frequencies are rather easy to implement with oversampling ADCs, where the relative low bandwidths of the biomedical signals would lead to large time constants and hence require external components. The anti-aliasing filter is implemented as a 2nd-order low-pass filter using an opamp in a Sallen&Key configuration [15]. To choose the cut-off frequency, the maximum signal frequency, the oversampling ratio and the needed signal attenuation have to be considered. Choosing an attenuation of more than 40 dB at half of the oversampling frequency and an oversampling ratio of 256, cut-off frequencies of 3 kHz for ECG/EEG and 15 kHz for EMG/EP ensure proper anti-aliasing functionality. The cut-off frequency can be configured externally via the serial configuration port.

Following the low-pass filter a symmetry stage converts the single-ended signal ($V_{\text{A}+}$) into a differential one using an opamp to obtain the inverted signal ($V_{\text{A}-}$). These signals are fed into the succeeding $\Sigma\Delta$-modulator. Additional care has to be taken to keep these signals stable at the modulators input because of its switching operation [16].

Therefore, an integrated low-pass RC-filter with $R = 5\ \mathrm{k}\Omega$ and $C = 2.5$ pF has been inserted between the symmetry stage outputs and the $\Sigma\Delta$-modulator inputs.

4 ADC

Analog-to-digital conversion for each channel is accomplished by a 2nd-order $\Sigma\Delta$-modulator and a decimation filter. The default oversampling ratio (OSR) of the $\Sigma\Delta$-modulator is 256, the OSR can be changed using the serial configuration port to a value of OSR = 256 $\pm$ 128 (in steps of 1). The sampling rate at the modulator's input is derived from the 12.8 MHz main clock using on-chip division by (n + 1). Like the OSR, n can be set by programming and has a value range from n = 0 to n = 255. For example, using an oversampling ratio of 256 and setting n to 0 or 249 data rates (DR) of 50 kHz and 200 Hz can be achieved, respectively.

The 2nd-order $\Sigma\Delta$-modulator is implemented as a discrete time 1-bit modulator. The circuit topology of the switched-capacitor $\Sigma\Delta$-modulator is derived from [17] and a modified version also used in the present topology has already been used in [4]. The sizing of the actual capacitors used for the switched capacitor integrators results in signal-transfer and noise-transfer functions (STF and NTF) [18]:

$$\mathrm{STF}(z) = \frac{1}{8z^2 - 14z + 7} \tag{5}$$

and

$$\mathrm{NTF}(z) = \frac{8z^2 - 16z + 8}{8z^2 - 14z + 7} \tag{6}$$

The required effective number of bits (ENOB) of the $\Sigma\Delta$-modulator is depending on the signal type: For EEG approx. 16 bits or a dynamic range (DR) of 98 dB are required whereas for ECG 12 bits or a dynamic range of 74 dB is sufficient.

Decimation and low-pass filtering of the modulator's output signal is accomplished using a comb filter having a $sinc^3(f)$ frequency response. The implemented filter uses decomposition and modulo arithmetic to minimize the system size [19].

5 Digital System Control and Interface

The communication to external components like a DSP or FPGA is realized by two independent serial ports. Therefore, the data and control paths are strongly separated. Fig. 4 shows an overview of the signal and control flows. Serial configuration port receives the control data and applies it to the sub-blocks even for the high speed dataport, setting up the data rate and the format of the transmission.

5.1 DC-Suppression Feedback Circuit

Offset suppression of each electrode is realized by an additional input of the AFE. If this feature is not needed (e.g. for ECG application) this input can be connected to the virtual ground node. For applications with very small amplitudes (e.g. EEG) the

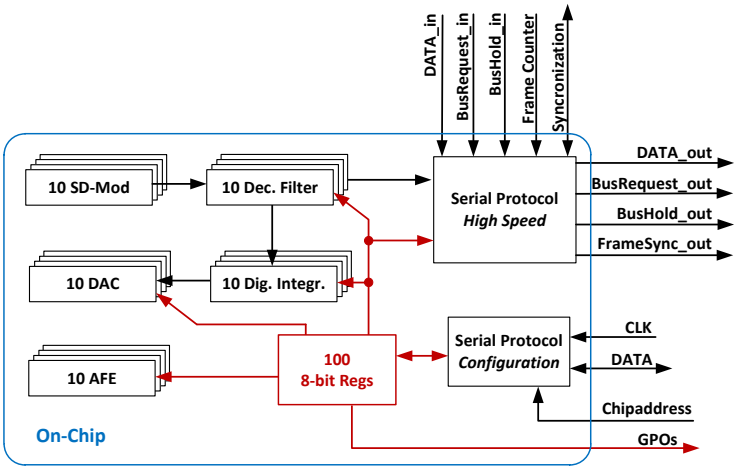


Fig. 4. Digital system control and interface

electrode offset should be removed allowing the required high amplification factor of the data path. In this case, as shown in (1), the on-chip DACs are used to generate the subtracting voltage.

The input of the DACs can be switched between a value which is written into the configuration register by the serial control port or the output of the on-chip digital integrator. This integrator sums up the bit-stream generated by the modulator and is configurable by the serial configuration port for different time constants. To avoid noise of the high resistive output node (100 kΩ) of the connected DAC and to smooth the non-linearities caused by value changing of the DAC, a capacitor in the range between 1 μF and 10 μF should be added between the V_{DAC} output and signal ground.

For the implementation of the DAC with a resolution of 12 bits a R-2R architecture [15] was chosen. The DAC occupies an area of 0.5 mm^2 and consumes only 33 μA.

5.2 Serial Configuration Port

The serial communication for setting up the system is realized by a synchronous bidirectional interface (see Fig. 4). The access to one of the hundred 8-bit registers is decoupled from other on-chip components and can be done asynchronously to the conversion clock and to the clock of the high speed serial dataport. Depending on the external interconnects the maximum bus speed can be used up to 100 kHz. This allows continuously to set up the DACs for offset cancellation and the external 16 General Purpose Outputs (GPOs) for controlling the possible external components during operation.

5.3 High Speed Serial Dataport

The data from the IC are read out by the high speed serial dataport which is synchronous to an applied clock signal as also shown in Fig. 4. The receiver is synchronized to a new data packet by the framesync signal. The rising edge of this signal determines the

new transmission, therefore, after a delay of one bit, the second bit is taken as valid data. Busrequest and bushold signals implement the handshake protocol allowing daisy chaining of 16 ICs, this number is determined by the 4-bit chip address.

The serial data from one IC includes a 22 x 16-bit wide packet which is sent for each sampled value of 10 channels. For a maximum data rate of 50k packets per second a minimum clock transmission frequency of 352 x 50 kHz = 17.6 MHz is needed. Hence, using the IC in daisy chained application the maximum operating frequency of this high speed port is specified to 53 MHz to allow the data transmission of 3 ICs.

Table 1 shows the implemented protocol for one data packet. The 1st word contains the start marker FF9h and the chip address which is set by external digital 4-bit port. The 2nd transmitted word is the frame counter value. The 3rd word consists of the upper 16-bit data of the first channel. The 4th data word (bit 15..12) consists of the 4 lower data bits of the first channel (channel <0>) and the actual DAC value for the same channel. The following 18 words are in the same scheme as for the first channel (3rd and 4th words).

Table 1. Implemented protocol for the high speed data-port

16 bit words	MSB		LSB	
	nibble		nibble	
	high	low	high	low
Word1	Fh	Fh	9h	chip address
Word2	framecounter			
Word3	upper 16 bits of channel <0> data			
Word4	lower 4 bits of data	DAC value for channel <0>		
Word5	upper 16 bits of channel <1> data			
Word6	lower 4 bits of data	DAC value for channel <1>		

6 Results

A prototype of the IC has been realized in a 0.35 μm CMOS process with a supply voltage of 3.3 V. To prevent coupling effects between digital and analog blocks, care has been taken to separate these parts in the layout. The power supply is likewise separated; additional on-chip stabilizing-capacitors are used for blocking off switching activities on the supply rails. These capacitors, which have been realized as poly-capacitors and Metal-Insulator-Metal (MIM) capacitors, also allow to fulfill the needed layer density specified by the foundry. Additional efforts have been taken to minimize resistive parasitics of the interconnects, this applies in particular to the supply rails and the connections to signal ground (V_{SGND}).

A special approach was taken for the chopper clock lines: To minimize coupling effects to the analog parts due to high frequency switching operation each chopper clock path has been placed between two paths tied to ground.

A microphotograph of the fabricated IC prototype is shown in Fig. 5a. The IC occupies an area of 47 mm^2 and the number of pins is 144. Using modern packaging technology the IC is encapsulated in a lead free Ball Grid Array (BGA) having a size of 13 mm x 13 mm. Fig. 5b shows a size comparison of the packaged IC to a eurocent coin.

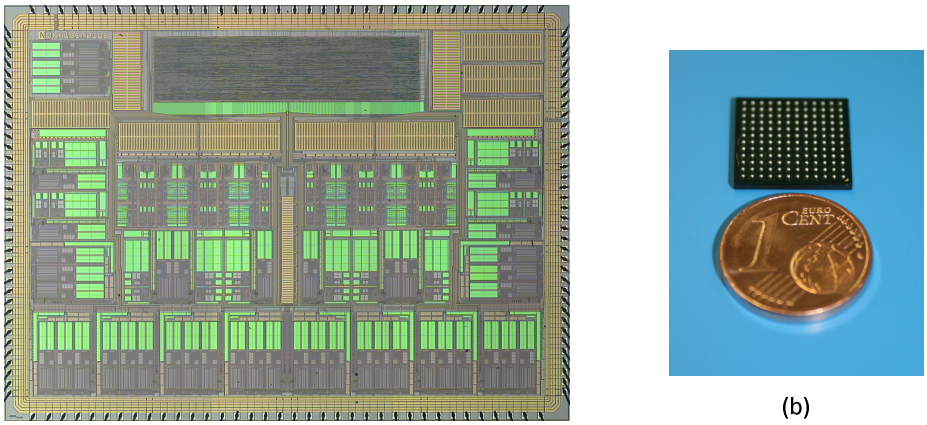


Fig. 5. (a) Microphotograph of the integrated circuit, (b) Photo of the integrated circuit in a BGA package

6.1 Analog Channel and CMRR Calibration

The measured overall gains of the analog channels (20, 80, 160 and 320) have a gain mismatch between the channels of not more than 1% . The input referred offset of the channels (with enabled chopper modulation) is less than 34 μV and 12 μV for the low-power and low-noise mode, respectively.

The DC-suppression circuit has been tested by applying a 0.1 Hz, 100 mVpp square-wave to the input of two IA channels. The outputs of both channels (one has the DC-suppression activated and the other deactivated) are shown in Fig. 6. The IA saturates after the rising or falling edges of the square wave. Due to regulation of the activated DC-suppression, the offset voltage is subtracted from the channel resulting in an offset free output.

The CMRR calibration of all 10 channels is shown in Fig. 7. The sweep of the implemented 64 calibration steps results in a channel CMRR of more than the specified 80 dB at 50 Hz for at least one calibration value.

The total input referred noise has been measured including the full channel, i.e. the AFE, $\Sigma\Delta$-modulator and decimation filter. Table 2 summarizes the noise performance of the IC. The power consumption of the IC varies between 28.7 mW in low-power mode and 209.8 mW in low-noise mode for data rates of 500 Hz and 50 kHz, respectively.

6.2 ADC and Digital IO

The ADC has been measured by applying a signal to the input of the channel. The measured peak SNDR for this setup is 75 dB which corresponds to an ENOB of 12.17 bits. An ENOB of more than 15.2 bits has been measured for the EEG bandwidth (0.05 Hz - 70 Hz).

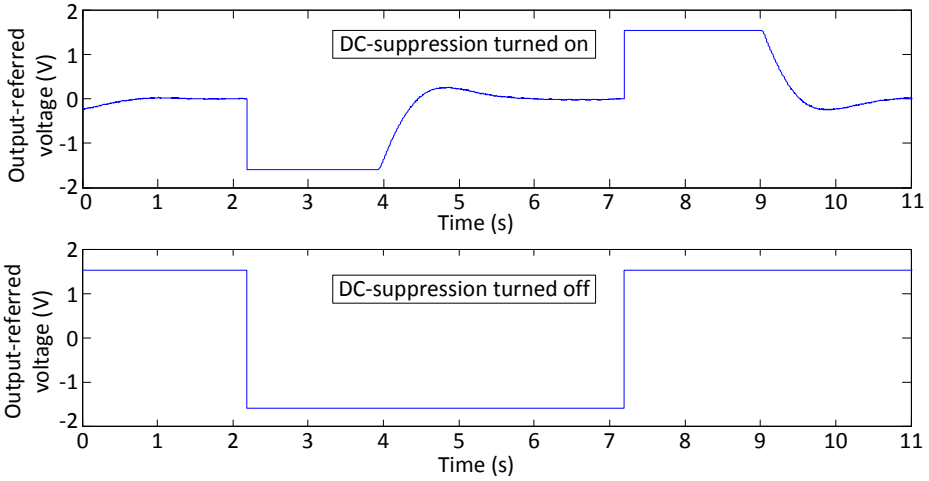


Fig. 6. IA output with DC-suppression turned on or off

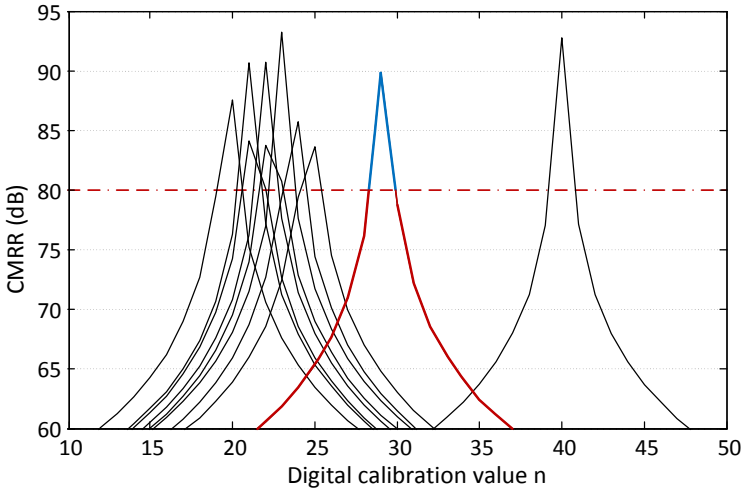


Fig. 7. CMRR at 50 Hz calibration of 10 channels

Table 2. Measured total input referred noise at ADC output

Integration bandwidth	Conditions/Remarks	Value μV_{rms}
0.05 Hz - 250 Hz (ECG)	Low-power, chopper enabled	0.19 - 0.29
	Low-power, chopper disabled	0.52 - 0.6
0.05 Hz - 70 Hz (EEG)	Low-noise, chopper enabled	< 0.11
0.01 Hz - 5 kHz (EMG)	Low-noise, chopper enabled	< 0.65
0.1 Hz - 3 kHz (EP)	Low-noise, chopper enabled	< 0.53

The digital communication ports are tested using an automated test setup [20]. The configuration port is tested by writing to all 100 registers with subsequent read-out of the written content and shows no missing data for a communication frequency of up to 100 kHz. The high speed serial bus was used for reading out of all channel measurements presented in this section and is operating properly for frequencies up to 55 MHz.

7 Experimental USB-Stick Application

The implemented integrated circuit is not only suitable for multi-electrode applications, where a small system size is mandatory, but also for mobile long-term biomedical signal acquisition. In this context, an experimental USB-stick application using the presented IC has been developed. A block diagram of the system is shown in Fig. 8.

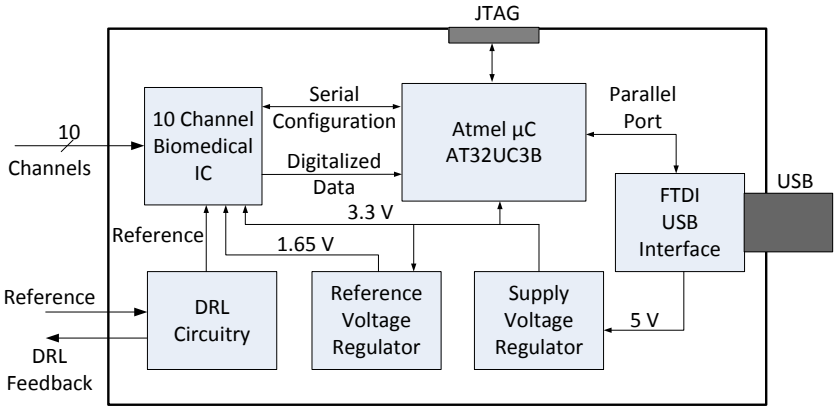


Fig. 8. Block diagram of the USB-stick application

The IC configuration input and data output is connected to an Atmel 32 bit microcontroller [21] which is used to comfortably control the system. The microcontroller itself is connected to a FTDI USB-interface chip [22] for communication with an PC. Additionally, the system includes supply and reference voltage regulators and support circuits to implement a driven-right-leg circuitry [12]. A small sized connector on the board is used for attaching electrode cables to the system. The USB-stick PCB has an area of 5 cm x 2.5 cm and is powered using the 5 V USB supply. Fig. 9 shows a photograph of the system. To facilitate the signal recordings on a PC or notebook using this stick, a graphical user interface for configuration and data acquisition has been developed.

The USB-stick systems allows to easily acquire high quality biomedical signals, it has been successfully employed in experimental EMG, ECG and EEG recording, see Fig. 10. Mobile signal acquisitions could be performed using a lightweight notebook.

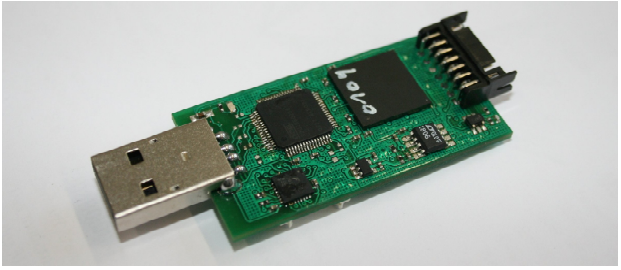


Fig. 9. Photo of the USB-Stick

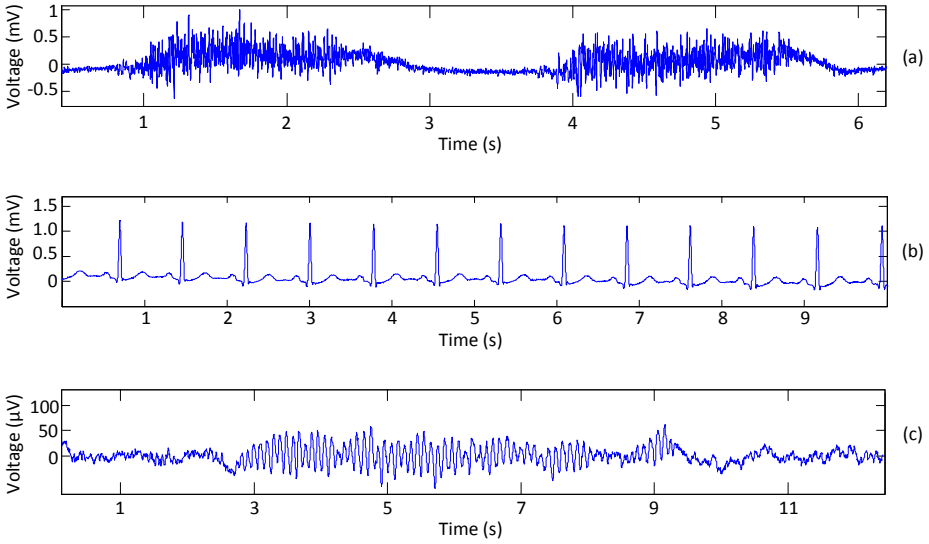


Fig. 10. Biomedical signals acquired using the USB-Stick: (a) EMG taken from the upper arm, (b) ECG measured between the lower arms, (c) EEG acquired from the occipital region

8 Summary and Conclusions

An integrated ten channel biomedical signal acquisition integrated circuit has been presented. Its main characteristics are summerized in Table 3. Each channel of the IC includes an analog front-end that is programmable with respect to noise and power and an $\Sigma\Delta$-analog-to-digital converter. In addition, the system includes both a CMRR calibration and a DC-suppression circuitry. The latter extends the input-referred DC-suppression range to a remarkable value of up to ± 1 V which largely exceeds the input offset range of other DC-coupled integrated biomedical amplifiers. The ICs digital interface is realized by means of a serial configuration port and a high-speed data output port. An experimental USB-Stick demonstrates the application of the presented IC.

Table 3. Main characteristics of the biomedical IC

Parameter	Conditions/Remarks	Value
Supply voltage		3.3 V
Current consumption	Low-power/Low-noise	8.7 mA/63.6 mA
Total input referred noise	EEG/ECG-bandwidth	$< 0.11/0.29\ \mu V_{\mathrm{rms}}$
CMRR	50/60 Hz	> 80 dB
Input common mode range	$V_{\mathrm{EL}} = V_{\mathrm{REF}}$	$\pm$ 1 V
Offset compensation range	HP & DC-suppr.	$\pm$ 1 V
Crosstalk		$<$ -60 dB
Die area		47 mm^2
Pin number		144

Acknowledgements. The authors would wish to acknowledge Natus (formerly Schwarzer) Corporation, Munich, for their support and very fruitful discussions. The authors would also like to thank Sven Vogel for his contributions to the design and Fabian Wagner for support with the measurements.

References

1. Martins, R., Selberherr, S., Vaz, F.A.: A CMOS IC for portable EEG acquisition systems. IEEE Trans. Instrum. Meas. 47(5), 1191–1196 (1998)
2. Ng, K.A., Chan, P.K.: A CMOS analog front-end IC for portable EEG/ECG monitoring. IEEE Trans. Circuits Syst. I, Reg. Papers 52(11), 2335–2346 (2005)
3. Desel, T., Reichel, T., Rudischhauser, S., Hauer, H.: A CMOS nine channel ECG measurement IC. In: 2nd International Conference ASIC (1996)
4. Fuchs, B., Vogel, S., Schroeder, D.: Universal application-specific integrated circuit for bioelectric data acquisition. Medical Engineering and Physics 24, 695–701 (2002)
5. Yazicioglu, R.F., Merken, P., Puers, R., Van Hoof, C.: A 200 μW Eight-Channel EEG Acquisition ASIC for Ambulatory EEG Systems. IEEE J. Solid-State Circuits 43(12), 3025–3038 (2008)
6. Martin, T., Jovanov, E., Raskovic, D.: Issues in wearable computing for medical monitoring applications: a case study of a wearable ECG monitoring device. In: The Fourth International Symposium on Wearable Computers, pp. 43–49 (2000)
7. Galjan, W., Naydenova, D., Tomasik, J.M., Schroeder, D., Krautschneider, W.H.: A portable SoC-based ECG-system for 24h x 7d operating time. In: IEEE Biocas 2008, Baltimore, pp. 85–88 (2008)
8. Scheer, H.J., Sander, T., Trahms, L.: The influence of amplifier, interface and biological noise on signal quality in high-resolution EEG recordings. Physiol. Meas. 27, 109–117 (2006)
9. Webster, J.G.: Medical instrumentation: application and design, 3rd edn. Wiley & Sons, New York (1998)
10. Van Helleputte, N., Tomasik, J.M., Galjan, W., Mora-Sanchez, A., Schroeder, D., Krautschneider, W.H., Puers, R.: A flexible system-on-chip (SoC) for biomedical signal acquisition and processing. Sens. Actuators A: Phys. 142(1), 361–368 (2008)
11. Hafkemeyer, K.M., Galjan, W., Tomasik, J.M., Schroeder, D., Krautschneider, W.H.: System-on-Chip Approach for Biomedical Signal Acquisition. In: 18th ProRISC Workshop, Veldhoven, pp. 26–29 (2007)

12. Winter, B.B., Webster, J.G.: Driven-right-leg circuit design. IEEE Trans. Biomed. Eng. 30, 62–66 (1983)
13. Meier auf der Heide, P., Bronskowski, C., Tomasik, J.M., Schroeder, D.: A CMOS operational amplifier with constant 68 phase margin over its whole range of noise-power trade-off programmability. In: Proceedings 33nd ESSCIRC 2007, Munich, pp. 452–455 (2007)
14. Tomasik, J.M., Hafkemeyer, K.M., Galjan, W., Schroeder, D., Krautschneider, W.H.: A 130nm CMOS Programmable Operational Amplifier. In: Proceedings NORCHIP 2008, Tallinn (2008)
15. Allen, P.E., Holberg, D.R.: CMOS Analog Circuit Design. Oxford University Press, New York (2002)
16. Maxim Integrated Products: Choosing the optimum buffer/ADC combination for your application. Application Note 1094 (2000)
17. Medeiro, F., Perez-Verdu, B., de la Rosa, J.M., Rodriguez-Vazquez, A.: Using CAD Tools for Shortening the Design Cycle of High-Performance $\Sigma\Delta$M: A 16.4bit 9.6kHz 1.71mW $\Sigma\Delta$M in CMOS 0.7μm Technology. International Journal of Circuit Theory and Applications 25, 319–334 (1997)
18. Fuchs, B.: Integrierte Sensorschaltungen zur EKG-und EEG-Ableitung mit praediktiver Signalverarbeitung. PhD thesis, Institute of Nanoelectronics, Hamburg University of Technology. Shaker Verlag, Aachen (2004)
19. Dijkstra, E., Nys, O., Piguet, C., Degrauwe, M.: On the use of modulo arithmetic comb filters in sigma delta modulators. In: IEEE Proc. ICASSP 1988, pp. 2001–2004 (1988)
20. Wagner, F., Jakobi, C., Tomasik, J.M., Hafkemeyer, K.M., Galjan, W., Schroeder, D., Krautschneider, W.H.: Design and Implementation of an Automated Test Environment for Signal-Acquisition ASICs. In: Semiconductor Conference Dresden (SCD), Dresden (2008)
21. Atmel Corporation, San Jose, USA, `http://www.atmel.com`
22. Future Technology Devices International Limited, Glasgow, UK, `http://www.ftdichip.com`

Movement Disorder Assessment and Attenuation Techniques for Removal of Tremor

Wesley Teskey, Mohamed Elhabiby, and Naser El-Sheimy

Department of Geomatics Engineering, University of Calgary
2500 University Dr. NW, Calgary, T2N 1N4, Alberta, Canada
{wjeteske,mmelhabi,elsheimy}@ucalgary.ca

Abstract. A weighted-frequency Fourier linear combiner (WFLC) filter is used for removal of tremor motion for data captured from movement disorders subjects with essential tremor (ET) and Parkinson's disease (PD). This technique is applied here in six degrees-of-freedom and data are filtered so that a comparison can be made before and after filtering to assess the extent to which the WFLC filter removed tremor. A wavelet spectral analysis is employed to determine the effectiveness of the filter in removing tremor in the 3-12 Hz band of interest. A Kalman filter is employed to improve data processing so that six degree-of-freedom tremor motion can be accurately rendered for subsequent filtering; such accurate rendering is needed so that full tremor motion can be adequately described. A Fourier coherence based technique is utilized so that relationships for interrelated tremors for the different six degrees-of-freedom can be identified. Much of the analysis shown is novel.

Keywords: Weighted-frequency fourier linear combiner (WFLC), Wavelets, Accelerometers, Gyroscopes, Essential tremor, Parkinson's disease, Kalman filter.

1 Introduction

1.1 Movement Disorder Assessment

In recent years there has been much focus on movement disorder tremor assessment using inertial sensors (accelerometers and gyroscopes [1] and [2]). Such assessment of tremor related disorders (focussing largely on size of tremor, frequency, axis of motion etc.) can help to create a standardized approach to assist medical professionals when diagnosing tremor; this approach can help to better understand the nature of the disorder under evaluation. As well, assessment of tremor can be used to evaluate the effectiveness of medication.

1.2 Movement Disorder Attenuation

Attenuation of tremor is another major focus of recent research conducted. It can take the form of an orthesis designed to actively remove tremor, or passive and active

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 57–73, 2013.
© Springer-Verlag Berlin Heidelberg 2013

feedback systems to dampen and mitigate tremor (such as a pen with a feedback system such that the tip moves so as to counteract tremor motion). Such attenuation can be quite useful because 90% of tremor patients report a disability[3].

1.3 Types of Disorders Evaluated

Tremors types evaluated for this research paper include essential tremor (ET) and Parkinson's disease (PD); these are among the most common types of tremor disorders (although, sufferers of these disorders can also exhibit other non-tremor related symptoms) [1]. ET is the more prevalent of these two, affecting 4% of people over age 65 [4]; while PD affects 1.5-2.5% of people older than 70 in the United States [5]. Tremor is generally regarded to have a frequency of 3-12 Hz [6].

2 Importance of Kalman Filter

When utilizing accelerometers for tremor assessment and attenuation, it is critical to consider that raw accelerometer data contain both lateral and rotational tremor components [1]. The former tremor type is measured directly by accelerometers and the latter is caused by rotation through the gravity field. Indeed, rotating an accelerometer through a gravity field (i.e. about a vector perpendicular to gravity) at a constant rate will cause the accelerometer signal to follow a generally sinusoidal trajectory with peaks at positive gravitational acceleration (generally approximately 9.81 m/s^2 depending on the local gravity field) and troughs at negative gravitational acceleration; this is illustrated in Figure 1. It follows that a rotational tremor (with no lateral motion component) will also register an accelerometer signal depicting tremor motion components. Such a rotational tremor is removed from accelerometer data in this research paper so that the lateral acceleration components can be evaluated independently of rotation. This is shown in Figure 2, an overall flow chart of the data processing. From this figure, it can be seen that processed accelerometer data (bottom half of the figure on the right hand side) is used to evaluate lateral tremor and processed gyroscope data (bottom half of the figure on the left hand side) is used to evaluate rotational tremor.

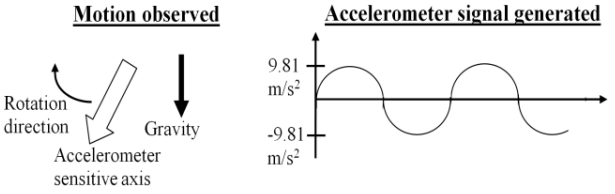


Fig. 1. An accelerometer rotated through the gravity field and the signal generated

To decipher lateral and rotational tremor motion (for six degree-of-freedom motion resolution), a Kalman smoother is employed. Strictly speaking, raw gyroscope data should be sufficient to remove rotational tremor components from accelerometer data by providing orientation information; however, due to signal noise, an accurate

solution generally involves extra information and data fusion to accurately obtain orientation data. For the Kalman smoother employed in this research paper, such data fusion uses the known start and end orientation of inertial sensors. As well, updates are performed using accelerometer data during relatively still motion signal portions to estimate orientation about the two horizontal axes (i.e. an orientation reading is performed using an accelerometer *gravity measurement* when the inertial sensors are stationary).

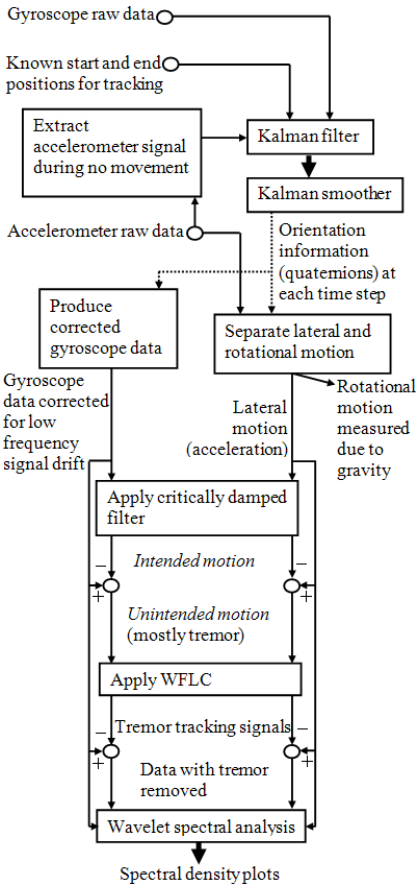


Fig. 2. Overall flow chart for data processing

After Kalman smoothing, the weighted-frequency Fourier linear combiner (WFLC) algorithm is used for further analysis. The application of WFLC for the evaluation of all six degrees-of-freedom of tremor motion is novel and is introduced in this research paper likely for one of the first times. The WFLC algorithm is generally regarded as most useful for both assessment and attenuation of tremor [1].

There are many advantages to using the WFLC technique for six degree-of-freedom tremor analysis. One is that it can allow for medical professionals studying tremor to determine the axis of motion for which tremor is most prevalent for

different patients. It also allows for one to determine the degree of correlation of different lateral and rotational tremors and their phase shift with regard to one another. Another feature important for tremor assessment, the overall signal shape (i.e. sinusoidal, zigzag etc.) for each axis of motion, can also be studied.

Six degree-of-freedom motion information is required to properly attenuate tremor for many applications; this is often the case when an external mechanism is used for attenuation (such as implied by the test setup utilized for the research conducted and presented here, where patients were told to simulate eating using a spoon). Such a motion was chosen for analysis because many tremor patients complained about to the fact that they often spilled their soup whilst trying to eat. Studying such a movement is also useful from the perspective of assessment in that significant tremor data is present during evaluation. Future applications stemming from the research carried out could see actuators between the spoon head and handle to mitigate tremor motion.

3 Mathematical Methods

3.1 Kalman Filter and Smoother

Kalman Filter. Raw data is first processed to determine sensor orientation for all six inertial sensors used in data collection (three accelerometer and three gyroscopes mounted on a rigid body). The state vector for the Kalman smoother is given as follows:

$$\bar{q} = [q_1, \quad q_2, \quad q_3, \quad q_4]^T \tag{1}$$

Where the first three elements of $\bar{q}$ *indirectly* give rotation magnitude about the x, y and z axes, respectively, and the last element gives the magnitude of overall orientation. The following equations can be used to determine the *Kalman filter a priori* values of the quaternion vector for subsequent time steps under evaluation (as found in [7]):

$$\bar{q}_{k+1} = \Phi_{k+1,k}\bar{q}_k \tag{2}$$

And

$$\Phi_{k+1,k} = I_{4,4} + \frac{1}{2}\Omega(\bar{\omega})\Delta t \tag{3}$$

Where Δt is time interval between data readings, $I_{4,4}$ is a four by four element identity matrix and the matrix $\Omega(\bar{\omega})$ is given by:

$$\Omega(\bar{\omega}) = \begin{bmatrix} 0 & -\omega_z & \omega_y & -\omega_x \\ \omega_z & 0 & -\omega_x & -\omega_y \\ -\omega_y & \omega_x & 0 & -\omega_z \\ \omega_x & \omega_y & \omega_z & 0 \end{bmatrix} \tag{4}$$

Elements ω_x, ω_y and ω_z (components of the vector $\bar{\omega}$) are gyroscope measurements for the x, y and z axes respectively. The *Kalman filter a priori* covariance matrix (P) for the quaternion state vector is found for subsequent time steps as follows:

$$P_{k+1} = \Phi_{k+1,k} P_k \Phi_{k+1,k}^T + \Delta t \left(\frac{1}{2}\right)^2 M C_{g,arw} M^T \tag{5}$$

Where

$$M = \begin{bmatrix} -q_4 & q_3 & -q_2 \\ -q_3 & -q_4 & q_1 \\ q_2 & -q_1 & -q_4 \\ q_1 & q_2 & q_3 \end{bmatrix} \tag{6}$$

And $C_{g,arw}$ is a three by three covariance matrix for gyroscope measurements populated with non-zero elements along only the main diagonal as in [7]. Values for the matrix are found using an angular random walk formulation as in [8], [9] and [10].

A standard *Kalman filter a posteriori* update procedure is used as in [11] and [12]. Updates are taken from known start and end orientations of the inertial sensors and accelerometer data measurements during periods of *relatively* stationary or limited motion (stationary or limited motion is determined from when accelerometer signals show low standard deviation and have a combined signal strength roughly equivalent to gravity). Such accelerometer *gravity measurements* can provide orientation information for two of the three axes of orientation (the two lying in the horizontal plane).

Kalman Smoother. After Kalman filtering has finished, a Rauch-Tung-Striebel (RTS) Kalman smoother is applied as given in [13] and [9]. This smoother has the effect of removing discontinuities in the processed data and improving data quality.

Once orientation data has been found and smoothed, it can be used to correct accelerometer signals for gravitational measurements. This is done by using a rotation matrix at each time step ($\hat{R}$) to transform accelerometer data ($\bar{a}$) (which lie in the *coordinate frame of the moving IMU (Inertial Measurement Unit)*) into a *consistent coordinate frame*. Such a *consistent coordinate frame* is fixed relative to the earth so that gravity ($\bar{g}$) can be subtracted from the signal. After gravity is removed, the remaining accelerometer data are transferred back into the *IMU coordinate frame*, as follows:

$$\bar{a}_t = \hat{R}^{-1}(\hat{R}\bar{a} - \bar{g}) \tag{7}$$

Where superscript -1 denotes matrix inversion and $\bar{a}_t$ is accelerometer data with only translational motion components remaining (and not influenced by gravitational acceleration). $\hat{R}$ can be found directly from Kalman smoothed quaternion values ($\hat{q}$), as depicted in [14] and [15]). As well, $\bar{g}$ (a three element vector) will have values of 0, 0 and $-g_m$, respectively, for its x, y and z axes (where g_m is the magnitude of gravitational acceleration) if the *consistent coordinate frame* chosen is that of the IMU start position as depicted in Figure 4 (a).

3.2 Application of the WFLC Algorithm

Critically Damped Filter. Processed gyroscope data (raw data with low frequency error drifts removed) and processed accelerometer data (with the gravitational influence on signals removed) are evaluated using a critically dampened filter. This

filter was found to be the best in estimating *intended motion* (i.e. motion with tremor components removed) when compared to a number of popular alternatives based on its ability to adequately track a signal without being influenced by signal components with significant tremor [3]. The filter effectively uses a least squares straight line fit of data with more recent data being given a higher weight (i.e. more influence on the fitting line parameters) than previous data [16].

WFLC Algorithm. The WFLC algorithm fits a series of sinusoidal signals (harmonic sines and cosines referenced to a fundamental frequency) to the data under evaluation [17]. In its early development, it was used for removing the tremor of a surgeon's hand during critical operations by using a feedback system and electric actuators within a surgical instrument to counteract tremor [18]. The algorithm also is quite useful for describing and removing movement disorder tremor because of its ease of implementation (i.e. simplistic mathematical iterations are utilized), zero phase lag real time filtering capabilities and its relative computational efficiency (utilizing just a small number of iterative computational steps).

3.3 Wavelet Spectral Analysis

Pre and post WFLC processed inertial data (both with gravitational effects on accelerometer data removed) are analyzed using wavelets to determine the frequency spectrum. The main advantages of using wavelets for such an application is the localization power in both frequency and time domain (so non tremor signal portions between trials can be easily negated) and the availability of numerous base functions, that can be used as mother wavelet function, leading to better signal modelling; as opposed to Fourier based analysis which is largely focussed on the use of sinusoidal functions for analysis.

A continuous wavelet transform was used to allow for a more thorough visual inspection of the signal evaluated than can be afforded using a discrete wavelet transform. The continuous wavelet transform is found as follows [19]:

$$w(p_j, t_k) = \frac{1}{\sqrt{p_j}} \int_{-\infty}^{\infty} s(t)\tilde{\psi}\left(\frac{t - t_k}{p_j}\right)dt \tag{8}$$

Where p_j is a scaling coefficient to allow for analysis of different frequencies of interest, t_k is a time shift parameter that allows for localization of the analysis, $s(t)$ is the inertial signal under evaluation at time t and $\tilde{\psi}$ is the mother wavelet analyzing function's complex conjugate. Wavelet scales selected (given as j in (8)) for evaluation span 1 to 64 (with corresponding *pseudo-frequencies* of 91.8 Hz and 1.4 Hz respectively; these *pseudo-frequencies* are found by scaling the wavelet center frequency [20]). Such a broad frequency spectrum for analysis allows for an in depth view of the signal under examination.

A coiflets wavelet of order three was used for evaluation because it matched closely with the data when compared to other possible wavelet candidate functions. The coiflets 3 mother wavelet is shown in Figure 3.

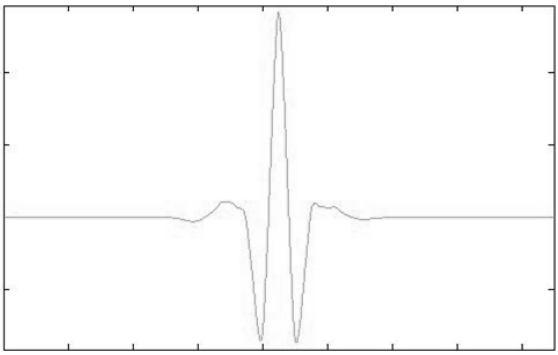


Fig. 3. The coiflets 3 mother wavelet

4 Experimental Methods and Results

4.1 Data Collection

Figure 4 depicts the manner in which data was collected from test subjects (note the IMU axes labels in Figure 4 (a)). Test subjects lifted an IMU out of a holster and simulated eating using a spoon attached to the IMU (simulating only one placement of food into their mouth). Upon completion of this task, the returned the IMU to the holster; ten such tests were carried out for each subject under evaluation. In this manner, both the start and end orientation of the IMU were known (relative to one another) which is important for Kalman filtering depicted in sub-section 3.1.

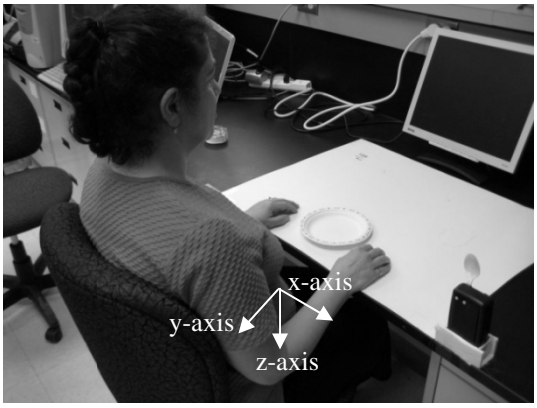


Fig. 4(a). A test subject prior to evaluation

Data logging took place at 130 Hz. The IMU used was manufactured by the Mobile Multi-Sensor Systems (MMSS) research group at the University of Calgary. A tri-axial accelerometer (LIS3L06AL from ST Microelectronics [21]) and three single-axis gyroscopes (XV-8100CV from Epson Toyocom [22]) were utilized.



Fig. 4(b). A test subject during evaluation

During experimentation, 11 controls (7 female), 9 ET patients (3 female) and 30 PD patients (20 female) were evaluated using testing that had received ethics approval from the Conjoint Health Research Ethics Board at The University of Calgary. The mean age of controls was 64.1, for ET patients it was 64.8 and for PD patients it was 66. A number of patients (2 ET and 27 PD) were on medication to help reduce tremor. For the ET patients, this medication was largely ineffective (based on conversation with the patients and results of the data analysis presented here) and for PD patients it had varying effectiveness depending on when they last took their medication and how large the dosage was.

Patients with the most *significant tremor* (8 of the ET patients and 9 PD patients) were evaluated separately for the analysis in the following sections of this research paper. These patients were selected based on a thresholding criteria that required their tremor to be one standard deviation in excess of the tremor measured for controls (for at least one of the six inertial signals evaluated), based on the mean absolute value of wavelet details coefficients at scale 18 (corresponding to approximately 5.1 Hz). A higher value for details coefficients suggested more tremor motion was present. Test subjects that did not pass this thresholding criteria produced data that largely resembled that of controls and therefore such data is not displayed in the following sections of this research paper, except in Section 4.4, were low tremor patients are referred to as having *limited tremor.*

4.2 Kalman Smoothing and WFLC

The Kalman smoothing algorithm was very effective in removing gravitational readings from accelerometer signals. A representative example of this is shown in Figure 5 for a control.

It is important to note that the raw data signal in Figure 5 has extended periods of time in which it is continuously a long distance from the zero acceleration mark. This is expected because gravitational readings can influence the sensors, depending on their orientation. When gravity's impact is removed, the signal remaining only has short durations when it is not near the zero accelerometer reading. This illustrates the

value of using the algorithm outlined because the remaining accelerometer data, for the most part, only have lateral motion information embedded within them, making subsequent data analysis significantly more useful than if raw data were used.

After accelerometer signals are corrected as depicted in Figure 5, they are analysed (along with processed gyroscope data) using a critically dampened filter. The goal is to estimate the *intended motion* of the test subjects under evaluation. The results for this are shown in Figure 6 for an ET subject. Generally, the critically dampened filter performed quite well and was a very efficient filter for estimating *intended motion* of test subjects. One issue that if unavoidable with such a filter is that if the influence of past measurements if kept high (in the least squares sense), the filter will lag the signal under evaluation; however, if the influence of past measurements is reduced, then the filter does not remove all tremor motion components. A balance needs to be struck to ensure both adequate tracking and tremor removal are achieved. The results given in Figure 6 are generally representative of results for all inertial data when a subject with high tremor is evaluated. In this case, a small amount of residual tremor remains in the processed signal; but the ability of the critically damped filter to track the signal adequately, so as to approximate *intended motion*, is reasonable.

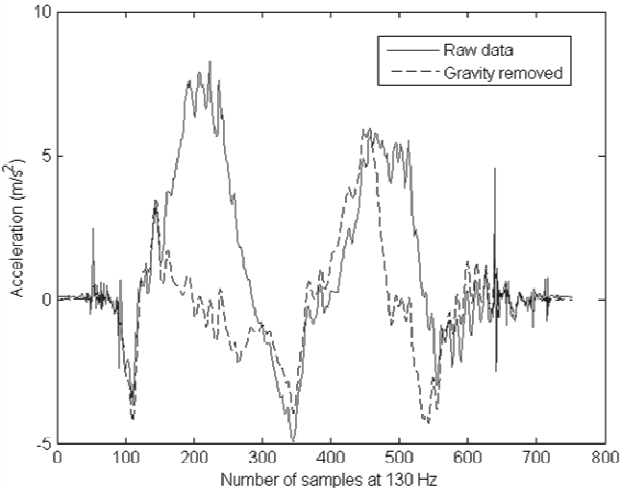


Fig. 5. An x-accelerometer signal before and after Kalman smoothing to remove gravitational readings

After the critically damped *intended motion* approximation is removed from the signal, the remaining signal portion can be evaluated using the WFLC algorithm. Application of this algorithm required a two stage iteration at each time step, the first iteration was used to find the fundamental frequency of the tracked signal and the second iteration was used to find the numerical weights for sinusoids tracking the signal. Multiple iterations were sometimes required at the same time step to allow the algorithm to sufficiently converge to a solution (particularly with the gyroscope data which had a large dynamic range). A representative sample for the WFLC motion approximation is shown in Figure 7; it is a close up of a signal portion of the data displayed in Figure 6.

It is clear from Figure 7 that the WFLC algorithm tracks tremor quite well. The result depicted is quite typical for all inertial data evaluated. Given such an approximation of the tremor, it is possible to very precisely pinpoint the frequency of the tremor observed and its magnitude. This is useful for assessment because it allows for medical professional to evaluate how tremor varies between patients and for the same patient before and after medication is taken.

For the purposes of attenuation, the WFLC algorithm provides the signal to be removed from motion to mitigate tremor. The WFLC algorithm is especially useful for attenuation because its zero phase lag property allows for accurate and real time tremor suppression which is critical for many attenuation applications.

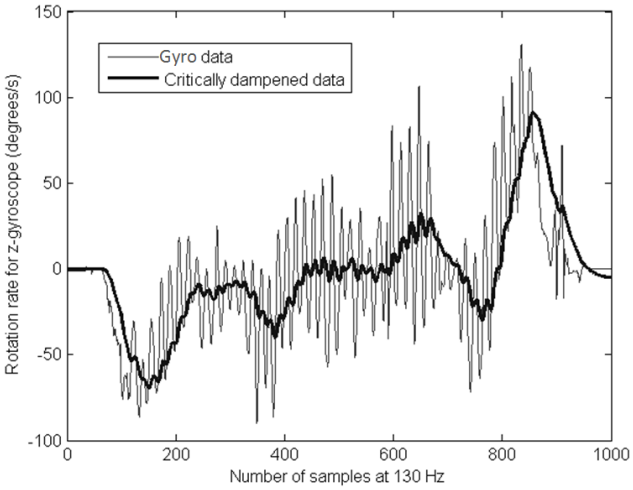


Fig. 6. A processed z-gyroscope signal before and after a critically damped filter is applied to approximate *intended rotational motion*

4.3 Wavelet Spectral Analysis

Signals can be compared before and after removal of WFLC tremor components to evaluate their frequency spectrum. Such an analysis can assist one in understanding what components of motion have been removed and whether tremor frequencies of interest have been adequately targeted (3-12 Hz). It also helps one to understand how useful the WFLC algorithm is for assessment by depicting how well the tremor motion can be tracked.

In Figure 8, the results of applying the continuous wavelet transform to the signal in Figure 6 (with the critically dampened portion of motion removed) are shown; lighter colours depict that more frequency content is present.

When comparing Figure 8 to Figure 6, it is clear that when a great deal of tremor motion is present, the amount of signal energy depicted in the 3-12 Hz frequency band increases significantly. Thus, Figure 8 validates the use of a coiflets 3 wavelet for the application undertaken (given that other inertial signals processed gave similar results).

The overall (population) results for the data analysed are given in Figures 9 and 10 for a representative accelerometer axis of motion and gyroscope axis of motion, respectively. Tellingly, the results for all three accelerometer axes of motion were very similar as were the results for all three gyroscope axes of motion. This tends to indicate that tremor acts along all axes of motion concurrently and also that it can be removed in a similar fashion (using the critically dampened and WFLC algorithms) for all of these axes. The results depicted were found by taking the mean magnitude of details coefficients at each wavelet scale for all test subjects of a particular group.

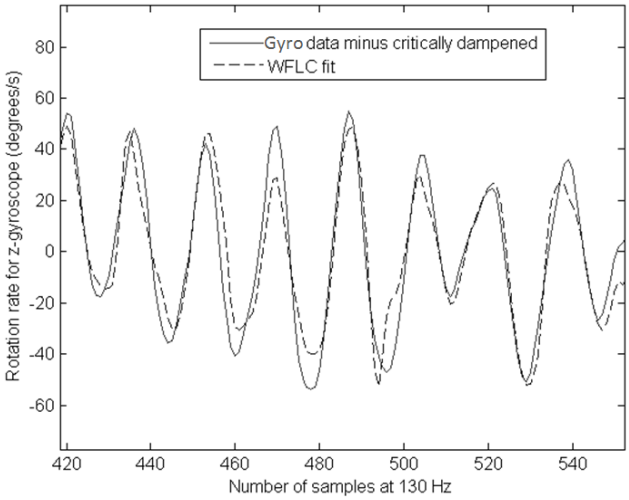


Fig. 7. A processed z-gyroscope signal portion and its WFLC approximation depicting rotational motion

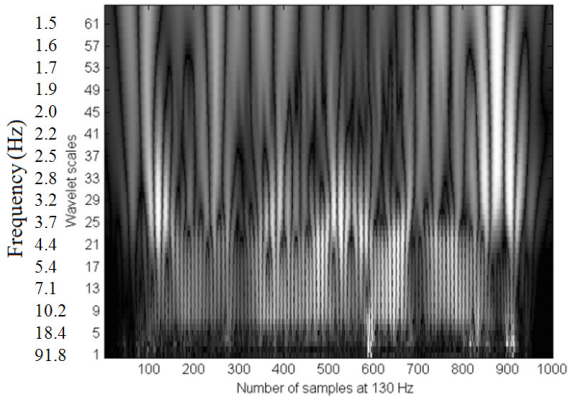


Fig. 8. Wavelet processed z-gyroscope data (only *unintended tremor motion* is processed for rotational movement)

It can be seen from the results in Figures 9 and 10 that the frequency of the tremor measured for both ET and PD patients was within the expected 3-12 Hz range (as

depicted by a bulge in the data displayed within this band). ET patients generally depicted more tremor that PD patients, which corresponds well to the fact that most ET patients were not on medications for treatment while most PD patient were.

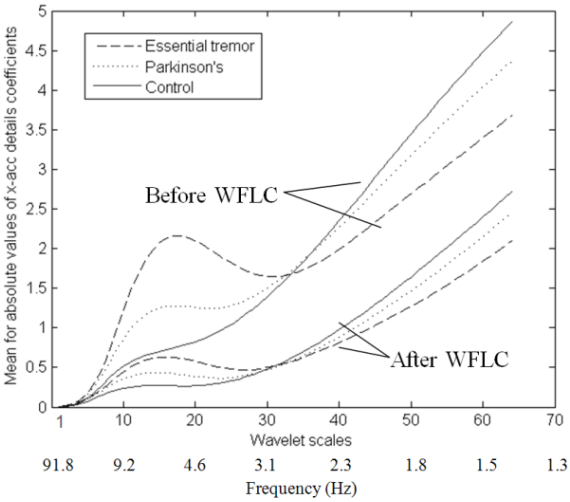


Fig. 9. X-accelerometer wavelet spectral analysis for lateral tremor

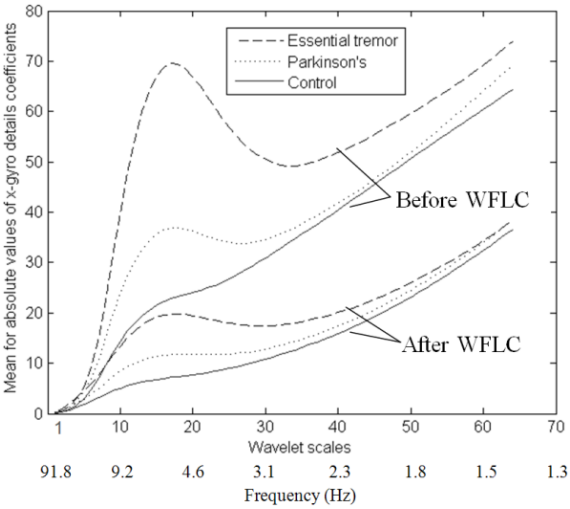


Fig. 10. X-gyroscope wavelet spectral analysis for rotational tremor

Tremor was reduced substantially for all three groups examined (ET, PD and control) when the WFLC algorithm was applied and it decreased in a proportionate manner, such that those with more tremor before processing also had more frequency content remaining in their motion after application the WLFC algorithm. Signal noise was also likely reduced inadvertently. One of the more significant results is that in the

3-12 Hz range, all three groups evaluated with the WFLC technique had less tremor after evaluation than the control group had before evaluation.

One unfortunate drawback of the analysis is that it seems to remove a lot of low and high frequency motion (as is evident in Figures 9 and 10) along with the 3-12 Hz frequency band of interest.

The high frequency motion likely represents jerky motion, such as when a subject inadvertently struck the IMU on the table during testing. Likely, for mechanical attenuation applications, some kind of thresholding criteria will need to be applied to ensure such signal spikes are not processed, because mitigation of such motion is likely unrealistic due to limits in the operation range of mechanical equipment.

The low frequency motion removed likely represents the inadequacy of the critically dampened filter in tracking *intended motion*. Determining which motion is *desired* and which is not is a very difficult research challenge that has been studied for many years [1]. Clearly, there still remains some work to be done to find and appropriate algorithm that is adequately computationally fast and has zero phase lag.

4.4 Determining Tremor Phase Shifts for Different Degrees-of-Freedom

In order to mitigate tremor to improve the daily lives of movement disorder patients, it is first necessary to classify tremor in such a manner so that its main axes of motion can be understood. To do this, one should examine the phase shifts among different tremor signals logged by inertial sensors. Since the x-accelerometer and y-gyroscope depicted the most tremor motion, these signals will be examined first. To carry this out, the same IMU as what is depicted in Figure 4 is utilized (as well as the same holster and test subjects), but with a laser placed on top of the IMU and with subjects asked to direct the laser at ten targets on a computer screen labeled 1 through 10 (this test is depicted in Figure 11). This test was much more suited to phase shift analysis because the orientation of the subject's limbs did not shift significantly during the course of a trial making it easier to determine the phase shifts among all tremors present in the six degrees-of-freedom evaluated. In the case of eating with a spoon (as evaluated earlier in this paper), the phase relationships studied tended to be non-static as subjects re-oriented their arm during the course of a trial. Although the appropriate phase relationships can still be studied in the case of eating with a spoon, the authors have chosen to use a simpler motion profile so as to explain the technique for the studying of phase relationships more clearly.

Peak coherence is used here to help obtain a value of frequency at which phase relationships should be studied. Peak coherence tends to depict the frequency of most relevance for studying phase relationships because it depicts when two signals are most strongly related (in terms of tremor). Coherence is defined in [23]and is a scaled version of the following cross-spectral term:

$$\hat{f}_{12}(\omega) = \frac{1}{2\pi LT}\sum_{l=1}^{L} d_{s1}(\omega, l)\widetilde{d_{s2}(\omega, l)} \tag{9}$$

Where $d_{s1}(\omega, l)$ and $d_{s2}(\omega, l)$ are discrete Fourier transforms for signal s_1 and s_2 respectively (the second discrete Fourier term is denoted as a complex conjugate given the notation directly above this parameter). The Fourier transforms operate over signal intervals of T length (chosen to be 130 samples). These Fourier transform are

defined fully in [23]. In total there are L intervals of length T samples (denoted $l = 1,2 \ldots L$). The term ω represents angular frequency. To properly scale the cross-spectral term $\hat{f}_{12}(\omega)$, auto-spectral terms $\hat{f}_{11}(\omega)$ and $\hat{f}_{22}(\omega)$ are used (these two terms are found by setting signal s_2 tos_1 and then setting s_1 to s_2, respectively, for the formuation given in (9)). The coherence values are thus found as follows:

$$\left|\hat{R}_{12}(\omega)\right|^2 = \frac{\left|\hat{f}_{12}(\omega)\right|^2}{\hat{f}_{11}(\omega)\hat{f}_{22}(\omega)} \tag{10}$$

Coherence in this case is a measure of the *closeness* of two signals. Phase shifts are evaluated at coherence peak frequencies (3-12 Hz) because this is where the most pronounced tremor often occurs (coherence peak frequencies also tend to coincide with auto-spectral peak frequencies as defined by $\hat{f}_{11}(\omega)$ and $\hat{f}_{22}(\omega)$).

By measuring phase shifts (based on discrete Fourier transforms as shown in (9)) it is possible to determine the interrelationship between two signals. In Table 1, a strong in phase relationship is depicted for x-translation data (obtained from integrating accelerometer data twice with respect to time) and y-rotation data (obtained when integrating gyroscope data once with respect to time). A right hand rule sign convention is used for rotation motion and axes are defined following what is given in Figures 4 and 11. Data sets used for coherence and phase analysis have been corrected through the use of Kalman filtering and gravity removal schemes as outlined in the top half of Figure 2.

Table 1. Population results of the mean and standard deviation for phase lag at peak coherence frequency of the y-rotation data behind the x-translation data

Subject type	Mean phase lag	Standard deviation of phase lag
Control	15.1°	17.1°
ET– *significant tremor*	18.5°	27.1°
PD – *significant tremor*	1.43°	12.8°
PD – *limited tremor*	1.80°	10.1°

It should be noted that there was one control, one ET patient and one PD patient with *significant tremor* that were removed from the data for the analysis depicted in Table 1 because they did not fit the overall pattern of motion observed. The data collected show results with generally low standards of deviation, meaning that likely a significant phase relationship has been observed. The magnitude of phase lag is close to zero, which would be expected for an x-axis lateral motion that is in phase with a y-axis rotational motion, as depicted in Figure 11. Given that such a motion exists even among controls, it likely has a biomechanics' mechanism that is partly responsible.

Knowing that an in phase tremor is acting along the x-axis and about the y-axis, it is possible to use this knowledge to design devices to mitigate tremor. For example, a cup could be designed with a handle that uses inertial feedback to move in a manner as to counteract tremor at a fluidic beverage so as to avoid spillage. Of course, the tremor in such a case would not be identical to what is observed in this research paper, but appropriate inertial data could be logged whilst patients were drinking beverages to determine the main axes along with mitigation should act. After such data was logged, the analysis presented in this research paper would provide insight into the motion observed.

The lateral x-axis and rotational y-axis tremor were not the only correlated tremors. Sometimes the z-axis rotational tremor was also in phase with these two tremors. This tended to be the case when subjects wrapped the palm of their hand under the IMU during testing as opposed to gripping the IMU from the vertical plane of the IMU facing them. Referring to Figures 4 and 11, this meant subjects were gripping the IMU like a door handle from its base rather than from its back and thus altering the axes along which tremor data was logged. As well, the y-axis lateral tremor was also sometime in phase with x-axis lateral tremor when right handed subjects wrapped their hand around the left portion of the IMU to obtain a better grip with their thumb (left is defined as the negative x-axis in Figures 4 and 11). As such, subjects were sometimes gripping the IMU partly from the left hand side as opposed to from the back, causing an alteration in the axes along which tremor was acting.

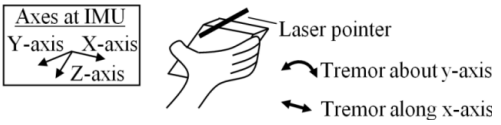


Fig. 11. The two largest tremors logged which were also in phase

The other major tremor correlation observed was between the z-axis lateral tremor and the x-axis rotational tremor (in the data, this tremor correlation showed up as being one hundred and eighty degrees out of phase when utilizing a right hand rule sign convention).

5 Summary and Conclusions

The analysis performed validated the use of the Kalman smoothing scheme depicted for removal of the gravitational influence on accelerometer signals. This was necessary so that accelerometer data would properly depict the lateral motion under consideration.

The most significant finding of this research paper is that a combination of a critically dampened filter and the WFLC technique adequately removed the tremor components of motion within the 3-12 Hz frequency band; this was possibly applied among the first times in six degrees-of-freedom for movement disorders in this research paper. The same procedure was applied to all motion axes with quality results in all cases.

Another significant finding of this research paper is that the wavelet analysis performed was well suited for the evaluation of a frequency spectrum. The coiflets 3 wavelets was quite capable of realizing tremor motion components, and provided a useful tool for identifying motion at frequencies of interest.

It was also shown that the phase relationship between different tremor signals could be verified. This is a very significant finding because in cases of strong coherence, such a phase relationship will display how subject motion relationships among different degrees of freedom are affected by different tremor types and this will influence the choice actuator used for tremor mitigation.

Acknowledgements. Thanks to Dr. Brian MacIntosh and Bruce Wright for their help with obtaining lab space, ethics approvals and the appropriate equipment. Thanks also to the following funding agencies: Alberta Innovates – Technology Futures (formerly the Alberta Ingenuity Fund), the Natural Sciences and Engineering Research Council of Canada (NSERC) and Geomatics for Informed Decisions (GEOIDE). Further thanks to volunteer subjects and the Mobile Multi-Sensors Systems (MMSS) research group at the University of Calgary.

References

1. Rocon, E., Belda-Louis, J.M., Sanchez-Lacuesta, J.J., Pons, J.L.: Pathological Tremor Management: Monitoring, Compensatory Technology and Evaluation. Technology and Disability 16, 3–18 (2004)
2. Rocon de Lima, E., Andrade, A.O., Pons, J.L., Kyberd, P., Nasuto, S.J.: Empirical Mode Decomposition: a Novel Technique for the Study of Tremor Time Series. Med. Bio. Eng. Comput. 44, 569–582 (2006)
3. Gallego, J.A., Rocon, E., Roa, J.O., Moreno, J.C., Koutsou, A.D., Pons, J.L.: On the use of inertial measurement units for real-time quantification of pathological tremor amplitude and frequency. In: Proceedings of the Eurosensors XXIII Conference. Procedia Chemistry, vol. 1, pp. 1219–1222 (2009)
4. Louis, E.D.: Essential tremor. Lancet Neurol. 4, 100–110 (2005)
5. Mansur, P.H.G., Cury, L.K.P., Andrade, A.O., Pereira, A.A., Miotto, G.A.A., Soares, A.B., Naves, E.L.M.: A Review on Techniques for Tremor Recording and Quantification. Critical Reviews in Biomedical Engineering 35(5), 343–362 (2007)
6. Elble, R.J., Koller, W.C.: Tremor. The John Hopkins University Press, Baltimore (1990)
7. Sabatini, A.M.: Quaternion-based extended Kalman filter for determining orientation by inertial and magnetic sensing. IEEE Transactions on Biomedical Engineering 53(7), 1346–1356 (2006)
8. El-Sheimy, N., Hou, H., Niu, X.: Analysis and modeling of inertial sensors using Allan variance. IEEE Transactions on Instrumentation and Measurement 57(1), 140–149 (2008)
9. Shin, E.: Estimation Techniques for Low Cost Inertial Navigation. Ph.D. Department of Geomatics Engineering, University of Calgary (2005)
10. Stockwell, W.: Angle Random Walk. Crossbow Technologies Inc., pp. 1–4 (2010), `http://www.xbow.com/pdf/AngleRandomWalkAppNote.pdf` (accessed July 13, 2010)
11. Chui, C.K., Chen, G.: Kalman Filtering With Real Time Applications, 2nd edn. Springer, Heidelberg (1991)
12. Grewal, M.S.: Kalmen Filtering: Theory and Practice. Prentice-Hall (1993)
13. Brown, R.G., Hwang, P.Y.C.: Introduction to Random Signals and Applied Kalman Filtering, 2nd edn. John Wiley and Sons Inc. (1992)
14. Altmann, S.L.: Rotations, Quaternions and Double Group. Dover Publications (1986)
15. Kuipers, B.J.: Quaternions and Rotation Sequences: A Primer With Applications to Orbits, Aerospace and Virtual Reality. Princeton University Press (1999)
16. Brookner, E.: Tracking and Kalman filtering made easy. John Wiley & Sons, Ltd. (1998)
17. Riviere, C.N., Reich, S.G., Thakor, N.V.: Adaptive Fourier Modeling for Quantification of Tremor. Journal of Neuroscience Methods 74, 77–87 (1997)

18. Riviere, C.N., Rader, R.S., Thakor, N.V.: Adaptive Canceling of Physiological Tremor for Improved Precision in Microsurgery. IEEE Transactions on Biomedical Engineering 45(7), 839–846 (1998)
19. Goswami, J.C., Chan, A.K.: Fundamentals of Wavelets: Theory, Algorithms and Applications. Wiley Series in Microwave and Optical Engineering. Wiley-Interscience (1999)
20. Matlab (computational environment) Help Documentation, scal2frq function (2008), http://matlab.izmiran.ru/help/toolbox/wavelet/scal2frq.html (accessed July 13, 2010)
21. LIS3L06AL MEMS Inertial Sensor Data Sheet, ST Microelectronics, pp. 1–17 (2006), http://www.st.com/stonline/products/literature/ds/11669.pdf (accessed July 13, 2010)
22. XV-8100CB ultra miniature size gyro sensor data sheet. Epson Toyocom, pp. 1–2 (2010), http://www.frank-schulte.com/media/Pdf/XV-8100CB_E06X.pdf (accessed July 13, 2010)
23. Halliday, D.M., Rosenberg, J.R., Amjad, A.M., et al.: A framework for the analysis of mixed time series/point process data – theory and application to the study of physiological tremor, single motor unit discharges and electromyograms. Prog. Biophys. Molec. Biol. 64(2), 237–278 (1995)

A New Approach for Diagnostic Investigation of Total Hip Replacement Loosening

Catherine Ruther, Ulrich Timm, Andreas Fritsche, Hartmut Ewald, Wolfram Mittelmeier, Rainer Bader, and Daniel Kluess

Department of Orthopaedics, University of Rostock
Doberaner Str. 142, 18057 Rostock, Germany
Department of General Electronics, University of Rostock
Justus-von-Liebig-Weg 2, 18059 Rostock, Germany
{catherine.ruther,andreas.fritsche,wolfram.mittelmeier,
rainer.bader,daniel.kluess}@med.uni-rostock.de,
{hartmut.ewald,ulrich.timm}@uni-rostock.de
http://www.forbiomit.med.uni-rostock.de

Abstract. Diagnosis of total hip replacement (THR) loosening using imaging often fails to provide reliable results. New implants instrumented with sensors shall conform to this challenge. Therefore, a novel concept of an in vivo sensor was tested. This simple mechano-acoustical sensor is integrated inside the total hip stem and enables detection of osseous fixation. The sensor is excited by an external coil and impinges inside the THR. The spring-back of the sensor can be detected by an extracorporeal coil, while the structure-borne sound is measured with a vibration sensor placed at the patient's leg. Experiments of a THR with one integrated sensor showed the possibility to differentiate between well fixed and loosened implants. The presented in vivo sensor system has a promising potential to detect loosening and to analyze osseointegration.

Keywords: Total hip replacement, Implant loosening, Sensor, Diagnostic investigation, Arthroplasty.

1 Introduction

Total Hip Replacement (THR) is one of the most successful treatments in orthopaedic surgery [1]. But there are still challenges, especially referring to implant loosening as a result of wear-particle induced osteolysis. Moreover, monitoring of implant fixation is an important factor in determining the longterm success rate of Total Hip Replacements [2]. Diagnosing implant loosening using imaging methods like radiographs or arthrography often provides poor sensitivities and specificities of about 80% [3]. Hence, researchers strive for supporting techniques or even attempt to replace ex vivo techniques by novel in vivo sensors. Accelerometers were used in vibrometry and showed that implant instability can be identified by several harmonics in the output signal of the sensors [4]. Besides miniaturized design and biocompatibility, a reliable power supply is a challenge in development of in vivo sensors. The technology of choice for vibrometry is wireless powering and data telemetry [5]. This is effective for implants where

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 74–79, 2013.
© Springer-Verlag Berlin Heidelberg 2013

both coils are positioned with small separation distance and fixed orientation. However, reproducible state cannot be achieved in many cases, caused by possible changes in the patients weight. The influence of variable distances between both coils leads to lower coupling factors and therefore poor signal transmission. A further problem is the use of only low frequencies, which affect the damping of the signals caused by eddy currents, especially in metal implants. This is an evident drawback for orthopaedic implants, especially for THR where biocompatible metal alloys are used. Due to the aforementioned reasons we present a novel and passive sensor concept without inductive coupling of two coils. The concept is characterized by an extracorporeal coil to excite permanent magnets inside the THR, which act as excitation unit. Detection of the loosening sensors can be realized by a second coil or a vibration sensor. As an extension to our publication of the first measurements with a detection coil [6], in this work we demonstrate first results of the detection of structure-borne sound by using a vibration sensor.

2 Methods

2.1 Principle of the Loosening Sensor

The presented acoustic method is based on the propagation of structure-borne sound induced by impingement of a magnetic body on a membrane placed inside the femoral hip stem. The magnetic body, fixed on a flat spring, is part of the sensor, which is designed as a simple mechanical oscillator for impinging inside the implant [7]. The magnetic poles allow oscillations using an external magnetic field impulse. Thus, motion of the oscillator is initiated using an external coil (Fig. 1). The excitation coil is placed outside in defined distances to the oscillator. It is realized with an iron core for a maximum concentration of the streamlines on the oscillator. An important requirement for the design of the excitation coil is to ensure the distance for sufficient magnetic excitation, which should be at least 100 mm.

The structure-borne sound, after impingement, varies depending on the osseointegration of the implant. On the one hand, a well osseous integrated implant leads to higher natural frequencies caused by the higher system stiffness. On the other hand, a loosened THR with soft tissue on the external surface results in lower natural frequencies caused by lower system stiffness. For detection of the structure-borne sound a vibration sensor is placed outside the patient.

2.2 In Vitro Measurements Using a Total Hip Stem

To analyze the feasibility of the acoustic principle, an oscillator was integrated in the proximal part of a custom cementless straight hip stem (CBH Shaft, Mathys, Switzerland) (Fig. 2). The hip stem was implanted in an untreated porcine femur, which was potted in a cylindrical set-up with epoxy resin. During the molding process, the flexible seating of the femoral condyles was ensured to model the flexible femur-tibia interface of the knee.

The hip stem was implanted using pressfit in the first test. Prior to a subsequent test, the hip stem was pulled out and the bore diameter was extended for modelling a

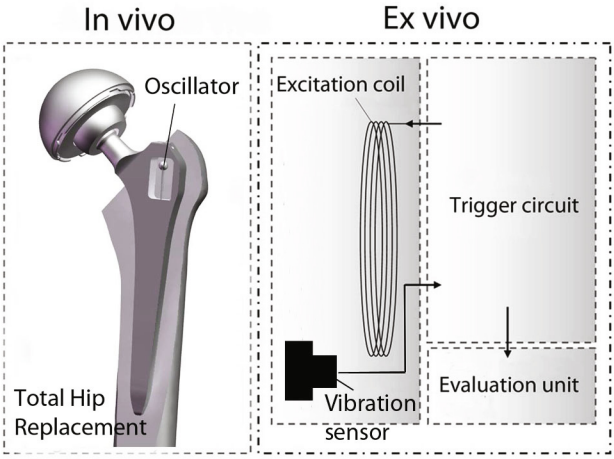


Fig. 1. System concept of the mechano-acoustical sensor realizing an excitation of the THR inside the hip stem and a wireless detection of implant loosening



Fig. 2. Total hip stem with one oscillator placed in the proximal part

completely loosened hip stem. Before testing, the porcine femur was placed in a water reservoir as a first approach to model soft tissue (Fig. 3). Furthermore, the set-up was placed on a marmoreal plate with elastic damper elements to avoid disturbance of external mechanical oscillations. A piezoelectric vibration sensor for broadband measurements between 1 and 24 kHz (Metra, Radebeul, Germany) with a high sensitivity of 100 mV/g was fixed outside the water reservoir to detect the transmitted oscillations. The excitation was realized by a coil outside the water reservoir placed at the level of the trochanter major with a distance of 50 mm to the porcine femur. The excited THR oscillated in its eigenmodes, in which mainly bending modes led to a sound emission of the implant to the surrounding bone and soft tissue.

3 Results

Evaluation of the resulting output signal was performed with Origin (OriginLab, Massachusetts, USA). The output signal revealed a good differentiation between the pressfit and the loosened hip stem (Fig. 4). The signal of the well fixed THR resulted in a higher dampening than the loosened THR. As a result the time in which the vibrational signal

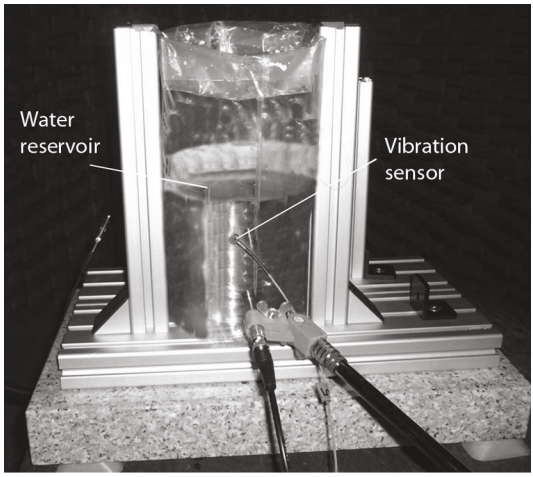


Fig. 3. Experimental set-up to measure the output vibration of the THR inside the water reservoir

reaches the complete deactivation of the excited state varies. A well fixed implant resulted in an oscillation time of 1.8 ms. In contrast to the stable implant, the THR with an unstable seating showed an oscillation time of 4.4 ms.

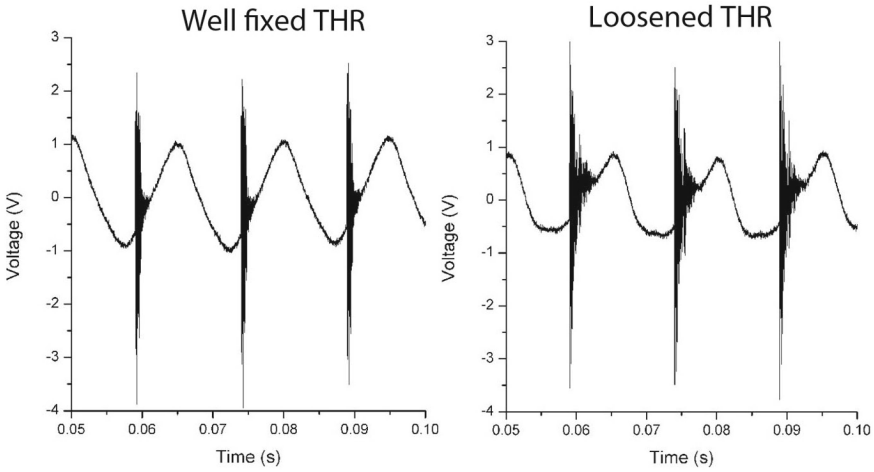


Fig. 4. Left: Time signal of the well fixed total hip stem. Right: Time signal of the loosened total hip stem.

In order to evaluate the natural frequency changing of the hip stem, a fast fourier transformation (FFT) was performed. The signal was filtered by a band pass to match the sensors measurement range. The FFT was carried out using a hanning window. For direct comparison of the spectra, both signals were normalized. In the frequency spectrum the loosened THR had its first eigenfrequency at 292Hz, while the first eigenfrequency of the well fixed THR was at 623Hz (Fig. 5).

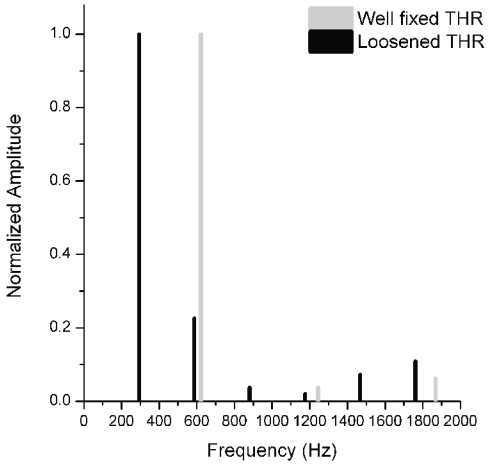


Fig. 5. Frequency spectrum of the loosened and the well fixed total hip stem

4 Discussion

With respect to the standard application of total hip replacements, unsatisfying results in loosening diagnosis increase the demand for more precise in vivo techniques. Inductive coupling based on radio frequency powering to provide energy supply of in vivo sensors causes problems, especially with regard to the coupling factor between the two coils. In order to prevent this problem, our approach uses the capability of a novel loosening sensor, which avoids a second coil inside the implant. The loosening sensor enables two different detection methods. The first detection method with a detection coil measuring the velocity after impingement of the oscillator was discussed in [6].

The detection of structure-borne-sound using the oscillator as internal sound source was presented as a second approach in this article. First results show, that a higher system stiffness with a well fixed THR has a higher resonance frequency than a loosened THR. Furthermore, a loosened THR shows much more resonance frequencies than a well fixed implant. A challenge of this principle is a reproducible sensor fixation at the patients leg to eliminate influences on the resonance frequency shift. Amplitude measurements can only be taken as an additional parameter and only in a normalized condition because of the soft tissue influence.

In this investigation, the resonance frequency shift could be analyzed clearly, which shows the usability of the described in vivo method. The robustness of the oscillators due to the simplicity of the assembly guarantees functionality after intraoperative impaction and sterilization of the implant. Moreover, the oscillators can be used in experimental applications to determine the quality of osseointegration of e.g. new coated implant materials.

5 Conclusions

In the present study, a new measurement method for in vivo diagnosis of THR loosening without inductive coupling based radio frequency powering was demonstrated. The described loosening sensor shows promising results in preliminary tests. Future work will include experiments on the sensitivity of the new loosening approach. Furthermore, analyses into the influence of soft tissue on the structure-borne sound are aspired.

Acknowledgements. This research project is granted by the German research foundation (DFG) under Reg.No. KL 2327-1.

References

1. Hailer, N.P., Garellik, G., Kaerrholm, J.: Uncemented and cemented primary total hip arthroplasty in the Swedish Hip Arthroplasty Register: Evaluation of 170, 413 operations. Acta Orthop. 81, 34–41 (2010)
2. Pastrav, L.C., Jaecques, S.V.N., Jonkers, I., van der Perre, G., Mulier, M.: In vivo evaluation of a vibration analysis technique for the peri-operative monitoring of the fixation of hip prostheses. J. Orthop. Surg. Res. 4, 1–10 (2009)
3. Temmermann, O.P., Raijmakers, P.G., Berkhof, J., Hoekstra, O.S., Teule, G.J., Heyligers, I.C.: Accuracy of diagnostic imaging techniques in the diagnosis of aseptic loosening of the femoral component of a hip prosthesis. J. Bone Joint Surg. 87, 781–785 (2005)
4. Li, P.L.S., Jones, N.B., Gregg, P.J.: Vibration analysis in the detection of total hip prosthetic loosening. Med. Eng. Phys. 18, 556–600 (1996)
5. Puers, R., Catrysse, M., Vandevoorde, G., Collier, R.J., Louridas, E., Burny, F., Donkerwolcke, M., Moulart, F.: A telemetry system for the detection of hip prosthesis loosening by vibration analysis. Sens. Act. 85, 42–47 (1999)
6. Ruther, C., Ewald, H., Fritsche, A., Mittelmeier, W., Bader, R., Kluess, D.: A new method for detection of total hip replacment loosening - Development and first results of a novel mechano-acoustical sensor. In: Proceedings of Biodevices 2011, pp. 70–73 (2011)
7. Ruther, C., Ewald, H., Biemann, A., Nierath, H., Bader, R., Kluess, D.: A New Concept for Non-Invasive Radiation-Free Detection of Implant Loosening. In: 56th Annual Meeting of the Orthopaedic Research Society, New Orleans, vol. 2413 (2010)

Wearable Monitoring Unit for Swimming Performance Analysis

Ana S. Silva, Antonio J. Salazar, Carla M. Borges, and Miguel V. Correia

INESC Porto - Instituto de Engenharia de Sistemas e Computadores
Faculdade de Engenharia da Universidade do Porto
R. Dr. Roberto Frias, 4200-465 Porto, Portugal
{amsilva,antonio.j.salazar,cmborges}@inescporto.pt,
mcorreia@fe.up.pt

Abstract. BIOSWIM (Body Interface System based on Wearable Integration Monitoring) is a joint multidisciplinary effort involving a number of Portuguese R&D teams. It seeks a pervasive monitoring solution for physiological and biomechanical signals from a swimmer under normal training conditions, both in and out of the water. A wearable inertial monitoring unit (WIMU) was developed in order to serve as the biomechanical data processing unit of the system. The preliminary version of the WIMU has a commercially available microcontroller and transceiver set, as well as a tri-axial accelerometer and a bi-axial gyroscope serving as sensors. Testing in and out of the water has provided promising data and contributed to design modifications. These also took into account input from athletes, trainers, and physicians. Future work includes the integration of the WIMU within the complete BIOSWIM swimsuit system, complemented by truly integrated EMG and ECG textile sensors and a chemical monitoring unit.

Keywords: Wearable, Monitoring, Biomechanical parameters, MEMS, Swimming analysis.

1 Introduction

Visual markers and inferred biomechanical data used to be the reference of choice for coaches and trainers when evaluating an athlete's performance. Nowadays, miniaturized low-powered accelerometers and gyroscopes, embedded within wearable monitoring systems, have closed the gap between an athlete's perception and that of trainers. Micro-electromechanical systems (MEMS) represent a viable alternative which, unlike many of its counterparts; can be integrated within wearable-water friendly solutions that can aid on swimming analysis, without burdening the athlete with cumbersome garments and/or equipment. Electronic textiles, interactive clothing, smart garments and wearable are becoming commonplace terms within the rehabilitation and sports community, revealing recent research trends within biomedical engineering and textile technology. During the past decade, researchers and engineers from a number of universities and companies, have proposed several monitoring solutions and mechanism for a wide variety of objectives, from technical applications to leisure. Although it seemed at first that the development momentum favored multimedia applications, the research

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 80–93, 2013.
© Springer-Verlag Berlin Heidelberg 2013

community promptly foresaw such technologies potential for healthcare and sports applications. Currently there exist literature for a large number of projects intended to measure physiological parameters remotely, such as MyHeart and Wealthy projects [1], VTAMN [2] and more recently BIOTEX [3] and Vital Jacket [4].

Although the wearable inertial monitoring unit (WIMU) is still at a prototype stage, preliminary data reveals the promise of such a device.

2 BIOSWIM Project

The BIOSWIM project gained inspiration from emerging wearable monitoring technology solutions and intends to develop a full body smart garment for bio-signals and movement related parameters monitoring, while considering both health and sports performance. Until recently, healthcare, rehabilitation and sports gathered physiological and mechanical data through cumbersome apparatus that affected the behavior of the target individual and sometimes even constrained the range of movement. These solutions were in their majority limited to a controlled laboratory environment, utilizing unpleasant methods for fixing sensors to their intended location. Although the medical and sports community benefitted from such data gathering approaches for decades, more pervasive approaches are nowadays possible. In addition, the need for a controlled environment and especial laboratory setups generally increased the associated research cost and limited the amount of time with the patient or athlete, seldom achieving to collect data that truly reflected the average behavior or performance of the individual in natural conditions. Taking into consideration the before mentioned, past experiences, known obstacles and limitations reported in related literature; a wearable solution was agreed upon which combined ease of use and seamless electronic integration for a more ubiquitous signal monitoring experience. The main goal of the BIOSWIM project is the development of wearable prototypes for measuring several biological parameters, critical for sport performance and health purposes, oriented for in-pool (or underwater depending of the sports) activity, although usable for regular sports and a number of additional applications. Moreover, the wearable system was designed with user comfort as a guiding factor, comprehending the best combination of textile material with embedded microelectronics for signal acquisition and data transmission.

Electromyography, cardiac rhythm, respiratory effort, oxygen consuming taxes, motion capture, wrist and arm accelerations and rotations, speed, hydrodynamics' pressure; constitute some of the parameters of interest that are to be collected using the wearable system. The before mentioned parameters can be utilized for a number of diverse calculations, models, and applications of varying nature. Fig. 1 provides a graphical reference of possible locations for the parameter monitoring sensors.

Although a wide array of ready to use commercially available sensors can be obtained at a relatively low cost; the specificity of some target parameters and the particular measurements conditions, i.e., in-pool environment, forced the design and development of new sensors adapted for this project. A worth mentioning related research, resulting from the BIOSWIM project, concerns new textile sensors for ECG

acquisition and respiratory frequency measurement. A comparison between electrodes based on different materials and structures was performed with promising results when performing vital signs acquisition under dry and wet conditions [5,6].

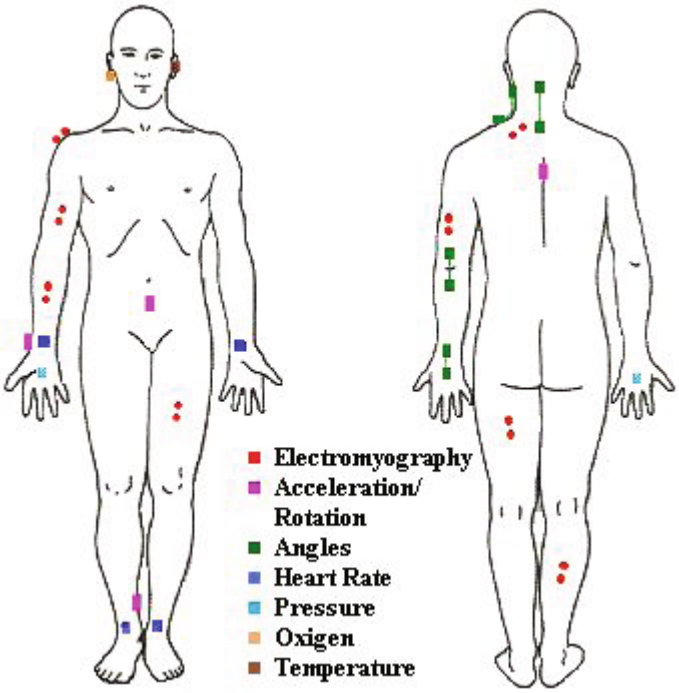


Fig. 1. BIOSWIM's monitoring objectives

This project represents a multidisciplinary effort of three research and development groups located within three well-known Portuguese institutions: the Centre of Textile Science and Technology of the University of Minho, the Institute for Systems and Computers Engineering of Porto (INESC Porto) and the Faculty of Sports of the University of Porto. The project proposal was submitted and approved by the Foundation of Science and Technology of Portugal and has been responsible for a number of academic projects and thesis at the undergraduate and graduate level, in areas such as electrical, biomedical and textile engineering. This multi-disciplinary and inter-institutional project is evidence of a growing trend in current research, which requires the collaboration of a number of specialties for achieving a common objective.

Although an immediate application of this wearable system is within sports, from competition athletes, to sport amateurs and body maintenance; the system itself can be adapted to a number of other applications. It is true that some current commercially available systems provide useful information for a number of activities, however when evaluating and improving an athlete's performance the number and variety of parameters to be monitored tends to require specialized (and therefore costly) equipment, especially at competitive levels. The overall goal for such scenarios can be at times simple, i.e. a faster free style 50 meter lap; however the movements, considerations and

adjustments an athlete must include in their actions, in order to improve their performance even a fraction of a second, can become a puzzling task. To this complex process one must add the present paradigms shift taking place within data gathering solutions. Not only the expensive laboratories and awkward equipments are being replaced by more sensible sized devices; even the biomechanical models are being updated to consider measurements (and measurement locations) that were not available (or reachable) some years ago. Beyond vital signs, other physiological and/or biomechanical signals can prove crucial for optimizing an athlete's effort and reducing fatigue when subjected to intensive training or competition challenges.

The development of wearable monitoring devices for sports practitioners can present several noticeable repercussions in the sport community, both in reference to the optimization of the training process of elite athletes, as well as to the promotion of safety regulations within rehabilitation and leisure activities. The idea behind the BIOSWIM project was to produce an wearable elite training evaluation station, allowing the most innovative monitoring, ambulatory registration, real-time visualization and post-exercise display of both physiological and biomechanical relevant data for training (heart-rate, respiratory frequency, oesophageal temperature, sweat, arm tri-axial acceleration, etc.). Important repercussions are expected both for practitioners, coaches and scientists, allowing increased safety in physical activity, an augmented objectivity - and efficiency - of the elite training process, and stress-free data collection for scientific research in neuro-physiology and biomechanics of sport.

3 Archictecture Overview

Three main sections are to be considered when referring to monitoring systems:

- Sensing section.
- Processing section.
- Transmitting section.

The mentioned sections can be found separate, intermixed or integrated depending on the design, but the objectives of their functions can be readily separated if needs be.

Sensing Section. Advances in micro-fluidics, material science, nano-structures, micro-electromechanical devices, bioelectrical interfaces, and others; have contributed to a new generation of wearable and implantable sensors and monitoring devices. Healthcare has greatly benefitted from the development of biosensors (also referred to as chemical sensors) [7] and physiological sensors. Such achievements have paved the way for truly pervasive monitoring strategies, which will benefit patients and reduce the load to healthcare facilities. From a sport monitoring perspective, non-invasive, minimally intrusive sensors are the preferred choice, and consideration of their positioning, calibration, noise, offset, deviation, etc., are concerns [8]. There exist a wide array of commercially available sensors and even more experimental devices and concepts waiting their turn.

Processing Section. In today's market, the competition to claim to be the lowest powered microcontroller is fierce. Depending of the complexity required by the application and the feature extraction methods to be applied, an array of Reduced Instruction Set Computer (RISC) or Advanced RISC Machine (ARM) architecture based microcontrollers offer different features which accommodate varying solutions. Based on the popular "motes" designs and further research in the area (internal reports [9]), there seems to be a preference for Texas Instrument MSP430 ultra-low power, Atmel's ATMEGA ultra-low power, and the Microchip's extreme-low-power (XLP) PIC microcontrollers. Using the power specifications, indicated on the datasheet of each microcontroller, as a base for comparison, can sometimes lead to problems and confusion; careful attention must be paid to the conditions in which each manufacturer measures their devices power consumption.

Transmitting Section. When referring to wearable monitoring systems, it is unavoidable to consider a wireless component for interfacing with the system; either be it for real-time or sporadic updating to a remote processing node, or for downloading the collected stored data, or even for transmitting the data from a sensor node to the on-body or remote processing unit. The presence of cables or the need for physical removal of the device for data download represents an alternative that while permissible at prototyping and troubleshooting stages, is impractical at more advance stages of design and implementation.

A number of alternatives exist for mid-range wireless communication including common protocols (GSM, WiMAX, UMTS, WLAN, etc.) and upcoming 4G mobile communication solutions. From a more local point of view the IEEE 802.15 Workgroup has introduced an array of solutions. Among the favorite standards one counts with the IEEE 802.15.1, known as Bluetooth, and the IEEE 802.15.4, also referred to as Zigbee. The number of low-power short-range transceivers in the market today is enough to overwhelm even experienced researchers. It seems every brand offers their particular RF solution, claiming low-power transmission; companies such as Texas Instrument, Atmel, Semtech, Maxim and Microchip (to mention a few), offer interesting and varying solutions.

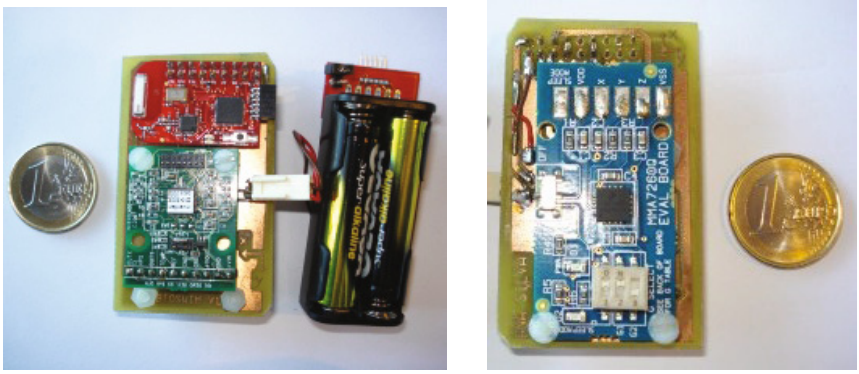
4 WIMU

For years, numerous devices and setups have been implemented in order to assist on swimming performance analysis. Many of these devices were based on video analysis, while others made direct measurement and signal capturing through awkward setups, generally uncomfortable for the swimmers and thus affecting their performance. Advances in a number of fields have allowed for compact wearable monitoring devices, greatly improving the data gathering process and closing the gap for a truly seamless biomechanical signal monitoring solution. Although there is a relatively reduced number of biomechanical signal monitoring systems being used for swimming performance analysis today (particularly when compared to the number of wearable monitoring devices for healthcare or even for land based sports), a shift on the approaches for swimming analysis is being noted. Different strategies have been applied by the mentioned

systems, however a common element seem to be their dependence on accelerometers. Some systems worth mentioning are the ETH Zurich Wearable Computing Laboratory's SwimMaster [10] and Imperial College BSN device [11].

4.1 Design Overview

The WIMU - Wearable Inertial Measurement Unit - was designed in the scope of the BIOSWIM project. The goal was to develop a MEMS-based wearable device for assessing biomechanical parameters of a swimming athlete. The inertial unit comprises a tri-axial accelerometer for linear acceleration and a bi-axial gyroscope for angular velocity measurements (however only one axes of the gyroscope was used in the current implementation stage); managed by a microcontroller for signal acquisition, packaging and wireless transmission. A photograph of the system can be observed in Fig 2. Both inertial sensors used are MEMS devices; micro-machined inertial sensors have a huge potential for applications in the biomedical field due to their small volume, portability, low power consumption and relative low cost.



(a) (b)

Fig. 2. Top 2(a) and bottom 2(b) view of the system

Commercially available evaluations boards were used to integrate the MEMS sensors into the WIMU. The accelerometer evaluation board consists of a small setup intended for evaluating the MMA7260QT Freescale Semiconductor accelerometer, adequate for fast prototype developing without the need of custom PCB designs. It also provides means for understanding the best mounting position and location of this accelerometer. Similarly, the IDG-300 Invensense gyroscope can be found integrated in an evaluation board along with the components necessary for application-ready functionality.

In order to integrate both the accelerometer and the gyroscope into the inertial unit, the Texas Instrument microcontroller based board, eZ430-RF2500, was selected from the many commercially available devices. The eZ430-RF2500 uses the MSP430F2274 microcontroller which combines a 16-MIPS performance with a 200 Ksps 10-bit analog-to-digital converter (ADC); paired with the CC2500 multi-channel radio-frequency (RF) transceiver, designed for low-power wireless applications. This board offers a

combination of hardware and software appropriate for fast prototyping of wireless projects, while offering low-power consumption. The eZ430-RF2500 can readily use the SimpliciTI wireless communication protocol, a low-power RF (2.4 GHz) protocol aimed for simple and small RF networks (proprietary of Texas Instruments). The SimpliciTI protocol claims to use a minimal set of microcontroller requirements, which in theory lowers the associated system cost. In this case, the network topology was configured so that the WIMU behaved as an end device (ED), transmitting data packets to a remote or base station. There is room for future expansion, since the protocol permits multiple end devices; therefore multiple WIMUs could be allocated at different body segments, in a truly body sensor network scheme.

4.2 WIMU Architecture

As mentioned above, the WIMU consists of a tri-axial accelerometer, a bi-axial gyroscope, a microcontroller based board and a power supply unit. However only one of the gyroscope axes was utilized during this initial stage of implementation. The architecture of the WIMU is presented in Fig. 3.

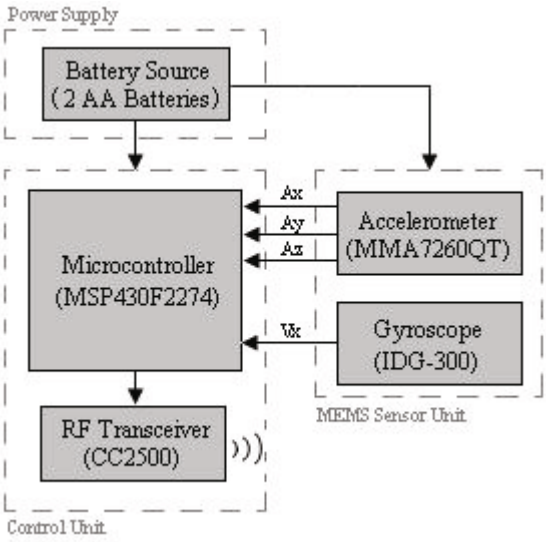
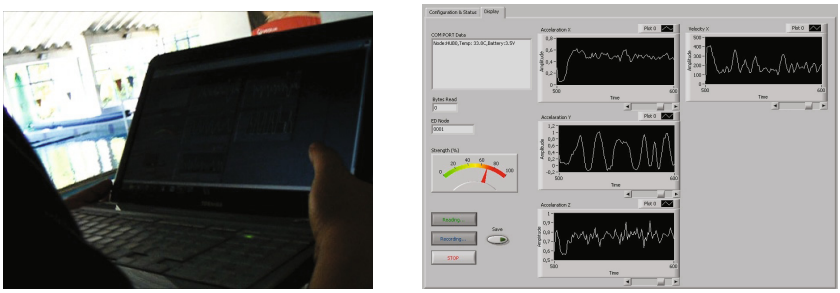


Fig. 3. WIMU's architecture

On the WIMU side, minimal pre-processing of the gathered data occurs, just enough to prepare the data for adequate packaging for transmission; however it should be pointed out that there is a potential for future signal conditioning at this stage. The acceleration and angular velocity signals are acquired and converted sequentially by the 10-bit ADC integrated in the MSP430F2274 micro-controller at a sampling rate of approximately 50 Ksps. The ADC was kept in a continuous loop, converting sequentially all sensor inputs. During the initial stage of implementation and testing, a simple broadcasting (from the WIMU point of view) time-window listening (from the receiver

point of view) scheme was used for the data gathering. Although this strategy introduces reduced timing errors and is not applicable for faster data packects transmission rates or for multi-location sensors synchronization, it proved to be effective for this stage of implementation providing a modest maximum rate of approximately 20 rpps (received packets per second). A more robust strategy which relies on time-stamping and buffer optimization packaging has since then been implemented showing stable data flow at around 100 rpps; however further testing on more realistic scenarios is still pending. Additional data processing operations at the WIMU level were left for post-processing analysis, in order to alleviate microcontroller resources and speed up the communication process. Careful attention should be paid to the receiving protocol data handling, since bottlenecks can occurr due inefficient procedures, such as time consuming conversions that could be performed in software instead of hardware). These types of issues do not become readily evident until attempting higher data packets transmission rates with more complex data packets.

The receiver or base station is composed by a laptop (see Fig. 4(a)), eZ430-RF2500 setup counterpart, and a custom software application for the acquisition, processing and visualization of the sensor packets coming from the WIMU. National Instruments' LabVIEW platform was used for the computer interface implementation as shown on Fig. 4(b). The application allows for COM port parameters configuration, swimmer's personal information and note taking recording, database storing for data, and visualization of incoming data and signal strength from the WIMU.



(a) (b)

Fig. 4. Data collection on the laptop through LabVIEWs user interface

The WIMU weights approximately 65.6 grams and measures 57x90.5x24 mm. The data collection was achieved at distances of up to 20 meters; however optimal signal strength was noticed after exceeding a distance of 10 meters depending of water conditions (agitation of the surface and existing obstacles). Two standard alkaline AA batteries insured the system could operate continuously during a couple of consecutive training sessions each of about two hours without any noticeable degrade in performance or signal intensity loss.

5 Experimental Results

The WIMU was located at the dorsal zone of the frontal plane, within the vertebral region at the inferior scapular section. This location was chosen in order to measure, in addition to the accelerations on the three axes, the longitudinal rotation of the trunk, as can be seen on Fig. 5.



(a) (b)

Fig. 5. (a) WIMU positioned at the upper back of the athlete. (b) The athlete swimming with the WIMU.

The longitudinal rotation is considered by some [12] as a useful factor when evaluating swimming performance, due to its correlation to the displacement velocity. Another important consideration was the comfort of the swimmer and avoiding restricting or modifying their average sequence of movements.

A number of preliminary tests were perform in laboratory settings, in order to optimize transmission protocols, verify waterproofness of the package and streamline data gathering protocols, both at the hardware and software level. A female athlete in her late teens served as the main tester at the initial stages. Before entering the pool, the swimmer was required to perform a number of flexibility related routines, in order to determine movement constraints. Once in the pool, she was asked to swim, submerge and perform various movements in order to determine if the WIMU's presence represented an obstruction to her movements. In both cases (outside and inside the pool) the swimmer reported that the unit did not affect her movements. Finally, the swimmer was told to complete several sets of laps using free-style (i.e. crawl) technique, then a number of laps with breaststroke, and finally butterfly. For all styles indicated the end of pool, the turn was perform through a vertical turnaround (i.e., stop-touch wall-turnaround), reversing direction without flipping under water. A second female swimmer with similar age, height and swimming experience collaborated in a comparison trial, however only free-style data was capture in such occasion.

Although data gathering for the crawl techniques laps did not required signal compensation strategies; the butterfly and breaststroke techniques did present some data gaps. The recorded data was later processed in Matlab applying compensation through cubic interpolation when required, followed by a simple moving average smoother

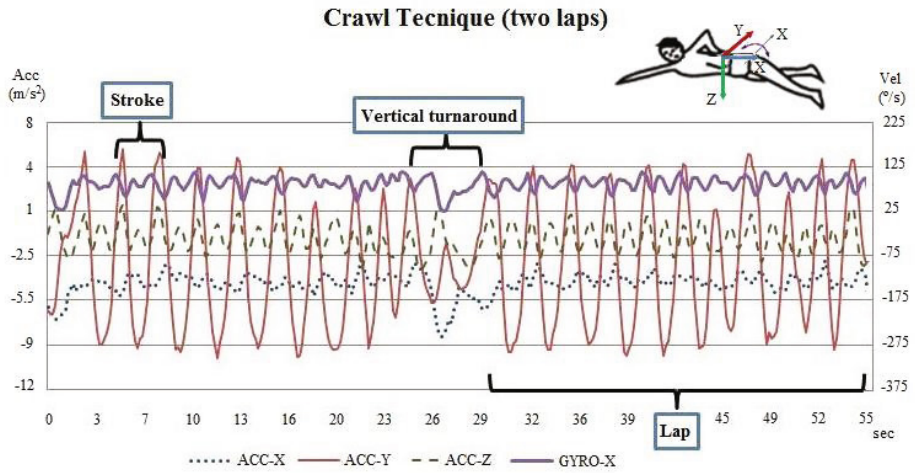


Fig. 6. Two laps using crawl technique signals. Gyroscope X-axis is located within the 25 and 125 °/s Vel lines.

based on window convolution; some results taken at a rate of approximately 7 rpps can be observed on Fig.(s) 6, 7 and 8.

It should be mentioned that the Fig.(s) 6 and 7 have a vertical scale in m/s^2 and °/s (for linear acceleration and angular velocity, respectively) which were obtained through calculations based on manufactures specifications. On Fig. 8, the different signals were computationally DC filtered and scaled for ease of comparison. All the mentioned figures have a horizontal scale in seconds.

Observing the graphs concerning each swimming technique, it was possible to differentiate between styles, each stroke, and start/end of lap. For example, free style can be readily distinguish from breaststroke and butterfly by the signal provided by the accelerometer in Y-axis or transverse axis. The referred signal presents large variations for the crawl technique, while for the other presented styles these variations are comparatively small. Alternatively, the butterfly and breaststroke technique can be differentiated from each other from the data produced by the accelerometers in the X and Z axes, or the longitudinal and anteroposterior axis respectively. At this point in time these observations are made from a qualitative point of view, since precise adjustments for reference compensation are still under study.

Fig. 8 shows the strokes of two female swimmers with similar age, weight, height, and swimming experience, offering a comparative view of an equivalent frame of time during a non-competitive (or relaxed) lap. Although both swimmers had similar lap times (close to 30 seconds for both) differing behavior can be observe. The visual perception of the swimmers during the measurements concurs with them presenting different styles. While swimmer one (1) had a calm, long and steady stroke, swimmer two (2) had a more intense and quick stroke.

It is still early in this project to produce conclusive results regarding feature extraction for performance analysis, and there is a clear need for accumulating data of diverse swimmers of different genders and competitive level; however these first steps

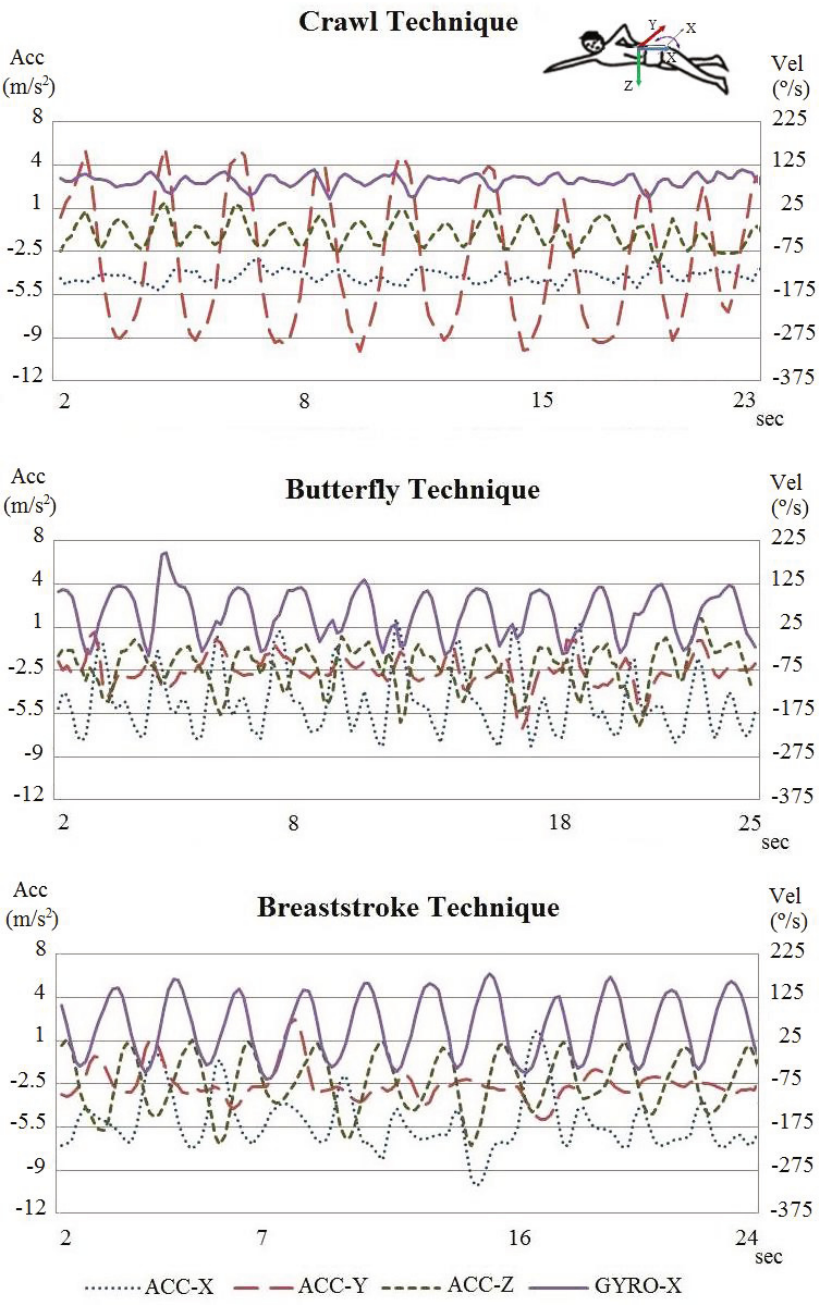


Fig. 7. Equivalent lap section captured signals for each performed technique

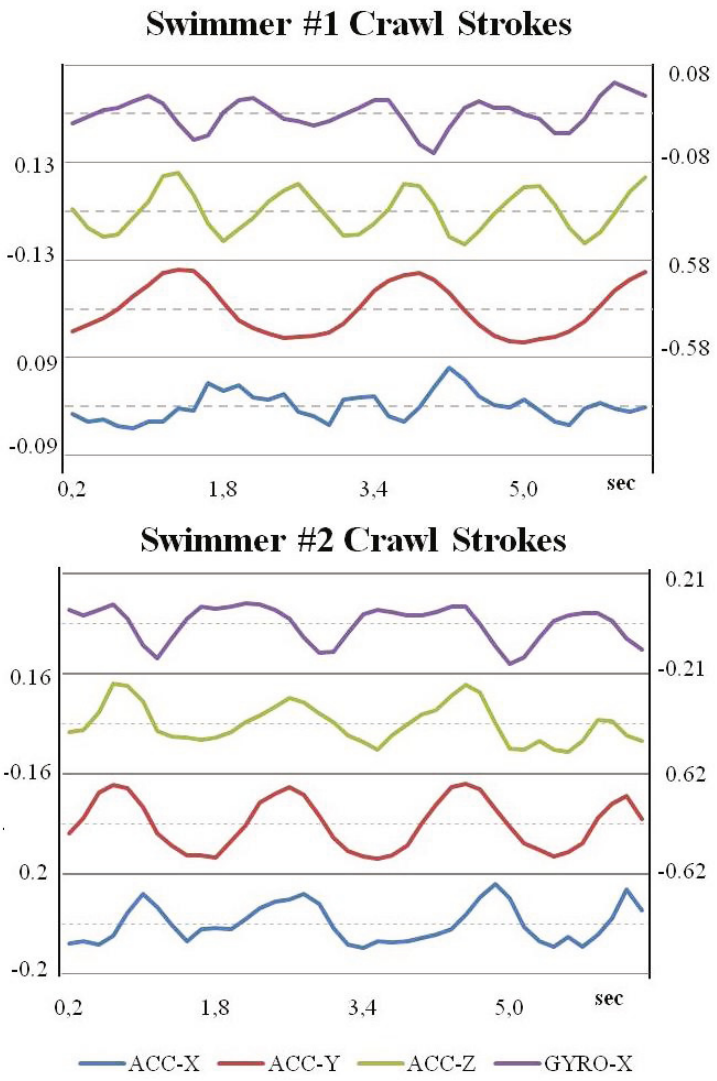


Fig. 8. Two swimmer strokes comparison. Signals from both swimmers correspond from top to bottom to: Gyro-X, Acc-Z, Acc-Y, Acc-X. Vertical axes have been scaled for ease of comparison.

seem quite promising. Several challenges present themselves when considering the data acquired and the references of the setup; in particular when considering the mobile nature of the individual, the curvature of the back, the inherit offset present in some sensors and possible error introduced by the acquisition chain. A more comprehensive analysis of the data gathered up to now and future collections, as well as feature extraction strategies for performance analysis will be presented in future works. In a parallel effort, biomechanical models are being developed that utilize the data in order to take advantage of their specificities. Although not its original objective, the WIMU has been used to capture signals from lower and upper limbs in a number of varying

activities, mostly for testing purposes. A particular case study of interest was the use of the WIMU during rehabilitation sessions for cases of cerebral stroke. The patients in such a case were undergoing rehabilitation terapy for upper limb movement restauration. The objective is the development and implementation of a home rehabilitation exercise assistant, utilizing the WIMU's signals for real-time audio-visual feedback.

6 Conclusions

In this paper, we present a prototype for swimming performance analysis monitoring based on accelerometers and gyroscopes, referred to as WIMU. The WIMU is one of the objectives of a much more ambitious multi-disciplinary effort of a group of Portuguese research and development teams, known as the BIOSWIM project. Such project seeks to characterize swimming through physiological and biomechanical signal capturing at points distributed throughout the athlete's body. Although the device presented is still at the prototype stage of development, it was capable of providing promising data under in-pool normal conditions. The current version of the WIMU serves as a basis for future implementations that will focus on wearability, energy harvesting, and integration within the BIOSWIM's project swimsuits.

Acknowledgements. The authors would like to acknowledge the Foundation for Science and Technology of Portugal for their support of the BIOSWIM project (PTDC/EEA-ELC/70803/2006) and some of the PhD students involved in this article (SFRH/BD/61396/2009 and SFRH/BD/60929 /2009). Additionally the authors would like to acknowledge the contribution of Barbara Mota and Catarina Araujo, main testing swimmers.

References

1. Pacelli, M., Loriga, G., Taccini, N., Paradiso, R.: Sensing Fabrics for Monitoring Physiological and Biomechanical Variables: E-textile solutions. In: 3rd IEEE/EMBS International Summer School on Medical Devices and Biosensors (2006)
2. Noury, N., Dittmar, A., Corroy, C., Baghai, R., Weber, J.L., Blanc, D., Klefstat, F., Blinovska, A., Vaysse, S., Comet, B.: VTAMN - A Smart Clothe for Ambulatory Remote Monitoring of Physiological Parameters and Activity. In: 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEMBS 2004 (2004)
3. Coyle, S., et al.: BIOTEX - Biosensing Textiles for Personalised Healthcare Management. IEEE Transactions on Information Technology in Biomedicine 14(2), 364–370 (2010)
4. Cunha, J.P., Cunha, B., Pereira, A.S., Xavier, W., Ferreira, N., Meireles, L.: Vital-Jacket: A wearable wireless vital signs monitor for patients' mobility in cardiology and sports. In: 4th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth (2010)
5. Silva, M., Catarino, A.P., Carvalho, H., Rocha, A., Monteiro, J.L., Montagna, G.: Textile sensors for ECG and respiratory frequency on swimsuits. In: Intelligent Textiles and Mass Customisation International Conference, Casablanca, Morocco, pp. 301–310 (2009)
6. Silva, M., Catarino, A.P., Carvalho, H., Montagna, G., Monteiro, J.L., Rocha, A.: Study of vital sign monitoring with textile sensors in swimming pool environment. In: Annual Conference of the IEEE Industrial Electronics Society, Porto, Portugal, vol. 35 (2009)

7. Patel, B., Anastassiou, C., O'Hare, D.: Biosensor Design and Interfacing. In: Yang, G.Z. (ed.) Body Sensor Networks. Springer, London (2006)
8. Yang, L., Shouqian, S.: Smart Sport Underwear Design. In: 7th International Conference on Computer Aided Industrial Design & Conceptual Design, pp. 1–3 (2010)
9. Salazar, A.J.: Wearable Monitoring Sports Systems. Technical report, Doctoral Programme in Electrical and Computer Engineering. Faculdade de Engenharia da Universidade do Porto. Porto, Portugal (2010)
10. Bächlin, M., Förster, K., Tröster, G.: SwimMaster: a wearable assistant for swimmer. In: 11th International Conference on Ubiquitous Computing, pp. 215–224 (2009)
11. Pansiot, J., Lo, B., Yang, G.Z.: Swimming Stroke Kinematic Analysis with BSN. In: 2010 International Conference on Body Sensor Networks, pp. 153–158 (2010)
12. Liu, Q., Hay, J.G., Andrews, J.G.: Body rolls and hand path in freestyle swimming: an experimental study. Journal of Applied Biomechanics 9, 238–253 (1993)

PART II

Bioinformatics Models, Methods and Algorithms

3D Protein Surface Segmentation through Mathematical Morphology

Virginio Cantoni, Riccardo Gatti, and Luca Lombardi

Computer Vision Lab, Department of Computer Engineering and Systems Science
University of Pavia, Via Ferrata 1, 27100, Pavia, Italy
{virginio.cantoni,riccardo.gatti,
luca.lombardi}@unipv.it
http://vision.unipv.it

Abstract. An important research activity in proteomics is the study of geometrical and topological aspects that characterize the surfaces of proteins or ligand molecules. The functionalities and the molecular interactions of an unknown protein can be, in a first approximation, inferred by comparing her geometrical shape with that of known molecules. The aim of this paper is to identify particular features of molecular 3D surfaces in a new and efficient way by using a set of mathematical morphological operators. The obtained segments constitute pockets and protrusions. Pockets are generally the concave and deepest spots were active sites take place. The protrusions set, on the other hand, is the complementary region of the pockets and therefore, can be used by docking algorithms in the study of protein-protein and protein-ligand interactions.

Keywords: Structural biology, Protein structure analysis, Protein-ligand interaction, Surface segmentation.

1 Introduction

In the last decade much work has been done on the detection and analysis of binding sites of proteins through bioinformatics tools. This is a preliminary, but important step that can reduced consistently the cost, in time and resources, of the subsequent experimental validation phase, which is applied only to the resulting subset loci. It is worth to point out that the effective identification of active sites is instrumental for structure-based drug development and design.

The various approaches proposed up-to-now are characterized by the solution of two subproblems: the protein representation and the matching strategies. Among the techniques that have been proposed up to now, we can quote: the geometric hashing of triangles of points on the SES and their associated physico-chemical properties [12]; a representation of the SES in terms of spherical harmonic coefficients [10]; a collection of spin-images [3]; a 'context shapes' representation [2]; a set of vertices of the triangulated solvent-accessible surface (SAS) [16].

Recently, a few packages for the process of detecting and characterizing candidate active sites are supplied on the web. The most known packages, in chronological sequence, are here shortly described.

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 97–109, 2013.
© Springer-Verlag Berlin Heidelberg 2013

The first, POCKET [13] has been developed in the early '90. The protein is mapped onto a 3D grid, and a grid point belongs to the protein if it is within 3 Å from an atom nucleus. The pockets consist of the set of grid points, in the solvent area for which a scanning along the x, y, or z-axes presents a sequence protein-solvent-protein. More recently LIGSITE [11] extends POCKET by scanning also along the four cubic diagonals (in fact, the POCKET's classification is dependent on the angle between the reference system and the protein). The solvent points that present a number of protein-solvent-protein events greater than a given threshold are classified as candidate active sites.

In the late '90 CAST [14] (updated with CASTp [2] in the early '00), based on 3D computational geometry, has been proposed. In this approach the protein is represented by a set of 3D tetrahedra having the vertices on the nucleus positions and is analyzed through convex-hull, alpha shapes and discrete flow theory. A tetrahedron having at least a facet crossing the solvent region is designated as 'empty tetrahedron'. Empty tetrahedra sharing a common triangle are grouped so 'flowing' towards neighbouring larger tetrahedra which act as sink. A pocket, which is a potential binding site, is a collection of empty tetrahedra. Pockets volume, mouth opening area and circumference are easily evaluated on this structure.

PASS [6], introduced in the early '00, is based on a purely geometrical method consisting in a sequence of steps: i) the protein surface, on the side of the solvent, is completely covered by probe spheres each one not contained in any others; ii) each probe is associated with a "burial" value, which corresponds to the number of atoms contained within a concentric sphere of radius 8 Å; iii) the probes with a "burial" value lower than a predefined threshold are eliminated; iv) the previous three steps are iterated (with step one applied only to surface's patches covered by the probes) until the regime, where no new buried probe can be added; v) a probe weight, which is dependent on the number of the neighbouring spheres and the extent to which they are buried, is assigned; vi) a shortlist of active site points (ASPs), ranked by the probe weight, is identified through the central probes that contain many spheres with high burial count.

Finally SURFNET-ConSurf [10] is based on a pocket-surface representation which combines geometrical features together with an evolutionary parameter based on the degree of conservation of the amino acids involved. Initially, through SURFNET [12], the clefts are detected by placing a sphere between all pairs of atoms such that the sphere just touches each atom of the pair, then this sphere is progressively reduced in size up to no further intersections with other atoms are present. The resulting sphere is retained only if its radius is greater than a minimum size predefined. Moreover, the regions that do not present highly conserved residues, as defined by the ConSurf-HSSP database [9], are removed, thus reshaping the cleft volumes. The remaining clefts are candidate active sites (in particular the largest ones).

In this paper we present a new method for segmentation of the SES with the aim of identifying respectively the concavities that can host a ligand or a protrusion of another protein and the protuberances that can match the inlet of others proteins. The paper is organized as follows: in section two we describe a new technique that, starting from the protein enlarged convex hull, propagates up to the SES to identify concave active sites; in section three a second backward propagation algorithm that detect the protuberances, starting from the peaks of the previous propagation is introduced. In section four a few

practical cases and some comparisons are given. Section five contains the conclusion and our near future subsequent activities.

2 Looking for Pockets and Tunnels

The details of the first half of this procedure are described in [7]. This segmentation is based on a propagation process (a Distance Transform (DT)) applied to the volume obtained subtracting the molecule to its Convex Hull (CH). Before presenting this process, a few preliminary definitions and statements are given.

The CH of a molecule is the smallest convex polyhedron that contains the molecule points. In R^3 the CH is constituted by a set of facets, that are triangles, and a set of ridges (boundary elements) that are edges. A practical O(n log n) algorithm for general dimensions CH computing is Quickhull [1], that uses less memory space and executes faster than most of the other algorithms (Fig. 1).

The protein and the CH are defined in a cubic grid V of dimension L x M x N voxels. Note that the grid is extended one voxel beyond the minimum and maximum coordinate of the SES in each orthogonal direction (in this way both SES and CH borders are always inside the V border). The voxel resolution adopted is 0.25 Å, so as to be small enough to ensure that, with the used radii in biomolecules atoms, any concave depression or convex protrusion is represented by at least one voxel.

Let us call R the region between the CH and the SES (the concavity volume [5]), that is:

$$R = CH \cap \overline{SES} \tag{1}$$

Let us call B_{CH} the set of the border voxels of CH, that is:

$$B_{CH} = CH - [CH \bullet K] \tag{2}$$

where $\bullet$ represent the erosion operator of mathematical morphology [15] and K the discrete unitarian sphere (in the discrete space, in 26 connectivity, a 3x3x3 cube).

Within the region R the following propagation is applied:

$$D_i = \left\{ \begin{matrix} 1\ if\ i \in B_{CH} \\ 0\ otherwise \end{matrix} \right\}$$
$A = B_{CH}$;
$N = (A \oplus K) \cap R$;
$E = N - A$;
while $E \neq \oslash$ *do*
    $\forall e \in E : d_e = min_{n \in n_e}(d_n + w_n)$;
    $A = N$;
    $N = (A \oplus K) \cap R$;
    $E = N - A$;
done

where:

- A represents the increasing set of voxels contained in R; E corresponds to the recruited set of near neighbors of A contained in R (i.e. the voxels reached by the last propagation step);
- $min_{n \in n_e}(d_n + w_n)$; is the minimum value among the distances d_e in the near neighbors belonging to D already defined, incremented by the displacement w_j between the locations (e, n): that is, if e and n have a common face $w_n = 1$; if e and n have a common edge $w_n = \sqrt{2}$; if e and n have a common vertex $w_n = \sqrt{3}$. In three dimensions, the total number of the near neighbor elements of p is 26: six of them that share one face and have distance equal to 1 from the voxel p, twelve neighbors that share only an edge and are at distance $\sqrt{2}$, and eight that share only a vertex and are at distance $\sqrt{3}$ always from voxel p. At each iteration, new voxels, inside R, are reached by the propagation process and the value they take is determined by the neighbor distance (from the convex hull) and the voxels distance from the neighbor involved; this in order to simulate an isotropic propagation process and the proper distance evaluation;
- $E = \oslash$ corresponds to the regime condition: no other changes are given and the connected component of R, adjacent to the border B_{CH}, is completely covered.

The values in D represent the distance of each voxel of A from B_{CH} and A corresponds to the connected component of R adjacent to the border. Having A, it is possible to easily identify and eventually remove the cavities C, that are the volumes completely enclosed in the macromolecule M:

$$C = CH - A - M \tag{3}$$

In order to separate the different pockets and tunnels the volume A must be partitioned into a set of disjoint segments $P_{SES} = \{P_1, \ldots, P_j, \ldots, P_N\}$, where N is the number of inlets. The partition must satisfy the following condition:

$$P_i \cap P_j = \oslash, i \neq j \tag{4}$$

$$P_1 \cup \ldots \cup P_j \cup \ldots \cup P_N = A \tag{5}$$

As can be easily realized, starting from the total set of convex hull facets, several waves are generated and propagation proceeds up to the complete coverage of the volume A: the connected component of R adjacent to the border. During the propagation phase two sets of salient points are identified: *local tops* LT and *wave convergence* WC points.

The LT set is exploited for the segmentation process. The cardinality of LT corresponds to N_{max} the maximum number of segments/inlets that can be considered. The effective number of segments, that determines obviously the number and the morphology of pockets and tunnels, is found out on the basis of two heuristic parameters: i) the minimum travel depth value of the local tops TD_{LT}; ii) an evaluation of the near tops pivoting effects PEs. The threshold TD_{LT} is introduced because the surface's irregularities and the digitalization process produce small irrelevant spurious cavities. Two thresholds are introduced on the basis of PEs taking into account morphological aspects

insight important cavities: the nearness of others, more significant, local tops (τ_1) and the relative values of the local-top travel-distance (τ_2). A detailed description of this first phase is given in [7].

The second phase is completely new. Let us assume that a pocket has at least the volume of a water molecule. Under this assumption we will identify the useful portion of A by:

$$B = A \bullet S_w \tag{6}$$

Where $\bullet$ represent the erosion operator and S_w the minimum sphere that contains a water molecule. The set of voxel E_A given in 7 constitutes the fine grains with a too strong concavity (Fig. 2):

$$E_A = A - B \tag{7}$$

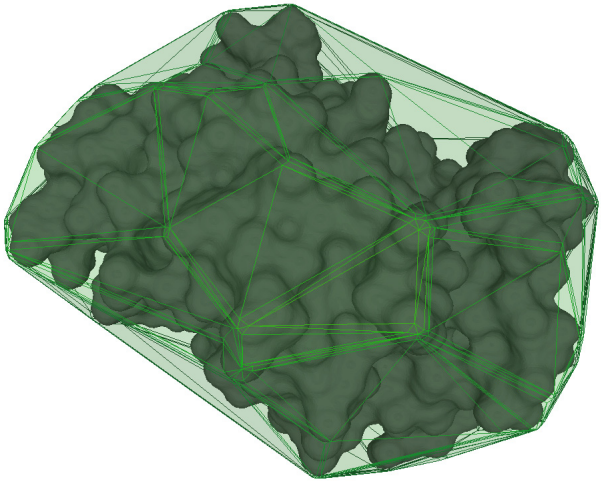


Fig. 1. SES of PDB ID 1MK5 and its extended CH

The results achieved in this phase are shown in figure 3. In general, the set LT is contained in the set E_A. Using as seed-points the set LT, a back-propagation toward B_ch is performed. At each step the new connected voxels having the same distance (d_e) are joined to the seed set ST. A candidate pocket is established active when ST maintains at each step at least a new voxel belonging to B.

During the propagation towards B_ch , if the new joined set of voxels is completely contained in E_A a bottleneck has been reached and the seed set ST in progress is anymore active. When two or more seed sets converge there are three possible cases:

- all the convergent sets are active: in this case a new active set is generated on the basis of the union of the new entry voxels;
- among the convergent sets there is at least one active set: this set continues the propagation (if there is more than one active set it is first applied the case i) recursively) including the convergent new entry voxels of the connected not active set(s);

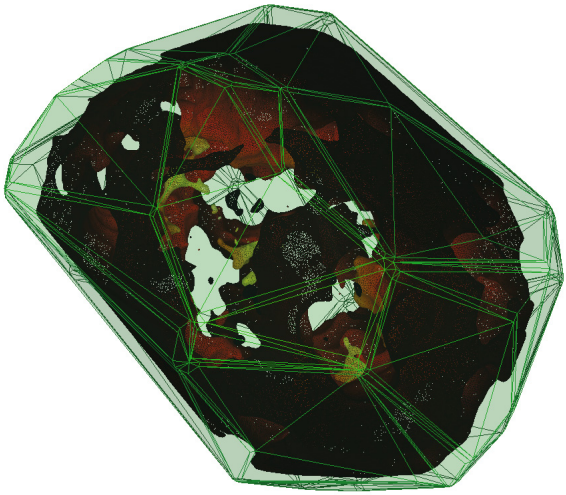


Fig. 2. The set E_A of PDB ID 1MK5

– all the convergent sets are not active: this means that some fine grains are joining together, and a new propagation seed composed by the union of the entry voxels is established if this union achieves a volume at least equal to the water molecule.

At the end of this operation, a consecutive and unique id number, is given at each pocket found, starting from the deepest one. In figure 3 it is shown, for the protein 1MK5, a 2D sketch representing the final result of this process.

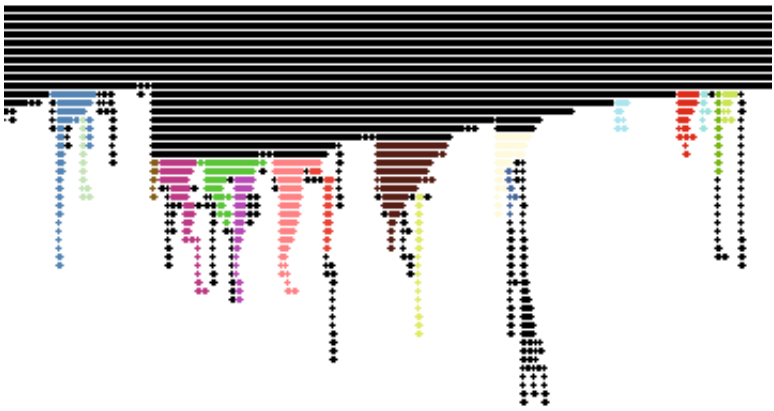


Fig. 3. A portion of the 2D sketch achieved by applying the pocket search algorithm on the protein 1MK5. The vertical clusters that are not associated to any pockets are black, meanwhile the ones representing pockets are colored.

2.1 Pockets Refinement

Each pocket is mainly characterized by a *local top*, a *mouth* and a *travel depth*. Each pocket is generally composed by a union of connected components inside the concavity volume.

The *local top* of a pocket is the deepest voxel which has the greatest travel depth from the nearest face of the protein's convex hull. In case that the local top is not a single voxel but is extended for more than one voxel, the local top is the central voxel of this connected component. The *mouth* is the set of adjacent voxels of the pocket closer to the convex hull. The *travel depth* (TD) is the total length of the pocket calculated as the difference between the mouth depth and the local top depth.

The pocket found after performing all the steps of chapter 2 have a particular characterization: each pocket covers a volume that ends to a wave convergence with another pocket. In this way it is assured that all reachable parts of a protein are associated with one pocket. However there are cases taht may require a refinement through a back propagation process. The target of the pocket refinement process is a precise volume evaluation, mantaining a good processing speed. The algorithm is the following:

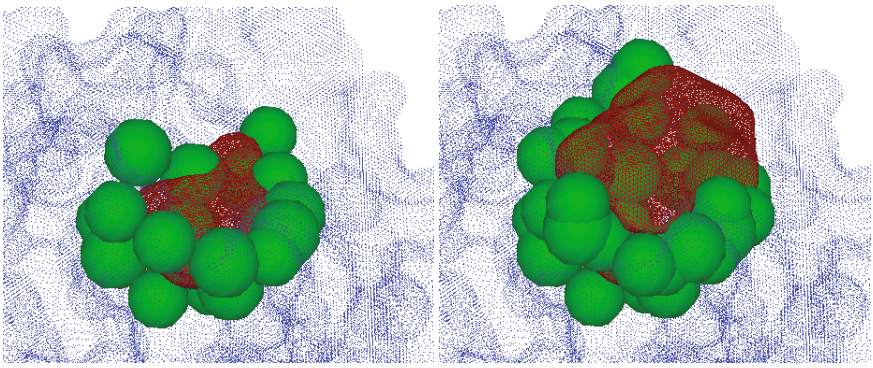
$$
\begin{aligned}
&d_C = d_{MAX} \\
&while\ d_C > d_{MIN}\ do \\
&\quad if(d_{M_i} = d_C)\ then \\
&\qquad V_{P_i} = V_{P_i} + \textstyle\sum v \in (M_{P_i} \circ K) : d_V \notin V_{P_k} \\
&\qquad M_{P_i} = \textstyle\sum((M_{P_i} \circ K) - M_{P_i}) : d_V > d_C \\
&\quad fi \\
&\quad d_C = d_C - 1 \\
&done
\end{aligned}
$$

where, at each iteration:

- d_C is the distance evaluated that range from d_{MAX} (the distance of the deepest local top of all the protein's pockets) to d_{MIN} (the distance of the nearest mouth of all protein's pockets);
- V_{P_i} and M_{P_i} represents respectively the total volume and the mouth surface of pocket i;
- d_V the distance between the voxel v and the protein's convex hull;
- K the discrete unitarian sphere (3x3x3).

Starting from the local tops, at each step the routine binds to a pocket an extra level of connected voxels by a dilation operation. When two or more pockets converge after the dilation operation the rule of "deepest pocket wins" is used. Notice that no shape descriptors are needed for the mouth evaluation. In fact all the connected components left in the concavity volume that are not associated with any pocket, are already bound to be compatible to the probe molecule thanks to the previous pockets search operations.

All the pockets generated fill the concavity volume. Pockets that have sides in common can also be joined together to form a unique big pocket. This join operation is very useful for protein-protein interactions where large areas of proteins are involved in the binding phase. A result of this refinement process is shown in figure 4.



(a) (b)

Fig. 4. Detail of pocket id 6 before (a) and after (b) the operation of refinement. In green are represented the atoms adjacent to the pocket surface (in red).

3 Looking for Protuberances

The objective here is to segment the SES to underline the protuberances. There is a duality relationship between this process and the previous one: pockets are SES segments mainly concave, meanwhile protuberances are mainly convex. Let us assume that protuberances we are looking for having a known maximum section area contained in a circle of radius r. Under this assumption we will identify a basis volume F by:

$$F = SES \bullet S_r \tag{8}$$

where $\bullet$ represents the erosion operator [15] and S_r the sphere of radius r. Let us call S_F the external surface of F. The set of voxel G given in 9 constitutes the working volume (Fig. 5) for our analysis:

$$G = SES - F \tag{9}$$

Within the region G in a similar way of the previous pocket search, the following propagation is applied:

$$D_i = \left\{ \begin{matrix} 1 \; if \; i \in S_F \\ 0 \; otherwise \end{matrix} \right\}$$
$N = (A \oplus K) \cap G;$
$E = N - A;$
$while \; E \neq \oslash \; do$
$\quad \forall e \in E : d_e = min_{n \in n_e}(d_n + w_n);$
$\quad A = N;$
$\quad N = (A \oplus K) \cap G;$
$\quad E = N - A;$
$done$

where:

- A represents the increasing set of voxels contained in G;
- E corresponds to the recruited set of near neighbors of A contained in G (i.e. the voxels reached by the last propagation step);
- the values in D represent the distance of each voxel of A from S_F.

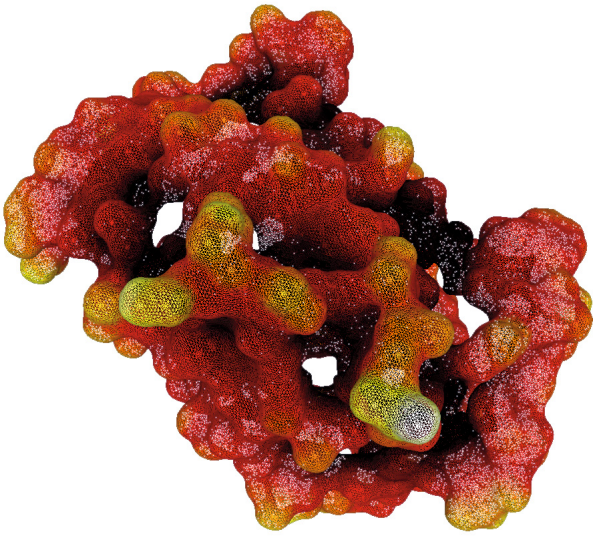


Fig. 5. Volume G for PDB ID 1MK5 build with a sphere S_r of radius 2.4 Å

Starting from the S_F's voxels (to which a common label is assigned), a new recursive scanning phase within G is applied, going toward SES. At each step, the new entry voxels are segmented in connected sets. When there is a one-to-one correspondence between a new segment set and a set of the previous step, its label is extended to the new segment. When a previous set is split in two or more segment sets a new label is generated for each one of them.

As in the propagation process for the search of pockets, during the propagation phase, two sets of salient points are identified: local tops LT and wave convergence WC points. Both these salient points are important for the docking analysis. A 2D sketch representing the final result is given in figure 6.

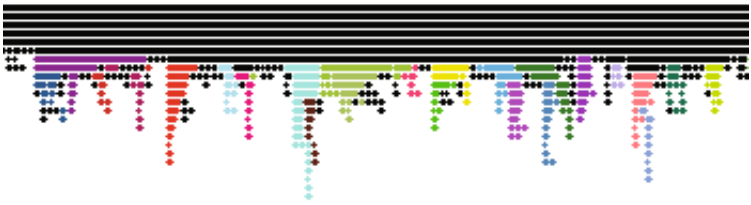


Fig. 6. A portion of the 2D sketch achieved by applying the protuberances search algorithm on the protein 1MK5. The vertical clusters that are not associated to any pockets are black, meanwhile the ones representing protuberances are colored.

4 Experiments and Comparisons

As an example, the proposed technique has been applied to the Apostreptavidin Wild-type Core-Streptavidin with Biotin structure (PDB ID: 1MK5). The analysis has been done with a resolution of 0.25 Å, which entails a van der Waals radius of more than five voxels to the smallest represented atoms. The SES is obtained from the van der Waals surface, after the execution of a closure operator, using a sphere with radius of 1.4 Å, approximately 6 voxels (corresponding to the conventional size of a water molecule), as structural element.

For what concerns the pockets detection the three parameters have been set as follows: the minimum travel depth of the local tops to TD_{LT} = 5 voxels; the nearness of others, more significant, local tops to τ_1= 150 voxels and the relative values of the local-top travel-distance to τ_2 = 2000 voxels. Moreover, the radius of the water molecule has been set to 6 voxels.

In figure 7 it is shown the final result of the segmentation process of the protein 1MK5 for the detection of pockets and tunnels. Note that among the 25 pockets that have been detected, 2 have a volume greater than 80 water molecules and have a travel depth of 26 voxels and mouth aperture of 8.343 and 30.547 respectively.

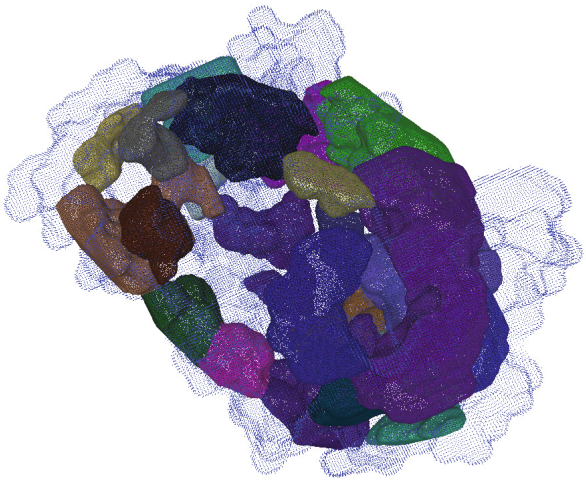


Fig. 7. Final result of the segmentation process of PDB ID 1MK5 for the detection of pockets

Figure 8 shows the final result of the segmentation process of the protein 1MK5 for the detection of protuberances. The parameter r representing the maximum section area has been set to 800 voxels. Note that among the 41 protuberances that have been detected 9 have a volume greater than 7 water molecules, and a base aperture of 22.968, 8.236, 16.000, 9.287, 8.444, 9.008, 7.469, 11.148, 9.684 voxels respectively.

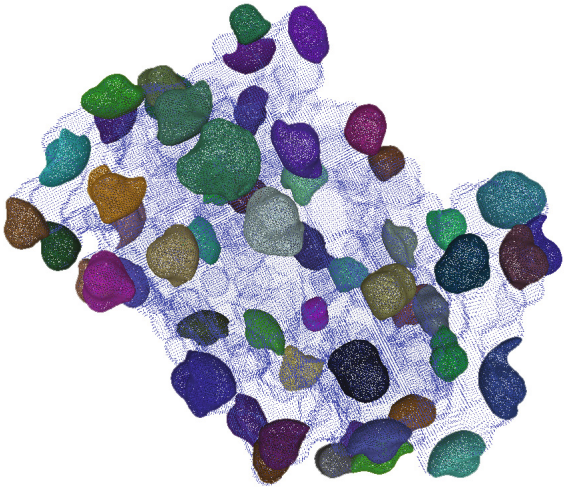


Fig. 8. Final result of the segmentation process of PDB ID 1MK5 for the detection of protuberances

All the packages quoted in this paper's introduction are related to just pockets and tunnels detection, at the authors knowledge there are not packages specialized for segmenting the SES on the basis of protuberances. Among the quoted packages the only one that was available and directly applicable has been CASTp. We compared the results obtained with this package to our one. The number of pocket selected has been 29 and 25 respectively for CASTp and our own. Let us first point out that all the main pockets of the quoted protein (the bigger and deeper ones) are detected in both cases. Moreover, for the main pockets, almost the same set atoms at the border of the SES delimiting the pockets. Nevertheless, in general the number of these atoms is higher in our solution (up to 20% in a few cases) and seems to cover in a complete way the pocket concavity. An example of this case is given in figure 9.

The results differ more for what concerns the smallest pockets. This is due to the thresholds to accept the concavity with a short travel depth as a possible active site. We have two thresholds on the basis of the travel depth and of a minimum concavity volume. Generally speaking CASTp accept more small concavities as pockets, but sometimes there are cases in which our volume constraint is satisfied and the concavity is not accepted by CASTp. This must not be a critical issues because [14] the binding sites are usually the pockets having the greatest volume. While CASTp includes empty volume internal to the protein, in our approach these are identified but not classified as pockets.

Referring to computational performance, our algorithm runs on an Intel Q6600 Processor with 4 GB of Ram. The analysis of pockets and protuberances on 1MK5 protein as been done in 58 seconds starting from the PDB file (this include the operations of creating the 3D matrix, the Convex Hull, all mathematical morphology operations, the triangulation of the voxels surface of each pocket/protuberance with a marching cube/mesh smoothing algorithm and so on). In fact besides the segmentation process for each detected segment (pocket or protuberance) a rich parameter set is computed to

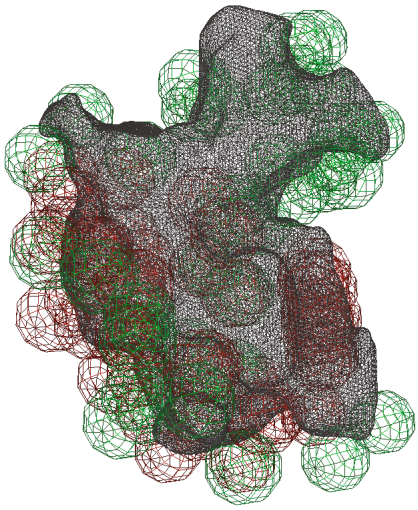


Fig. 9. Wireframe of the main binding site of PDB ID 1MK5. In red atoms detected by CASTp. Our software detects both the red and green atoms.

guide the analysis of possible match, such as volume, surface to volume ratio, mouth (base) aperture, travel depth, and many others (a full list is given in [8]). It is important to note that all the algorithms presented in this paper are already thought to be simply implemented into parallel architectures.

5 Conclusions

In this paper we present a new approach for the segmentation of SES of proteins in order to support the search of dual active sites (i.e. pockets and protuberances) that can be morphologically arranged together. This is a preliminary step for important structural biology application. The results achieved look very promising and in comparison to others solutions presented in literature it seems to add something not only from the computational point of view. Now we have started an extensive experimentation phase to validate our solution from the best practice point of view.

References

1. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The Quickhull Algorithm for Convex Hull. ACM Transactions on Mathematical Software 22(4), 469–483 (1996)
2. Binkowski, A.T., Naghibzadeh, S., Liang, J.: Castp: Computed atlas of surface topography of proteins. Nucl. Acids Res. 31(13), 3352–3355 (2003)
3. Bock, M.E., Garutti, C., Guerra, C.: Effective labeling of molecular surface points for cavity detection and location of putative binding sites. In: Proc. of CSB, San Diego, vol. 6, pp. 263–744 (2007)
4. Bock, M.E., Garutti, C., Guerra, C.: Cavity detection and matching for binding site recognition. Theoretical Computer Science 408, 151–162 (2008)

5. Borgefors, G., Sanniti di Baja, G.: Analyzing Nonconvex 2D and 3D Patterns. Computer Vision and Image Understanding 63(1), 145–157 (1996)
6. Brady, G.P., Stouten, P.F.W.: Fast prediction and visualization of protein binding pockets with PASS. J. Comput-Aided Mol. Des. 14, 383–401 (2000)
7. Cantoni, V., Gatti, R., Lombardi, L.: Segmentation of SES for Protein Structure Analysis. In: Proc. of the 1st International Conference on Bioinformatics, Valencia, pp. 83–89 (2010)
8. Cantoni, V., Gatti, R., Lombardi, L.: Proteins Pockets Analysis and Description. In: Proc. of the 1st International Conference on Bioinformatics, Valencia, pp. 211–216 (2010)
9. Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T., Bental, N.: The consurf-hssp database: the mapping of evolutionary conservation among homologs onto pdb structures. Proteins 58(3), 610–617 (2005)
10. Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A., Thornton, J.M.: A Method for Localizing Ligand Binding Pockets in Protein Structures. PROTEINS: Structure, Function, and Bioinformatics 62, 479–488 (2006)
11. Huang, B., Schroeder, M.: LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Structural Biology 24, 6–19 (2006)
12. Laskowski, R.A.: Surfnet: a program for visualizing molecular surfaces, cavities and intermolecular interactions. J. Mol. Graph. 13(5), 323–330 (1995)
13. Levitt, D.G., Banaszak, L.J.: Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J. Mol. Graph. 10(4), 229–234 (1992)
14. Liang, J., Edelsbrunner, H., Woodward, C.: Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. 7(9), 1884–1897 (1998)
15. Serra, J.: Image analysis and mathematical morphology. Academic Press (1983)
16. Shulman-Peleg, A., Nussinov, R., Wolfson, H.: Recognition of Functional Sites in Protein Structures. J. Mol. Biol. 339, 607–633 (2004)

Simple Tool for Semi-automated Evaluation of Yeast Colony Images

Jan Schier and Bohumil Kovář

Institute of Information Theory and Automation of the ASCR
Pod vodárenskou věží 4, CZ-182 08 Prague 8, Czech Republic
{schier,kovar}@utia.cas.cz

Abstract. A tool for semi-automated evaluation of images of yeast colonies is presented in the paper. The input data for the tool is a batch of Petri dish images, possibly in nested directories, the output is written to a text file in a simple CSV (comma-separated values) format. The tool is designed to evaluate, for each dish image, the relative coverage of the dish and the number of colonies contained in the dish. It is intended for use in a research laboratory, where a general-purpose imaging setup is used to take the images, which are then processed off-line.

In the paper, the characteristic features of the yeast colony images are described and the processing flow and the user interface of the tool are reviewed. The performance of the counting is evaluated using a test set of 245 images with typical relative coverage of less than 10%.

The tool is available for download from our web page.

1 Introduction

Yeast, namely *Saccharomyces cerevisiae*, is often used as a *model organism* in biological research. This is mainly due to its many favourable properties, such as the short generation time, easy genetical manipulation, as well as thanks to its established use in industry.

When performing experiments with *S. Cerevisiae* yeast, part of the process is growing the yeast colonies in Petri dishes, and evaluating the growth parameters, such as the coverage of the dish and the number of colonies. Traditionally, the colonies have been counted manually, using either manual colony counter or a counting raster. Examples of the manual counters are the Colony counter SC6 by Stuart[1] or the EW-14211-02 Hand-Held Electronic Colony Counter from Cole-Parmer[2]. However, the hand counting is considered rather laborious and time-consuming process, error prone due to the fatigue and eye-strain of the technician. Also, it does not provide any figures on the coverage of the dish or the radii of the colonies.

There are also a number of automated colony counters available – it is of interest that such system has been described as early as in 1974 [1]. An automated counter usually consists of a single-purpose chamber with an illumination and camera system, where the dishes are placed either manually one-by-one or by an automated plate-handler,

[1] www.stuart-equipment.com
[2] www.coleparmer.com

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 110–125, 2013.
© Springer-Verlag Berlin Heidelberg 2013

and of a computer running an image-processing program that performs thresholding, segmentation, and statistical evaluation of the image.

An automated counter is definitely the system-of-choice for a commercial lab. Our system, however, has been developed for a the cooperating university research lab. The requirements were not only that it performs adequately for typical image variations, but also – quite importantly – to reuse the imaging equipment already available in the lab (to achieve a low-cost solution), and to provide an off-line batch processing of a set of images taken beforehand in the darkroom. Additionally, the tool should operate with minimum possible manual intervention of operator, while allowing for manual correction of the counting errors (multiple detections of a single colony or missed colonies). Finally, it should provide not only the count of colonies, but also evaluate the total coverage of the dish.

Finally, the open-source framework ImageJ [2] should be noted in the context of existing tools. It is a feature-rich tool for image processing, with focus in medical and biological applications. In the field of our interest, it provides many algorithms for image thresholding and segmentation. In our view, however, it is targeted more generally and to prepare a specialized application which shields the user of the complexity of the underlying system would require an effort similar to development of our solution. We have decided to use MATLAB with the Image processing toolbox, because of our previous experience with that environment.

In the paper, we describe the processing flow of our tool and present the performance data that we have achieved. The paper is structured in the following way: in Section 2, the characteristics of typical images are reviewed. In Section 3, the components of the system are discussed. Section 5 describes the experiments and performance results. Finally, Section 6 concludes the paper.

2 Characteristics of Typical Images

As has been stated, the requirement on our system was that it should be able to process batches of images taken with a simple general-purpose imaging system; the laboratory is using a camera/copy stand with lightning units by Kaiser Fototechnik[3] (see Fig. 1). The dishes are manually placed roughly into the center of the image, on a mat black background. No dish holder or stopper is used.

Examples of typical images of yeast colonies are shown in Figure 2. The images are characterized by the following properties:

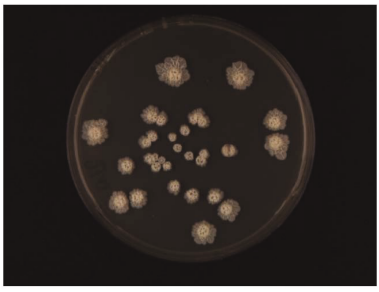
- dark background to increase image contrast
- illumination by two linear light sources along the short edges (this is not important for processing, though)
- variations of dish position – cf. Figure 2a and 2b
- varying diameter of the dish image and varying intensity of the dish background (both depend on the adjustment on the imaging setup in the particular session)

The colonies selves are characterized by differing types of morphologies (smooth or fluffy), colony size dispersion (even within single dish) and by colonies often touching

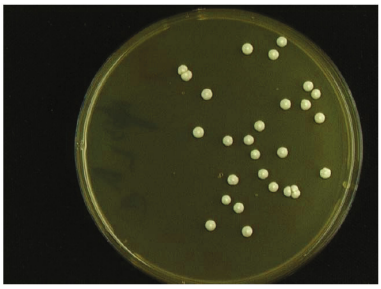
[3] www.kaiser-fototechnik.de



Fig. 1. General-purpose imaging setup used with the proposed colony counter



(a) Fluffy and touching colonies, diverse radii, dish centered.



(b) Smooth colonies, dish touching edge of image.

Fig. 2. Examples of typical images of yeast colonies

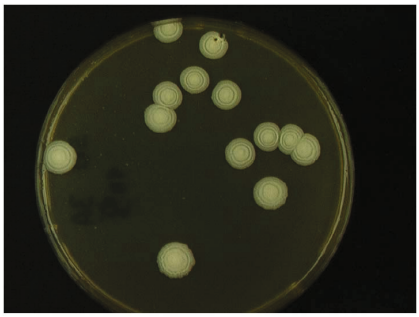
each other. Let us note that the morphology and size of colonies depends both on their age and on the growth conditions.

3 System Description

Counting of colonies typically consists of several operations, which follow from the image characteristics[3,4]. This operations include image thresholding, dish localization, and the counting self. In this paper, we divide the processing into two basic parts – *image preprocessing* and *colony counting*.

3.1 Image Preprocessing

The goal of this step is to determine the exact position of the dish in the image, to determine the area of the dish, and to extract the image of colonies and measure it's surface, so that the relative coverage can be determined. The image is checked for any faults in this step, such as missing parts (see Fig. 3a) or corrupted background (Fig. 3b). Let us note that while the missing part makes it impossible to determine properly the total area of the dish, the corrupted background can be eliminated to some extent.



(a) Missing part of the dish.



(b) Corrupted background.

Fig. 3. Examples of faulty images

The processing flow of the preprocessing is outlined in Fig. 4 on the following page. It starts with operations related to dish localization: the image is first binarized. A threshold that is used in this process is computed using samples from the image corners. Let these samples be $\mathbf{S}_1,\ldots,\mathbf{S}_4$ (see Fig. 5 on the next page), with the mean values $\bar{\mathbf{S}}_1,\ldots,\bar{\mathbf{S}}_4$ and the standard deviations $\sigma_1,\ldots,\sigma_2$. The background threshold θ_b is derived from the sample with minimum mean value:

$$i = \arg\min_i(\bar{\mathbf{S}}_i) \quad , \quad \theta_b = \bar{\mathbf{S}}_i + 3\sigma_i + \text{offset}$$

Note: offset is used to ensure that the threshold will be set above the peaks of the background sample. In our implementation, we use offset equal to 5.

Next step is to project the binary image along the horizontal and vertical axis. The projections P_x and P_y of the binary image $\mathbf{I}_b$ are given as the pixel sum in the given direction:

$$P_x = \sum_j \mathbf{I}_{b,j} \quad ,$$
$$P_y = \sum_i \mathbf{I}_{b,i} \quad ,$$

where i is the row index and j is the column index.

The dish is, in these projections, manifested as a broad 'lobe', or peak (see Fig. 6 on page 115). The partly uncovered background – if present – would show up as another peak (cf. the graphs in the bottom row of Fig. 6 on page 115), however, the

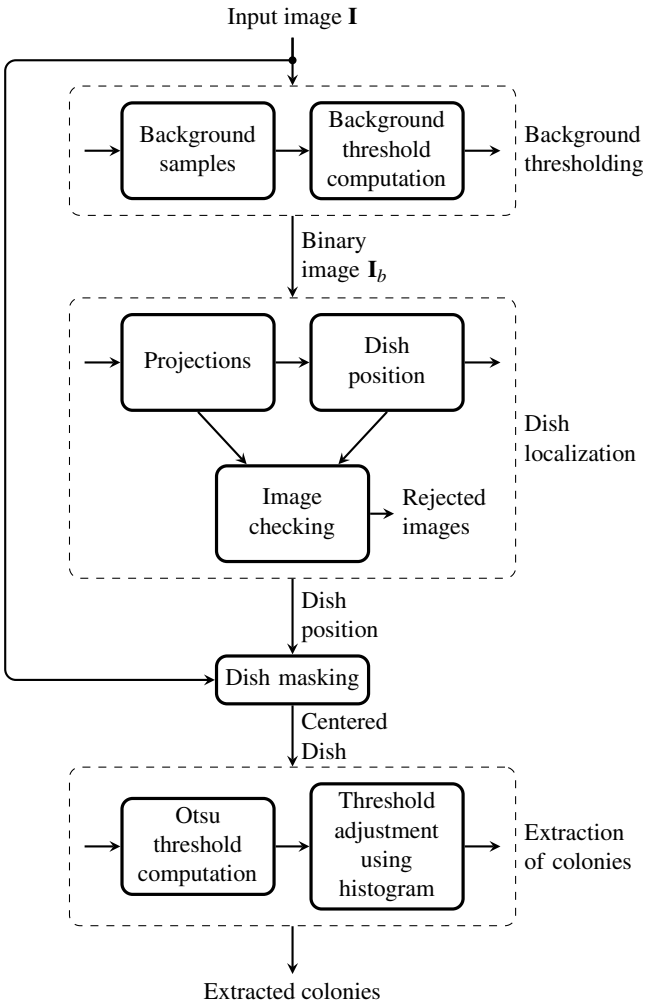


Fig. 4. Processing flow of dish preprocessing



Fig. 5. Position of image background samples

lobe corresponding with the dish is the broadest one. Additional check for uniformity of estimated diameters in horizontal and vertical direction is applied to eliminate heavily corrupted images.

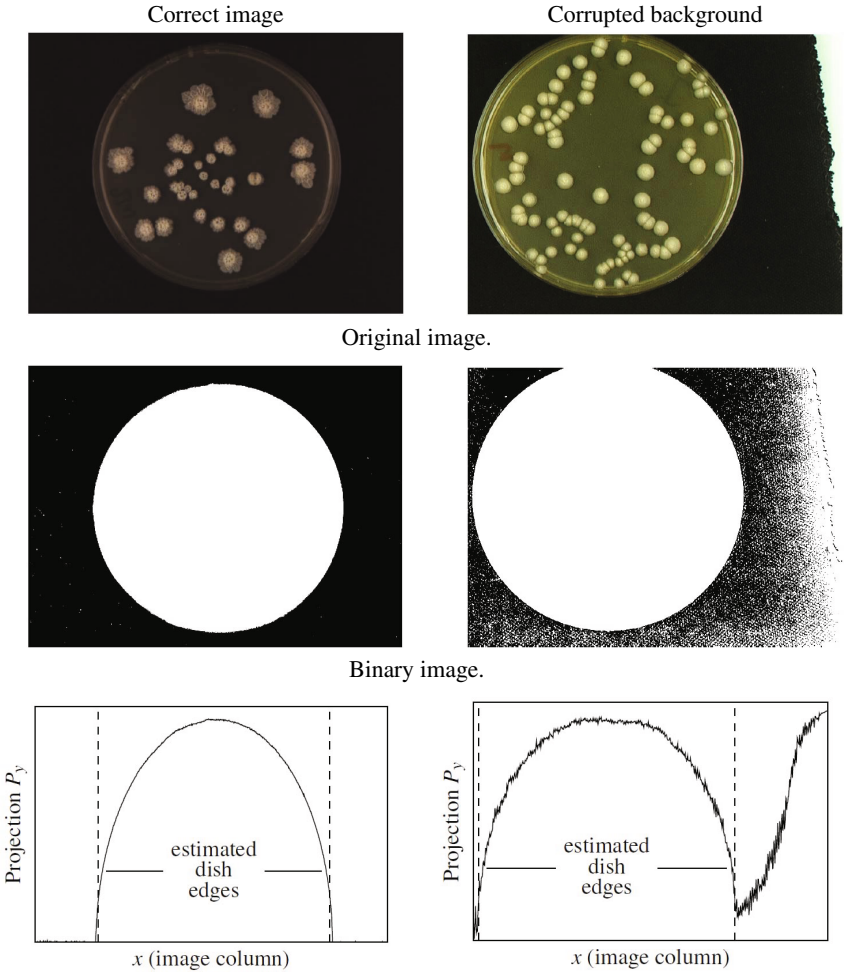


Fig. 6. Computation of dish projection P_y – example

Finally, the dish is approximated by a circle with the center and diameter derived from the edges determined in the dish localization step. The result of the preprocessing is a centered image of the dish, with eliminated background. An illustrative example is given in Fig. 7 on the following page.

The last step of the image preprocessing is the *extraction of colonies*, which is used to assess the total covered area, in order to determine the relative coverage of the dish. A new threshold is used in this step, located between the brightness levels of the dish background and of the colonies, as close as possible to the background.

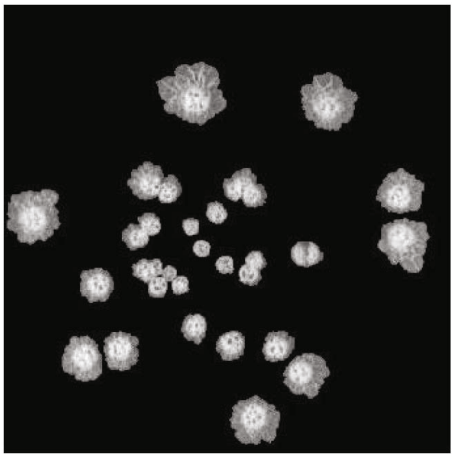


Fig. 7. Dish image after preprocessing

The most straightforward solution is to use the well-established Otsu thresholding [5]. However, the threshold based purely on this method may be set too high, if the colony edge is fading gradually. To avoid this, we have introduced a *histogram-based correction* to the method (see Fig. 8 on the next page and Fig. 9 on page 118).

The principle of this correction is following: first, by limiting the histogram by the Otsu threshold θ_O, we cut-off the part corresponding with (brighter) colonies. The peak of the remaining part (Fig. 9) corresponds with the dish background. The threshold should be placed just above the right edge of this peak.

More formally, let $H = [h_1, \ldots, h_O]$ be the histogram of the pixels below the Otsu threshold θ_O. Let H_{MA} be histogram filtered by a moving average filter:

$$H_{MA} = [h_{MA,1}, \ldots, h_{MA,O}], \quad h_{MA,i} = (\sum_{j=1}^{N} h_{i-j+1})/N, \quad i-j+1 < 0 : h_{i-j+1} = 0.$$

Let $\Delta_{H_{MA}}$ be the first difference of the filtered histogram. Then, the dish threshold with correction θ_d is given as

$$\theta_d = \max(i), \quad \Delta_{H_{MA}}(i) < -s_d,$$

where s_d is a user-selected setpoint. In our implementation, we use $N = 4$ and $s_d = 100$, with good thresholding performance.

3.2 Colony Counting

In this step, the centered image of the colonies, resulting from preprocessing, is used to find the individual colonies.

The simplest approach to counting is to correlate the contour of the colonies with a circular pattern, assuming that the maxima of correlation are located at the centers of colonies. This has been the first method we have tested. During experiments, we

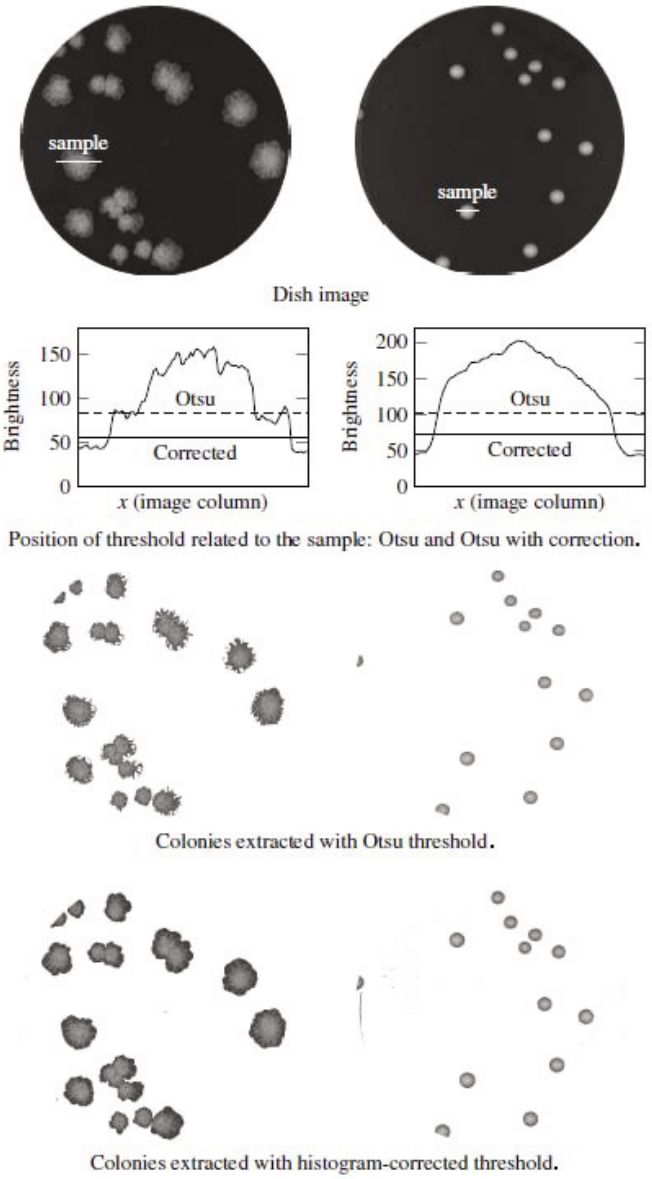


Fig. 8. Dish thresholding - pure Otsu method and a histogram-based correction

have experienced poor performance in the case of big clusters (the error rate of our implementation was reaching up to 30%) and sensitivity to mismatch between the radius of the circular pattern diameter and the radius of the colony (and, hence, sensitivity to dispersion of colony radii).

An interesting approach to the counting of bacteria colonies has been described in [3]: to find colonies, a number of shape and structure criteria are used on the

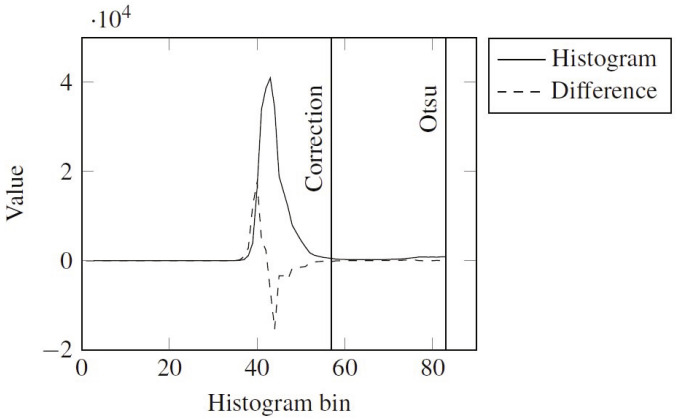


Fig. 9. Position of Otsu threshold and of threshold with correction relative to histogram

image pixels, including e.g. mean object radius, roundness of an object, compactness and asymmetry, radial monotony of fall-off in intensity, etc. These criteria are combined into shape and structure quality parameters and evaluated using fuzzy logic. The method works best under the assumptions of well-defined circular shape and monotone intensity fall-off of the colonies, which are well satisfied for bacteria (the case treated in the work of Marotz), but not necessarily in our case of the yeast colonies (for examples of possible morphologies see e.g. [6,7]).

A popular method to detect circular objects is the circular Hough transform (for definition see e.g. [8]). Based on our experiments, if multiple yeast colonies are overlapping or touching, it tends to detect (incorrectly) adjacent colonies as a single object. Also, following the results presented in [9], the output of the Hough transform may be rather noisy.

Another possibility is to use the methods based on *radial symmetry* – the radial symmetry transform has been introduced in [10].

Fast Radial Transform. In our solution, we use the Fast radial transform – a modified version of the radial symmetry transform, first presented in [9]. It is a transform that maps the original image to the transformed image according to its contribution to radial symmetry of the gradients at distance $n \in N$ (N is the set of radii) away from each point. For full details of the method see the original description.

First, image gradient $g_{i,j}$ at each point (i, j) is calculated. Then, the *positively-* and *negatively affected pixels* are determined. An affected pixel is defined as the point either *in-* or *counter* the direction of the gradient vector $g_{i,j}$, at a distance n pixels away from the point at coordinates (i, j). The coordinates of affected pixels are given by

$$p_+(i,j) = (i,j) + \text{round}\left(\frac{g_{i,j}}{\|g_{i,j}\|} n\right) \quad ,$$

$$p_-(i,j) = (i,j) - \text{round}\left(\frac{g_{i,j}}{\|g_{i,j}\|} n\right) \quad .$$

The gradient matrix **g**, together with the coordinates of affected pixels, is used to determine the *orientation* and *magnitude* projection images $\mathbf{O}_n$ and $\mathbf{M}_n$ for the given radius n:

$$\begin{aligned} O_n(p_+(i,j)) &= O_n(p_+(i,j)) + 1 \quad , \\ O_n(p_-(i,j)) &= O_n(p_-(i,j)) - 1 \quad , \\ M_n(p_+(i,j)) &= M_n(p_+(i,j)) + \|g_{i,j}\| \quad , \\ M_n(p_-(i,j)) &= M_n(p_-(i,j)) - \|g_{i,j}\| \quad . \end{aligned}$$

The radial symmetry at radius n is defined by convolution of matrices

$$\mathbf{S}_n = \mathbf{F}_n * \mathbf{A}_n \quad ,$$

where the elements

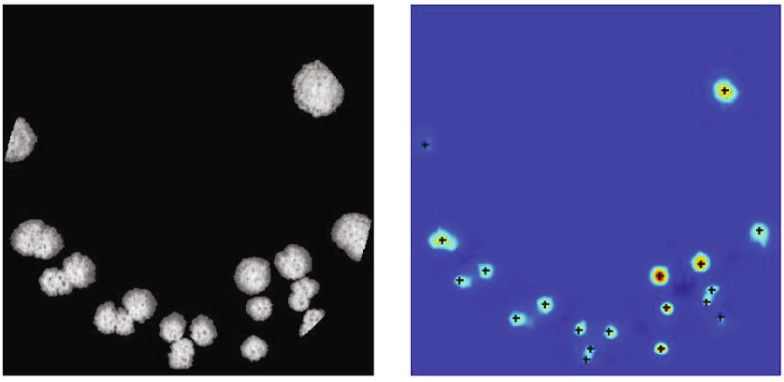
$$F_n(i,j) = \frac{M_n(i,j)}{k_n} \left(\frac{|\tilde{O}_n(i,j)|}{k_n}\right)^{\alpha} ,$$

and

$$\tilde{O}_n(i,j) = \begin{cases} O_n(i,j) & \text{if } O_n(i,j) < k_n \quad , \\ k_n & \text{otherwise} \quad . \end{cases}$$

Here, $\mathbf{A}_n$ is a two-dimensional Gaussian, α is the so-called radial strictness parameter and k_n is a scaling factor used to normalize $\mathbf{M}_n$ and $\mathbf{O}_n$ (for settings used in our implementation see 5.1). Projection images $\mathbf{M}_n$ and $\mathbf{O}_n$ are initially set to zero.

An example of a symmetry map for a dish is given in Figure 10. Note that in our implementation, the symmetry map is only computed for the foreground pixels.



(a) Preprocessed image of colonies. **(b)** Symmetry map S with the centers of colonies.

Fig. 10. Illustration of function of the radial symmetry algorithm

Colony Radius Estimation. Let us recall that there are often colonies with different radii contained in a single dish (see Sec. 2). Hence, prior to the radial transform computation, the set of radii N of the colonies in the dish is to be estimated, using the following procedure:

- the equivalent diameter d and eccentricity ε is computed for each object in the image. The diameter is derived from area A (number of pixels) of an object. The eccentricity equals to the eccentricity of an ellipse with the same second moments as the object. It ranges from $\varepsilon = 0$ for a circle to $\varepsilon = 1$ for a line segment.
- Nearly circular objects (with eccentricity $\varepsilon < \theta_\varepsilon$, where θ_ε is the eccentricity threshold), are selected.
- Min and max diameter $d_{\min}$ and $d_{\max}$ of the objects in the set of circular objects are determined. The interval between them is divided to ν equidistant subintervals. The set of radii N for the fast radial transform is then given by:

$$\Delta = d_{\max} - d_{\min}$$
$$N = \{d_{\min} + [0, \ldots, \nu] \cdot \Delta/\nu\}/2$$

Estimation of Colony Centers. The centers of the colonies are represented by local maxima of the symmetry matrix S, which are detected using the the `nonmaxsuppts()` function described in [11]. The output of this function are the coordinates of the centers.

4 Counting Tool

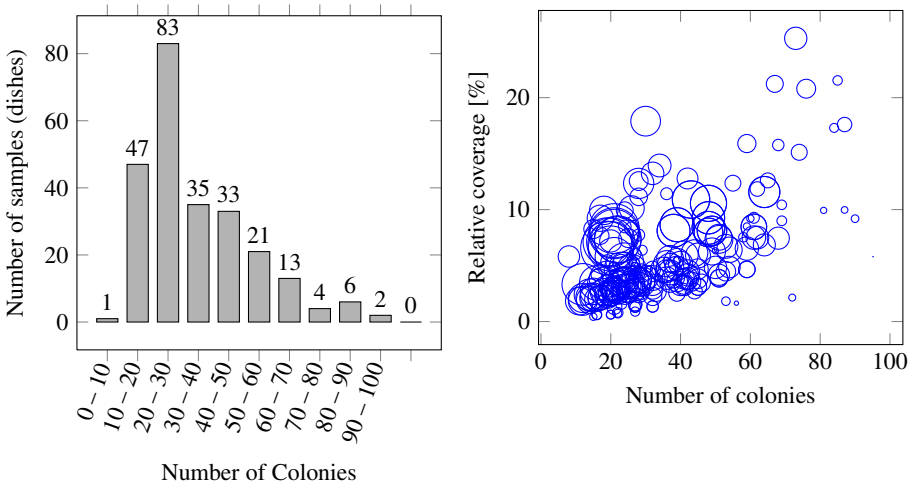
The processing flow, as described in Section 3, has been implemented in a counting tool (Fig. 11 on the next page). It performs fully automatic thresholding of the image, localization of the dish and counting of colonies. The tool provides an environment for selection of the directory tree with images to process, of the file to store the counting results, and of the counting mode: in the semi-automatic mode, a simple point-and-click editor can be used for manual correction of the system output, in the fully-automatic mode, the output of the system is directly stored into the result file. This tool was used also to obtain the reference results used in this paper: using manual correction, we obtained the ground-truth counts of colonies in Petri dishes, which were used to evaluate the counting error rates.

5 Experiments and Results

To evaluate the system recognition performance, it has been tested using a set of 245 images, containing colonies with different morphology and relative coverage of the dish. The distribution of the test set in the terms of frequencies of the number of colonies in the dish is shown in Fig. 12a. Figure 12b shows the distribution of relative coverage of the dish and mean diameter of colonies (marker size) in comparison with the number of colonies in the dish. In the set, the minimum and maximum counts were 8 and 95 colonies per dish, when 80% of dishes contained less than 50 colonies. The minimum and maximum coverage were 0.43% and 25.3%, respectively, and 80% of dishes had coverage up to 8.26%.



Fig. 11. Counting application



(a) Histogram of dish sample distribution.

(b) Relative dish coverage related to the number of colonies. Diameter of markers represents mean radius of the colonies in the dish.

Fig. 12. Distribution of the test set

5.1 Algorithm Settings

The following settings were used in our experiments:

– Parameters of the fast radial transform:

$$k_n = 6, \quad \alpha = 4$$

Let us note that setting $\alpha = 2$ (see Sec. 5.1) could be better choice – the response of the radial transform on fluffy colonies with less circular shape would be higher

and the computational demands of the method would be somewhat lower. However, all experimental data presented later have been obtained for $\alpha = 4$. We have performed a test with $\alpha = 2$ on a randomly selected sample of 15 images. The counting performance has been slightly improved, however, the system was more sensitive to the background noise along the dish edge (giving numerous false positives in some cases). The problem of elimination of the reflections on the dish edge has been treated e.g. in [4]; we perform only simple reduction of the dish radius, which should cover the dish lip region.

– Parameters of the non-maxima suppression `nonmaxsuppts()` [11]:

$$\text{thresh} = 4, \quad \text{radius} = 0.8 \cdot \min(N),$$

where N is the set of radii for the fast radial transform.

– Construction of N: eccentricity threshold θ_ε (see Section 3.2) is initially set to $\theta_\varepsilon = 0.25$. The number of equidistant intervals ν is set to $\nu = 4$. At least five colonies of given eccentricity must be in the image, else the eccentricity is increased by step of 0.1 (we start from stricter requirement on circularity of colonies, to eliminate clusters. If there are not enough colonies of the given circularity, we loosen the criterion and accept those with less circular shape).
Let us note that the distribution of colony diameters should also be considered, however, this is not implemented in the tool at the moment.

5.2 Typical Detection Errors

Figure 13 illustrates typical detection errors of fast radial transform. A colony may be missed if it touches other colony or multiple colonies, so that they form a cluster. This detection error is almost absent if the colonies touch only at one point, creating a chain-like structure. More colonies touching each other form a structure in which their shape is distorted and internal colonies cease to be circular. The circular outer border of colony located in such cluster could be too short for the proposed method to work properly.

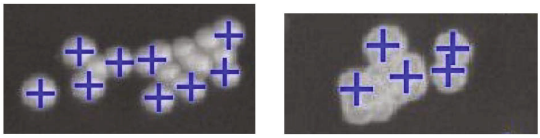


Fig. 13. Example of typical detection error of fast radial transform

5.3 Counting Performance

The counting errors for various number of colonies are summarized in Tab. 1 on the facing page. The relatively high counting error for the dishes with the colony counts greater than 60 is given by two factors. First, with the increasing number of colonies increases also the probability that there will be colonies touching each others (Fig. 13). Second, the number of samples with this high density of coverage was relatively low, thus increasing the evaluation error (Fig. 12a on the previous page).

Table 1. Dependence of the counting errors on the number of colonies in the dish. The system average counting error is under 4%.

♯ colonies	samples	missed [%]	false [%]
0 – 17	21	1.47	0
18 – 20	19	2.99	0
21 – 23	25	3.31	0
24 – 26	29	4.00	0
27 – 29	27	3.28	0
30 – 40	35	4.17	0
41 – 49	30	4.00	0
50 – 60	29	3.87	2.00
> 60	30	5.36	3.22

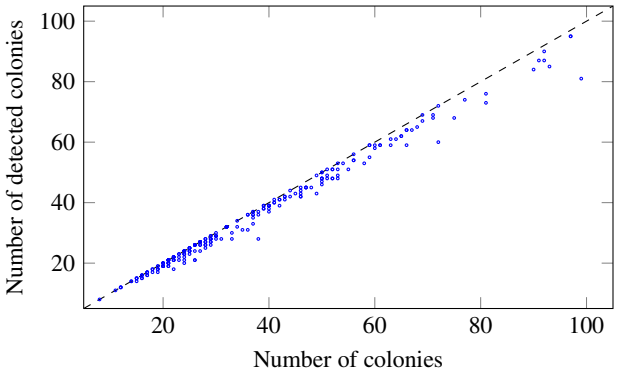
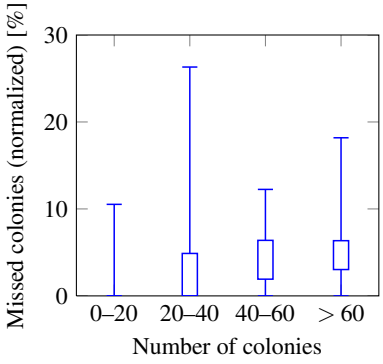


Fig. 14. Algorithm recognition performance related to the number of colonies



# colonies	samples
0–20	40
20–40	123
40–60	52
>60	30

Fig. 15. Relative number of missed colonies dependent on the number of colonies

Figure 14 on the preceding page shows the recognition performance related to the number of colonies in the dish. The dishes, where some colonies have been missed fall below the dashed line.

Another view on the algorithm performance is provided by Fig. 15, which presents a box-whisker plot of relative counting error (missed colonies) in several groups of the colony counts. The numbers of samples per group are given in the attached table.

6 Conclusions

In the paper, we have presented the tool for evaluation of yeast colony images, targeted towards batch processing of images prepared with a general-purpose imaging system. The processing flow of the system has been described. The performance of the tool has been tested on a set of 245 images with different degrees of coverage. The distribution of the data in the test set and the performance of the system have been discussed. The average counting error is below 4%. It is difficult to compare the performance with the commercial solutions, since the performance data of these systems are not available.

There are still performance challenges, especially in resolution of clustered fluffy colonies. The resolution of clustered colonies is rather common problem, reported also in other works [4]. The performance, however, has been acceptable for our target laboratory user.

The application has been implemented in a Matlab. This tool has been successfully deployed in the cooperating biology research laboratory and is available for download[4].

Acknowledgements. This research has been supported from the TA01010931 "System for support of the FISH method evaluation" project of the Technology Agency of the Czech Republic.

We would like to thank the staff of the Yeast Colony Group, Department of Genetics and Microbiology, Faculty of Sciences, Charles University, who have introduced us into the problem and provided us with the sample images from their experiments.

References

1. Goss, W.A., Michaud, R.N., McGrath, M.B.: Evaluation of an automated colony counter. Appl. Microbiol. 27, 264–267 (1974)
2. Burger, W., Burge, M.J.: Digital Image Processing: An Algorithmic Introduction using Java. Springer, Heidelberg (2007)
3. Marotz, J., Lübert, C., Eisenbeiss, W.: Effective object recognition for automated counting of colonies in petri dishes (automated colony counting). Computer Methods and Programs in Biomedicine 66, 183–198 (2001)
4. Chen, W.B., Zhang, C.: An automated bacterial colony counting and classification system. Information Systems Frontiers 11, 349–368 (2009)
5. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66 (1979)

[4] `http://zs.utia.cz/index.php?ids=results&id=yeastcolcount&lang=eng`

6. Granek, J.A., Magwene, P.M.: Environmental and genetic determinants of colony morphology in yeast. PLoS Genetics 6, e1000823 (2010)
7. Šťovíček, V., Váchová, L., Kuthan, M., Palková, Z.: General factors important for the formation of structured biofilm-like yeast colonies. Fungal Genetics and Biology: FG & B 47, 1012–1022 (2010)
8. Ballard, D.H., Brown, C.M.: Computer Vision. Online Book (2003)
9. Loy, G., Zelinsky, A.: Fast radial symmetry for detecting points of interest. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 959–973 (2003)
10. Reisfeld, D., Wolfson, H., Yeshurun, Y.: Context-free attentional operators: The generalized symmetry transform. International Journal of Computer Vision 14, 119–130 (1995)
11. Kovesi, P.D.: MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering, The University of Western Australia (2005), http://www.csse.uwa.edu.au/~pk/research/matlabfns/

Reconstructing Gapless Ancestral Metabolic Networks

Esa Pitkänen[1], Mikko Arvas[2], and Juho Rousu[1]

[1] Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014, Helsinki, Finland
first.last@cs.helsinki.fi
[2] VTT Technical Research Centre of Finland
P.O. Box 1000, FI-02044 VTT, Otaniemi, Finland
first.last@vtt.fi

Abstract. We present a method for inferring the structure of ancestral metabolic networks directly from the networks of observed species and their phylogenetic tree. In particular, we aim to minimize the number of mutations on the phylogenetic tree, whilst keeping the ancestral networks structurally feasible, or *gapless*. In gapless metabolic networks all reactions are reachable from external substrates such as nutrients.

To this end, we introduce the gapless minimum mutation problem: finding parsimonious phylogenies of gapless metabolic networks when the topology of the phylogenetic tree is given, but the content of ancestral nodes is unknown. This formulation can be extended also to infer reactions that are missing from the input metabolic networks due to errors in annotation transfer, for example.

The gapless minimum mutation problem is shown to be computationally hard to solve even approximatively. We then propose an efficient dynamic programming based heuristic that combines knowledge on both the metabolic network topology and phylogeny of species. Reconstruction of each ancestral network is guided by the heuristic to minimize the total phylogeny cost. We experiment by reconstructing phylogenies generated under a simple random model and derived from KEGG for a number of fungal species.

1 Introduction

Modelling of metabolism is essential in a variety of applications of biotechnology and medicine including bioprocess development [25], study of metabolic diseases [30] and drug target identification [15]. Global characteristics of cellular metabolism by metabolic networks have been studied intensively by a variety of computational approaches, including metabolic reconstruction (see [24] for a recent survey), metabolic flux analysis [21], 13C isotopic tracing [26] and structural analysis of metabolic networks [17].

The structure of the metabolic network is a major contributor to the phenotypes that an organism manifests. Importantly, the metabolic network structure is known to constrain the phenotypes the network can realize [21], thus the network structure is also likely to be conserved in evolution [33].

Recently, the increasing number of fully sequenced genomes has enabled comparative genomics analysis of metabolic network evolution (for review see [5]). Many computational approaches have concentrated on deriving rigorous measures of biological

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 126–140, 2013.
© Springer-Verlag Berlin Heidelberg 2013

network and pathway similarity [29], thus enabling construction of phylogenies of networks with distance-based methods. Particularly methods for metabolic network and pathway comparison have been developed [8,31,7,18].

Distance-based methods do not immediately yield predictions on the contents of *ancestral networks*, however. Knowledge on ancestral networks is important as it may shed light on the evolutionary mechanisms that have generated the observed networks. An approach complementary to distance-based methods often used to give insight on ancestral node contents in phylogenetic trees is *maximum parsimony*, where one tries to find ancestral objects which minimize the total number of evolutionary changes required to explain the observed data. The maximum parsimony principle has been utilized in many domains, for instance in the analysis of sequence data [10,28,6,32] as well as gene regulatory networks [4].

A direct application of maximum parsimony methods to biological network data generally results in structurally infeasible ancestral networks. For instance, consider the two metabolic networks Y and Z with a common immediate ancestor X shown in Fig. 1. A parsimonious scenario (top right) includes an ancestral network where a metabolite is required by the network but cannot be produced, suggesting that the pathway $m \to c$ cannot operate and the network is infeasible.

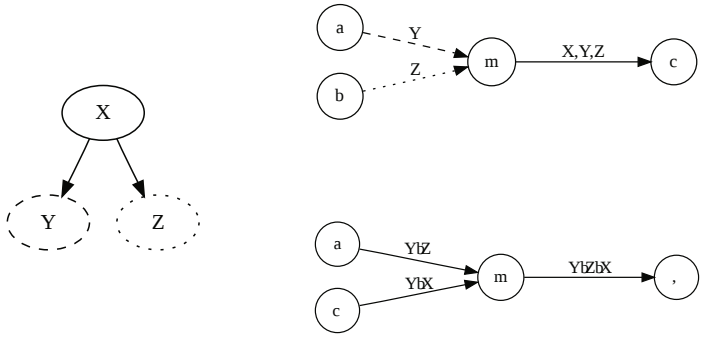


Fig. 1. Left: a small example phylogeny for three metabolic networks X, Y, Z. Right: two parsimonious scenarios involving metabolic pathways $a \to m, b \to m$ and $m \to c$. Metabolites and pathways are shown as circles and arrows, respectively. Pathways are labeled with organisms which have the pathway. In upper scenario, only pathway $m \to c$ is assigned to ancestor X, thus leaving metabolite m without a producing pathway. In lower scenario, both pathways from Y and Z are assigned to X.

Graph evolution models taking into account dependencies imposed by the network structure have been proposed. For instance, Mithani *et al.* gave a Markov process for simulating metabolic network evolution under a neighbor dependency model where appearance of a reaction depends on the fraction of neighboring reactions already present in the network [20,19]. However, they reported results only for relatively small metabolic networks.

In this paper, we introduce a computational method for reconstructing ancestral metabolic networks in a given phylogenetic tree. The method combines the maximum parsimony principle with the requirement that the resulting networks are plausible in

terms of network connectivity, thus contributing towards bridging the gap between the structural and phylogenetic analysis of metabolic networks. Our method builds phylogenies of metabolic networks where the ancestral nodes of the phylogeny adhere to structural network constraints: The networks are required to be free of *reaction gaps*, that is, reactions whose substrates cannot be produced from external metabolites. The choice of external metabolites reflects the estimated metabolic environment: the organisms are assumed to have them available in abundance and possess the necessary transports. To this purpose, computational methods have been developed to identify a set of minimal nutrients, given metabolic network structure [14,3].

In Section 2, we formulate the *gapless minimum mutation* problem where the topology of the phylogenetic tree is taken as input and the problem is to infer the structure of the ancestral networks so that the total phylogeny cost is minimized. We show the problem to be computationally hard to even approximate and go on to propose an efficient heuristic algorithm, which solves the problem well in practise. In Section 3, we experiment by reconstructing random and fungal phylogenies. Section 4 ends the paper with conclusions.

2 Methods

We are interested in metabolic networks that are functional in the sense that the network is able to produce substrates of all its reactions from some given set of source metabolites. Such networks are termed *gapless*, with a precise definition given below.

A metabolic network can be described as a binary string $N \in \{0,1\}^m$, where each $N_i = 1$ states that the reaction r_i, drawn from a collection of reactions $\mathcal{R}$, is in the network. We use the shorthand $r_i \in N$ when $N_i = 1$. Further we assume a set of metabolites $\mathcal{M}$ is consumed and produced by the m reactions. The set of substrate and product metabolites of a reaction r_i are given by $S(r_i) \subset \mathcal{M}$ and $P(r_i) \subset \mathcal{M}$, respectively.

To see how a string N encodes metabolic network connectivity, note that N induces a directed bipartite graph $G(N) = (V, E)$, with a node $v_r \in V$ for each reaction $r \in N$. Additionally each metabolite $b \in S(r_i) \cup P(r_i)$ for every $r_i \in N$ contributes a node $v_b \in V$. Edges $(v_r, v_b) \in E$ and $(v_b, v_r) \in E$ are added whenever reaction r produces or consumes metabolite b, respectively. Figure 2 shows this graph representation implicitly encoded by a string N for five reactions.

We first define gaplessness in terms of reactions that are *reachable* from a set of source metabolites S [22].

Definition 1. *Let N be a metabolic network and $S \subseteq \mathcal{M}$ be a set of* source *metabolites.*

- *A reaction $r \in N$ is reachable from S in N if all its substrates $S(r)$ are reachable from S in N.*
- *A metabolite $b \in \mathcal{M}$ is reachable from S in N if either $b \in S$ or $b \in P(r)$ for some reaction $r \in N$ that is reachable from S in N.*

A gapless metabolic network *N under S is a metabolic network where all reactions are reachable from S in N.*

We often omit an explicit mention of the source set S if it is clear in the context, saying only that a metabolic network is gapless. If a reaction $r \in N$ is not reachable from S, we say that r is a *reaction gap* (under S). In addition, we say that a metabolic network is gapped, if it contains at least one reaction gap. Figure 2 illustrates these concepts.

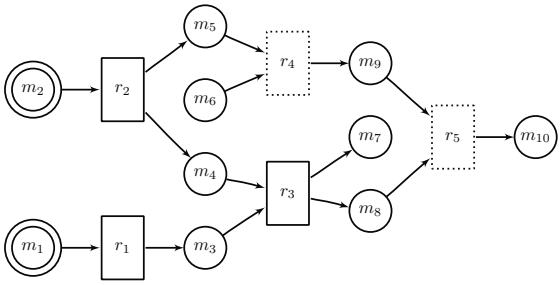


Fig. 2. Example metabolic network with reactions $N = \{r_1, \ldots, r_5\}$ and metabolites $\{m_1, \ldots, m_{10}\}$. When $S = \{m_1, m_2\}$ (double circles), reactions r_4 and r_5 (dotted rectangles) are reaction gaps in N. However, the network N under $S' = \{m_1, m_2, m_6\}$ would be gapless, because then reaction r_4 would be reachable. On the other hand, $N' = \{r_1, r_2, r_3\}$ is gapless under $S = \{m_1, m_2\}$.

2.1 Gapless Minimum Mutation Problem

We next introduce the computational problem of finding a gapless phylogeny when the tree topology and input taxa are given.

Problem 1. Gapless Minimum Mutation problem (GMM). Given a reaction collection $\mathcal{R}$, $m = |\mathcal{R}|$, a rooted binary tree $T = (V, E)$, labeling $L(u) \in \{0,1\}^m$ specifying a metabolic network for each leaf node and source metabolites S, find a labeling for each internal node of T such that

(1) $c = \sum_{(u,v)\in E} d(L(u), L(v))$ is minimized and
(2) $L(u)$ is a gapless metabolic network under S for each internal node $u \in V$,

where d is Hamming distance.

The equivalent problem defined for binary strings without the gapless constraint (2), Minimum Mutation problem, can be solved in polynomial time with the Fitch algorithm because each character position can be solved independently of each other [10,13]. However, in contrast to the Minimum Mutation problem, the character positions in GMM are not necessarily independent of each other: setting a certain $L_i(u) \in \{0,1\}$ may impose constraints on other positions $j \neq i$ due to the gapless constraint. Note that the taxa contained in the leaves may or may not correspond to gapless metabolic networks.

Theorem 1. *Deciding whether a solution with cost $c \leq k$ to Gapless Minimum Mutation problem exists given $k \in \mathbb{N}$ is NP-complete.*

Proof. Given a solution to GMM, we can both compute the cost c and check that each network $L(u)$ is gapless in polynomial time [22], hence the problem is in NP.

To show that the problem is NP-hard, we reduce the well-known NP-complete Minimum Set Cover problem [12] to Gapless Minimum Mutation problem. Let X be a finite set and C be a collection of subsets of the set X. In the Minimum Set Cover problem, we ask for the smallest subset $C' \subseteq C$ such that every element of X belongs to at least one member of C'. To create an instance of GMM, we first set up a reaction collection $\mathcal{R}$ with two groups of reactions, one group for items in X, another for sets in C. Specifically, let $\mathcal{R}$ to contain a reaction r_i for each $x_i \in X$ with $S(r_i) = \{b_i\}$ and $P(r_i) = \{c_i\}$, and a reaction q_j for each $c_j \in C$ with $S(q_j) = \{a_j\}$ and $P(q_j) = \{b_i \mid x_i \in c_j\}$. In addition, let $\mathcal{R}$ contain a reaction x with $S(x) = \{c_i \mid x_i \in X\}$ and $P(x) = \{m\}$.

To set up a phylogenetic tree, let $T = (V, E)$ be a binary rooted tree with $V = \{v_1, v_2, v_3\}$, $E = \{(v_3, v_1), (v_3, v_2)\}$ and root v_3. Finally, let input metabolic networks at leaves be $L(v_1) = L(v_2) = \{r_i \mid x_i \in X\} \cup \{x\}$ and the set of source metabolites be $S = \{a_j \mid c_j \in C\}$.

A minimal solution to GMM will contain at node v_3 a network with r_i and x, and a minimal number of reactions q_j to make $L(v_3)$ gapless. Existence of x will ensure that $L(v_3) \neq \emptyset$ when $|C| = |X|$. To transform a solution to GMM back to a solution to Minimum Set Cover, let set C' contain c_i for each q_i assigned to node v_3. The reduction can be done in polynomial time. In reduction, each reaction q_i is assigned to v_3 if and only if c_i appears in the optimal solution, assuming without loss of generality a unique set cover solution. If there is no solution to the set cover instance, also GMM is unsolvable as it is not possible to fix $L(v_3)$ to be gapless. Figure 3 shows the reduction from an example set cover instance $X = \{x_1, \ldots, x_5\}$ and $c_1 = \{x_1, x_2, x_3\}, c_2 = \{x_1, x_2, x_4\}, c_3 = \{x_3, x_5\}$.

Since the problem is both in NP and NP-hard, the claim follows. □

Theorem 2. *Gapless Minimum Mutation cannot be efficiently α-approximated unless $P = NP$.*

Proof. We show next that the reduction described above is actually an approximation-ratio preserving reduction and we can thus exploit the inapproximability of the set cover problem. Set cover problem has been shown to be hard to approximate within a logarithmic factor unless $P = NP$ [27,1].

Assume that there is an α-approximation algorithm for GMM for some $\alpha > 1$. We can thus obtain a solution of cost $\leq \alpha OPT$ where OPT is an optimal solution cost of GMM. Given a set cover instance with optimal size k, we obtain a GMM instance of optimal cost $2k$ in polynomial time with the above reduction. Solving the instance approximatively we get a solution of cost at most $2\alpha k$. This yields an approximative solution $\hat{k} \leq \alpha k$ to the set cover problem, thus contradicting the assumption. □

Often we have gapped metabolic networks as taxa to begin with. For instance, initial networks may be a result of function assignment by annotation transfer, where the resulting structure of the draft metabolic network is not of concern. However, these networks should also be functional and therefore we can attempt to fix them gapless while

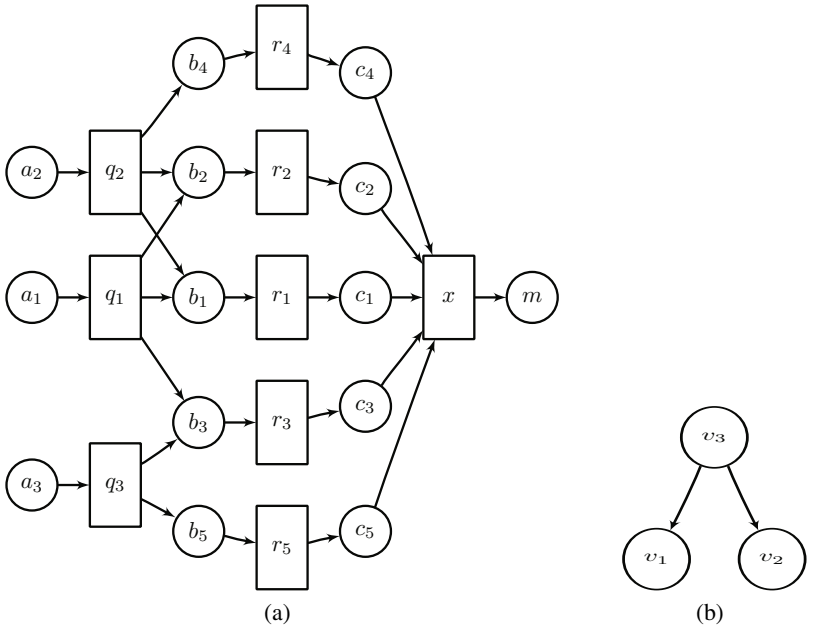


Fig. 3. Example reduction of a Minimum Set Cover instance with $X = \{x_1, \ldots, x_5\}$ and $c_1 = \{x_1, x_2, x_3\}, c_2 = \{x_1, x_2, x_4\}$ and $c_3 = \{x_3, x_5\}$ to Gapless Minimum Mutation problem. Left: reactions in $\mathcal{R}$. Right: tree T.

finding ancestral nodes. To do this, we can extend the tree T such that for each leaf u we add an edge (u, u') and assign $L(u') \leftarrow L(u)$. Solving GMM in the modified tree thus finds a solution where nodes u' retain the original input networks but gaps in internal nodes u are fixed. We provide an example of such situation in experiments, where we utilize gapped networks derived from a metabolic database.

2.2 Local Adjustment Algorithm

To overcome the computational complexity, we next propose an algorithm that solves the Gapless Minimum Mutation problem in two phases. In the first phase, the assignments corresponding to the minimum mutation cost are computed with the Fitch algorithm [10,13]. In the second phase, the tree is traversed top-down and a gapless metabolic network is assigned at each node by filling the gaps remaining after the Fitch pass. The algorithm relies on estimates on how much filling each gap would increase the total phylogeny cost, and attempts to choose gap-filling reactions which increase the cost as little as possible. The cost increase estimates are computed by the algorithm for each ancestral network.

In Algorithm 1 at line 2, we first compute *equality sets* $F_i(v) \subseteq \{0, 1\}$ for each internal node v and position i with the Fitch algorithm. In a binary tree, equality set $F_i(v)$ is defined for each position i as

Algorithm 1. `GaplessMinimumMutation`: Local adjustment algorithm for gapless minimum mutation

```
1:  Input: tree T, taxa L
2:  F ← Fitch(T, L)
3:  for all internal nodes v in top-down order do
4:      if v is root then
5:          for all i ∈ {1, ..., m} do
6:              L_i(v) ← 0 if 0 ∈ F_i(v); otherwise L_i(v) ← 1
7:      else
8:          for all i ∈ {1, ..., m} do
9:              if F_i(v) = {0, 1} then
10:                 L_i(v) ← L_i(pa(v))
11:             else
12:                 L_i(v) ← 0 if 0 ∈ F_i(v); otherwise L_i(v) ← 1
13:     (D, M) ← MinDist(v, S)
14:     L(v) ← L(v) ∪ MinFill(v, S, D, M)
```

$$F_i(v) = \begin{cases} \{L_i(v)\} & \text{iff } v \text{ is leaf} \\ F_i(x) \cap F_i(y) & \text{iff } v \text{ is not leaf and } F_i(x) \cap F_i(y) \neq \emptyset \\ F_i(x) \cup F_i(y) & \text{iff } v \text{ is not leaf and } F_i(x) \cap F_i(y) = \emptyset \end{cases},$$

where x and y are the children of an internal node v [13].

The tree is traversed in a top-down pass and an initial labeling $L(v)$ is decided according to the Fitch top-down phase (Algorithm 1, lines 3–12). Each initial labeling is then adjusted so that it satisfies the gaplessness constraint. For each internal node v, the algorithm calls the subroutine MinFill, which returns a gapless network containing $L(v)$ and *gapfill* reactions (Algorithm 2). Particularly, if $L(v)$ is already gapless, it is returned as such. MinFill attempts to satisfy the gaplessness criterion by backtracking from each reaction not reachable from sources S in network $L(v)$ and iteratively adding reactions to the fill set Γ. Subprocedure terminates when either all reactions have been reached or the algorithm notices that all gaps cannot be filled.

A heuristic is used to guide the backtracking phase by considering reaction assignments in the parent and child nodes. In particular, the algorithm attempts to choose reactions to the fill set Γ such that the parsimony cost increase is minimized. To do this, we first compute distances d_f that provide an estimate for each reaction how much the parsimony cost will increase compared to the optimal Minimum Mutation solution, if the reaction is added to the network. Formally, the distance d_f is a lower bound to the increase in parsimony cost that is the result of adding reaction r_i and other reactions required to make r reachable to v:

$$d_f(v, r_i) = \max_{m \in S(r_i)} \min_{q \in Pr(m)} d_f(v, q) + d_\delta(v, r_i)$$

where $S(r_i)$ are the substrates of reaction r_i and $Pr(m)$ are the reactions producing m. To introduce some notation, let $G_i(w, v) = \{L_i(w)\}$ iff $pa(v) = w$ and $G_i(w, v) = F_i(w)$ iff $w \in ch(v)$, where the parent and children of v are denoted by $pa(v)$ and $ch(v)$, respectively. Cost function $d_\delta(v, r_i)$ specifies the increase in

parsimony cost when adding reaction r_i only and is defined as $d_\delta(v, r_i) = \delta(v, pa(v), r_i) + \sum_{c \in ch(v)} \delta(v, c, r_i)$ with

$$\delta(v, w, r_i) = \begin{cases} 1 & \text{if } G_i(w, v) = \{0\} \text{ and } L_i(v) = 0 \\ -1 & \text{if } G_i(w, v) = \{1\} \text{ and } L_i(v) = 0 \\ 0 & \text{if } G_i(w, v) = \{0, 1\} \text{ and } L_i(v) = 0 \\ 0 & \text{if } L_i(v) = 1 \end{cases}$$

Note that at root v, we set $\delta(v, pa(v)) = 0$.

Parsimony cost increase d_δ at each reaction is always non-negative. To see this, we can enumerate the values for d_δ at an internal node given G_i at parent and children (Table 1). Symmetric cases are omitted from the table, where a boldface **0** signifies that this combination would yield $L_i(v) = 1$ in the bottom-up phase of the algorithm and thus the reaction would already be in the network.

Table 1. Values of heuristic d_δ for different parent and children assignments

Parent	Left child	Right child	d_δ	Parent	Left child	Right child	d_δ
0	0	0	3	1	0	0	1
0	0	1	1	1	0	1	**0**
0	0	0,1	2	1	0	0,1	0
0	1	1	**0**	1	1	1	**0**
0	1	0,1	**0**	1	1	0,1	**0**
0	0,1	0,1	1	1	0,1	0,1	**0**

Subroutine MinDist is called from Algorithm 1 to compute distances d_f for all r with dynamic programming (Algorithm 3). If for some required substrate there are only producers r_i that have $d_f(v, r_i) = \infty$, the algorithm fails as the required substrate cannot be reached from S. In such cases, source set S needs to be expanded or reaction collection $\mathcal{R}$ revised. To avoid loops, $\epsilon > 0$ is added to distances d_f, ensuring that distances strictly increase when traversing the network.

Algorithm 2. `MinFill`

Input: v, S, D, M
$\Gamma \leftarrow \emptyset$; $Q \leftarrow \{r \in v \mid D(r) = \infty\}$
while $|Q| > 0$ **do**
  $r \leftarrow \texttt{pop}(Q)$
  for all $o \in S(r)$ **do**
    if $o \notin M$ **then**
      $q \leftarrow \texttt{argmin}_{q \in Pr(o)} D(q)$
      if $D(q) = \infty$ **then**
        `return` "Impossible to find gapfilling set"
      if $q \notin \Gamma$ and $q \notin v$ **then**
        $\texttt{push}(Q, q)$
        $\Gamma \leftarrow \Gamma \cup \{r\}$
Return Γ

Algorithm 3. `MinDist`: compute distances $d_f(v, S)$

```
 1: Input: v, S
 2: M ← S; Q ← min-heap
 3: for all r ∈ {r ∈ R | S(r) ⊆ S} do
 4:    insert(Q, r, 0)
 5: for all r ∈ R do
 6:    D(r) ← 0 if r ∈ Q; D(r) ← ∞ otherwise
 7: while |Q| > 0 do
 8:    r ← extract-min(Q)
 9:    for all o ∈ P(r) do
10:       if o ∉ M then
11:          M ← M ∪ {o}
12:          for all q ∈ {q ∈ R | D(q) = ∞ ∧ S(q) ⊆ M} do
13:             D(q) ← D(r) + δ_f(v, q) + ϵ
14:             insert(Q, q, D(q))
15: return (D, M)
```

Theorem 3. *Local adjustment algorithm returns gapless metabolic networks $L(u)$ for each internal node u or reports failure in $O(nm \log m)$ time, where n is the number of species and $m = |\mathcal{R}|$ is the number of reactions.*

Proof. As each reaction is inserted at most once to the heap, Algorithm 3 takes $O(m \log m)$ time to compute distances d_f assuming that both heap operations `insert` and `extrac t-min` take $O(\log m)$ time. Further we assume that $|P(r)|$ is bound by a constant, which is reasonable as typically in enzymatic reactions $|P(r)| = 1, \ldots, 3$. In Algorithm 2 each reaction is inserted at most once to the queue Q, hence the subroutine takes $O(m)$ time. Algorithm 1 makes a single call to subroutine `Fitch` which takes $O(nm)$ time. Loop at line 11 is executed also $O(nm)$ times. The time complexity of the algorithm is dominated by the $O(n)$ calls to Algorithm 2, resulting in total $O(nm \log m)$ time. Figure 4 illustrates the operation of the algorithm. □

Note that it is possible to have an instance of GMM where the algorithm fails to find a gapless solution although such solution exists. As an example, consider two networks with a single reaction $L(v_1) = L(v_2) = \{r\}$, where $S(r) = \{m_1\}, P(r) = \{m_2\}$ and $S = \emptyset$, and the tree of Fig. 3. Then, the optimal solution is $L(v_3) = \emptyset$. However, as the algorithm does not attempt to remove reactions from the initial Fitch solution $L(v_3) = \{r\}$, the network remains gapless. Such cases are avoided by carefully selecting the reaction collection and source metabolites. In practice, one must ensure that the source metabolite set is large enough for each reaction to be reachable in network $L(u) = \mathcal{R}$.

3 Experiments

We experimented with two datasets. First, we generated random phylogenies consisting of gapless metabolic networks under a simple model of metabolic network evolution. Second, we derived metabolic networks for 16 fungal species from the KEGG database [16].

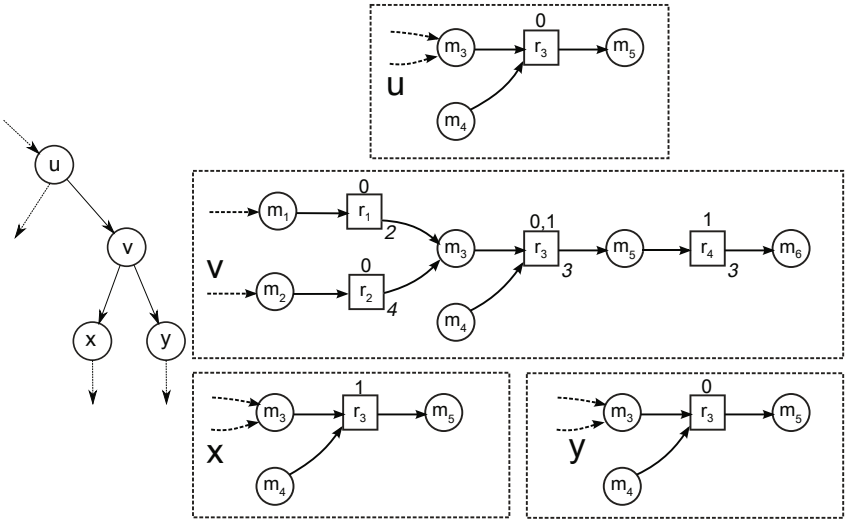


Fig. 4. Operation of Algorithm 1 to solve the Gapless Minimum Mutation problem. Left: A part of a phylogenetic tree. Right: Subnetworks of nodes u, v, x, y. State at node v after a call to `MinDist`: example values of d_f and reaction assignments shown at bottom right corner and on top of each reaction of v, respectively. Metabolite m_4 assumed to be a source. Dashed edges indicate parts of networks not shown. Distances for r_3: $d_\delta(r_3, v) = d_\delta(r_3, u) + d_\delta(r_3, x) + d_\delta(r_3, y) = 1 - 1 + 1 = 1$ and thus $d_f(r_3, v) = \max(\min(d_f(r_1, v), d_f(r_2, v)), 0) + d_\delta(r_3, v) = \max(\min(2, 4), 0) + 1 = 2 + 1 = 3$.

3.1 Random Phylogenies

To generate a phylogenetic tree, we first started with a gapless network containing 300 random reactions from KEGG and then simulated evolution by randomly adding or removing one reaction at time. Only additions and deletions, or reaction *flips*, that preserved gaplessness were allowed. Probability of both the addition and deletion was 0.5. The reaction to be added or deleted was chosen uniformly from the set of reactions whose addition or deletion preserved gaplessness. After each flip, a speciation event occured at a fixed probability 0.005 resulting in a new branch in the phylogenetic tree. The process was terminated after a tree of 30 nodes was generated resulting in networks of 249 ± 22 reactions at each node[1].

The generated taxa were reconstructed by the Fitch algorithm and Algorithm 1. Prior to reconstruction, the input taxa were randomly perturbed to simulate effects of annotation errors. Specifically, each reaction present in the taxa was deleted with the probability $p_d = 0, 0.01, 0.02, 0.05, 0.1$. Table 2 shows the reconstruction error measured as the average Hamming distance between node labels in reconstruction and generating taxa. Further, the average number of gapped and gapfill reactions at each node are shown. For instance, with random deletion probability $p_d = 0.1$, the reconstruction errors measured, 76.5 and 73.2 for Fitch's and our algorithm, respectively, were

[1] We use the $\pm$ notation to indicate standard deviations.

Table 2. Reconstructing random phylogenies when errors were introduced to data by deleting reactions from taxa with a fixed probability p_d. Columns FitchError and GaplessError give the reconstruction error measured as the average Hamming distance between reconstructed network and generating taxa at an internal node. Columns Gaps and Fills show the average number of gapped reactions and gapfill reactions added by our algorithm. Results are averages over 25 repeats. Standard deviations given in parentheses.

p_d	FitchError	GaplessError	Gaps	Fills
0.0	52.3 (4.12)	50.3 (3.80)	0.18 (0.11)	0.15 (0.09)
0.005	53.3 (3.49)	51.2 (3.47)	0.35 (0.22)	0.24 (0.14)
0.01	54.8 (3.93)	52.4 (4.09)	0.55 (0.21)	0.35 (0.13)
0.02	57.3 (4.84)	54.9 (4.60)	0.94 (0.35)	0.59 (0.20)
0.05	63.3 (4.06)	60.3 (4.04)	1.82 (0.60)	1.11 (0.27)
0.1	76.5 (4.14)	73.2 (4.00)	3.65 (1.00)	2.37 (0.53)

statistically different from each other (paired $t(48) = 2.87$, $p = 0.006$). Moreover, the simulated trees contained on the average 3.65 gaps in each network. Our algorithm added 2.37 gap-filling reactions to each network on the average. Regardless of parameter p_d, gapless reconstruction of each instance took about 27 seconds on a standard desktop computer running a Python implementation our algorithm.

3.2 Fungal Phylogenies

To experiment with a more real-world scenario, we constructed metabolic networks for 16 fungal species corresponding to 17 carbohydrate metabolism pathways [16] from KEGG gene-reaction links. As shown in Table 3, the process resulted in a high number of gapped reactions in these initial networks largely due to incomplete annotations and stoichiometry in KEGG enzymes and reactions.

As input, we provided our algorithm with the initial networks and a phylogenetic tree of the species ([11], Fig. 5). Because we had gapped networks to begin with, we added a new internal node for each species and an edge to the node as described earlier. Many microorganisms and free living fungi in particular can synthesize all their cellular components from inorganic sources, given a source of energy and carbon such as glucose [9]. However, many fungi require one or more vitamins such as thiamine or biotin. To model the fungal metabolic environment, a fixed set of 1518 source metabolites were used containing fungal energy and carbon sources, cofactors such as ATP and NAD and metabolites needed to account for the large number of otherwise isolated subgraphs in the KEGG universal metabolic network.

Table 3 lists the number of reactions added to fill gaps in the internal nodes corresponding to species. On the average, 39.4 ± 3.9 reactions were used to fill 72.5 ± 7.0 gaps divided into 7.2 ± 0.9 graph components at each node. The optimal minimum mutation cost was 1082; our algorithm achieved gapless minimum mutation cost of 1789.

To give an example of how the GMM result can provide insight into metabolic network evolution and aid reconstruction curation efforts, Fig. 6 shows five reactions from the internal node N22 that is the parent of species nodes mgr and ncr (Fig. 5). Two

Table 3. Species in fungal dataset. Columns Reactions, Gaps and Fills give the number of all reactions and gapped reactions in the initial networks, and reactions added by the algorithm to fill gaps, respectively.

Abbr	Species	Reactions	Gaps	Fills
ago	*Ashbya gossypii*	272	64 (23.5%)	37 (+13.6%)
afm	*Aspergillus fumigatus*	468	80 (17.1%)	50 (+10.7%)
ani	*Aspergillus nidulans*	382	74 (19.4%)	42 (+11.0%)
aor	*Aspergillus oryzae*	396	70 (17.7%)	42 (+10.6%)
cgr	*Candida glabrata*	270	70 (25.9%)	35 (+13.0%)
cne	*Cryptococcus neoformans*	336	68 (20.2%)	35 (+10.4%)
dha	*Debaryomyces hansenii*	326	66 (20.2%)	34 (+10.4%)
fgr	*Fusarium graminearum*	474	86 (18.1%)	43 (+9.1%)
kla	*Kluyveromyces lactis*	292	68 (23.3%)	36 (+12.3%)
mgr	*Magnaporthe grisea*	390	74 (19.0%)	41 (+10.5%)
ncr	*Neurospora crassa*	416	74 (17.8%)	38 (+9.1%)
dpch	*Phanerochaete chrysosporium*	410	90 (22.0%)	44 (+10.7%)
sce	*Saccharomyces cerevisiae*	332	80 (24.1%)	36 (+10.8%)
spo	*Schizosaccharomyces pombe*	278	68 (24.5%)	35 (+12.6%)
uma	*Ustilago maydis*	296	76 (25.7%)	47 (+15.9%)
yli	*Yarrowia lipolytica*	362	92 (25.4%)	43 (+11.9%)

reactions (KEGG ids R02933 and R02957) remained gapped after the first pass. Three reactions were predicted by the algorithm to fill these particular gaps (R01184, R01481, R01483). All filling reactions were used in parent but were absent from children, thus addition of each reaction increased the parsimony cost by one.

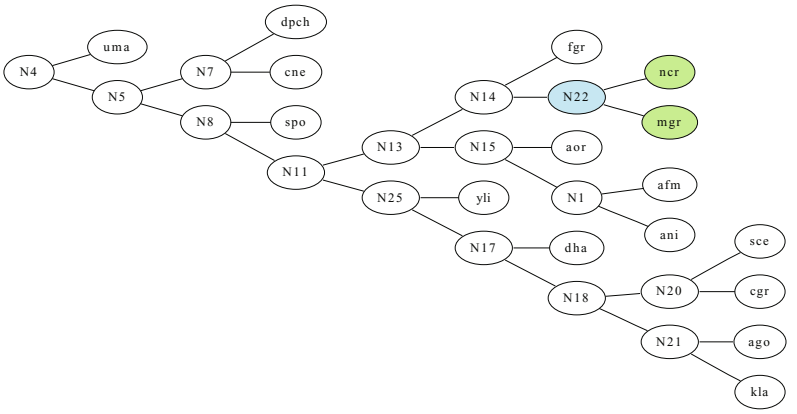


Fig. 5. A phylogeny tree for the 16 fungal species rooted at node N4. Shaded nodes indicate the two species and their immediate ancestor N22 highlighted in Fig. 6.

Even though KEGG reaction-gene links were missing for reactions R01184 and R01481, algorithm predictions were supported by homologous genes [2] found for both reactions in *M. grisea*. Further, a homologous gene was found also in *N. crassa* for reaction R01184. No gene was found to support the predicted existence of reaction R01483,

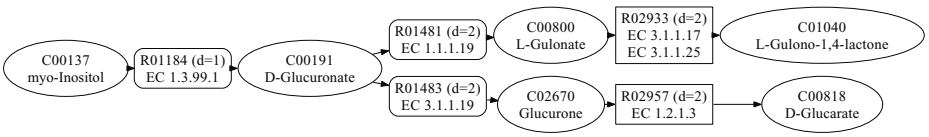


Fig. 6. A subnetwork of the metabolic network reconstructed at node N22 of the phylogenetic tree shown in Fig. 5. Reactions and metabolites are drawn as rectangles and ellipses, respectively. Rounded rectangles show the three reactions added by the algorithm to fill the gapped reactions R02933 and R02957. Distances d_f given in parentheses as d.

warranting further study. For the two gapped reactions (R02933 and R02957), homologues were found in both organisms, supporting KEGG data.

4 Conclusions

In this paper, we introduced a maximum parsimony algorithm for reconstructing gapless ancestral metabolic networks for a given phylogenetic tree. Furthermore, the method can be used to suggest gapless variants of draft metabolic networks of observed species given as input. The algorithm minimizes the number of mutations in the phylogenetic tree while maintaining the gaplessness of the ancestral networks. Thus, the algorithm can be used both to elucidate network structures in ancestral nodes and to fill in gaps in draft metabolic networks in a evolutionarily plausible manner. We argue that such approach, where the reconstruction networks are required to be gapless, improves prediction performance over the unconstrained case. This work extends the method of [23], where gapless reconstructions of individual metabolic networks were inferred from sequence data, to take into account the phylogenetic context of the reconstructed network.

The proposed algorithm was found to perform well in practice despite the computational complexity of the underlying problem. In experiments with random data, the algorithm was able to recover the original data from perturbed input more accurately than the baseline method that did not enforce gaplessness in reconstructed networks. Moreover, the algorithm yielded explanations to the question why a given reaction (enzyme) appears in an ancestral network by suggesting the required reactions that render the reaction gapless. This is especially important when we attempt to uncover the evolutionary history leading into observed networks.

While we experimented only with a simple random model of evolution, the framework introduced here lends itself to more realistic models. Exploring this direction is left as future work, though we note the importance of incorporating sequence data with the joint metabolic network/phylogenetic tree topology driven approach presented here. To this end, dealing with inaccuracies and omissions in the underlying metabolic reaction databases presents an additional challenge.

Acknowledgements. We would like to thank Pasi Rastas and Esko Ukkonen for insightful discussions. This work was financially supported by Academy of Finland grant 118653 (ALGODAN), in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886-PASCAL2, and by the Academy

of Finland postdoctoral researcher's fellowship 127715 (as part of the Finnish Centre of Excellence in White Biotechnology - Green Chemistry, Project No. 118573). This publication only reflects the authors' views.

References

1. Alon, N., Moshkovitz, D., Safra, S.: Algorithmic construction of sets for k-restrictions. ACM Trans. Algorithms 2(2), 153–177 (2006)
2. Arvas, M., Kivioja, T., Mitchell, A., Saloheimo, M., Ussery, D., Penttilä, M., Oliver, S.: Comparison of protein coding gene contents of the fungal phyla Pezizomycotina and Saccharomycotina. BMC Genomics 8(1), 325 (2007)
3. Borenstein, E., Kupiec, M., Feldman, M.W., Ruppin, E.: Large-scale reconstruction and phylogenetic analysis of metabolic environments. PNAS 105(38), 14482–14487 (2008)
4. Bourque, G., Sankoff, D.: Improving gene network inference by comparing expression time-series across species, developmental stages or tissues. J. Bioinform. Comput. Biol. 2(4), 765–783 (2004)
5. Caetano-Anollés, G., Yafremava, L., Gee, H., Caetano-Anollés, D., Kim, H., Mittenthal, J.: The origin and evolution of modern metabolism. The International Journal of Biochemistry & Cell Biology 41(2), 285–297 (2009)
6. Clemente, J.C., Ikeo, K., Valiente, G., Gojobori, T.: Optimized ancestral state reconstruction using sankoff parsimony. BMC Bioinformatics 10(51) (2009)
7. Clemente, J., Satou, K., Valiente, G.: Phylogenetic reconstruction from non-genomic data. Bioinformatics 23(2), e110 (2007)
8. Dandekar, T., Schuster, S., Snel, B., Huynen, M., Bork, P.: Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochem. J. 343(Pt 1), 115–124 (1999)
9. Deacon, J.: Fungal biology. Wiley-Blackwell (2006)
10. Fitch, W.M.: Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20, 406–416 (1971)
11. Fitzpatrick, D., Logue, M., Stajich, J., Butler, G.: A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology 6(1), 99 (2006)
12. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman (1979)
13. Gusfield, D.: Algorithms on Strings, Trees, and Sequences. Cambridge University Press (1997)
14. Handorf, T., Christian, N., Ebenhöh, O., Kahn, D.: An environmental perspective on metabolism. Journal of Theoretical Biology 252(3), 530–537 (2008)
15. Jamshidi, N., Palsson, B.O.: Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Systems Biology 1(26) (2007)
16. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y.: Kegg for linking genomes to life and the environment. Nucleic Acids Res. 36, D480–D484 (2008)
17. Lacroix, V., Cottret, L., Thebault, P., Sagot, M.F.: An introduction to metabolic networks and their structural analysis. IEEE Transactions on Computational Biology and Bioinformatics 5(4), 594–617 (2008)
18. Mano, A., Tuller, T., Bj, O., Pinter, R.Y.: Comparative classification of species and the study of pathway evolution based on the alignment of metabolic pathways. BMC Bioinformatics 11(suppl. 1), S38 (2010)

19. Mithani, A., Preston, G., Hein, J.: A bayesian approach to the evolution of metabolic networks on a phylogeny. PLoS Computational Biology 6(8) (2010)
20. Mithani, A., Preston, G.M., Hein, J.: A stochastic model for the evolution of metabolic networks with neighbor dependence. Bioinformatics 25(12), 1528–1535 (2009)
21. Palsson, B.: Systems biology: properties of reconstructed networks. Cambridge University Press, Cambridge (2006)
22. Pitkänen, E., Rantanen, A., Rousu, J., Ukkonen, E.: Finding Feasible Pathways in Metabolic Networks. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 123–133. Springer, Heidelberg (2005)
23. Pitkänen, E., Rantanen, A., Rousu, J., Ukkonen, E.: A computational method for reconstructing gapless metabolic networks. In: Proceedings of the 2nd International Conference on Bioinformatics Research and Development (BIRD 2008). CCIS, vol. 13. Springer, Heidelberg (2008)
24. Pitkänen, E., Rousu, J., Ukkonen, E.: Computational methods for metabolic reconstruction. Current Opinion in Biotechnology 21(1), 70–77 (2010)
25. Raman, K., Chandra, N.: Flux balance analysis of biological systems: applications and challenges. Briefings in Bioinformatics 10(4), 435–449 (2009)
26. Rantanen, A., Rousu, J., Jouhten, P., Zamboni, N., Maaheimo, H., Ukkonen, E.: An analytic and systematic framework for estimating metabolic flux ratios from 13 C tracer experiments. BMC Bioinformatics 9(1), 266 (2008)
27. Raz, R., Safra, S.: A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In: Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, pp. 475–484 (1997)
28. Sankoff, D.: Minimal mutation trees of sequences. SIAM J. Appl. 28, 35–42 (1975)
29. Sharan, R., Ideker, T.: Modeling cellular machinery through biological network comparison. Nature Biotechnology 24, 427–433 (2006)
30. Sigurdsson, M.I., Jamshidi, N., Jonsson, J.J., Palsson, B.O.: Genome-scale network analysis of imprinted human metabolic genes. Epigenetics 4(1), 43–46 (2009)
31. Tohsato, Y., Matsuda, H., Hashimoto, A.: A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp. 376–383 (2000)
32. Tuller, T., Birin, H., Gophna, U., Kupiec, M., Ruppin, E.: Reconstructing ancestral gene content by coevolution. Genome Res. 20(1), 122–132 (2010)
33. Wagner, A.: Evolutionary constraints permeate large metabolic networks. BMC Evolutionary Biology 9(1), 231 (2009)

Discretized Kinetic Models for Abductive Reasoning in Systems Biology

Gabriel Synnaeve[1], Katsumi Inoue[2], Andrei Doncescu[3],
Hidetomo Nabeshima[4], Yoshitaka Kameya[5], Masakazu Ishihata[5], and Taisuke Sato[5]

[1] E-Motion Team at INRIA 38334, Grenoble, France
gabriel.synnaeve@gmail.com
[2] National Institute of Informatics, Tokyo, Japan
ki@nii.ac.jp
[3] LAAS-CNRS 31007, Toulouse, France
andrei.doncescu@laas.fr
[4] University of Yamanashi, Yamanashi, Japan
nabesima@yamanashi.ac.jp
[5] Tokyo Institute of Technology, Tokyo, Japan
{kameya,ishihata,sato}@mi.cs.titech.ac.jp

Abstract. The study of systems biology through inductive logic programming (ILP) aims at improving the understanding of the physiological state of the cell by reasoning with rules and relations instead of ordinary differential equations. This paper presents a method for enabling the ILP framework to deal with quantitative information from some experimental data in systems biology. The method consist in both discretizing the evolution of concentrations of metabolites during experiments and transcribing enzymatic kinetics (for instance Michaelis-Menten kinetics) into logic rules. Kinetic rules are added to background knowledge, along with the topology of the metabolic pathway, whereas discretized concentrations are observations. Applying ILP allows for abduction and induction in such a system. A logical model of the glycolysis and pentose phosphate pathways of *E. Coli* is proposed to support our method description. Logical formulae on concentrations of some metabolites, which could not be measured during the dynamic state, are produced through logical abduction. Finally, as this results in a large number of hypotheses, they are ranked with an expectation maximization algorithm working on binary decision diagrams.

1 Introduction

Nowadays, systems biology represents the key field to explain the functionality of life science. To analyze a biological system it is necessary to find out new mathematical models allowing to explain the evolution of the system in a dynamic context or to deal in a simple manner with the complex situations where the human experience overtakes mathematical reasoning [1]. Many physical and biological phenomena may be represented on an analytical form using dynamical system. Our case study is based on wet biology experiment consisting in applying a pulse of glucose in a small bio-reactor containing *E.Coli* that led to building an ordinary differential equations (ODEs) based

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 141–154, 2013.
© Springer-Verlag Berlin Heidelberg 2013

simulator. We used high performance liquid chromatography to measure some metabolites concentrations and some others had to be estimated, using a simulated annealing algorithm, since no experimental results were available. So, knowing completely the evolutions of metabolites concentrations of this system, we applied our approach to show its correctness. For that, we took only steady-state values of metabolites concentrations and ran our model.

Several attemps have been done for logic-based approaches to analyze biochemical pathways (interaction of metabolites, signaling networks) in Systems Biology. They use action languages [2], abduction [3–6], SAT [7], inductive logic programming [8] or answer set programming [9]. All these previous approaches are based on qualitative modeling, and none of them can handle continuous domains appropriately. Temporal logic combined with the representation of kinetic models in stochastic logic programming (SLP) [10] have a similar goal using different means: the authors modeled the kinetics of biochemical systems by continuous time Markov chains as input to SLP where we took an approach to discretize (through continuous HMM) concentrations of metabolites first and then use them combined with a logical translation of ODEs-based kinetics as input to ILP. The goal of this research is to incorporate continuous values and kinetics within the logic-based approach to metabolic pathways. In particular, we enhance an abductive framework proposed in [11], which consists of abductive hypothesis generation and statistical hypothesis evaluation, by enabling us to handle real-valued data obtained from measurement in observations.

For that, we now propose a *loop for learning about a metabolic pathway from experiments* in which we have to (each step corresponds to a section, as in Fig. 1):

1. clusterize continuous concentrations of metabolites over time into discrete levels and discrete timesteps.
2. use them in an ILP-based model of the pathway, in conjunction with a set of knowledge-generating rules, here in the example describing Michaelis-Menten kinetics.
3. sort the resulting abduced facts or inducted rules with our defined metrics.
4. use this ranking for enhancing our knowledge base and *goto* the beginning of this process.

In this paper, we show how this "closed loop" architecture can be applied to an inverse problem: given the measured concentrations of some metabolites in a steady state, we compute the concentrations of metabolites before the dynamic transition to this steady state based on the kinetic modeling. We worked with the beginning of an automated framework (see Fig. 2 for a practical data-centric circuit) to deal with different real world pathways and experiments. It is mainly composed of four tools:

- The combination of an implementation of continuous HMMs [12, 13] with PYTSDISC to discretize experimental values.
- KEGG2SYMB, using the KEGG API, that transform pathways from KEGG [14, 15] into symbolic models.
- SOLAR, a consequence finding system working on Skipping Ordered Linear tableaux [16], which is *complete* for finding minimal explanations, to conduct abduction or induction.

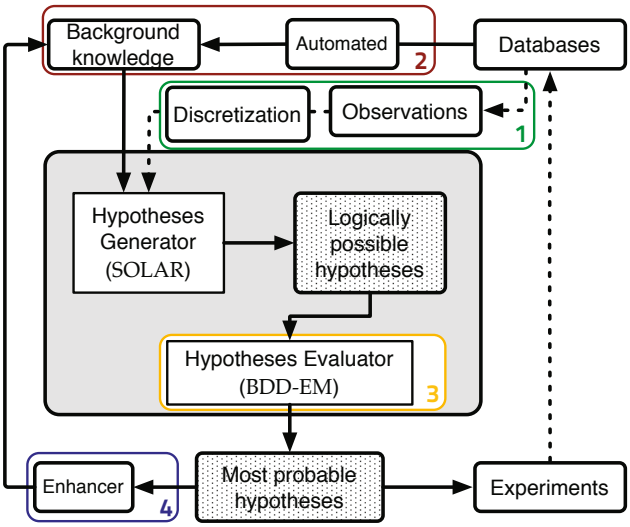


Fig. 1. Overview of the complete process

– BDD-EM, an implementation of the expectation-maximization algorithm on binary decision diagrams [11, 17] to rank hypotheses.

We chose to illustrate this method on the conjunction of glycolysis and pentose phosphate pathways for *E.Coli*, simplified the model by keeping 16 relevant reactions and discretized experimental values (16 values) as in section 1. We added the three Michaelis-Menten based rules and the three constraints of unicity for the levels as in section 2. We had 15 unknown levels of concentrations of metabolites before the transition to the steady state (yielding 15×3 levels = 45 abducibles). SOLAR, used for abduction, outputs 98 hypotheses that cover all these metabolites. With such a number, picking the right hypotheses should be done in an automated way as we did in section 3.

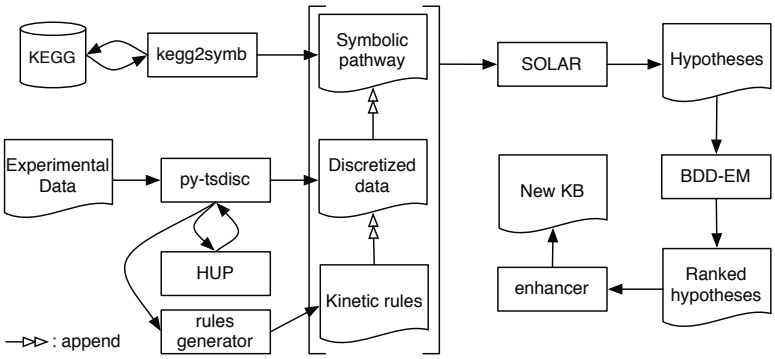


Fig. 2. Data-centric schema of the process

2 Discretization of Time Series from Experiments

In our modeling, we first introduce discrete concentration levels to filter what are the relevant changes of concentration of the metabolites, in regard to hypotheses generation from ILP. We need to be able to infer hypotheses that have a certain level of generality and, for that, we should use intervals instead of single real values. This could have been done with an interval constraints approach [18], but we currently choose a discretization approach. Although this gives us less freedom in the logic part as levels are fixed (as if we have fixed intervals), levels can be handled just as symbols in a logical model of pathways.

Discretizing time series is a research field in which many works [19, 20] have been conducted recently. Our practical problem is that we want to have a statistically relevant (unsupervised) discretization for N metabolites concentrations over time. We also discretize the values of K_m (Michaelis-Menten constants, see (1)), for each reaction, with the same levels. For that purpose, we use a probabilistic model, used in speech recognition and time series analysis: continuous hidden Markov model (HMMs) [21]. We can therefore compute an appropriate number of levels (that was three for *E.Coli*) in regard to a Bayesian score such as Bayesian Information Criterion (BIC) [22] or as the Cheeseman-Stutz score [23] or as the variational free energy. This process can be achieved through the following methods all described in [24], respectively: maximum likelihood estimation or maximum a posteriori estimation or through a variational Bayesian method.

We use continuous (Gaussian) HMMs with parameter tying[1]. This is a solution to the problem of sharing the same symbolic levels in all the logic models in order to be able to assign the level of a compound to another and be dealing with the same real values behind the scene. We first prepare N continuous HMMs (one for each metabolite), where each state variable takes a concentration level, and each output variable takes a measurement of concentration and follows a univariate Gaussian distribution[26]. All the HMMs share a state space as well as the parameters in the output variables (i.e. means and variances), so that they produce discrete levels that are corresponding. These relevant discretized levels of concentration are computed through the expectation-maximisation (EM) algorithm with maximum a posteriori (MAP) estimation [12] or through the variational Bayes EM (VB-EM) [13, 24]. We prefer this last method as it is shown [24] that variational free energy provides a more accurate approximation of the marginal log-likelihood than BIC or the Cheeseman-Stutz score.

Then, we use a simple round-mean aggregation of them for time-sampling. We set a maximal number of time steps and look for the better fitting width and alignment for equal-width time intervals. We are currently developping a different process in the direction of discretization of our time series from molecular biology experiments that

[1] Parameter tying is a notion often used in HMMs for speech recognition [21] and recently in statistical relational learning [25]. In our case, the mean and the variance for $X_t^{(n)}$, the output variable at time t in the HMM for the n-th metabolite ($n = 1, \ldots, N$), are tied with the mean and the variance for $X_{t'}^{(n')}$, respectively ($n \neq n'$ and $t \neq t'$). In contrast, to allow the concentration levels of the N compounds to change in different ways, the *transition parameters* for $P(S_{t+1}^{(n)} \mid S_t^{(n)})$ are *not* tied.

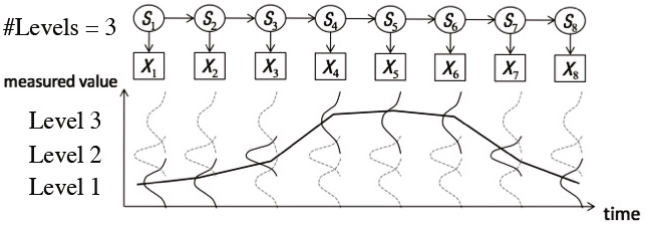


Fig. 3. 3-state continuous HMM discretizing one experimental time series, where X_t is the measurement of concentration at time t and S_t is the hidden state that indicates the corresponding discretized level

will discretize time and levels simultaneously but current results are already useable (see Table 1 and Fig. 5) and that is what we based the work presented here on.

3 Modeling of the Pathways of *E.Coli*

To obtain an understanding of the central metabolism, a logical model has been developed according to a kinetic model including the glycolysis and the pentose phosphate pathway for *Escherichia coli* [27]. The Fig.4 shows the simplified pathway that we modelized logically with relations `reaction(Substrate, Enzyme, Product, Km).`

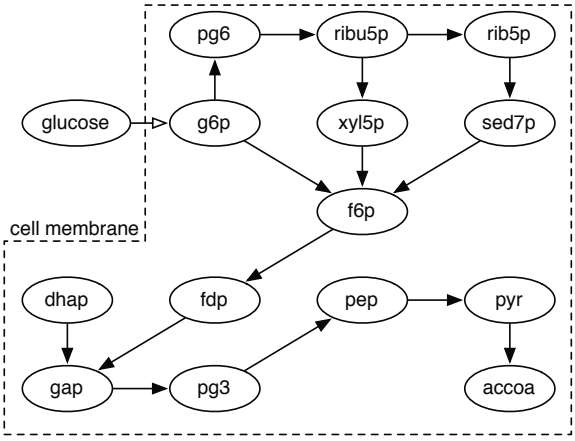


Fig. 4. Simplified glycolysis and pentose phosphate pathways for *E.Coli*

The metabolic networks dynamics are in their enzymatic part ruled by the combination of classical kinetics: essentially Michaelis-Menten, Hill and allosteric ones. If we limit our modeling to these kinetics, we can highly simplify their mathematical handling, and that is what we did. We chose to use only Michaelis-Menten kinetics, because we had a pathway simple enough and that it is the more general representation for a non-linear allosteric regulation system. It assumes that the two enzyme binding equilibria are fast when compared to the interconversion of enzyme + substrate (ES)

and enzyme + product (EP) compounds. That assumption appears reasonable considering that the dynamics of the experiment were happening in less that a minute: this implies that the effects of genetic regulation of the enzymes included are negligible and so the maximum reaction rates represent the amount and catalytic activity of enzymes.

$$E + S \rightleftharpoons^{k_1}_{k_{-1}} ES \rightarrow^{k_2} E + P$$

$$Michaelis - Menten\ equation : \frac{d[P]}{dt} = V_m \frac{[S]}{[S] + K_m} \quad (1)$$

If both the substrate (S) and the product (P) are present, neither can saturate the enzyme. For any given concentration of S the fraction of S bound to the enzyme is reduced by increasing the concentration of P and *vice versa*. For any concentration of P, the fraction of P bound to the enzyme is reduced by increasing concentration of S. When we have $S \leftrightarrows P$, we just have to consider reactions for both directions. We consider a time discretization of the chemical rate equation for a reation between a substrate and a product with respective stoechiometric coefficient s and p:

$$s.S \rightarrow p.P \ : \ rate = \frac{1}{p} \times \frac{d[P]}{dt} \longrightarrow_{disc.time} \frac{1}{p} \times \frac{\Delta[P]}{\Delta T} \quad (2)$$

$$(1)\ and\ (2) \implies p \times rate = V_m \frac{[S]_T}{[S]_T + K_m}$$

$$\approx \frac{[P]_{T+timestep} - [P]_T}{(T + timestep) - T}$$

We chose to work with a constant timestep :

$$\implies [P]_{T+1} = V_m \frac{[S]_T}{[S]_T + K_m} + [P]_T \quad (3)$$

We can note that the Michaelis-Menten constants (`Km`) are homogenous to a concentration. We can then state `conc(Km, Level, Time)` in our modeling to set them, where `conc` stands for concentration. The experimental response observations of intracellular metabolites to a pulse of glucose were measured in continuous culture employing automatic stopped flow and manual fast sampling techniques in the time-span of seconds and milliseconds after the stimulus with glucose. The extracellular glucose, the intracellular metabolites: glucose-6-phosphate (g6p), fructose-6-phosphate (f6p), fructose1-6bisphosphate (fdp), glyceraldehyde3phosphate (gap), phospho-enolpyruvate (pep), pyruvate (pyr), 6phosphate-gluconate (6pg), glucose-1-phosphate (g1p) as well as the cometabolites: atp, adp, amp, nad, nadh, nadp, nadph were measured using enzymatic methods or high performance liquid chromatography. All the steady-state concentrations *measurements* of the *E.Coli* experiment and their corresponding discrete levels are summarized in Table 1.

Inductive Logic Programming, used for induction or abduction [28], allows to deal with discrete levels (symbols) and qualitative rules [8]. Given the background knowledge B and an observation E (example), the task of ILP is to find an hypothesis H such that:

Table 1. Concentrations (mM/L) of the Metabolites and their discretized levels for steady states

#	Metab.	Conc.	Lvl	#	Metab.	Conc.	Lvl
1	glucose	0.055	0	2	g6p	3.480	2
3	f6p	0.600	0	4	fdp	0.272	0
5	gap	0.218	0	6	pep	2.670	2
7	pyr	2.670	2	8	6pg	0.808	1
9	g1p	0.653	0	10	amp	0.955	1
11	adp	0.595	0	12	atp	4.270	2
13	nadp	0.195	0	14	nadph	0.062	0
15	nad	1.470	1	16	nadh	0.100	0

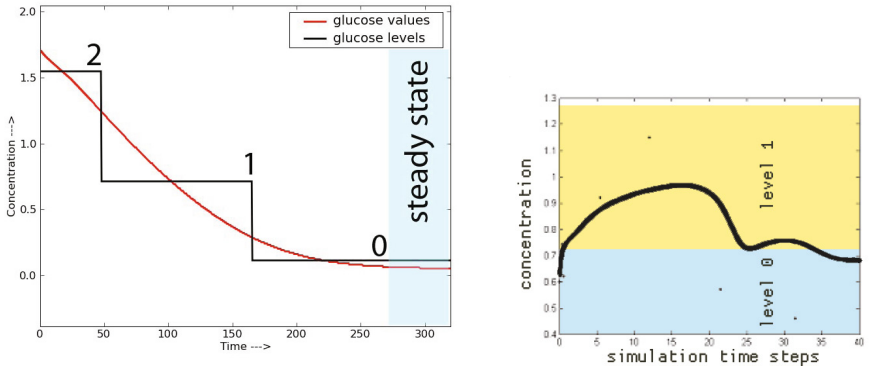


Fig. 5. Top: Discretization of the concentration of glucose in the Glycolysis Pathway of *E.Coli* after an initial pulse. Bottom: Simulated evolution of the concentration of fructose-6-phosphate during the whole experiment.

- $B \wedge H \models E$ and
- $B \wedge H$ is consistent

Inverse entailment [29–31] enables us to compute H through deduction by using:

- $B \wedge \neg E \models \neg H$ and
- $B \not\models \neg H$

We are here interested in abducing what happens during the dynamical transition based on observations from Table 1. Inverse entailment for abduction is studied in [29] in which abductive computation can be realized by the consequence finding procedure SOL. In this case, both E and H are sets of literals, so both $\neg E$ and $\neg H$ are clauses.

This approach can be further extended for inducing general hypotheses in [31], which is generalized from [30], to allow B, E and H for full clausal theories. $\neg H$ is constructed by a method called *CF-induction*, which computes *characteristic clauses* of $B \wedge \neg E$, selects its subset CC called the *bridge formula* from it, and generalizes $\neg CC$. CF-induction then realizes sound and complete hypotheses finding from *full clausal theories*, and not only definite clauses but also non-Horn clauses and integrity constraints can be constructed as H.

SOLAR can be used as an abductive procedure to infer a hypothesis H in the form of a set of literals. Our logical model is based on the simplified Michaelis-Menten equation **(3)** which has here been represented by three background clauses using the `conc(Compound, Level, Time)` predicate. If we make the approximations for extreme values in:

$$[P]_{T+1} = V_m \frac{[S]_T}{[S]_T + K_m} + [P]_T \quad (3)$$

With only 3 levels, as we have in our discretization of *E.Coli* experiments, we will get the following simple rules:

- $\underline{[S] \ll K_m} \Rightarrow \frac{\Delta[P]}{\Delta T} = \frac{V_m}{K_M} \Rightarrow [P]_{T+1} = [P]_T$ `reaction(S, P, Km)` $\wedge$ `conc(S, 0, T)` $\wedge$ `conc(Km, 2, T)` $\wedge$ `conc(P, L, T)` $\rightarrow$ `conc(P, L, T+1)` The concentration of the product will not change between T and T+1 if the reaction is *very* slow.
- $\underline{[S] \simeq K_m} \Rightarrow \frac{\Delta[P]}{\Delta T} = \frac{V_m}{2} \Rightarrow [P]_{T+1} = V_m/2 + [P]_T$ `reaction(S, P, Km)` $\wedge$ `conc(S, L, T)` $\wedge$ `conc(Km, L, T)` $\wedge$ `conc(P, L2, T)` $\rightarrow$ `conc(P, L2, T+1)` The concentration change of the product between T and T+1 is not big enough to switch from one level to another. This is an approximation and a handy consequence of our discretization (using a logarithmic scale on real values).
- $\underline{[S] \gg K_m} \Rightarrow \frac{\Delta[P]}{\Delta T} = V_m \Rightarrow [P]_{T+1} = V_m + [P]_T$ `reaction(S, P, Km)` $\wedge$ `conc(S, 2, T)` $\wedge$ `conc(Km, 0, T)` $\wedge$ `conc(P, L, T)` $\rightarrow$ `conc(P, 2, T+1)` If the reaction is very quick, it will result in transforming all the substrate into product in one time step.

If we had more than three levels, we would either need more rules (they can be automatically generated) or a general procedure for handling our kinetic model. This last one is a current implementation issue related to SOLAR. Another way to deal with more levels being currently explored consist in the automated generation of kinetics rules w.r.t. the discretization. Furthermore, we made some simplifications in the pathways to be able to use only Michaelis-Menten kinetics, another research topic is to extend our modeling to reactions ruled by other types of kinetics.

We also added constraints about the unicity of levels at a given time to reduce the number of hypotheses while keeping consistency:

- $\neg$`conc(S, 0, T)` $\vee$ $\neg$`conc(S, 1, T)`
- $\neg$`conc(S, 0, T)` $\vee$ $\neg$`conc(S, 2, T)`
- $\neg$`conc(S, 1, T)` $\vee$ $\neg$`conc(S, 2, T)`

Now we set the observations for the 6 metabolites (#2 $\rightarrow$ #7) from Table 1, which have been possibly affected by the stimulus with glucose, and the abducibles as those literals of the form `conc(_,_,0)`. Using SOLAR, we get 98 hypotheses as:

```
H76 = conc(g6p,2,0) ∧ conc(adp,2,0) ∧ conc(fdp,0,0) ∧
conc(dhap,0,0)
∧ conc(gap,0,0) ∧ conc(glucose,2,0) ∧ conc(pg3,2,0) ∧
conc(pep,2,0)
∧ conc(atp,0,0) ∧ conc(pyr,2,0)
```

4 Ranking Hypotheses

[17] proposed the BDD-EM algorithm that is an implementation of the expectation maximization algorithm working on binary decision diagram, allowing it to deal with boolean functions. [11] have applied the BDD-EM algorithm to rank hypotheses obtained through abduction. To rank our $H_1, \ldots, H_n$ hypotheses by probability, we consider the finite set of ground atoms $\mathcal{A}$ that contains all the values that can take our `conc(Compound, Level, Time)` and `reaction(Substrate, Product, Km)`. Each of the elements of $\mathcal{A}$ is a boolean variable. One of its subsets is the subset of abducibles Γ composed of all the possible values of `conc(Compounds, Level, 0)`. With $\theta_i = P(A_i)\ for\ A_i \in \mathcal{A}$, we have to maximize the probability of the disjunction of hypotheses helped with the background knowledge B: $F = (H_1 \vee \cdots \vee H_n) \wedge ground(B)$ to set the good θ parameters (by the BDD-EM algorithm). F can still be too big to be retained as a BDD, so an optimisation F' of its size is obtained through the use of the minimal proofs for B and each H_i. Then, the BDD-EM algorithm computes the probabilities of ground atoms in $\mathcal{A}$ that maximizes the probability of F'. Finally, the probabilities of each hypotheses used for the ranking are computed as the products of the probabilities of literals appearing in each H_i.

To sort our 98 abduced hypotheses, we ran the EM algorithm on the BDDs corresponding to our hypotheses 10,000 times with random initializations. Note that if the comparison of these probabilities with each other is relevant, they should not be taken as absolute probabilities. The 10 most probable abduced hypotheses are the following:

Table 2. 10 most probable hypotheses

Hyp. #	Probability	Abduced conc. levels at T=0
H76	≈ 1.000	g6p: 2, adp: 2, f6p: 0, fdp: 0, dhap: 0, gap: 0, glucose: 2, pg3: 2, pep: 2, atp: 0, pyr: 2
H41	0.822	the same as H76 except pg3: 0
H56	0.625	the same as H76 except g6p: 0
H70	0.553	the same as H76 except atp: 2
H13	0.515	the same as H56 except adp: 0
H90	0.455	the same as H70 except pg3: 0
H82	0.442	the same as H76 except dhap: 2
H43	0.369	the same as H76 except pyr: 1
H9	0.364	the same as H41 except dhap: 2
H68	0.346	g6p: 0, adp: 0, f6p: 0, fdp: 0, dhap: 0, gap: 0, glucose: 2, pg3: 2, pep: 2, atp: 2

These hypotheses are corresponding to our biological knowledge that pyruvate is a bottleneck [32] and that the glucose that is totally consumed (e.g. top plot of Fig. 5 from simulation) was in high concentration at the beginning of the experiment (pulse). It goes along with the very general reaction of glycolysis: $glucose + 2\mathrm{ADP} + 2\mathrm{P} + 2\mathrm{NAD}^+ \rightarrow 2\ pyruvate + 2\mathrm{ATP} + 2(\mathrm{NADH}, \mathrm{H}^+) + 2\mathrm{H_2O}$. Also, for some metabolites, such as fructose-6-phosphate, the levels found through abduction are corresponding to the

output of the simulation (e.g bottom plot of Fig.5) with the same low level (0) before and after the dynamic transition.

5 Discussion

5.1 Enhancing the Knowledge Base

Increasing our knowledge about a system is considered as an iterative process: at first, we consider the background knowledge combined with the observations as our knowledge base. Then we produce hypotheses and we need to use an algorithm to enhance (update) our knowledge base with some of the discovered hypotheses, here: abducibles. Ideally, we would re-run the hypothesis finding process until we cannot find anything new. This is particularly important when working with complex chained reactions and multiple time steps as it can enable deeper learning. This idea of revising the knowledge base is already found in [33] with a nonmonotonic approach, but their revision method stays in a qualitative modeling and do not take quantitative aspects into account.

Here, it is needed to pick hypotheses that are consistent with the background knowledge and with each others. For example, if we apply a greedy algorithm (as **Algorithm 1**) that picks hypothesis in decreasing probability order such that the hypothesis add some knowledge and that our enhanced knowledge stays consistent, it prevents from abducing other discoverables than the ones contained in H76. For instance we cannot find concentrations at T=0 for ribu5p, rib5p, sed7p, xyl5p, because if they were abduced, the resulting hypotheses would become inconsistent with H76. Note also that the abducibles added into the knowledge base may reduce the computational cost of later iterations of abduction/induction, but it is comparable to discard some branches of exploration. We could have chosen to pick a combination of hypotheses that discovers more abducibles by penalizing the solutions including too few different abducibles with a scoring function inspired by the BIC [22]: $score = -2\ln(error) + \lambda \cdot f(k, n)$ with k being the number of chosen hypotheses, n the number of abducibles, f a function that indicates the structural complexity of the combination of hypotheses (decreasing with the increase of n and increasing with the increase of k) and $error$ the product of the probabilities of chosen hypotheses. We assume here that we can use their relative significations in $error$ by unbiasing the score with a λ parameter. So that the goal of such an algorithm would be to discover all abducibles while minimizing this score.

5.2 A Finer Discretization

It is important to have as many levels as needed to be able to handle the different K_m (Michaelis Menten constant): $K_m(\mathrm{ATP}) = 0.4\ mmol/L \longleftrightarrow K_m(\mathrm{HCO_3^-}) = 26\ mmol/L$. With M levels, for example: `1 2 3 4 5 6 7`. When the need for $N < M$ levels arises, we do a projection, for example:

small: {`1, 2`} ; medium: {`3`} ; big: {`4, 5, 6, 7`}

Here, if we have only the 3 previously presented rules (Michaelis-Menten specific approximations). We center `medium` on the discrete level of the relevant K_m by doing

Algorithm 1. An algorithm to enhance the knowledge base: most probables firsts

```
knowledge ← knowledge_base
sorted_hypotheses ← sort(hypotheses)
while length(discoverable) > 0 &&
length(sorted_hypotheses) > 0 do
   tmp ← sorted_hypotheses.pop()
   if contains(tmp, discoverable) && consistent(tmp, knowledge) then
      knowledge.enhance(tmp)
      discoverable.remove(tmp)
   end if
end while
```

With the explicit functions *length*, *pop* (destructive), and:

- *sort* sorts the hypotheses by decreasing probability.
- *contains* is a function that returns statements of first argument contained in the second.
- *consistent* performs consistency checking of two theories and return True if they are consistent.
- *enhance* adds statements that are not yet present in the considered ("self", "this") knowledge.
- *remove* deletes statements from argument present in the considered ("self", "this") object (could make use of *contains*).

`medium(A) :- level(A)==level(Km)` *in the context of a reaction* with a substract of product `A` and constant `Km`. This means that the discrete level that will correspond to `medium` for the current application of the three 3-levels-based rules will be the level taken by K_m. We made a context-dependent projection.

Another way to deal with numerous levels is to change our modeling to allow for one (or many) more general(s) rule than the 3 rules showed above and resulting from approximations. This is what we try to achieve by introducing a new built-in predicate `compute` in SOLAR. As we take V_m and K_m from the literature, we can write it:
`reaction(X, E, Y, Km, Vm), concentration(X, LX, T), concentration(Y, LY, T) ⇒ compute(LX, LY, Vm, Km, LXX, LYY), concentration(X, LXX, T+1) concentration(Y, LYY, T+1)`

And in the other direction (from T to $T-1$):
`reaction(X, E, Y, Km, Vm), concentration(X, LXX, T), concentration(Y, LYY, T) ⇒ compute(LX, LY, Vm, Km, LXX, LYY), concentration(X, LX, T-1) concentration(Y, LY, T-1)`

`compute` would be executing the mathematical computations in the Java part of SOLAR while using a mathematical expression of the kinetics. For instance, with the simplified Michaelis-Menten equation:

$$[P]_{T+1} = V_m \frac{[S]_T}{[S]_T + K_m} + [P]_T \quad (3)$$

The discretization part can handle as many levels as needed, as shown in Fig .6 representing the evolution of the concentration of 4 metabolites in the Glycolysis Pathway of *S.Cerevisiae* after an initial pulse of glucose.

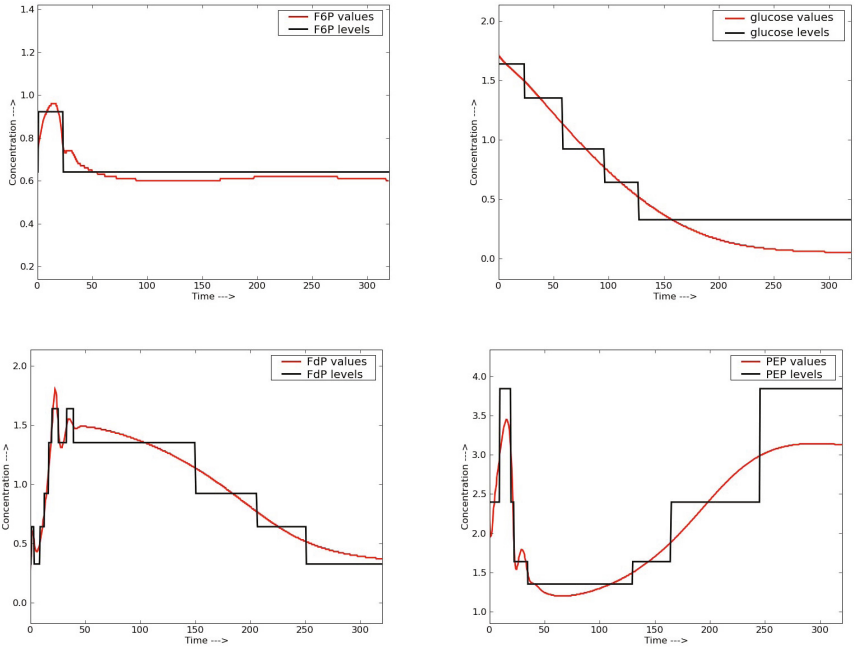


Fig. 6. Discretization with 7 common levels of the concentration of 4 metabolites in the Glycolysis Pathway of *S.Cerevisiae* after an initial pulse of glucose

6 Conclusions

As we found that our results (for time T=0) agreed with existing background knowledge in biology and our ODEs-based simulator, this paper showed a method to deal with the kinetics of metabolic pathways with a symbolic model (i.e. Fig 1). We explained how to discretize biology experiments into relevant levels to be used with ILP and logic programs in the large. Moreover, based on these discretization of concentration into levels, we explained our process to transform Michaelis-Menten analytical kinetics equation into logic rules, the authors are not aware of any previous work in this direction. Therefore the originality of the work is given by the capacity of a logical model to find the dynamic response of micro-organism when a pulse of glucose has been made. We think that this approach improves the accuracy of the metabolic flux analysis. Allowing for other kinds of kinetic modeling (two substrate and/or two products reactions) would enable us to work with more complete models.

As in [5], this approach tries to study the behaviour of many ordinary differential equations while considering a symbolic model with its advantages whereof the statistical evaluation of hypotheses. The process of statistically evaluating hypotheses, thanks

to BDD-EM [11], is seen as a good method to find relevant knowledge among the large quantity of processed data. The practical validity of this full process (including discretization) has been shown by the results of this paper while working in a well-known theoretical framework [28, 31]. We strongly believe that the use of time series discretization and a kinetic modeling to enable ILP to deal with ODE will yield great results. We also prefer to consider knowledge discovery as an iterative loop where one must review his knowledge base in the light of new findings (i.e. add "New KB" next turn in Fig. 2).

Still, our modeling can be improved, and time and concentration discretization could be finer. Experiments dealing with more than 3 levels and many time steps will be lead on the Glycolysis and Pentose Phosphate pathways of another bacteria, *Saccharomyces Cerevisiae* (yeast), with both real world data from experiments and simulated data. More experiments with enhancing and updating the knowledge base on this dataset is necessary to get more accurate results. A more global approach of discretizing experimental data and using it in conjunction with automatically generated symbolic pathways extracted from KEGG [14, 15] can be applied regardless of the model chosen for infering new knowledge. This approach can be generically applied to turn quantitative results from systems biology into qualitative (symbolic) ones.

References

1. Kitano, H.: Systems biology toward system-level understanding of biological systems. Science 295, 1662–1664 (2002)
2. Baral, C., Chancellor, K., Tran, N., Tran, N., Joy, A., Berens, M.: A knowledge based approach for representing and reasoning about signaling networks. In: Proc. of the 12th Int. Conf. on Intelligent Systems for Molecular Biology, pp. 15–22 (2004)
3. Juvan, P., Demsar, J., Shaulsky, G., Zupan, B.: Genepath: from mutations to genetic networks and back. Nucleic Acids Res. 33 (2005)
4. King, R., Whelan, K., Jones, F., Reiser, P., Bryant, C., Muggleton, S., Kell, D., Olivier, S.: Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427, 247–252 (2004)
5. King, R., Garrett, S., Coghill, G.: On the use of qualitative reasoning to simulate and identify metabolic pathways. Bioinformatics 21, 2017–2026 (2005)
6. Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A., Muggleton, S.: Application of abductive ILP to learning metabolic network inhibition from temporal data. Machine Learning 64, 209–230 (2006)
7. Tiwari, A., Talcott, C., Knapp, M., Lincoln, P., Laderoute, K.: Analyzing Pathways Using SAT-Based Approaches. In: Anai, H., Horimoto, K., Kutsia, T. (eds.) Ab 2007. LNCS, vol. 4545, pp. 155–169. Springer, Heidelberg (2007)
8. Doncescu, A., Yamamoto, Y., Inoue, K.: Biological systems analysis using Inductive Logic Programming. In: IEEE International Symp. on Bioinf. and Life Science Computing (2007)
9. Dworschak, S., Grell, S., Nikiforova, V., Schaub, T., Selbig, J.: Modeling biological networks by action languages via answer set programming. Constraints 13, 21–65 (2008)
10. Fages, F., Soliman, S., France, I.R.: Model Revision from Temporal Logic Properties in Computational Systems Biology. In: De Raedt, L., Frasconi, P., Kersting, K., Muggleton, S.H. (eds.) Probabilistic Inductive Logic Programming. LNCS (LNAI), vol. 4911, pp. 287–304. Springer, Heidelberg (2008)

11. Inoue, K., Sato, T., Ishihata, M., Kameya, Y., Nabeshima, H.: Evaluating abductive hypotheses using and EM algorithm on BDDs. In: Proc. of IJCAI 2009, pp. 815–820. AAAI Press (2009)
12. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing 2, 291–298 (1994)
13. Ji, S., Krishnapuram, B., Carin, L.: Variational bayes for continuous hidden markov models and its application to active learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 522–532 (2006)
14. Kanehisa, M., Goto, S.: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30 (2000)
15. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y.: KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, 480–484 (2008)
16. Nabeshima, H., Iwanuma, K., Inoue, K.: SOLAR: A consequence finding system for advanced reasoning. In: Proc. of the 11th International Conference TABLEAUX 2003. LNCS (LNAI), vol. 2786, pp. 257–263 (2003)
17. Ishihata, M., Kameya, Y., Sato, T., Minato, S.: Propositionalizing the EM algorith by BDDs. Technical report, TR08-0004, Dept. Comp. Sc., Tokyo Instute of Technology (2008)
18. Benhamou, F.: Interval Constraint Logic Programming. In: Podelski, A. (ed.) Constraint Programming: Basics and Trends. LNCS, vol. 910, pp. 1–21. Springer, Heidelberg (1995)
19. Geurts, P.: Pattern Extraction for Time Series Classification. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 115–127. Springer, Heidelberg (2001)
20. Keogh, E., Lin, J., Fu, A.: HOT SAX: efficiently finding the most unusual time series subsequence. In: 5th IEEE International Conference on Data Mining (2005)
21. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. of the IEEE 77, 257–286 (1989)
22. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6, 461–464 (1978)
23. Cheeseman, P., Stutz, J.: Bayesian classification (autoclass): Theory and results. In: Advances in Knowledge Discovery and Data Mining, pp. 153–180. The MIT Press (1995)
24. Beal, M.: Variational Algorithms for Approximate Bayesian Inference. PhD thesis, Gatsby Comp. Neurosc. Unit, University College London (2003)
25. De Raedt, L.: Logical and Relational Learning. Springer, Heidelberg (2008)
26. Kameya, Y., Synnaeve, G., Doncescu, A., Inoue, K., Sato, T.: A bayesian hybrid approach to unsupervised time series discretization. In: International Conference on Technologies and Applications of Artificial Intelligence, pp. 342–349 (2010)
27. Chassagnole, C., Rodrigues, J., Doncescu, A., Yang, L.T.: Differential evolutionary algorithms for in vivo dynamic analysis of glycolysis and pentose phosphate pathway in Escherichia Coli. A. Zomaya (2006)
28. Mooney, R.: Integrating abduction and induction in machine learning. In: Working Notes of the IJCAI 1997 Workshop on Abduction and Induction in AI, pp. 37–42 (1997)
29. Inoue, K.: Linear resolution for consequence finding. Artificial Intelligence 56, 301–353 (1992)
30. Muggleton, S.: Inverse entailment and progol. New Generation Computing 13, 245–286 (1995)
31. Inoue, K.: Induction as consequence finding. Machine Learning 55, 109–135 (2004)
32. Peters-Wendisch, P., Schiel, B., Wendisch, V., Katsoulidis, E., et al.: Pyruvate carboxylase is a major bottleneck for glutamate and lysine production by corynebacterium glutamicum. Molecular Microbiol. Biotechnol. 3 (2001)
33. Ray, O., Whelan, K., King, R.: A nonmonotonic logical approach for modelling and revising metabolic networks. In: IEEE Complex, Intelligent and Software Intensive Systems (2009)

Quasi-Steady State Approximations and Multistability in the Double Phosphorylation-Dephosphorylation Cycle

Guido Dell'Acqua and Alberto Maria Bersani

Dipartimento di Scienze di Base e Applicate all'Ingegneria
Sapienza University of Rome, Via Antonio Scarpa 16, 00161 Rome, Italy
{dellacqua,bersani}@dmmm.uniroma1.it

Abstract. In this paper we analyze the double phosphorylation-dephosphorylation cycle (or double futile cycle), which is one of the most important biochemical mechanisms in intracellular reaction networks, in order to discuss the applicability of the standard quasi steady-state approximation (sQSSA) to complex enzyme reaction networks, like the ones involved in intracellular signal transduction. In particular we focus on what we call "complex depletion paradox", according to which complexes disappear in the conservation laws, in contrast with the equations of their dynamics. In fact, in common literature the intermediate complexes either are ignored or are supposed to rapidly become negligible in the quasi steady-state phase, differently from what really happens, as shown studying the cycle without any quasi-steady state approximation. Applying the total quasi steady-state approximation (tQSSA) to the double phosphorylation-dephosphorylation cycle, we show how to solve the apparent paradox, without the need of further hypotheses, like, for example, the substrate sequestration.

1 Introduction: Michaelis-Menten Kinetics and Quasi Steady-State Approximation

Michelis-Menten kinetics [1,2,3,4] represents a fundamental milestone in biochemistry, as it gives a very good approximation of the dynamics of the different enzymes involved. Its formulation considers a reaction where a substrate S binds an enzyme E reversibly to form a complex C. The complex can then decay irreversibly to a product P and the enzyme, which is then free to bind another molecule of the substrate.

This process is summarized in the scheme

$$E + S \underset{d}{\overset{a}{\rightleftarrows}} C \xrightarrow{k} E + P, \tag{1}$$

where a, d and k are kinetic parameters (supposed constant) associated with the reaction rates. The fundamental step is modeling all of the intermediate reactions, including binding, dissociation and release of the product using mass action and conservation laws. This leads to an ordinary differential equation (ODE) for each involved complex and substrate, where the concentration variation for each reactant is proportional to the reactant concentrations. We refer to this as the full system.

From now on we will use the same symbols for the names of the enzymes and their concentrations. For (1) the equations are

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 155–172, 2013.
© Springer-Verlag Berlin Heidelberg 2013

$$\frac{dS}{dt} = -a(E_T - C)S + dC \tag{2}$$

$$\frac{dC}{dt} = a(E_T - C)S - (d+k)C\,. \tag{3}$$

with initial conditions

$$S(0) = S_T, \quad C(0) = 0, \tag{4}$$

and conservation laws

$$E + C = E_T, \quad S + C + P = S_T\,. \tag{5}$$

The initial conditions give the concentrations of S and C at the beginning of the reaction, and their dynamics is described by the ODEs, while E and P can be deduced from S and C via (5). Here E_T is the total enzyme concentration, assumed to be free at time $t = 0$. Also the total substrate concentration, S_T, is free at $t = 0$. This is called Michaelis-Menten (MM) kinetics [3,4,5,6]. Let us observe that the system (2–3) admits only one asymptotic solution. This solution is given by $C = S = 0$, so that from (5) we get asymptotically $P = S_T$ and $E = E_T$. This means that all the substrate eventually becomes product due to the irreversibility, while the enzyme eventually is free and the complex concentration tends to zero.

Assuming that the complex concentration is approximately constant after a short transient phase leads to the usual Michaelis-Menten (MM) approximation, or *standard quasi steady-state approximation* (sQSSA). It leads to an ODE for the substrate while the complex is assumed to be in a quasi-steady state (i.e., $\frac{dC}{dt} \approx 0$):

$$\frac{dS}{dt} \approx -kC \approx -\frac{V_{max}S}{K_M + S}\,, \quad S(0) = S_T\,, \tag{6}$$

where

$$V_{max} = kE_T\,, \quad K_M = \frac{d+k}{a}\,. \tag{7}$$

Segel and Slemrod [7] showed that a necessary condition for the validity of the sQSSA approximation is

$$E_T \ll K_M + S_T$$

which is valid when the enzyme concentration is much lower than either the substrate concentration or the Michaelis constant K_M [5,7]. This condition is usually fulfilled for *in vitro* experiments, but often breaks down *in vivo* [8,9]. However, to simulate physiologically realistic *in vivo* scenarios, one faces the problem that the MM approximation is no longer valid, as mentioned above. Hence, even though the kinetic constants such as K_M are identical *in vivo* and *in vitro*, they need to be implemented in an approximation which is valid for the system under investigation.

Approximations such as the *total QSSA* (tQSSA) [10,11,12], which is valid for a broader range of parameters covering both high and low enzyme concentrations, have been introduced recently. It arises by introducing the total substrate

$$\overline{S} = S + C.$$

(2)–(3) then become

$$\begin{aligned}\frac{d\overline{S}}{dt} &= -kC\\ \frac{dC}{dt} &= a[C^2-(E_T+\overline{S}+K_M)C+E_T\overline{S}]\ .\end{aligned} \tag{8}$$

Assuming that the complex is in a quasi steady-state $\left(\frac{dC}{dt}\approx 0\right)$ yields the tQSSA

$$\frac{d\overline{S}}{dt}\approx -kC_-(\overline{S}),\quad \overline{S}(0)=S_T, \tag{9}$$

where

$$C_-(\overline{S})=\frac{(E_T+K_M+\overline{S})-\sqrt{(E_T+K_M+\overline{S})^2-4E_T\overline{S}}}{2} \tag{10}$$

is the only biologically allowable solution of $\frac{dC}{dt}=0$ in (8). Tzafriri [12] showed that the tQSSA is valid whenever

$$\varepsilon_{Tz}:=\frac{K}{2S_T}\left(\frac{E_T+K_M+S_T}{\sqrt{(E_T+K_M+S_T)^2-4E_TS_T}}-1\right)\ll 1, \tag{11}$$

where $K=\frac{k}{a}$, and that this is always roughly valid in the sense that

$$\varepsilon_{Tz}\le\frac{K}{4K_M}\le\frac{1}{4}. \tag{12}$$

The parameter K is known as the Van Slyke-Cullen constant.

Most of current literature uses the sQSSA also to describe the network of enzyme reactions involved in the intracellular signal transduction. However, *in vivo* the reactions are coupled in complex networks or cascades of intermediate, second messengers with successive reactions, competition between substrates, feedback loops etc. In some cases approximations of such scenarios have been carried out within the MM scheme, not only without any examination of its applicability, but also neglecting the complexes involved in the reactions (see for example [13,14,15]). Other authors [16] make use of conservation laws that account for the presence of the complexes. Nevertheless the asymptotic values of the reactants do not yet correspond to the values obtained integrating numerically the full systems. In order to explain this apparent incoherence we must underline that the sQSSA, as every QSSA, represents the system dynamics after a (in general short) transient phase, during which the substrates are partially bound and the complexes begin to form. Consequently, since the QSSA is applied considering the complexes substantially constant, the total concentration of free and activated substrate(s) will be considered constant, but its value, due to the presence of complexes, cannot coincide with the initial substrate value, when all the complex concentrations

were equal to zero. Setting S_T as initial value for the total amount of (inactive and activated) substrate concentrations in the sQSSA naturally brings to wrong conclusions, since the system is forced to fulfill a conservation law that implicitly neglects all the complex concentrations. This is clearly shown in Fig. 6.

In this paper we study the double phosphorylation-dephosphorylation, on one side showing that the use of the sQSSA brings to wrong asymptotic values, on the other showing that the use of the tQSSA brings to correct predictions for the asymptotic concentrations of all the reactants, because the total variables take simultaneously into account substrates and complexes. In particular, we show that the sQSSA can predict bistability for large value ranges, whereas the full system shows monostability.

2 The Covalent Modification Cycle

A case where it is important to consider the contribution from intermediate complexes is the ubiquitous mechanism of covalent modification, such as the cycle of phosphorylation and subsequent dephosphorylation of an enzyme. This reaction is very important in every intracellular pathway, because the process of phosphorylation and dephosphorylation is one of the most important to activate and inactivate enzymes. The mechanism, which provides the building blocks of the MAPK cascade, consists of a substrate S, which can be modified, for example by phosphorylation, to the form P. Vice versa, P can be transformed, e.g. by dephosphorylation, back to S. The scenario investigated in the ground-breaking work [17] assumes that the enzymes follow the Michaelis-Menten reaction mechanism, given by

$$\begin{aligned} S + E_1 &\underset{d_1}{\overset{a_1}{\rightleftarrows}} C_1 \xrightarrow{k_1} E_1 + P \\ P + E_2 &\underset{d_2}{\overset{a_2}{\rightleftarrows}} C_2 \xrightarrow{k_2} E_2 + S \end{aligned} \tag{13}$$

The reaction is governed by the coupled ODEs

$$\begin{aligned} \frac{dS}{dt} &= -a_1 E_1 \cdot S + d_1 C_1 + k_2 C_2 \\ \frac{dC_1}{dt} &= a_1 E_1 \cdot S - (d_1 + k_1) C_1 \\ \frac{dC_2}{dt} &= a_2 E_2 \cdot P - (d_2 + k_2) C_2 \end{aligned} \tag{14}$$

with initial conditions

$$S(0) = S_T\,, \qquad C_i(0) = 0\,, \quad i = 1,2$$

and conservation laws

$$S_T = S + C_1 + C_2 + P, \tag{15}$$

$$E_{i,T} = E_i + C_i\,, \quad i = 1,2\,. \tag{16}$$

In the MAPK pathway, the upstream kinase (denoted MKKK, i.e., MAP kinase kinase kinase; for example Raf), when activated, phosphorylates the immediately downstream target, which is also a kinase (MAPKK, i.e., MAP kinase kinase, for example MEK) successively on two specific sites, eventually activating it. This last double-phosphorylated kinase (MAPKK-PP) acts on the MAPK (for example ERK) through specific phosphorylation events on two distinct sites. The activated MAPK is then responsible for further downstream signalling. The activated cascade is shut down by the reverse action of specific phosphatases [18,19], whose outcome is the time modulation of the signal, probably through the regulation of the active kinase (for example, transient versus sustained activation). Moreover, the phosphatase controls the steady state level of activated MAPK, which, in turn, controls downstream processes.

The double phosphorylation, as well as double dephosphorylation of MAPK, was recently modeled taking into consideration the competition between the pools of MAPK with different phosphorylation states [13,14]. We model this process by assuming that competition holds for both the phosphorylation as well as the dephosphorylation processes as in [13]. In presence of a reaction cycle, it is natural to expect that the complexes are continuously depleted and created and that in a stationary state their concentrations cannot tend to zero. This fact was already observed in [17] in the case of the phosphorylation-dephosphorylation cycle. Nevertheless most of current literature, when applying the sQSSA to complex schemes, imposes implicitly the depletion of all the complexes, seriously affecting the conservation laws and consequently the asymptotic values of the reactant concentrations. However, as observed in the previous section, even when the conservation law for substrates takes into account the complexes, the application of the sQSSA corresponds to ignore the initial, rapid transient phase, where complexes begin to be created and the total amount of free (inactive, monophosphorylated and double phosphorylated) substrates is no more equal to S_T. On the other hand, the tQSSA cannot produce this situation, because the substrates and the complexes formed by them are included in the same (total) variables and the system cannot distinguish between free substrates and bound ones. For this reason, differently from the sQSSA, the tQSSA saves the conservation laws and produces the same asymptotic values as the full system.

The most dramatic consequence is that, when applied to well-known mechanism, like, e.g., the (double) phosphorylation-dephosphorylation cycle, or the MAPK cascade, the sQSSA predicts phenomena which do not appear when the mechanisms are studied by means of the full system of equations describing the systems or by means of the tQSSA. For example, in [20] it is shown that the supposed depletion of the complexes in the sQSSA brings to uncorrect asymptotic values of the inactive and active substrate in the Goldbeter-Koshland cycle, as predicted in [17]. On the other hand, the tQSSA reproduces in a very satisfactorily way not only the asymptotic values but also the dynamics of the reactants.

In [21] it is shown that in several mechanisms (like, e.g., the antagonistic toggle switch) the sQSSA can yield bistability even when the system, described by the full system of equations, is not bistable. In [22] and [20] it is shown that when the system undergoes oscillations any QSSA risks to fail, because the central hypothesis for the quasi steady-state assumption is a substantial equilibrium for the complexes, which

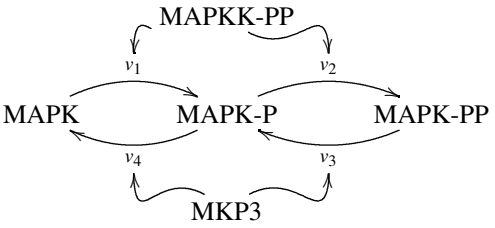
cannot be guaranteed in presence of the substrate oscillations. Flach and Schnell tested their considerations on the van Slyke-Cullen mechanism [23], while Pedersen et al. studied the MAPK cascade with feedback, as in [24]. In [25,26] it is shown that the tQSSA reproduces zero-order ultrasensitivity in the Goldbeter-Koshland cycle, according to what was predicted in [17], while the sQSSA, for a wide range of parameter values, is not able to yield ultrasensivity whenever it is expected by the theory.

As already remarked, the main reason for the failures of the sQSSA lays in the fact that the complexes, far from asymptotically going to zero, are not accounted in the conservation laws. This implies the prediction, by the sQSSA, of lower total concentrations than what is expected or experimentally observed. This fact induced some authors [27,28,29], either to re-discover the sequestration of the substrates by the kinases or the phosphatates, or to postulate the existence of some substrate sequestration mechanisms, by means of competition or inhibition, made by other enzymes. Actually, the use of the tQSSA not only does not need any additional sequestration hypothesis, but also correctly accounts for the exact asymptotic concentration values, of inactive, active and bound substrates. Let us remark that one of the main advantages of any QSSA is the simplification of the mathematical scheme describing the enzyme reactions, which allows us to capture many qualitative and quantitative features that could not observed by means of the full system.

In the next section we will compare the application of the sQSSA and the tQSSA to the study of the double phosphorylation-dephosphorylation mechanism, showing that, while the latter brings to the same results as the full system, the former brings to consistent errors, mainly in the asymptotic concentration values, predicting bistability for large value ranges, whereas the full system (and the tQSSA) show monostability.

3 The Equations

We want to study the scheme



(17)

where the double-phosphorylated kinase (MAPKK-PP) acts on the MAPK (for example ERK) through specific phosphorylation events on two distinct sites, while the phosphatase (MKP3) acts with a reverse action on MAPK-PP, inactivating it.

Several authors have modeled this reaction. In [16] it is supposed that both phosphorylation and dephosphorylation happen in only one step. In [14] the authors describe the dephosphorylation reaction as a two step mechanism. In [30] both phosphorylation and dephosphorylation are assumed to be two step reactions. We will focus our analysis on the models described in [16] and [14].

The reaction, as described in [16], can be summarized as follows

$$\begin{aligned} M+E &\underset{k_{-11}}{\overset{k_{11}}{\rightleftarrows}} M-E=C_1 \xrightarrow{k_{12}} M_p+E \\ M_p+E &\underset{k_{-21}}{\overset{k_{21}}{\rightleftarrows}} M_p-E=C_2 \xrightarrow{k_{22}} M_{pp}+E \\ M_{pp}+F &\underset{k_{-31}}{\overset{k_{31}}{\rightleftarrows}} M_{pp}-F=C_3 \xrightarrow{k_{32}} M_p+F \\ M_p+F &\underset{k_{-41}}{\overset{k_{41}}{\rightleftarrows}} M_p-F=C_4 \xrightarrow{k_{42}} M+F \end{aligned} \tag{18}$$

where M, M_p and M_{pp} respectively represent the inactive, the monophosphorylated and the double phosphorylated MAPK, E and F are respectively the kinase MAPKK-PP and the phosphatase MKP3 and C_i are the intermediate complexes.

Using the law of mass action, the full system of equations governing the dynamics of the system is therefore

$$\begin{aligned} \frac{dM}{dt} &= -k_{11}ME + k_{-11}C_1 + k_{42}C_4 \\ \frac{dM_p}{dt} &= -k_{21}M_pE + k_{-21}C_2 + k_{-41}C_4 - k_{41}M_pF + k_{32}C_3 + k_{12}C_1 \\ \frac{dM_{pp}}{dt} &= -k_{31}M_{pp}F + k_{-31}C_3 + k_{22}C_2 \\ \frac{dC_1}{dt} &= k_{11}ME - (k_{-11}+k_{12})C_1 \\ \frac{dC_2}{dt} &= k_{21}M_pE - (k_{-21}+k_{22})C_2 \\ \frac{dC_3}{dt} &= k_{31}M_{pp}F - (k_{-31}+k_{32})C_3 \\ \frac{dC_4}{dt} &= k_{41}M_pF - (k_{-41}+k_{42})C_4 \end{aligned} \tag{19}$$

with initial conditions

$$M(0)=M_T, \quad M_p(0)=M_{pp}(0)=0, \quad C_i(0)=0, \tag{20}$$

where $i=1,\ldots,4$ and conservation laws

$$M+M_p+M_{pp}+C_1+C_2+C_3+C_4=M_T, \tag{21}$$

$$E+C_1+C_2=E_T, \quad F+C_3+C_4=F_T\,. \tag{22}$$

Setting $K_i=\dfrac{k_{-i1}+k_{i2}}{k_{i1}}$, $i=1,\ldots,4$, the sQSSA implies that

$$C_1=\frac{ME}{K_1}, \quad C_2=\frac{M_pE}{K_2}, \quad C_3=\frac{M_{pp}F}{K_3}, \quad C_4=\frac{M_pF}{K_4} \tag{23}$$

which give

$$\begin{aligned}
\frac{dM}{dt} &= -\tfrac{k_{12}}{K_1} M E + \tfrac{k_{42}}{K_4} M_p F \\
\frac{dM_p}{dt} &= -\tfrac{k_{22}}{K_2} M_p E - \tfrac{k_{42}}{K_4} M_p F + \tfrac{k_{32}}{K_3} M_{pp} F + \tfrac{k_{12}}{K_1} M E \\
\frac{dM_{pp}}{dt} &= \tfrac{k_{22}}{K_2} M_p E - \tfrac{k_{32}}{K_3} M_{pp} F
\end{aligned} \tag{24}$$

where

$$E = \frac{E_T}{1 + \frac{M}{K_1} + \frac{M_p}{K_2}}, \quad F = \frac{F_T}{1 + \frac{M_{pp}}{K_3} + \frac{M_p}{K_4}}$$

and

$$M(0) = M_T, \quad M_p(0) = M_{pp}(0) = 0\,.$$

Let us observe that, when we apply the sQSSA, we set all the complexes constant. This means that the conservation law becomes

$$M + M_p + M_{pp} = constant \tag{25}$$

The constant in (25) is, in general, different from M_T but the application of both the sQSSA and the conservation law since the beginning of the reaction naturally brings to the equality $constant = M_T$. This leads to the complex depletion paradox: the application of the sQSSA implies that, while the complexes are related to the substrates by the equations (23), they are implicitly set equal to zero, because of (25). The consequences are that the sQSSA predicts asymptotic values for the different substrate species which are higher than those ones predicted by the full system.

Let us introduce the *total* QSSA (tQSSA) by setting

$$\overline{M} = M + C_1, \quad \overline{M}_p = M_p + C_2 + C_4, \quad \overline{M}_{pp} = M_{pp} + C_3\,. \tag{26}$$

In terms of these new variables, supposing that the complexes are in a quasi steady-state, the set of equations (19) becomes

$$\begin{aligned}
\frac{d\overline{M}}{dt} &= k_{42} C_4 - k_{12} C_1\,, \\
\frac{d\overline{M}_p}{dt} &= -k_{22} C_2 - k_{42} C_4 + k_{32} C_3 + k_{12} C_1\,, \\
\frac{d\overline{M}_{pp}}{dt} &= -k_{32} C_3 + k_{22} C_2\,.
\end{aligned} \tag{27}$$

with same initial conditions of (20) and conservation law

$$\overline{M} + \overline{M}_p + \overline{M}_{pp} = M_T \tag{28}$$

and where the complexes are given in function of the total substrates by

$$\begin{aligned}
(\overline{M} - C_1)(E_T - C_1 - C_2) - K_1 C_1 &= 0 \\
(\overline{M}_p - C_2 - C_4)(E_T - C_1 - C_2) - K_2 C_2 &= 0 \\
(\overline{M}_{pp} - C_3)(F_T - C_3 - C_4) - K_3 C_3 &= 0 \\
(\overline{M}_p - C_2 - C_4)(F_T - C_3 - C_4) - K_4 C_4 &= 0
\end{aligned} \tag{29}$$

The above system of equations has been obtained using the QSSA ($\frac{dC_i}{dt} = 0$), (21) and (22) in the last 4 equations of (19).

Let us remark that, in this case, we do not observe any complex depletion paradox, because the initial conditions are given on the total substrates, no matter if the substrates are free or bound. Consequently, even after the application of the quasi steady-state approximation, the conservation law is fully respected, without any additional condition on the complexes. Thus the tQSSA yields the same asymptotic values for all the reactants (complexes included) as the full system.

Similarly, the reaction described in [14] can be summarized as follows:

$$\begin{aligned}
M + E &\underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} M - E = C_1 \xrightarrow{k_2} M_p + E \\
M_p + E &\underset{k_{-3}}{\overset{k_3}{\rightleftarrows}} M_p - E = C_2 \xrightarrow{k_4} M_{pp} + E \\
M_{pp} + F &\underset{h_{-1}}{\overset{h_1}{\rightleftarrows}} C_3 \xrightarrow{h_2} C_4 \underset{h_{-3}}{\overset{h_3}{\rightleftarrows}} M_p + F \\
M_p + F &\underset{h_{-4}}{\overset{h_4}{\rightleftarrows}} C_5 \xrightarrow{h_5} C_6 \underset{h_{-6}}{\overset{h_6}{\rightleftarrows}} M + F
\end{aligned} \tag{30}$$

which can be translated, as before, into differential equations by the law of mass action:

$$\begin{aligned}
\frac{dM}{dt} &= -k_1ME + k_{-1}C_1 + h_6C_6 - h_{-6}MF \\
\frac{dM_p}{dt} &= -k_3M_pE + k_{-3}C_2 + h_3C_4 - h_{-3}M_pF - h_4M_pF + k_2C_1 + h_{-4}C_5 \\
\frac{dM_{pp}}{dt} &= -h_1M_{pp}F + h_{-1}C_3 + k_4C_2 \\
\frac{dC_1}{dt} &= k_1ME - (k_{-1} + k_2)C_1 \\
\frac{dC_2}{dt} &= k_3M_pE - (k_{-3} + k_4)C_2 \\
\frac{dC_3}{dt} &= h_1M_{pp}F - (h_{-1} + h_2)C_3 \\
\frac{dC_4}{dt} &= h_{-3}M_pF + h_2C_3 - h_3C_4 \\
\frac{dC_5}{dt} &= h_4M_pF - (h_{-4} + h_5)C_5 \\
\frac{dC_6}{dt} &= h_5C_5 + h_{-6}MF - h_6C_6
\end{aligned} \tag{31}$$

with initial conditions

$$M(0) = M_T, \quad M_p(0) = M_{pp}(0) = 0, \quad C_i(0) = 0, \tag{32}$$

where $i = 1, \ldots, 6$ and conservation laws

$$M + M_p + M_{pp} + C_1 + C_2 + C_3 + C_4 + C_5 + C_6 = M_T, \tag{33}$$

$$E + C_1 + C_2 = E_T, \quad F + C_3 + C_4 + C_5 + C_6 = F_T. \tag{34}$$

Setting

$$K_i = \frac{k_{-i} + k_{i+1}}{k_i}, i = 1,2, \quad H_i = \frac{h_{-i} + h_{i+1}}{h_i}, i = 1,4$$

the sQSSA approximation implies that

$$C_1 = \frac{ME}{K_1}, \quad C_2 = \frac{M_p E}{K_2}, \quad C_3 = \frac{M_{pp} F}{H_1},$$

$$C_4 = \frac{1}{h_3}\left[h_2 \frac{M_{pp}}{H_1} + h_{-3} M_p\right] F, \quad C_5 = \frac{M_p F}{H_4}, \quad C_6 = \frac{1}{h_6}\left[h_5 \frac{M_p}{H_4} + h_{-6} M\right] F$$

which give

$$\begin{aligned} \frac{dM}{dt} &= -\tfrac{k_2}{K_1} ME + \tfrac{h_5}{H_4} M_p F \\ \frac{dM_p}{dt} &= -\tfrac{k_4}{K_2} M_p E - \tfrac{h_5}{H_4} M_p F + \tfrac{h_2}{H_1} M_{pp} F + \tfrac{k_2}{K_1} ME \\ \frac{dM_{pp}}{dt} &= \tfrac{k_4}{K_2} M_p E - \tfrac{h_2}{H_1} M_{pp} F \end{aligned} \tag{35}$$

where

$$E = \frac{E_T}{1 + \frac{M}{K_1} + \frac{M_p}{K_2}}$$

and

$$F = \frac{F_T}{1 + \frac{M_{pp}}{H_1}(1 + \frac{h_2}{h_3}) + H_5 M_p + \frac{h_{-6}}{h_6} M}$$

with $H_5 = \frac{h_{-3}}{h_3} + \frac{1}{H_4} + \frac{h_5}{h_6 H_4}$ and

$$M(0) = M_T, \quad M_p(0) = M_{pp}(0) = 0.$$

Let us introduce again the tQSSA by setting

$$\overline{M} = M + C_1 + C_6, \quad \overline{M}_p = M_p + C_2 + C_4 + C_5, \quad \overline{M}_{pp} = M_{pp} + C_3.$$

In terms of these new variables, supposing that the complexes are in a quasi steady-state, the set of equations (31) becomes

$$\begin{aligned} \frac{d\overline{M}}{dt} &= -k_2 C_1 + h_5 C_5, \\ \frac{d\overline{M}_p}{dt} &= k_2 C_1 - k_4 C_2 + h_2 C_3 - h_5 C_5, \\ \frac{d\overline{M}_{pp}}{dt} &= k_4 C_2 - h_2 C_3. \end{aligned} \tag{36}$$

with same initial conditions of (20) and conservation law

$$\overline{M} + \overline{M}_p + \overline{M}_{pp} = M_T \tag{37}$$

and where the complex are given in function of the total substrates by

$$\begin{array}{l}
(\overline{M} - C_1 - C_6)(E_T - C_1 - C_2) - K_1 C_1 = 0 \\
(\overline{M}_p - C_2 - C_4 - C_5)(E_T - C_1 - C_2) - K_2 C_2 = 0 \\
(\overline{M}_{pp} - C_3)(F_T - C_3 - C_4 - C_5 - C_6) - H_1 C_3 = 0 \\
h_2 C_3 + h_{-3}(\overline{M}_p - C_2 - C_4 - C_5)(F_T - C_3 - C_4 - C_5 - C_6) - h_3 C_4 = 0 \\
(\overline{M}_p - C_2 - C_4 - C_5)(F_T - C_3 - C_4 - C_5 - C_6) - H_4 C_5 = 0 \\
h_5 C_5 + h_6(\overline{M} - C_1 - C_6)(F_T - C_3 - C_4 - C_5 - C_6) = 0
\end{array} \tag{38}$$

As before, the above system of equations has been obtained using the QSSA ($\frac{dC_i}{dt} = 0$), (33) and (34) in the last 6 equations of (31).

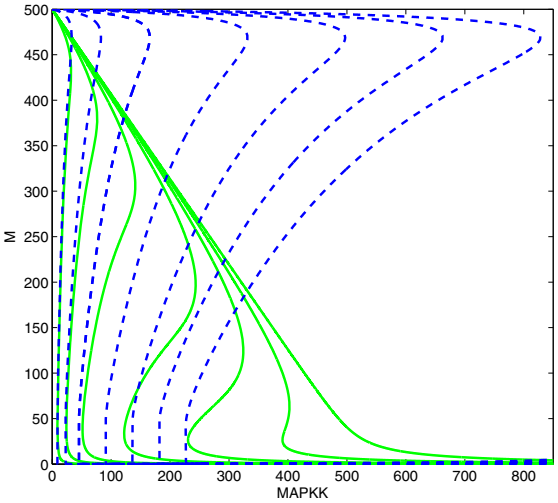


Fig. 1. Stationary branches of the inactive MAPKK (M) in the full system (19) (solid) and in its sQSSA approximation (24) (dashed), obtained varying the initial concentration of the kinase MAPKK, for different values of the initial concentration of the phosphatase: MKP3=20,50,100,200,300,400,500 (left-right); kinetic parameters as in Table 1 and $M_T = 500$

4 Numerical Simulations and Figures

We have numerically integrated (19) and (31), as well as their M-M approximations (24) and (35), with the standard MATLAB stiff integrator ODE15S. The set of kinetic parameters shown in Table (2) was taken from [14], as well as the value of M_T. The set of kinetic parameters shown in Table (1) is almost the same, except in the two-step dephosphorylation phase, where we have eliminated the parameters related to the

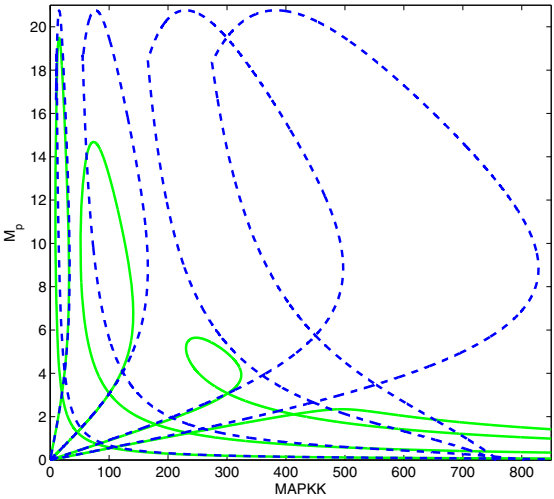


Fig. 2. Stationary branches of the mono-phosphprylated MAPKK (M_p) in the full system (19) (solid) and in its sQSSA approximation (24) (dashed), obtained varying the initial concentration of the kinase MAPKK, for different values of the initial concentration of the phosphatase: MKP3=20,100,300,500 (left-right); kinetic parameters as in Table 1 and $M_T = 500$

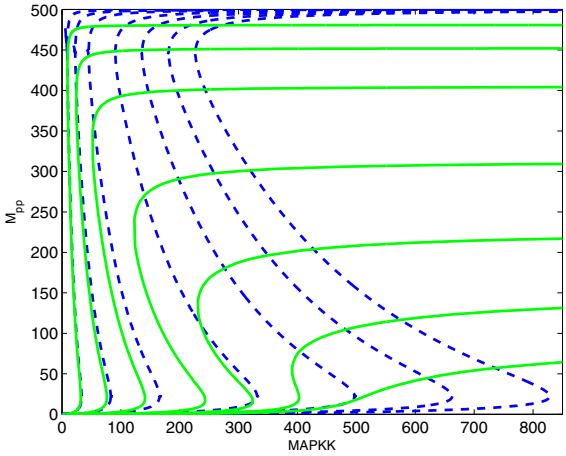


Fig. 3. Stationary branches of the double phosphorylated MAPK (M_{pp}) in the full system (19) (solid) and in its sQSSA approximation (24) (dashed), obtained varying the initial concentration of the kinase MAPKK, for different values of the initial concentration of the phosphatase: MKP3=20,50,100,200,300,400,500 (left-right); kinetic parameters as in Table 1 and $M_T = 500$

second step. Doing so, we tried to find the qualitative differences between a one-step and a two-step dephosphorylation process.

The results are shown in Fig(s). 1–7. Note that the asymptotic values of the tQSSA are not shown because, as already remarked, the tQSSA gives the same asymptotic

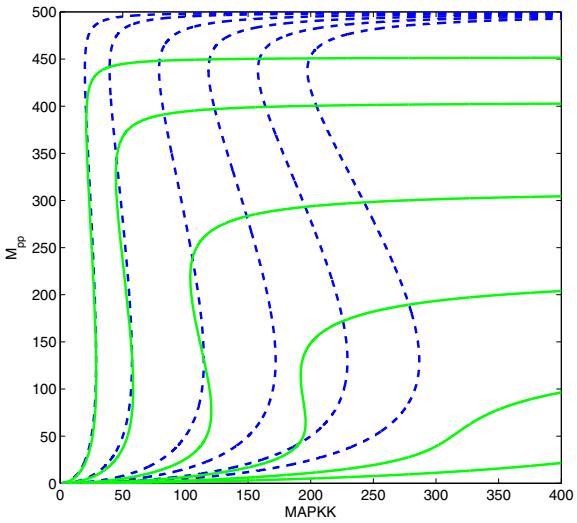


Fig. 4. Stationary branches of the double phosphorylated MAPK (M_{pp}) in the full system (31) (solid) and in its sQSSA approximation (35) (dashed), obtained varying the initial concentration of the kinase MAPKK, for different values of the initial concentration of the phosphatase: MKP3=50,100,200,300,400,500 (left-right); kinetic parameters as in Table 2 and $M_T = 500$

Table 1. Kinetic parameters of reaction (18)

k_{11}	k_{-11}	k_{12}	k_{21}
0.02	1	0.01	0.032
k_{-21}	k_{22}	k_{31}	k_{-31}
1	15	0.0045	1
k_{32}	k_{41}	k_{-41}	k_{42}
0.0092	0.01	1	0.5

values of the full system. To find the stationary branches of the three substrate forms (M, M_p and M_{pp}) we have used an iterative algorithm, that can be explained briefly as follows. First, when looking for stationary points, i. e., when we set all the first members of (19) equal to zero, it is easy to express all variables in function of the complexes C_1 and C_2. Then, we begin the iteration on a stable branch, finding the solution of the full system of ODEs via numerical integration (performed with the standard MATLAB stiff integrator ODE15S). At the subsequent step, to find the solution of the two equations for C_1 and C_2 on the same stable branch, we use as starting guess in the MATLAB's routine FSOLVE the solution obtained at the previous step, and we repeat iteratively this procedure. As expected, stability problems appeared near threshold values, where stable and unstable stationary branches cross. Fine tuning mechanisms of input values of the iterative scheme allowed us to reconstruct efficiently and satisfactorily all the three stationary branches. The computational costs for the simulations and the iterative algorithm, in terms of CPU time, are in general absolutely low, of the order of seconds for the former, and of few minutes for the latter.

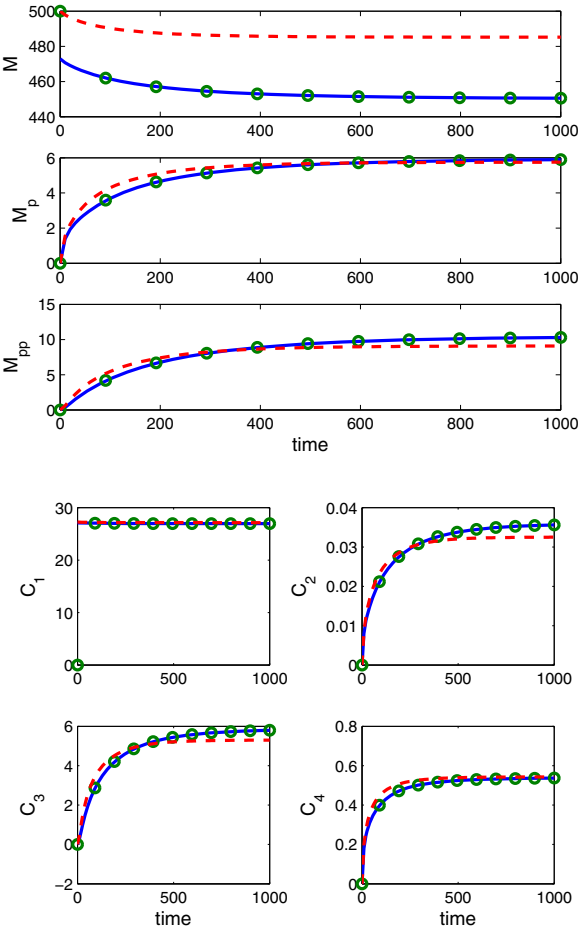


Fig. 5. Dynamics of the system (19) with MAPKK=30, MKP3=20, kinetics parameters as in Table 1 and $M_T = 500$: full system (circles), tQSSA approximation (solid), sQSSA approximation (dashed)

The results show clearly the limits of the sQSSA when applied to cycles, where the central role of the intermediates cannot be neglected. Indeed, when the total amount of enzymes is sufficiently low the sQSSA approximation shows discrete results (first two branches on the left of Fig(s). 1–4), while when the total concentration of enzymes grows the sQSSA is completely wrong both for the asymptotic values and, more importantly, for the range of bistability. Note that for $MKP3 = 400,500$ the sQSSA predicts a wide range of bistability, while the dynamics of the system is in a monostable regime. In Fig. 5 we show the dynamics of all the substrate and complex concentrations. The sQSSA fails its prediction of the behavior of M, because the initial conditions bring to ignore the presence of the complexes, while, as shown in the plot of C_1, complexes cannot be neglected, even if the initial kinase and phosphatase concentrations are much less than M_T. In Fig(s). 6–7 we show the meaning of the complex depletion paradox:

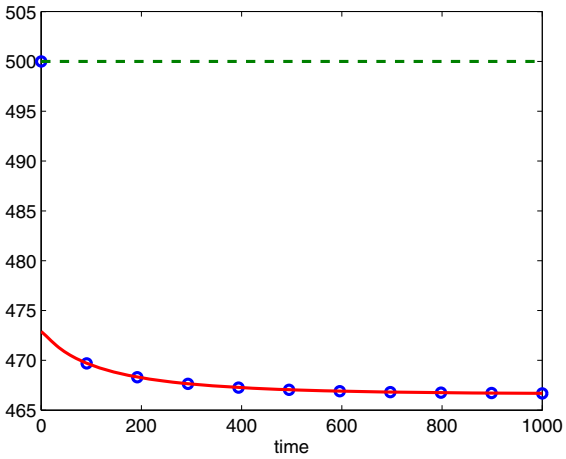


Fig. 6. Dynamics of $M+M_p+M_{pp}$ in the reaction (18): full system (circles), tQSSA approximation (solid), sQSSA approximation (dotted). Here MAPKK=30, MKP3=20, kinetic parameters as in Table 1 and $M_T = 500$. Differently from the full system and the tQSSA, the sQSSA predicts at any time a constant value M_T for $M+M_p+M_{pp}$ that would imply $C_1+C_2+C_3+C_4 = 0$.

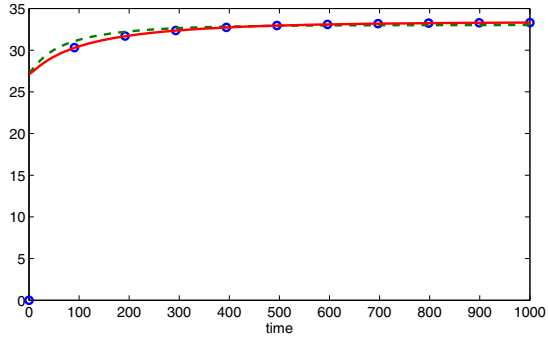


Fig. 7. Dynamics of $C_1+C_2+C_3+C_4$ in the reaction (18): full system (circles), tQSSA approximation (solid), sQSSA approximation (dashed). Here MAPKK=30, MKP3=20, kinetics parameters as in Table 1 and $M_T = 500$. In contrast with the previous figure, the sQSSA predicts at any time a value of $C_1+C_2+C_3+C_4$ different from zero.

ignoring the initial, fast transient phase, the sQSSA imposes as initial condition for the sum of the substrate concentrations $M+M_p+M_{pp}$ the value M_T of the initial concentration of the unphosphorylated substrate. Actually, even in a fast transient phase the substrates begin to be bound in the complexes and the initial condition for the quasi steady-state phase should be set less than M_T. Thus, using (25), the total contribution of substrates remains constant, with value M_T. Nevertheless, as shown in Fig. 7, relations (23) give at any time a concentration value different from zero for the total complexes $C_1+C_2+C_3+C_4$, in contrast with the previous figure, where the fact that $M+M_p+M_{pp}$ is always equal to M_T would imply that $C_1+C_2+C_3+C_4 = 0$ at any time. On the other

Table 2. Kinetic parameters of reaction (30), taken from [14]

k_1	k_{-1}	k_2	k_3
0.02	1	0.01	0.032
k_{-3}	k_4	h_1	h_{-1}
1	15	0.045	1
h_2	h_3	h_{-3}	h_4
0.092	1	0.01	0.032
h_{-4}	h_5	h_6	h_{-6}
1	0.5	0.086	0.0011

hand, in the tQSSA, considering as initial condition for the total substrate $\overline{M}(0) = M_T$ we obtain a very good approximation of the true (and in general unknown) initial condition after the transient phase, because in this phase we can reasonably suppose that only the binding of M in the complex C_1 has happened, without affecting the value of $\overline{M}$.

5 Conclusions

Many intracellular reactions are governed by threshold mechanisms. The wrong determination of the asymptotic reactant concentrations can heavily affect the mathematical model of the system and its predictions. Though the sQSSA represents a very important tool for the qualitative and quantitative study of single enzyme–substrate reactions, it has shown to be in general inadequate, for several reasons, when applied to complex reaction networks, like those ones occurring inside the cell. In this paper we show that, in the case of the phosphorylation-dephosphorylation cycle, the application of the tQSSA or, better, of the full system of equations is in general much more appropriate. One of the main advantages of the sQSSA is the simplification of the mathematical scheme describing the enzyme reactions, which allows us to capture many qualitative and quantitative features that could not observed by means of the full system. Again, as shown in this paper, the application of the tQSSA brings to predictions that are much more accurate than the sQSSA ones, quantitatively and qualitatively. Thus, the use of the total substrates shows to be much more than a mere variable change, bringing to a correct explanation and prediction of the sequestration mechanism, which is not allowed by the sQSSA. Let us also remark that, as shown in [20], in presence of mechanisms, like feedbacks, generating oscillations, any QSSA is inadequate to approximate the full system, since substrate oscillations imply kinase and phosphatase oscillations, which are in contrast with the main quasi steady-state assumption. In this cases, in order to obtain realistic predictions, the reactions can be described only by the full system, though it contains a much greater number of equations and variables.

Our aim is to extend our investigations to more complex reactions, governing the cell functioning. In fact, it is well known that the mathematical description of the double phosphorylation–dephosphorylation mechanism is a common feature not only of the MAPK but of any reaction involving a double-step activation and the corresponding double-step deactivation. Even if we expect, in general, much more involved

phenomena, we think that our mathematical tools will be able to model, explain and predict their main characteristics.

The results that we present in this paper do need, of course, a validation through experimental data. For this purpose, we have got off the ground a collaboration with personnel from ISMAC–CNR, the Institute for Macromolecular studies (Genova, Italy), in order to test our predictions.

References

1. Henri, V.: Recherches sur la loi de l'action de la sucrase. C. R. Hebd. Acad. Sci. 133, 891–899 (1901)
2. Henri, V.: Über das gesetz der wirkung des invertins. Z. Phys. Chem. 39, 194–216 (1901)
3. Michaelis, L., Menten, M.L.: Die kinetik der invertinwirkung. Biochem. Z. 49, 333–339 (1913)
4. Briggs, G.E., Haldane, J.B.S.: A note on the kinetics of enzyme action. Biochem. J. 19, 338–339 (1925)
5. Segel, L.: On the validity of the steady state assumption of enzyme kinetics. Bull. Math. Biol. 50, 579–593 (1988)
6. Bisswanger, H.: Enzyme Kinetics. In: Principles and Methods, Wiley-VCH (2002)
7. Segel, L.A., Slemrod, M.: The quasi steady-state assumption: a case study in perturbation. Siam Rev. 31, 446–477 (1989)
8. Straus, O.H., Goldstein, A.: Zone behavior of enzymes. J. Gen. Physiol. 26, 559–585 (1943)
9. Sols, A., Marco, R.: Concentrations of metabolites and binding sites, implications in metabolic regulation. Curr. Top. Cell. Regul. 2, 227–273 (1970)
10. Laidler, K.J.: Theory of the transient phase in kinetics, with special reference to enzyme systems. Can. J. Chem. 33, 1614–1624 (1955)
11. Borghans, J., de Boer, R., Segel, L.: Extending the quasi-steady state approximation by changing variables. Bull. Math. Biol. 58, 43–63 (1996)
12. Tzafriri, A.R.: Michaelis-Menten kinetics at high enzyme concentrations. Bull. Math. Biol. 65, 1111–1129 (2003)
13. Hatakeyama, M., Kimura, S., Naka, T., Kawasaki, T., Yumoto, N., Ichikawa, M., Kim, J.H., Saito, K., Saeki, K.M., Shirouzu, M., Yokoyama, S., Konagaya, A.: A computational model on the modulation of mitogen-activated protein kinase (mapk) and akt pathways in heregulin-induced erbb signalling. Biochem. J. 373, 451–463 (2003)
14. Markevich, N.I., Hoek, J.B., Kholodenko, B.N.: Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. J. Cell Biol. 164, 353–359 (2004)
15. Chickarmane, V., Kholodenko, B.N., Sauro, H.M.: Oscillatory dynamics arising from competitive inhibition and multisite phosphorylation. J. Theor. Biol. 244, 68–76 (2007)
16. Ortega, F., Garces, J.L., Mas, F., Kholodenko, B.N., Cascante, M.: Bistability from double phosphorylation in signal transduction. kinetic and structural requirements. FEBS J. 273, 3915–3926 (2006)
17. Goldbeter, A., Koshland, D.E.: An amplified sensitivity arising from covalent modification in biological system. Proc. Natl. Acad. Sci. 78, 6840–6844 (1981)
18. Camps, M., Nichols, A., Arkinstall, S.: Dual specificity phosphatases: a gene family for control of map kinase function. FASEB J. 14, 6–16 (2000)
19. Zhan, X.L., Wishart, M.J., Guan, K.L.: Nonreceptor tyrosine phosphatases in cellular signaling: regulation of mitogen-activated protein kinases. Chem. Rev. 101, 2477–2496 (2001)
20. Pedersen, M.G., Bersani, A.M., Bersani, E.: Steady-state approximations in intracellular signal transduction – a word of caution. J. Math. Chem. 43, 1318–1344 (2008)

21. Sabouri-Ghomi, M., Ciliberto, A., Kar, S., Novak, B., Tyson, J.J.: Antagonism and bistability in protein interaction networks. J. Theor. Biol. 250, 209–218 (2008)
22. Flach, E.H., Schnell, S.: Use and abuse of the quasi-steady-state approximation. IEE Proceed. Syst. Biol. 153, 187–191 (2006)
23. van Slyke, D.D., Cullen, G.E.: The mode of action of urease and of enzymes in general. J. Biol. Chem. 19, 141–180 (1914)
24. Kholodenko, B.N.: Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. Eur. J. Biochem. 267, 1583–1588 (2000)
25. Ciliberto, A., Capuani, F., Tyson, J.J.: Modeling networks of coupled anzymatic reactions using the total quasi-steady state approximation. PLoS Comput. Biol. 3, 463–472 (2007)
26. Pedersen, M., Bersani, A.M.: The total quasi-steady state approximation simplifies theoretical analysis at non-negligible enzyme concentrations: Pseudo first-order kinetics and the loss of zero-order ultrasensitivity. J. Math. Biol. 60, 267–283 (2010)
27. Bluthgen, N., Bruggeman, F.J., Legewie, S., Herzel, H., Westerhoff, H.V., Kholodenko, B.N.: Effects of sequestration on signal transduction cascades. FEBS J. 273, 895–906 (2006)
28. Legewie, S., Schoeberl, B., Blüthgen, N., Herzel, H.: Competing docking interactions can bring about bistability in the mapk cascade. Biophys. J. 93, 2279–2288 (2007)
29. Xing, J., Chen, J.: The Goldbeter-Koshland switch in the first-order region and its response to dynamic disorder. PLOS ONE 3, e2140 (2008)
30. Salazar, C., Hofer, T.: Kinetic models of phosphorylation cycles: A systematic approach using the rapid equilibrium approximation for protein-protein interactions. Biosystems 83, 195–206 (2006)

Improving Latent Semantic Analysis of Biomedical Literature Integrating UMLS Metathesaurus and Biomedical Pathways Databases

Francesco Abate, Elisa Ficarra, Andrea Acquaviva, and Enrico Macii

Dep. of Control and Computer Engineering, Politecnico di Torino
Cso Duca degli Abruzzi 24, 10129 Torino, Italy
{francesco.abate,elisa.ficarra,andrea.acquaviva,
enrico.macii}@polito.it

Abstract. The increasing pace of biotechnological advances produced an unprecedented amount of both experimental data and biological information mostly diffused on the web. However, the heterogeneity of the data organization and the different knowledge representations open the ways to new challenges in the integration and the extraction of biological information fundamental for correctly interpreter experimental results.

In the present work we introduce a new methodology for quantitatively scoring the degree of biological correlation among biological terms occurring in biomedical abstracts. The proposed flow is based on the latent semantic analysis of biomedical literature coupled with the UMLS Metathesarurs and PubMed literature information. The results demonstrate that the structured and consolidated knowledge in the UMLS and pathway database efficiently improves the accuracy of the latent semantic analysis of biomedical literature.

Keywords: Bioinformatic, Latent semantic analysis, Text mining, Biological pathway.

1 Introduction

The pace of genetical and biological research has surprisingly spread in the last few years. The development of new technologies (DNA/RNA next generation sequencing) and the improvement of the old one (microarray for gene expression at lower costs) have being supporting the activity of scientists and researchers. Hundred of experimental results supported by the biotechnological enhancement have highlighted the mechanisms behind the complex interactions among biological entities as well as brought to light new unknown phenomena. Nevertheless, if on one side this rising technological development opens the way to amazing scenarios in the biomedical research, the problem of handling the great amount of new information calls for the need of developing new computational infrastructures. Specifically, the capability of extracting relevant knowledge from genetical and biomedical studies has become one of the major thrusts in bioinformatic research. Nowadays, the main sources of information about the biomedical research are mainly biomedical literature databases and classification systems such as ontologies and thesaurus. Most of these systems are distributed on the Web and thus

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 173–187, 2013.
© Springer-Verlag Berlin Heidelberg 2013

they are extremely easy to access and opened to the entire scientific community. However, in this scenario, the scientist have to face the twofold drawback of 1) *retrieving* the documents containing the information that he/she is looking for and 2) *correlating* the concepts in order to better extract the knowledge they provide. Both the operation are extremely time-consuming and the degree of correlation among biological concepts strongly depends on the scientist knowledge background. Therefore, an automated tool for the computation of a quantified score between biological terms according to literature information is strongly necessary in order to support the analysis of experimental results occurring during the research activity and to refine the design of the experiments by focusing on the most relevant features. The correlation score must be the result of a statistical co-occurrences analysis of biological terms both in the same article and in the related ones. Moreover, the tool must be able to integrate as many as possible sources of information, distributed on the Web and collected in different formats and structures, in order to guarantee the best accuracy and reliability for the computed biological correlation score. Section 2 overviews the state of the art where the proposed tool is contextualized. Section 3 shows the methodology used to both retrieve the biological pathway information and extract the knowledge by the overall information base. Section 4 demonstrates the effectiveness of the proposed approach in computing the biological correlation score among biomedical terms. In Section 5 the conclusions and the opening of the future work are faced.

2 Background

To improve the quality of the search over public literature databases, many search engines have been developed to classify and correlate the huge amount of paper that a search may return. Most of this tools improves the search engine capability adopting ontologies and thesaurus as information base. *GoPubMed* [1] is a search engine tools that make the search on *PubMed* (PubMed) repository categorized and ranked. In details, it applies text-mining algorithm in order to classify abstracts retrieved from PubMed according to terms coming from GO and *Medical Subject Headings* (MeSH) vocabulary [4]. A further extension to GoPubMed is *GoGene* [2]. The scope of this tool is to rank a list of genes, occurring in a document set after a PubMed search, according to terms of GO and MeSH. After performing a document search by means of GoPubMed, a text mining algorithm looks for gene names, GO and MeSH terms over the collection of paper abstracts. Finally, it returns a list of genes associated with concepts of GO and MeSH. Even if the association is performed computing a correlation score between a single gene and a term, the scope of the score is not to express a measure of biologic relationship between either two terms, or two gene, or a gene and a term, but it is an a posteriori statistical analysis that allows to map a gene into a ontology term. Furthermore, the problem of computing a quantified semantic correlation score among biological concepts according to the information coming from ontologies has been addressed in [6]. In this work authors developed a tool for measuring the semantic similarity of two Gene Ontology terms [5]. The algorithm exploits the graph-based hierarchical topology of GO and it aggregates the semantic contribute of the ancestors providing a numeric value that reflects the distance among GO terms from the biological meaning point of view. Nevertheless, this tools is limited in foreclosing a set of

biological topics related to the genetical and biological field and the base of information is strongly limited to the GO vocabulary. For instance, the medical term "*Parkinson* is not present in GO, therefore the correlation between this disease and a certain gene is uncomputable.

A different approach for providing a quantified measure among biological terms, based on the *Vector Space Model* theory is proposed in [7]. In this work, the authors apply the *Latent Semantic Indexing* (LSI) algorithm [8] to a set of document abstracts automatically retrieved through PubMed web services. The LSI algorithm is strongly based on the decomposition, by means of *Singular Value Decomposition* (SVD) algorithm, of a term-by-document occurrences matrix X in three sub-matrix U, Σ and V where U and V contain the left and right singular vectors of X and the matrix Σ is a diagonal matrix containing the singular values. The correlation score about two terms in the X matrix is calculated by applying the *cosine product* of the vector of U-*by*-Σ matrix corresponding to the correlating terms. Transforming the X matrix into the U-*by*-Σ matrix allows to consider not only the occurrences of terms but also the co-occurrences, exploiting the conceptual links among terms occurring in the different documents, but belonging to the same context. The analysis performed by the LSI algorithm is further coupled with a pre-processing phase based on the *Metamap* program, provided by NIH [9]. The algorithm behind Metamap allows to extract biomedical concepts from document text so that each abstract is translated into a list of biomedical concepts, namely a list of UMLS Concept Unique Identifiers (CUI) [10]. Moreover, the authors emphasize that the resulting correlation scores, namely *Semantic Correlation Score, SRS* allow the comparison among biological concepts belonging to the same retrieved document abstract set but it looses accuracy when comparing terms and concepts belonging to different experimental run and document set. Therefore, the SRSs go out of the scope of calculating an universal correlation score as the validity of the comparison is limited to a specific experimental run. In order to overcome this limitation, authors analyze the density function of the statistical distribution of the correlation scores within a single experimental run and divide the density function in percentiles. Consequently, the evaluation through percentiles allow to evaluate the correlation between genes and biological concepts independently of the experimental run, because it provides an information on the frequency distribution of the SRSs in the overall CUI set occurring in an experimental run.

Nevertheless, the effectiveness of the methodology introduced in [7] strongly depends on the reliance of the information in the retrieved abstract documents. Scientific papers, and specifically those from PubMed as far as the biomedical field is concerned, are certainly the most consulted source of information by scientist and biologist because they contain the more updated state of the art in science and medicine. However, the nature itself of biomedical literature is characterized by a significant level of uncertainty. In fact, the scientific publications, reporting the state of the art on scientific topics, are continually evolving and, as a consequence, the addicted argumentations might be often contradicting. Therefore, a biological correlation score computed by considering documents coming exclusively from the biomedical literature is affected by a certain level of uncertainty. For this reason, integrating the set of documents coming from literature with *trusted* information is a mandatory step in order to enhance

the computational accuracy of the biological correlation among biological concepts. In order to get the scope, a twofold question must be answered: firstly, what are the *trusted* source of informations and to what extend a source of information can be considered as *trusted*; secondly, once the *trusted* information have been collected, what is the more effective way to integrate this information in the Latent Semantic Analysis algorithmic infrastructure?

2.1 *Trusted* Information Sources

It is worth underlining that several sources of biomedical information, mainly ontologies and thesaurus, are universally accepted as *trusted* source of information. GO, for instance, correlates gene and gene products by means of annotations. GO curators coordinate the development of the ontology rooted in the experimental literature through an annotation-driven process and, moreover, the accuracy of the relations and terms definitions are continually checked [11]. Moreover, genetic and metabolic pathways are a graph-based representation of the interaction of genes and proteins in a biological process.

The annotations related to new discovered pathways are collected in several pathway databases. These databases play a key role in the genetic and biologic research field. Examples of pathways databases that have been developed in the last years are the Kyoto Encyclopedia of Genes and Genomes (KEGG) [12], BioGRID [13], HumanCyc [14]. On the base of pathways information, a number of knowledge based tools have been developed. In particular, *Pathway Commons* is a projects that aims at integrating all the distributed databases and the pathway information under a common web interface [15], providing a single access point to multiple pathway data sources. *Pathway Commons* runs on *cPath* software engine [16], an open source framework that makes the aggregation of custom pathway data sets easy. The key feature of cPath is mainly the ability to support standard pathways exchange format as PSI-MI [17] and BioPax [18], as well as providing all the information content in XML format through web service API. Moreover, cPath stores external link records matching the biological entity with Gene Ontology terms if the references are present in the PSI-MI file.

2.2 Integration of *Trusted* Information in the LSI Algorithm

LSI is particularly effective for text clustering applications because it exploits co-occurences among terms in order to gather *latent* associations [8]. This feature makes the LSI particularly successful in handling synonyms and polysemys. However, LSI considers all the words in the documents set as a *bag of words*, where all the terms are equally weighted. In [19], authors introduce the "*sprinkling*", an efficient technique that enhances text classification accuracy taking into account document class labels. Labeling the documents helps LSI promoting inferred latent associations between words conceptually belonging to the same class. Considering the LSI algorithm, the documents labeling results in creating a new occurrences matrix where new terms are added. Figure 1 shows how *sprinkling* technique influences the occurrence matrix content during the LSI algorithm computation. The class-labelled terms corresponds to extra rows with non-zero value only in the labelled documents, and all-zeros otherwise. The term rows

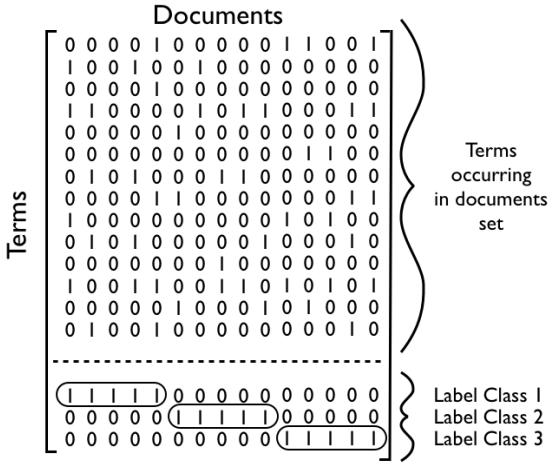


Fig. 1. Modified Occurrences Matrix with *sprinkling* label classes

added by the *sprinkling* artificially create new co-occurrences among terms belonging to the same document class making explicit the implicit associations. Therefore, the efficiency in promoting implicit associations makes the *sprinkling* process a good candidate to integrate *trusted* information in the LSI algorithm in order to promote implicit associations among terms embedded in the document set. Starting from LSI and the *sprinkling* idea we developed a new customized methodology in order to integrate the associations coming from the *trusted* biological pathway with the information included in a set of documents retrieved from PubMed web services.

3 Methodology

The complete flow of the propose method is mainly composed by three phases depicted in Figure 2: *Document Abstract Analysis*, *Semantic Analysis and Trusted Info Integration*, *Results Presentation*. The following sub-sections describe the methodologies behind each phases in more details.

3.1 The *Document Abstract Analysis* Phase

The initial set of documents is composed by abstract retrieved through PubMed web services. In particular, we consider the case the scientist is interest in quantify the biological correlation score among a biological process and a specified gene. The set of keyword used to query the PubMed web-service is composed by the biological process and the gene name (initial keyword set), and a list of synonyms and terms correlated to both the gene and the biological process (expanded keyword set), resulting from querying the UMLS database[10]. In detail, in order to get the more complete set of synonyms and correlated terms, we exploit the fact that ULMS database store the relations among concepts belonging to almost 37 different biomedical database and ontologies [10]. Hereafter, we select the relevant relations in order to download the more related

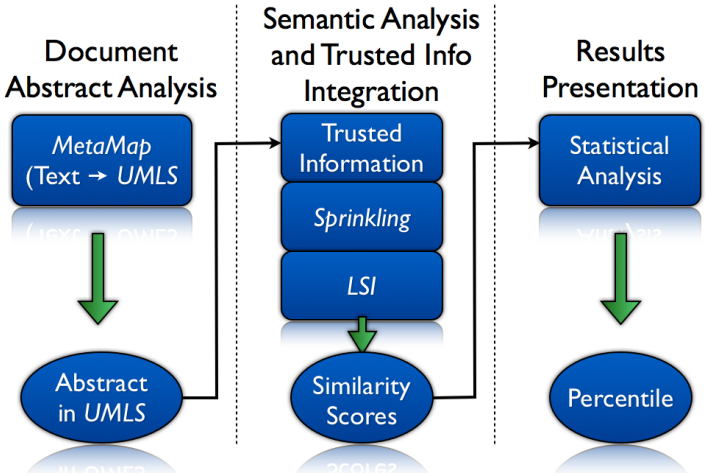


Fig. 2. Complete Semantic Analysis Flow

abstract set, exploiting the methodology introduced in [7]. It is worth noting that the more the retrieved abstracts relate both the biological process and the gene, the more the resulting biological score is accurate. In fact, the presence of unrelated with the specified terms increases the noising information that in turn alter the result accuracy.

Moreover, in the *Document Abstract Analysis* phase, the retrieved set of document abstracts from PubMed written in *Natural Language* are translated in a *Concept Language* exploiting Metamap capability. In fact, Metamap allows to map natural language text in *Concept Unique Identifiers* (CUI). Pre-processing the abstracts by means of Metamap makes the semantic analysis more accurate in that it reduces the ambiguities among terms that express the same concept with different sentences and words, and it allows to reduce non-relevant terms filtered by *Semantic Type* [7]. Therefore, at the end of *Document Abstract Analysis* phase, the set of abstract documents from PubMed translated in Concept Language is returned forming the set of documents on whom the semantic analysis is performed.

3.2 The *Semantic Analysis and Trusted Info Integration* Phase

This phase is the main core of the proposed flow and it consists of 1) retrieving the *Trusted Information*, 2) applying the *sprinkling* technique in order to integrate the information from the document abstract, 3) performing the latent semantic analysis in order to get the most accurate *SRS list*.

Trusted Information. The proposed tool considers genetical pathways as *trusted* source information. In fact, genetical pathways are a graphical-based representation of the gene involved in a specific biological pathways. Thus, they provide a direct link between a gene and the corresponding biological process and in this sense they are good candidate as additional source of information to the biomedical abstracts set. Moreover, genetical pathways are constantly subjected to a review process and consequently they can be considered as *trusted*. In order to automatically access genetical pathways information,

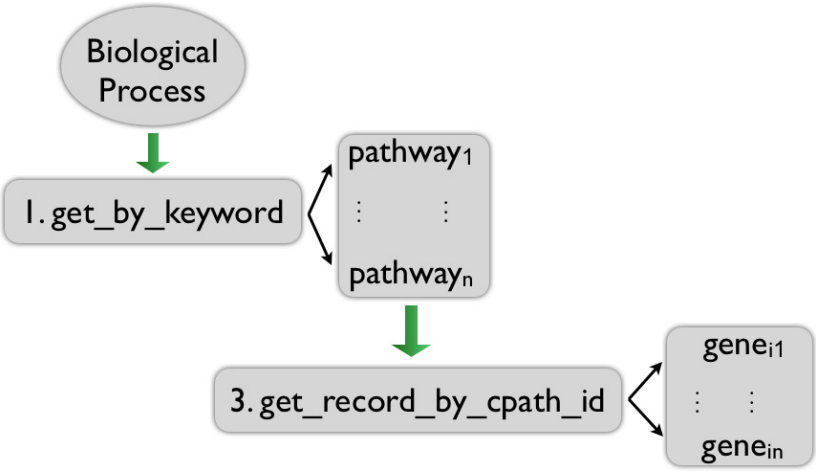


Fig. 3. *GetPathwayInformationByBiologicalProcess* procedure to get *trusted* information

a software client for querying *Pathway Commons* web-service has been developed. In details, pathways information are extracted through the *GetPathwayInformationByBiologicalProcess* procedure shown in Figure 3. *GetPathwayInformationByBiologicalProcess* returns from Pathway Commons the list of pathways where the biological process specified as input occurs in the pathway description or in the labeling title. Hereafter, for each pathway in the list, the *cpath_id*(i.e. the unique pathway identifier according to cPath framework [16], is extracted from the pathway description and it is used to access the complete set of genes involved in the pathway.

It is worth noting that the presented method achieves a twofold scope: firstly, the link between the specified gene and biological process is confirmed consulting a *trusted* source of information; secondly, a list of gene *trustly* involved in the specified biological process are collected. The latter information is particularly tailored in enhancing the co-occurrences of genes correlated with the specified gene and biological process, increasing the overall accuracy during the semantic analysis.

Applying *Sprinkling* to Semantic Analysis. In this phase, the information extracted from pathway database are integrated in the semantic analysis process of PubMed abstracts. The term-by-document occurrence matrix is the main data structure processed by the LSI algorithm and, consequently, it is the main access point to integrate additional information. The term-by-document occurrence matrix is a two-dimensional occurrence matrix in which the columns represent the dimensional space of a corpus of documents and the rows represent the dimensional space of the terms occurring into the documents. However, the abstract set and the retrieved pathways present a different structure of the informational content. The abstract set is composed by documents already suitable for the latent semantic analysis algorithm, whereas the pathways are represented as a connected graph. Therefore, in order to integrate pathways content in the semantic analysis flow, the information coming from pathways are firstly extracted and translated into a new document containing the *trusted* information. In detail, the created

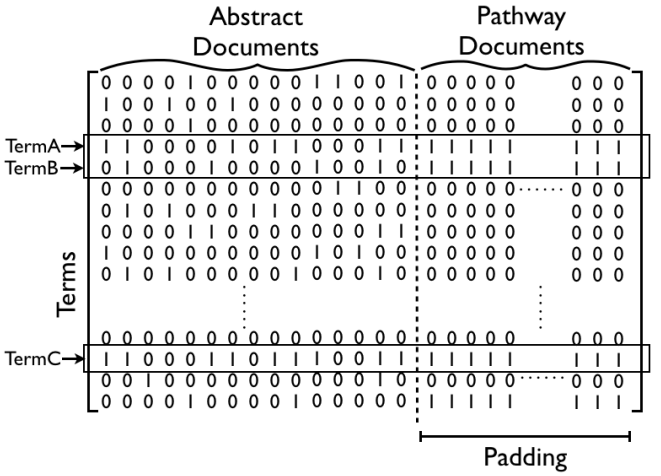


Fig. 4. Occurrences Matrix adding pathway information. The dashed line separate the abstract document set by document pathways.

trusted document contains the list of gene names involved in the retrieved pathways as well as the list of retrieved biological pathways. Nevertheless, adding a single document to the overall document set weakly affects the semantic analysis in terms of accuracy because poorly relevant from the statistical point of view. Thus, the *trusted document*, containing the extracted pathway information, is replicated. We refer at this procedure as *document padding*. Figure 4 shows a trivial term-by-document occurrences matrix sprinkled with additional document padding. The number of instances of the padded document, namely NPD, is obtained according the following equation:

$$NPD = \lceil p \cdot NAD \rceil \tag{1}$$

where NAD is the total number of PubMed abstract document and p is a padding factor spanning from 0 to 1. Consistently with this equation, the number of padded document spans from zero to the same number of document abstracts. Padding the overall document set with *trusted* information affects the singular values of the occurrence matrix. In fact, the singular values reflect on the distribution of the prevalent concepts occurring in the document set. The higher the singular value score, the greater is the *presence* of the concept in the overall document corpus. It is worth noting that increasing the p factor makes the *trusted concepts* included in the *trusted documents* predominant with respect to the overall concepts.

The analysis of the singular values is fundamental to evaluate how the *trusted document* padding affects on the overall document corpus by coherently tuning the p factor value. In the LSI algorithm, the singular values, obtained after applying the *SVD*, represent the attendance of the concepts in the analyzed documents set. The more a concept is relevant and frequent in the document set the higher is the singular value corresponding to it. Moreover, during the latent semantic analysis the singular values computed by SVD are set in the diagonal Σ matrix in ascending order, and consequently the first singular values represent the more relevant concepts in the analyzed corpus. Figure 5

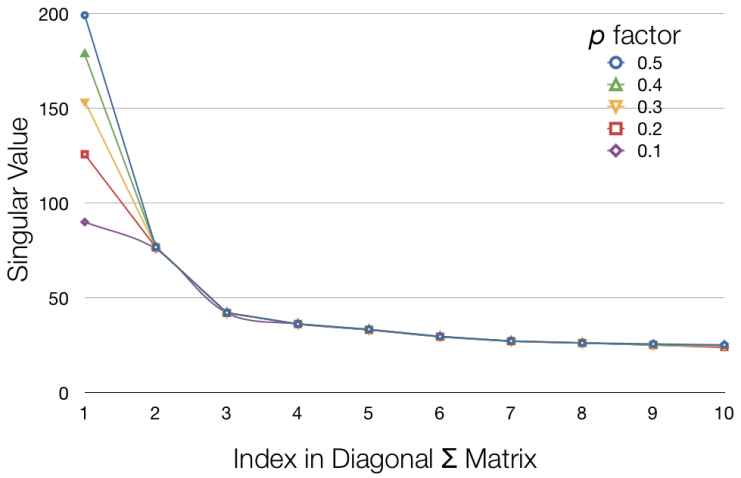


Fig. 5. Singular values obtained launching the latent semantic analysis for computing the biological correlation score between AKT1 gene and Angiogenesis biological process. The ordinate reports the singular value score and the abscissa represents the index in the diagonal singular value matrix.

plots the first ten singular values resulting from the analysis of abstract document set for the computation of the correlation score among the *Angiogenesis* biological process and *AKT1* gene term. The singular values are obtained with five different p factor values. When the document padding is increased by tuning the p factor, the concepts in the *trusted* documents become more relevant respect to the concepts in the abstract document set. When the p factor increases even more the concepts in the *trusted document* set become so relevant that the concepts in the abstract document set are overcome. As a consequence, the singular values corresponding to the *trusted document* set are boosted up and they occupy the first positions in the diagonal Σ matrix. This is the reason why, as shown in Figure 5, the greater the p factor value, the higher the first singular value score. Therefore, with greater p factor values the most relevant concepts in the overall document set correspond mainly to the concepts in the *trusted document* set.

The alteration on the singular values of the occurrence matrix by the trusted document padding consequently reflects on the correlation score between the biological terms. Consider the scenario depicted in Figure 4: *TermA* and *TermB* present a weak correlation in the document abstract set because quite linearly independent, whereas they are strongly correlated in the *trusted document* set. The document padding increases the parallelism of the two terms vector, consequently enhancing the correlation score. Furthermore, both *TermA* and *TermB* are correlated with *TermC* creating an indirect correlation due to co-occurrence. If the *TermC* occurs even in the *trusted document* set the resulting correlation score between *TermA* and *TermB* is further enhanced. Moreover, when a term occurs both in the document abstract set and in the *trusted documents* set, the term conceptually belongs to the *trusted concepts* as well. If two biological terms belong to the *trusted concepts* their correlation score is higher because conceptually linked to the predominant concepts in the overall documents content. Specifically, the

biological correlation score between the biological process and the gene term is influenced by trusted information by a twofold effect: in the case the terms corresponding to biological process and gene term occur both in the document abstract set and in the *trusted documents* the score is directly affected; moreover, the correlation score is further enhanced by the presence of terms correlated with both the biological process and gene terms because occurring both in the document abstract set and in the *trusted document* set. Table 1 shows the correlation scores between *Angiogenesis* and *AKT1* gene, both in SRSs and in Percentile. The reported results show that increasing the p factor value implies a corresponding increase in the correlation score. Therefore, it is possible to assert that the proposed methodology, provides an effective correlation score combining the information extracted in the document abstract set with the *trusted* information extracted from pathways databases. Moreover, the influence of the *trusted* information on the semantic text analysis is tuned by the p factor parameter that directly affects the correlation results.

Table 1. SRS and Percentile values corresponding to p factor

p **Factor**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
SRS	15.7	17	18.1	18.8	19.5
Percentile	76	78	82	83	84

3.3 The *Results Presentation* Phase

An evaluation criterion allowing the comparison among each gene in the genes set and each biological process is fundamental in order to focus the attention only on those genes correlated with the biological processes of interest behind a certain degree of correlation. Therefore, the correlation measurement between genes and biological processes must express an universally valid score allowing the comparison among different correlation analysis. Actually, SRS provides a percentage measurement to compare genes and biological process corresponding to terms belonging to the same occurrence matrix and thus it quantifies the semantic relationship of biological terms during the same semantic text analysis (i.e. a single execution of the tool). Different analyses are characterized by different sets of documents and concepts. Consequently, SRSs values computed on different occurrences matrices are not directly comparable. In order to define an evaluation criterion that allows to compare SRSs resulting from different semantic analysis, and therefore different occurrences matrices, we analyzed the SRS distribution on different occurrences matrix. The SRS Distribution (SRSD) is defined as the density function that reports the percentage of occurrences, or frequencies, of a certain score for the possible SRSs corresponding to all the CUIs occurring within the correlation analysis. Furthermore, the SRSD has been divided into hundred percentile intervals of equivalent area depending on the frequency percentage of the SRS in the density function. We define each interval as Biological Correlation Index (BCI) expressing that the higher the BCI, the greater the degree of biological correlation between the biological process and the genes. The evaluation through BCI allows to evaluate the correlation between genes and biological process independently of the experimental

run, because it provides an information on the frequency distribution of the SRSs in the overall terms occurring in an experimental run.

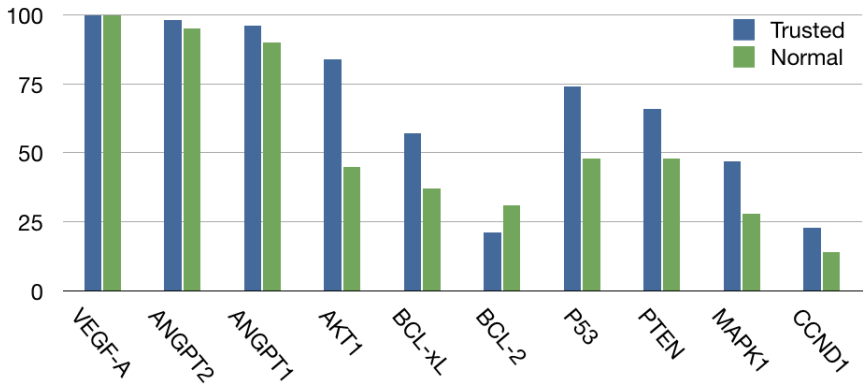
4 Experimental Results

The following section reports the correlation scores among a set of three biological processes (*Angiogenesis*, *Apoptosis*, and *Signal Transduction*) and ten gene terms (*VEGFA*, *ANGPT2*, *ANGPT1*, *AKT1*, *BCL-xL*, *BCL-2*, *P53*, *PTEN*, *MAPK1*, *CCND1*) applying the proposed flow. Figure 6 resumes the results and demonstrate the effectiveness of integrating the information coming from the PubMed document abstract set with the information from pathway database. The biological correlation scores obtained integrating trusted information (*Trusted INFO*) are compared with the scores obtained without integrating any *trusted* information (*Normal INFO*). Moreover, p factor has been set to *0.5*. This choice implies that the *trusted document* set has been padded in order to equal almost the half of the number of abstract document set and to emphasize the effect of the integration of *trusted* information on the overall document set. Furthermore, the biological correlation score is reported in the *BCI* format in order to allow the comparison among the scores of same set of genes computed in different biological process [7].

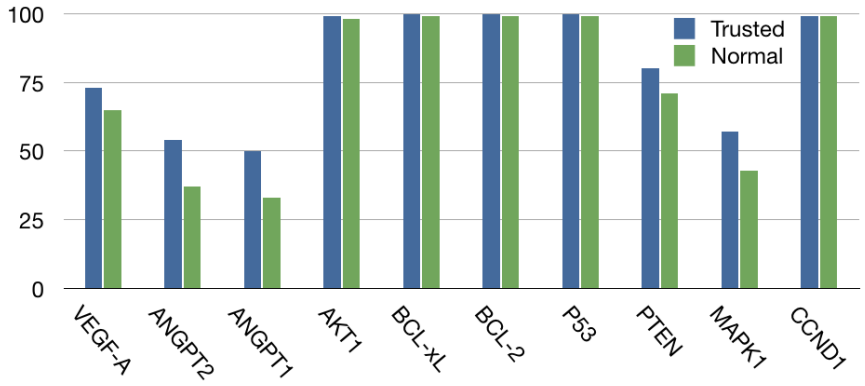
In order to have a better understanding of the resulting biological correlation scores, the functional definition of the gene set according to NCI Thesaurus is listed below:

- *VEGFA*: This gene is involved in the regulation of blood vessel growth.
- *ANGPT2*: This gene encodes angiopoietin-2 protein and it plays a role in both angiogenesis and neovascularization.
- *ANGPT1*: This gene plays a role in both vascular development and angiogenesis.
- *AKT1*: This gene is involved in signal transduction and negative regulation of apoptosis.
- *BCL-xL*: This protein plays a role in the modulation of apoptosis.
- *BCL2*: This gene is involved in apoptotic regulation. Overexpression of this gene promotes the pathogenesis of B-Cell lymphomas, due to anti-apoptotic activity.
- *P53*: This protein plays a role in the regulation of both the cell cycle and apoptosis.
- *PTEN*: This gene plays a role in signal transduction and apoptosis. It is also involved in the regulation of cell cycle progression.
- *MAPK1*: This gene plays a role in signal transduction and positive regulation of the cell cycle.
- *CCND1*: This gene plays a role in the regulation of mitotic events.

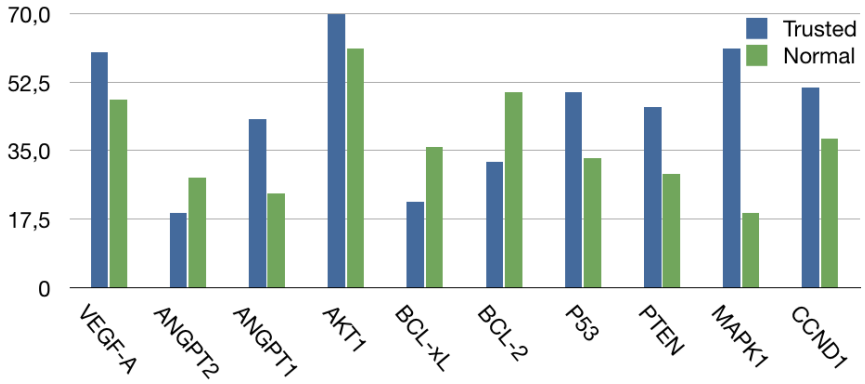
From the same *NCI Thesaurus* source, Angiogenesis is defined as the "*blood vessel formation*" and, specifically for the case of tumor angiogenesis, it is "*the growth of blood vessels from surrounding tissue to a solid tumor*". It is worth noting that the results concerning the biological correlation scores between angiogenesis and the gene set coherently reflect the NCI Thesaurus definitions. In fact, *VEGFA*, *ANGPT1* and *ANGPT2* present the almost same correlation scores with the *Angiogenesis* biological process both in the case of *Trusted INFO* and in the case of *Normal INFO*. The small difference in the results depends on the fact that in both the cases the biological correlation scores presents highest BCI, therefore the *Trusted INFO* poorly affect the accuracy of



(a) Angiogenesis.



(b) Apoptosis.



(c) Signal Transduction.

Fig. 6. Experimental Results Expressed in BCI score

the computation. Moreover, the correlation score of *AKT1* gene term with *Angiogenesis* process is highly enhanced by the integration of *Trusted INFO*. This gene is specifically involved in the *Signal Transduction* and in the *Apoptosis* process, whereas it is less

typical of *Angiogenesis* biological process respect to *VEGFA*, *ANGPT1*, and *ANGPT2* genes even if correlated with it. Our results reflects this assertion because *AKT1* gene directly appears in the *Angiogenesis* pathway and it is directly correlated with this biological process according to the *trusted information* set, but its correlation according to the information coming from the document abstract set is minor compared with *VEGFA*, *ANGPT1*, and *ANGPT2* genes. Coherently, this gene presents a lower biological correlation score compared with *VEGFA*, *ANGPT1*, and *ANGPT2* genes even if, correctly, its score has been significantly boosted by adding *Trusted INFO*.

Moreover, the integration of *Trusted INFO* enhances the biological correlation score of *P53*, *PTEN*, *MAPK1* and *CCND1*. These genes are generally involved in *Apoptosis*, *Signal Transduction* and *Cell Cycle* biological processes that are indirectly related with *Angiogenesis* process as well. The *AKT1* gene, that is mainly involved in *Signal Transduction* and *Apoptosis*, also occurs in the *Angiogenesis* pathway. The biological correlation of the AKT1 gene with *P53*, *PTEN*, *MAPK1* and *CCND1* genes and, on the other side, the occurrence of this gene in the *trusted* document concerning *Angiogenesis* process creates an indirect link between *P53*, *PTEN*, *MAPK1* and *CCND1* genes and the same *Angiogenesis* process. It accordingly results in a biological correlation score enhancement.

Furthermore, the introduction of *Trusted* information positively increases *ANGPT1*, *ANGPT2* and *MAPK1* genes scores. These genes are generally typical of *Angiogenesis* and *Signal Transduction* but these processes are sometime related to *Apoptosis* and thus they co-occur in the document abstact set. Therefore, even if the *Apoptosis* pathway do not include this gene set, the indirect biological correlation between terms in the *Trusted* information and terms occurring in the document abstract set, belonging to the *Apoptosis* semantic area, increases the resulting value of the biological correlation score. However, all of them present a score significatively lower than the one of *Apoptosis* genes *AKT1*, *BCL-xL*, *BCL2*, and *P53*.

According to the results concerning *Signal Transduction* process, the biological correlation score of *MAPK1* gene is particularly interesting. This gene is strongly related to *Signal Transduction*, but this biological processes is quite generic and it interact with many biological phenomena typical of the cell cycle. Considering only the document abstract set as information base, the occurrence of *MAPK1* gene is not always frequently guaranteed, thus resulting in a low correlation score. However, integrating the *Trusted* information the score is almost triplicated. The introduction of *Trusted* information creates the missing conceptual links of the *MAPK1* gene with terms semantically correlated with the biological process of interest, resulting in an increase of the final biological correlation score. This result remarks the effectiveness of the proposed approach in enhancing accuracy of measuring the biological degree of correlation.

Furthermore, integrating *Trusted* information in the document abstract set makes the results more accurate also reducing the biological correlation score. *BCL2* gene, for instance, is functionally specific of *Apoptosis* biological process. On the wave of this assertion, this gene further decreases the correlation score in the case of both *Angiogenesis* and *Signal Transduction* when applying *Trusted* information. The expected behavior would be at most of unaltered result because the *Trusted* information do not addict any further information concerning the correlation score between *BCL2* and the

two biological processes. However, integrating *Trusted* information increases the correlation score of those terms related to the biological processes and their correlation scores overcomes the *BCL2* one, that in turn becomes lower. Similar consideration are also valid about the biological correlation score between *BCL-xL* gene and *ANGPT2* gene and *Signal Transduction* biological process.

5 Conclusions

The proposed semantic analysis tool provides a framework for measuring the biological correlation score among biomedical terms. The score is the results of a text mining analysis performed on the abstracts of the most relevant scientific publications with the support of the completeness of the UMLS Metathesaurus. The knowledge extracted by biomedical literature is further integrated with the information coming from sources generally known as trustworthy. In order to satisfy this requirement, public biological pathways databases have been chosen and the embedded information are first retrieved and then correlated in the overall semantic analysis flow. Moreover, in order to integrate the knowledge coming from an heterogeneous base of information, a new version of the latent semantic analysis, based on the sprinkling technique, has been developed. The results remark the efficiency of the proposed approach in enhancing the accuracy of the biological correlation score computation by combining the knowledge extracted by scientific document abstract set with the pathways information.

As future step the complete automation of the tool will allow to retrieve the document abstract set in an automated and accurate manner. Moreover, the integration of information coming from other trusted sources will be tailored and developed in order to better improve the overall accuracy and robustness of the proposed semantic analysis method.

References

1. Doms, A., Schroeder, M.: GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Research (2005)
2. Plake, C., Royer, L., Winnenburg, R., Hakenberg, J., Schroeder, M.: GoGene: gene annotation in the fast lane. Nucleic Acids Research (2009)
3. PubMed., `http://www.ncbi.nlm.nih.gov/pubmed/`
4. MeSH, Medical Subject Headings (MeSH) Fact sheet. National Library of Medicine (2005)
5. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nature Genetics (2000)
6. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F.: A New Method to Measure the Semantic Similarity of GO Terms. Bioinformatics (2007)
7. Abate, F., Ficarra, E., Acquaviva, A., Macii, E.: An Automated Tool for Scoring Biomedical Terms Correlation Based on Semantic Analysis. In: International Conference on Complex, Intelligent and Software Intensive Systems (2010)
8. Gliozzo, A.M., Strapparava, C.: Domain Kernels for Text Categorization. In: Ninth Conference on Computational Natural Language Learning (2005)
9. Aronson, A.R.: Effective Mapping of Biomedical Text to the UMLS. Metathesaurus: The MetaMap Program. In: AMIA Fall Symposium (2001)

10. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research (2004)
11. Hill, D.P., Smith, B., McAndrews-Hill, M.S., Blake, J.A.: Gene Ontology annotations: what they mean and where they come from. Bioinformatics (2008)
12. Kanehisa, M., Goto, S.: Kegg: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research (1999)
13. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Research (2006)
14. Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., Karp, P.D.: Computational prediction of human metabolic pathways from the complete human genome. Genome Biology (2004)
15. Pathway Commons (2007), `http://www.pathwaycommons.org`
16. Cerami, E.G., Bader, G.D., Gross, B.E., Sander, C.: cPath: open source software for collecting, storing, and querying biological pathways. Bioinformatics (2006)
17. Hermjakob, H., et al.: The HUPO PSI's molecular interaction format community standard for the representation of protein interaction data. Natural Biotechnology (2004)
18. BioPAX: Biological Pathways Exchange (2007), `http://www.biopax.org`
19. Chakraborti, S., Mukras, R., Lothian, R., Wiratunga, N., Watt, S., Harper, D.: Sprinkling: Supervised Latent Semantic Indexing. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 510–514. Springer, Heidelberg (2006)

PART III

Bio-inspired Systems and Signal Processing

Towards a Patient-Specific Model of Lung Volume Using Absolute Electrical Impedance Tomography (aEIT)

Suzani Mohamad Samuri[1], George Panoutsos[1], Mahdi Mahfouf[1], G.H. Mills[2], M. Denaï[3], and B.H. Brown[4]

[1] Department of Automatic Control and Systems Engineering, University of Sheffield Mappin Street, Sheffield, S1 3JD, U.K.
[2] Department of Critical Care and Anaesthesia, Northern General Hospital, Sheffield, U.K.
[3] School of Science and Eng., Teesside University, Middlesbrough, U.K.
[4] Department of Medical Physics, Royal Hallamshire Hospital, Sheffield, U.K.
{s.mohamad-samuri,g.panoutsos,m.mahfouf, b.h.brown}@sheffield.ac.uk, Gary.Mills@sth.nhs.uk, M.Denai@tees.ac.uk

Abstract. Electrical Impedance Tomography (EIT), and in particular its application to pulmonary measurement, has been the subject of intensive research since its development in the early 1980s by Barber and Brown. One of the relatively recent advances in EIT is the development of an absolute EIT system (aEIT) which can estimate absolute values of lung resistivity and associated lung volumes. In this paper we present a new approach based on Computational Intelligence (CI) modelling to model the 'Resistivity - Lung Volume' relationship that will allow more accurate lung volume estimations using data from eight (8) healthy volunteers measured simultaneously via the Sheffield aEIT system and a Spirometer. The developed models show an improved accuracy in the prediction of lung volumes, as compared with the original Sheffield aEIT system. However the inter-individual differences observed in the subject-specific modelling behaviour of the 'Resistivity-Lung Volume' curves suggest that a model extension is needed, whereby the modelling structure auto-calibrates to account for subject (or patient-specific) inter-parameter variability.

Keywords: Electrical impedance tomography (EIT), ANFIS, Data-driven modelling, Lung imaging, Non-invasive monitoring.

1 Introduction

Electrical Impedance Tomography (EIT) has been a topic of interest for researchers including clinicians due to its ability to offer a non-invasive, radiation-free and continuous monitoring of the lung function. EIT aims to generate cross-sectional images of the studied subjects based on measurements of surface electrical potentials resulting from an excitation with small alternating currents via an array of equally-spaced electrodes attached to the surface of the thorax at about 4-5 cm above the xyphoid process [1-2].

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 191–204, 2013.
© Springer-Verlag Berlin Heidelberg 2013

EIT has been used to generate images of various parts of the human body; nonetheless, the pulmonary ventilation measurement has always been regarded as the area which can benefit the most from the development of EIT [3-4]. Brown et al (1985) suggested the use of EIT in lung imaging and ventilation monitoring in what was the first summary of possible clinical applications for this technique [5]. For most of the recent EIT studies, the focus has been on the changes in impedance with time (relative/functional EIT), instead of the absolute values. The new approach of absolute impedance tomography takes this a step further, by not only looking at the changes in impedance during the respiratory cycles, but also producing absolute (as opposed to relative) values of impedance that can be compared to normal or reference values. Indeed, a multi-frequency system was developed in the mid 1990's to calculate the absolute impedance [6] and subsequently calculate an absolute volume measurement of the lung.

The method of determination of lung absolute resistivity [7] is based on a 3-D finite difference model of the thorax developed from CT cross sections of a normal subject [8] and scaled to take into account the geometry of the chest (circumference and ellipse ratio) of a particular subject. The elements in the model were assigned fixed resistivity values in the range 1-80 Ω-m depending on their anatomical location (fat, muscle, bone, blood or lung) in the CT images. The modeled data are then compared with the real measurements over a pre-determined region of interest for values of the lung resistivity between 3 and 80 Ω-m. The value of lung resistivity, which minimizes the mean difference between these data sets, is returned as the value of the absolute lung resistivity, an EIT image is reconstructed by filtered back projection [9]. As lung resistivity is a function of the frequency of the applied current, at high frequency, when the capacitive reactance of the cell membranes are reduced virtually to zero, the lungs consist of just two equivalent electrical components; air with almost infinite resistivity and lung tissue with an almost homogeneous resistivity determined by that of the intracellular and extra-cellular fluids [10]. If the resistivity is known, then it becomes possible to calculate both lung density and air volume using the Cole equation [11].

A significant amount of research has so far been devoted to investigating the feasibility of EIT to assess the level of lung ventilation in comparison with the volume of air measured with spirometer. Harris *et al.* (1987) showed the proportional relationship between the lung volume change and lung resistivity using EIT imaging and confirmed the system's ability to assess the level of lung ventilation [12]. Their work identified a close correlation between an impedance index computed from dynamic resistivity images and volume of inspired air measured by a spirometer. In 1988 research confirmed a high correlation ($r > 0.95$) between the change in lung impedance and volume of air inspired in four healthy subjects while at rest and on a bicycle ergo-meter [13]. The study demonstrated that real-time EIT (relative) ventilation measurements of lung volume were possible to an accuracy of +/- 10% of the spirometer values. The posture of the subject was again an area of interest, with data recorded in five subjects in both seated and supine positions, showing an impedance variation of between -3.8% and +9.5% from the former to later posture. Nicolas *et al.* [14] elicited a parametric model of the relationship between EIT and total lung volume with the aim at facilitating inter individual comparisons of EIT images by providing volumetric scale in place of the usual arbitrary units scale. The

lung volume changes predicted by the model were compared to the volume changes measured by Spirometry. The model was able to predict the lung volume changes with 9.3% to 12.4% accuracy. These studies confirmed the fact that there exists a significant correlation between the variable derived by EIT and lung volume changes measured with spirometer and it is possible to model associated relationship.

In this paper, the relationship between absolute resistivity from the aEIT system and the lung volume measured from Spirometry is studied based on the data from 8 healthy volunteers. Three data-driven models were developed; 1) the first model (AEIT) 'tries' to mimic the behaviour of the system in producing the absolute lung air volume from absolute resistivity data, while 2) the second model (EITSPIRO) was developed for mapping the relationship between Spirometry-based lung air volume and absolute resistivity. Finally, 3) a general model of EITSPIRO with an additional input (patient-specific) was introduced. Comparisons between the general model with and without the extra input are also performed. The Adaptive Neural-Fuzzy Inference System (ANFIS) network design [15] is used in all of the modelling work.

2 Study Protocol

2.1 Subjects

A total of eight (8) healthy subjects (males) participated in this study. The subjects' height, weight, circumference and ellipse ratio were measured and recorded. The studied subjects' information is shown in Table 1.

Table 1. The anthropometric information of the subjects

Subject	Gender	Height (cm)	Weight	Circum (cm)	Ellipse ratio
1	M	170	80	95	1.51
2	M	170	67	88	1.38
3	M	171	60	83	1.58
4	M	186	113	109	1.39
5	M	184	93	104	1.53
6	M	191	79	95	1.53
7	M	168	66	90.5	1.36
8	M	171	62.5	88	1.45

2.2 Equipments and Tools

A disposable tape measure was used to measure the subjects' chest circumference. "Mitutoyo Absolute Digmatic" callipers were used to measure the subjects' chest; measurements were taken of chest width and depth in order to calculate an ellipse ratio. The aEIT data were acquired via the Sheffield Mk 3.5 absolute EIT system. The Mk3.5 aEIT (Figure 1) uses eight AgCl ECG type electrodes to inject small alternating currents at 30 frequencies typically within the range 2 kHz to 1.6 MHz and records the resulting potentials at a rate of 25 frames.s^{-1}. The computer user interface to control the Mk3.5 system is written in MATLAB, and is able to display real-time images.



Fig. 1. The Sheffield EIT Mk 3.5 system

2.3 Spirometry and EIT Measurements

Eight (8) skin electrodes were attached around the circumference of the chest, and connected to the EIT data acquisition unit. Ideally the electrodes were attached in a horizontal plane 5cm above the xyphoid process, and equally spaced around the circumference. The subjects were simultaneously breathing through the spirometer tube (SensorMedics) while attached to the Mk 3.5 aEIT system. The data were measured using adjacent drive and receive combinations of electrodes, connected to the data acquisition unit. A 60 sec recording of data were performed involving quiet breathing and maximum inspiration and expiration manoeuvres in sitting position. The acquired EIT data were then resampled using MATLAB according to the spirometry sampled data. The spirometry sampled data represent the instantaneous changes in lung volume (relative to residual volume).

3 Data-Driven Modelling Using ANFIS

Neural-Fuzzy modelling falls under the umbrella of Computational Intelligence (CI) modelling and can be used as *the* non-linear paradigm for mapping the inputs to the outputs of the environment under consideration. This non-linear mapping can be learned from process data using various algorithms. The architecture used in this study is the Adaptive Neural-Fuzzy Inference System (ANFIS) consisting of a set of Takagi-Sugeno-Kang (TSK)-type fuzzy IF-THEN rules [15-17]. This type of a model was proposed in an effort to develop a systematic approach to generate fuzzy rules from a given input-output data set. A typical fuzzy rule in a Sugeno fuzzy model has the following form:

$$\text{IF } x \text{ is A and } y \text{ is B THEN } z = f(x,y) \tag{1}$$

Where A and B are fuzzy sets in the antecedent, while $z = f(x,y)$ is a crisp function in the consequent. The ANFIS architecture is used as the main facet for the modelling structure in order to map the aEIT data. In the first part of the 'Results' section (4.1) only the aEIT data recordings are used to develop a data-driven model of the 'Lung

Resistivity – Lung Volume' (Lung R-V) relationship. This model will map the non-linear relationship 'Lung R-V' by 'imitating' the aEIT system's physical equations [6]. In the second part of the 'Results' in Section (4.2) the Spirometry recordings are used along the aEIT data in an effort to 'bypass' the aEIT system's lung estimations using physical equations, hence attempting to predict directly lung volumes using the Spirometry data for a reference. Figure 2 shows an example of the 'Lung Resistivity – Lung Volume' modelling structure. The Spirometry data (relative lung volume) were converted to absolute lung volume data by estimating the Residual Volume (RV) of all eight volunteers using Body Plethysmography in the Royal Hallamshire Hospital, Sheffield UK.

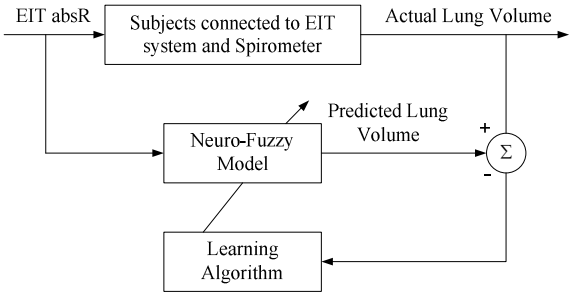


Fig. 2. Model structure for ANFIS 'Lung Resistivity – Lung Volume' model. EITabsR=EIT absolute resistivity. LV Spiro= spirometry lung volume, EIT absLV= EIT absolute lung volume and $\varepsilon(t)$ = error between actual and predicted lung volume.

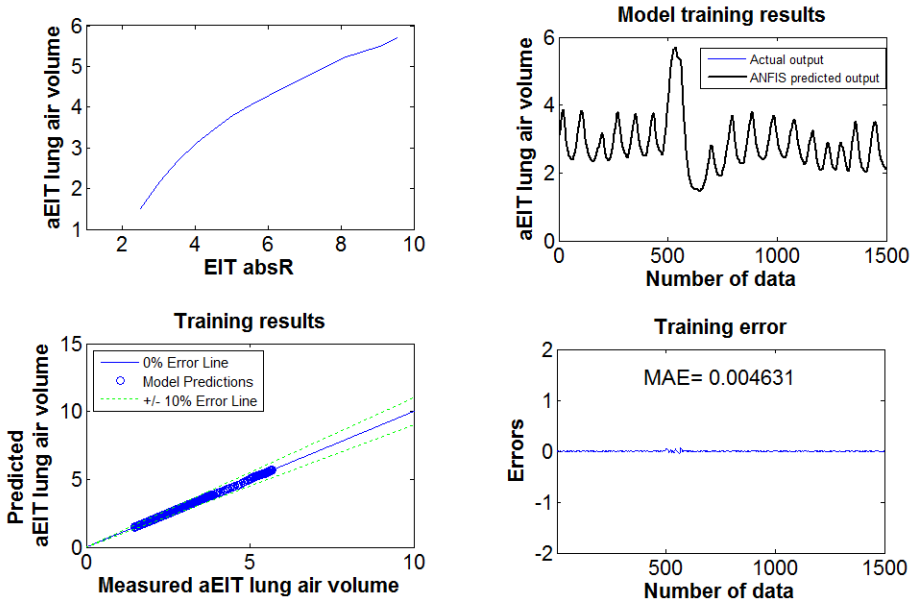
4 Results and Discussions

4.1 AEIT Models Training and Testing Results

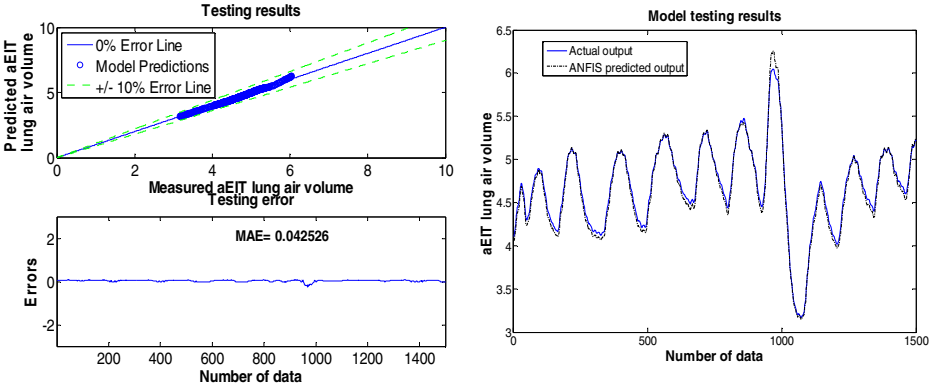
The first type of model built in this study is the AEIT model which is based on data of absolute EIT resistivity and absolute lung air volume (as measured by the aEIT system) obtained from 8 healthy subjects in a sitting position. The aEIT mk3.5 system uses a number of non-linear equations to infer absolute lung volumes from resistivity data. An example of such equations includes the Cole-Cole equation (to link a frequency spectrum to resistivity data), the Nopp model (to link lung tissue resistivity as a function of lung volume) [6-7], and a number of population mean models of lung weight based on gender and body height. Most of these equations are empirical/theoretical and introduce uncertainties and inaccuracies in the final estimations of lung volumes. The objective of the first modelling exercise is to mimic the behaviour of the physical/theoretical/empirical equation based on data from these 8 healthy subjects. The 8 subject-specific AEIT models results are shown in Figure 3.

Table 2 shows the modelling performance results. The root mean square error (RMSE), mean absolute error (MAE%), the correlation coefficient (Cor.) and the standard deviation of the error (eSD) were used as the performance indices. As shown in Figure 3, the AEIT model can predict the absolute lung air volume with a good accuracy

in training (99.8%) and testing (95%). The current modelling results show that ANFIS is a good modelling method to learn the relationship between the absolute resistivity and absolute lung air volume as currently described in the aEIT system. However, such a model would inherit all the inaccuracies of the aEIT system in the estimation of lung volumes, as shown in [18-20] and detailed in the following paragraph.



(a)



(b)

Fig. 3. (a) The AEIT model training results (b) The AEIT model testing results

Even though the AEIT models lead to good performances, in reality the models also inherit the errors of the aEIT system when we compare the lung volumes are compared to the real Spirometry data. Figure 4 shows an example of the absolute EIT lung air volume as compared to the real lung air volume measured via Spirometry.

Table 2. (a) The models training fit results (b) The models testing fit results

(a)

	RMSE	MAE%	eSD	Cor
S1	0.0026	0.0635	0.0026	1
S2	0.0081	0.149	0.0081	0.9999
S3	0.0043	0.102	0.0043	1
S4	0.0206	0.2779	0.0206	0.9999
S5	0.0152	0.4739	0.0152	0.9998
S6	0.0343	0.5429	0.0343	0.9993
S7	0.0054	0.2561	0.0054	1
S8	0.0039	0.068	0.0039	1
mean ± SD	0.012 ± 0.0	0.242 ± 0.2	0.012 ± 0.0	1.000 ± 0.0

(b)

	RMSE (mean ± SD)	MAE%(mean ± SD)	Cor(mean ± SD)
S1	0.487 ± 0.32	0.27 ± 0.13	0.889 ± 0.13
S2	0.244 ± 0.27	0.192 ± 0.19	0.966 ± 0.07
S3	0.159 ± 0.16	0.159 ± 0.16	1 ± 0.00
S4	0.319 ± 0.24	0.29 ± 0.24	0.990 ± 0.01
S5	0.279 ± 0.4	11.27 ± 8.1	0.996 ± 0.0
S6	0.070 ± 0.1	9.478 ± 5.3	0.999 ± 0.0
S7	0.408 ± 0.7	9.539 ± 7.5	0.769 ± 0.4
S8	0.089 ± 0.0	5.012 ± 2.8	0.997 ± 0.0

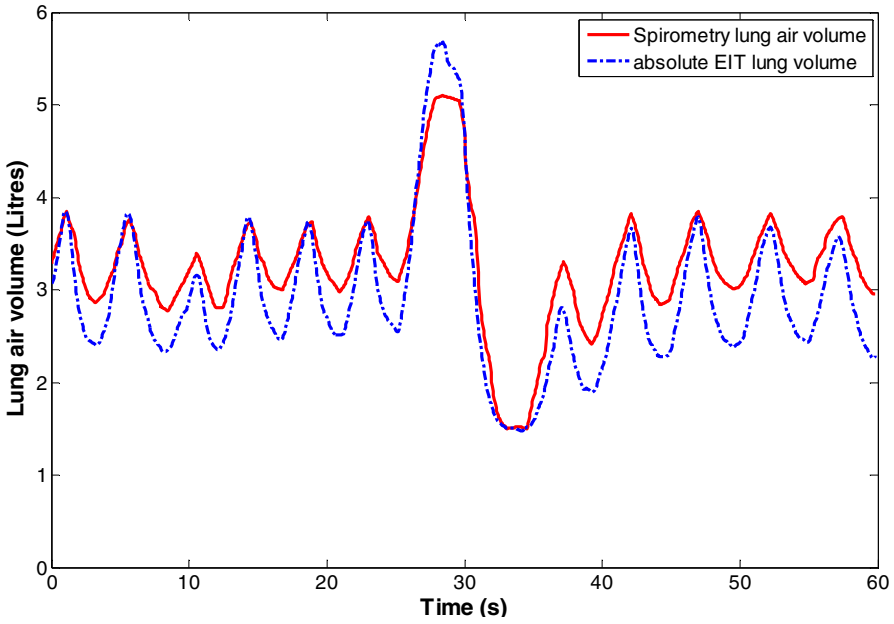


Fig. 4. The plot of actual absolute EIT lung volume and actual spirometry lung air volume

As can be seen in Figure 4, there is a clear difference between the aEIT estimated lung volume and the real lung volume measured using Spirometry (20.7% average error for all eight subjects). Hence, it can be concluded that the ANFIS model is capable of mapping such non-linear behaviour accurately, but some inherited modelling errors (included in the aEIT equations) do not allow for a more accurate lung volume modelling. It is however, possible to 'bypass' the aEIT equations and attempt to model the lung volume directly from volumetric measurement data using Spirometry as described in the next section.

4.2 EITSPIRO Models Training and Testing Results

The second type of model elicited in this study is the EITSPIRO model which is based on data of absolute EIT resistivity and Spirometry lung air volume obtained from 8 healthy subjects in a sitting position. The models are designed to predict the spirometry lung air volume directly from the absolute EIT resistivity as obtained from the EIT Mk 3.5 system. The 8 subject-specific EITSPIRO models results are summarised in Table 3 and one representative example of one subject is shown in Figure 5. Root mean square error (RMSE), mean absolute error (MAE%), correlation coefficient (Cor.) and standard deviation of the error (eSD) were used as the performance indices.

Table 3. (a) The models training fit results. (b) The models testing fit results.

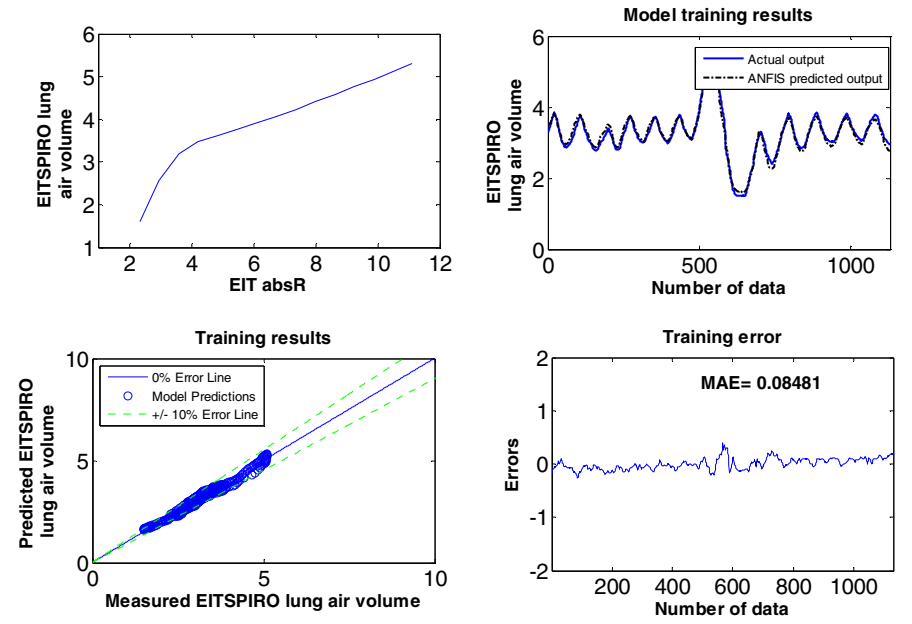
(a)

	RMSE	MAE%	eSD	Cor
S1	0.0951	1.6843	0.0952	0.9968
S2	0.2325	4.8183	0.2326	0.9718
S3	0.1025	1.1356	0.1026	0.9962
S4	0.6068	13.6506	0.607	0.9417
S5	0.1553	4.1812	0.1554	0.9867
S6	0.3583	6.4761	0.3585	0.821
S7	0.0951	2.6752	0.0952	0.9893
S8	0.1245	2.7523	0.1245	0.9714
mean ± SD	0.221 ± 0.2	4.672 ± 4.0	0.221 ± 0.2	0.959 ± 0.1

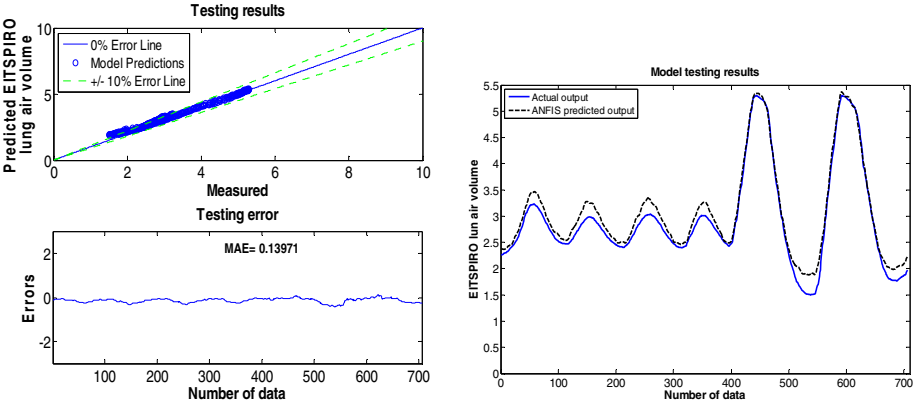
(b)

	RMSE (mean ± SD)	MAE%(mean ± SD)	Cor(mean ± SD)
S1	0.655 ± 0.42	14.17 ± 8.56	0.931 ± 0.09
S2	4.094 ± 3.39	51.77 ± 13.47	0.837 ± 0.13
S3	0.733 ± 0.42	16.73 ± 9.68	0.925 ± 0.08
S4	1.465 ± 0.63	36.39 ± 16.07	0.864 ± 0.10
S5	0.953 ± 0.28	20.13 ± 7.71	0.915 ± 0.09
S6	1.267 ± 0.51	28.47 ± 9.33	0.912 ± 0.04
S7	2.178 ± 1.78	25.84 ± 12.56	0.828 ± 0.13
S8	2.381 ± 2.19	24.9 ± 8.63	0.856 ± 0.12

The EITSPIRO modelling results show that this modelling structure can predict the lung air volume with a good accuracy in training (95.3%). When testing a subject-specific model on a different subject generalisation the average performance is approximately 72.7% accuracy. While this is an acceptable performance it clearly demonstrates the effect of inter-individual differences and the need for subject-specific models (or patient-specific in the case of clinical use).



(a)



(b)

Fig. 5. (a) The EITSPIRO model training results (b) The EITSPIRO model testing results

Figure 6 shows the lung R-V relationship for the eight subjects as predicted by the ANFIS model. There is a common trend between the different subjects (resembling the Nopp model/equation) [7], but it also shows how different this behaviour can be between subjects. Figure 7 shows the lung volume of one healthy subject as measured by Spirometry, estimated by aEIT and predicted by the ANFIS model. It is clear that the ANFIS model predicts best the subject's lung volume as compared with the aEIT

system. The advantage using of the ANFIS model is apparent, however to be able to implement the proposed modelling framework on a real system the model would need a form of calibration, every time it is used, in order to account for the inter-parameter variations between subjects.

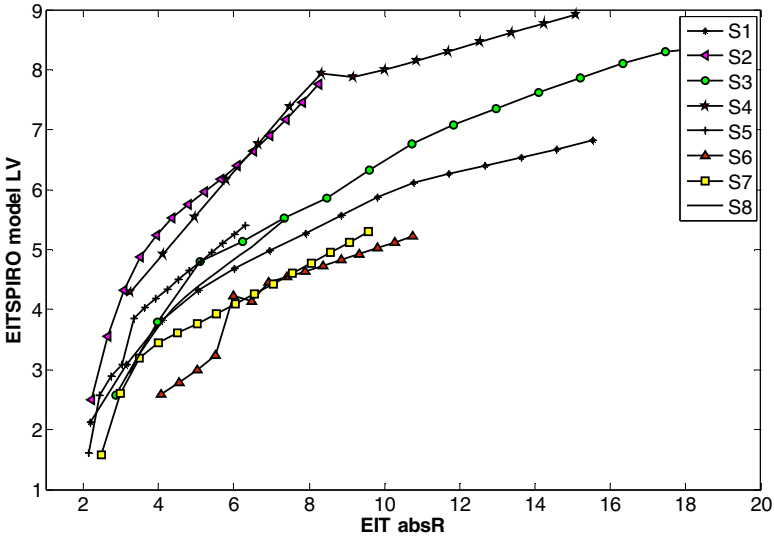


Fig. 6. The plot of lung R-V relationship for the eight subjects as predicted by the ANFIS model

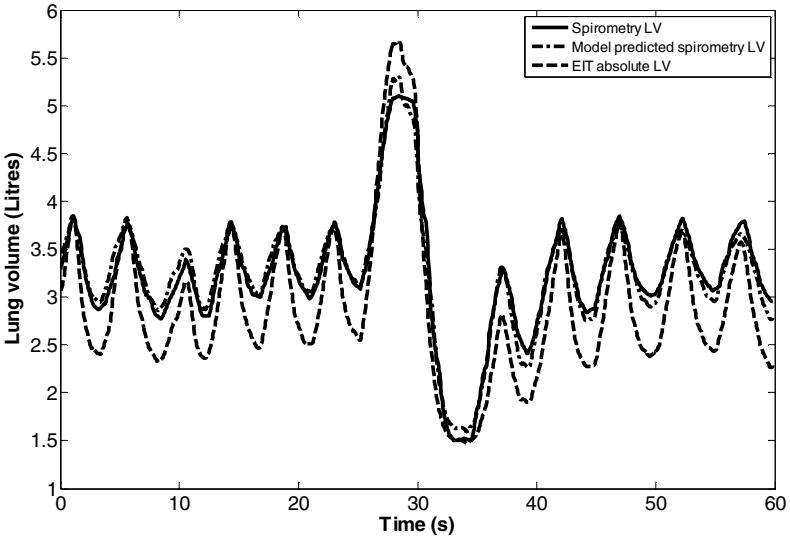


Fig. 7. The plot of EITSPIRO model predicted lung air volume, actual absolute EIT lung volume and actual spirometry lung air volume

5 'Generic' Modelling of EITSPIRO

The performance of the 'individual' models presented in the previous section can be considered as satisfactory, however, in practice, it may prove difficult to calibrate such models in-situ without the required expertise. It would be more appropriate to try to develop a 'general model' that has the capability of predicting the lung volume with a good accuracy and a good consistency. A significant amount of research conducted previously revealed that among the anthropometric information of the subjects (eg: height, weight, body surface area, etc), height is correlated best with lung volume [21]. In this section, a more 'generic' model of absolute EIT resistivity and Spirometry lung air volume was developed with an addition of an extra input: 'body height'. The authors hypothesise that by including the subject's body height as an additional input, the model would strongly improve the performance of the general EITSPIRO model by correctly mapping the inter-parameter differences between subjects.

5.1 General EITSPIRO Model Training and Testing Results

The general model built in this study is based on data of absolute EIT resistivity, height and Spirometry lung air volume obtained from 8 healthy subjects in sitting position. Similarly to the previous models in this paper, ANFIS was selected as the structure of this general model. For a total of 2396 modelling data, 50% were randomly chosen as the training data and 25% as the checking data. A further 25% of the data were used for testing the model. Figure 8 shows the model structure for the 'generic' EITSPIRO model with the additional input of body height.

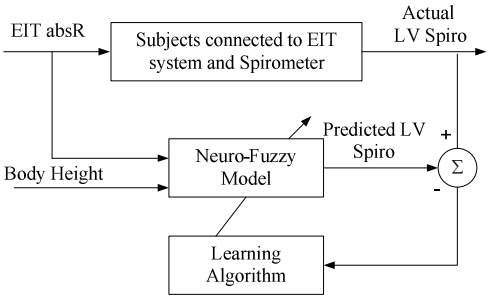


Fig. 8. Model structure for general EITSPIRO model with body height

The initial fuzzy system structure was decided upon using subtractive clustering method (Chiu, 1994) [22]. During training, the mean squared errors for the training set and the checking set were monitored continuously and ANFIS automatically chooses the model parameters associated with the minimum checking error just prior to the point where the error for the checking set stars to increase. This is to prevent the model from over-training (over-fitting the data).

Table 4 shows the general modelling performance results with and without the body height as an extra input. Root mean square error (RMSE), mean absolute error (MAE%), correlation coefficient (Cor.) and standard deviation of the error (eSD) were used as performance indicators. It can be seen from the results of model

performance that the EITSPIRO generic model with the input height can predict the absolute lung air volume with a good accuracy in training (92.5%), checking (92.4%) and testing (92%) as compared to the EITSPIRO general model without the height which led to an accuracy in training (86.4%), checking (86.4%) and testing (86.3%). In studies conducted in [17-18] on the assessment and improvement of absolute lung volume estimation and localisation using absolute EIT (aEIT), it was shown that the absolute EIT system is able to provide good estimates of lung volumes and lung air localisation but with average accuracy of just 88%. Therefore, in this case, being able to provide an estimate of lung volume with a maximum of 8% error (as provided by the ANFIS EITSPIRO generic model), the authors confirm that the elicited model with the additional information of body height, is judged as being 'very good' for clinical use providing further clinical data (testing on more subjects) confirm the initial modelling results from this paper.

Table 4. The general model training, checking and testing results

	Training		Checking		Testing	
	with Height	without Height	with Height	without Height	with Height	without Height
RMSE	0.38	0.68	0.39	0.68	0.39	0.68
MAE (%)	7.49	13.65	7.56	13.65	8.02	13.69
Correlation	0.95	0.84	0.95	0.84	0.96	0.84
eSD	0.4	0.68	0.39	0.68	0.39	0.68

As an addition to the performance of the EITSPIRO generic model, Figure 9 shows the 3D-surface (simulated) of the model which represents the mapping of inputs 'body height' and 'resistivity' to the output 'lung volume'. As should be expected, and from theoretical/clinical knowledge, the higher the resistivity, the higher the lung volume will be [23-24].

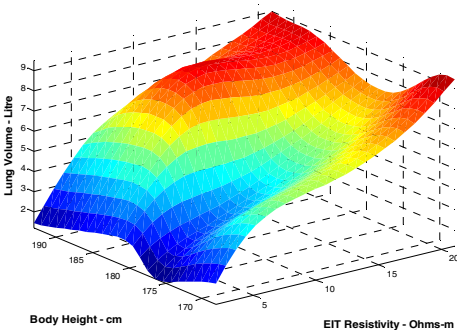


Fig. 9. 3-D surface view of the ANFIS EITSPIRO generic model

6 Conclusions and Future Work

Electrical Impedance Tomography was developed in the early 1980s and it has since shown real potential to being exploited for clinical use (bedside monitoring in the Intensive Care Unit - ICU). Recent developments in the field of absolute EIT demonstrate how one may use it to estimate absolute values of lung volumes which are

key to any on-line EIT based monitoring system. However, the current system can be further improved, particularly in the area of lung volume estimation accuracy. In this specific research study a Neural-Fuzzy modelling structure is used to model the relationship between the lung absolute resistivity and lung volume (lung R-V). Data recordings were used from eight (8) healthy subjects in a sitting position in order to train the models. It was shown that the modelling structure can model very accurately the aEIT lung volume estimation, although this method forces the model to 'inherit' the inaccuracies associated with the aEIT theoretical and empirical equations. In a different approach, it was also shown how one can model the lung R-V by 'bypassing' the physical equations and directly model the lung volume using real volumetric measurements and Spirometry (to record relative volume) and Body Plethysmography (to record lung Residual Volume). To our knowledge this is the first data-driven model capable of describing the behaviour of lung Resistivity-Volume in the absolute EIT system. The developed models show a very good agreement between the real data and the model predictions, however high inter-individual differences were also noted, although, on an individual basis, each ANFIS model (patient-specific) outperforms the current aEIT system's lung volume estimations. In clinical science, inter-patient parameter variability is endemic; this is why it is of the opinion of the authors that an extension to the presented approach is needed, whereby the model auto-calibrates to account for inter-individual differences between patients. In this paper the inclusion of body height as an extra input variable further improves the prediction results. The inclusion of more parameters (that are easily measured) in the modelling structure may lead to a truly general model that is able to account for inter-individual differences. Such a new modelling structure should be able to classify the 'patient-type' based on the R-V behaviour curves and adjust the predictions accordingly.

Acknowledgements. The authors would like to acknowledge financial support from the UK-EPSRC under Grant Number EP/F02889X/1.

References

1. Barber, D.C., Brown, B.H.: Applied potential tomography. Journal of Physics E: Scientific Instruments 17, 723–733 (1984)
2. Brown, B.H.: Electrical impedance tomography EIT: a review. Journal of Medical Engineering & Technology 27(3), 97–108 (2003)
3. Panoutsos, G., Mahfouf, M., Brown, B.H., Mills, G.H.: Electrical Impedance Tomography (EIT) In Pulmonary Measurement: A Review of Application and Research. In: Proceedings of the Fifth IASTED International Conference on Biomedical Engineering, BioMED 2007, pp. 221–230 (2007)
4. Denai, M.A., Mahfouf, M., Mohamad-Samuri, S., Panoutsos, G., Brown, B.H., Mills, G.H.: Absolute electrical impedance tomography (aEIT) guided ventilation therapy in critical care patients: Simulations and future trends. IEEE Transactions on Information Technology in Biomedicine 14(3), 641–649 (2010)
5. Brown, B.H., Barber, D.C.: Applied potential tomography: possible clinical applications. Clinical Physics & Physiological Measurement 6(2), 109–121 (1985)
6. Brown, B.H., Leathard, A.D.: Measured and expected Cole parameters from electrical impedance tomographic spectroscopy images of the human thorax. Physiological Measurement 16(Supplement 3A), 57–67 (1995)

7. Brown, B.H., Primhak, R.A., Smallwood, R.H., Milnes, P., Narracott, A.J., Jackson, M.J.: Neonatal lungs-can absolute lung resistivity be determined non-invasively? Med. & Biological Eng. & Computing 40, 388–394 (2002)
8. Zubal, I.G., Harrell, C.R., Smith, E.O., Rattner, Z., Gindi, G., Hoffer, P.B.: Computerized three-dimensional segmented human anatomy. Medical Physics 21(2), 299–302 (1994)
9. Barber, D.C., Seagar, A.D.: Fast reconstruction of resistance images. Clinical Physics & Physiological Measurement 8(Supplement A), 47–54 (1987)
10. Barber, D.C., Borsic, A.: Electrical Impedance Tomography: Methods, History and Application. Institute of Physics (2005)
11. Brown, B.H., Mills, G.H.: Indirect measurement of lung density and air volume from Electrical Impedance Tomography EIT data. In: World Congress on Medical Physics and Biomedical Engineering, Seoul, Korea (2006)
12. Harris, N.D., Suggett, A.J., Barber, D.C., Brown, B.H.: Applications of applied potential tomography APT in respiratory medicine. Clin. Phys. Physiol. Meas. 8, 155–165 (1987)
13. Harris, N.D., Suggett, A.J.: Applied potential tomography: a new technique for monitoring pulmonary function. Clinical Physics & Physiological Measurement 9(Supplement A), 79–85 (1988)
14. Coulombe, N., Gagnon, H., Marquis, F., Skrobik, Y., Guardo, R.: A parametric model of the relationship between EIT and total lung volume. Physiological Measurement 26, 401–411 (2005)
15. Jang, J.: ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Trans. on Systems, Man and Cybernetic 23, 665–685 (1993)
16. Sugeno, M., Kang, G.T.: Structure Identification of Fuzzy Model. Fuzzy Sets and Systems 28(1), 15–33 (1988)
17. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modelling and control. IEEE Trans. Systems, Man, and Cybernetics 15(1), 132–166 (1985)
18. Panoutsos, G., Mills, G.H., Wang, A., Mahfouf, M., Brown, B.H.: Initial comparisons of absolute electrical impedance tomography (EIT) lung volume estimates with Spirometry. Proceedings of the ARS Meeting Anaesthetic Research Society, British Journal of Anaesthesia 98(2), 294 (2007)
19. Panoutsos, G., Tunney, D.R., Mills, G.H., Al-Jabary, T., Mahfouf, M., Brown, B.H.: Electrical Impedance Tomography: an evaluation of its ability to detect changes in lung volume and expansion during single lung ventilation. Proceedings of the ARS Meeting Anaesthetic Research Society, British Journal of Anaesthesia 100(4), 584 (2008)
20. Mohammad-Samuri, S., Denai, M.A., Panoutsos, G., Mahfouf, M., Linkens, D.A., Meekings, T., Mills, G.H., Brown, B.H.: The Sheffield Mk3.5 Absolute Resistivity aEIT System - Review of Recent Updates and Future Trends. In: 10th International Conference on Biomedical Applications of Electrical Impedance Tomography (EIT), Manchester, UK, June 16-19 (2009)
21. Hepper, N.G.G., Fowlwer, W.S., Helmholz Jr., H.F.: Relationship of Height to Lung Volume in Healthy Men. Journal of the Americam College of Chest Physicians 37, 314–320 (1960)
22. Chiu, S.: Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent and Fuzzy Systems 2(3) (1994)
23. Momamad-Samuri, S., Mahfouf, M., Denaï, M., Ross, J.J., Mills, G.H.: Absolute EIT Coupled To a Blood Gas Physiological Model For The Assessment of Lung Ventilation In Critical Care Patients. In: 21st Meeting of European Society for Computing and Technology in Anaesthesia in Intensive Care (ESCTAIC), Amsterdam, Netherland, October 6-9 (2010)
24. Hinz, J., Hahn, G., Neumann, P., Mohrenweiser, P., Hellige, G., Burchardi, H.: End-expiratory Lung Impedance Change Enables Bedside Monitoring of End-Expiratory Lung Volume Change. Journal of Intensive Care Medicine 29(1), 37–43 (2003)

Auditory Processing Inspired Robust Feature Enhancement for Speech Recognition

Hari Krishna Maganti and Marco Matassoni

Fondazione Bruno Kessler, Center for Information Technology - IRST
via Sommarive 18, 38123 Povo, Trento, Italy

Abstract. The performance of Mel-frequency cepstrum based automatic speech recognition system significantly degrade in noisy environments. In this article, the feasibility of utilizing the bio-inspired auditory features to improve noise robustness is investigated. The features are based on auditory characteristics, which include gammatone filtering and modulation spectral processing to emulate the mechanisms performed in the cochlea and middle ear aimed to improve robustness in human ear. The robust noise resistant features that emulate cochlea frequency resolution are extracted by gammatone filtering. And then a long-term modulation spectral processing, which preserves speech intelligibility in the signal is performed. Compared and discussed are the features based on the performance on Aurora5 database, comprising the meeting recorder digit task recorded with four different microphones in a hands-free mode at a real meeting room and living room and office room simulated data corrupted with different levels of additive noises. The performance of these features is also investigated for CHiME challenge, aiming at speech separation and recognition in noise background that has been collected from a real family room using binaural microphones. The experimental results show that the proposed features provide considerable improvement with respect to the standard feature extraction techniques for both the versions of the database.

1 Introduction

A significant trend in ubiquitous computing is to facilitate the user to communicate and interact naturally with concerned applications. Speech is an appealing mode of communication for such applications. The human-machine interaction using automatic speech processing technologies is a diversified research area, which has been investigated actively [1,2,3].

Speech acquisition, processing and recognition in a non-ideal acoustic environments are complex tasks due to presence of unknown additive noise, reverberation and interfering speakers. Additive noise from interfering noise sources, and convolutive noise arising from acoustic environment and transmission channel characteristics contribute to a degradation of performance in speech recognition systems. This article addresses the problem of robustness of automatic speech recognition (ASR) systems due to convolutive noise by modeling techniques performed by cochlea in human auditory processing system.

The influence of additive background noise on the speech signal can be expressed as

$$y(t) = x(t) + n(t) \tag{1}$$

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 205–218, 2013.
© Springer-Verlag Berlin Heidelberg 2013

where $y(t)$ is the degraded speech signal, $x(t)$ represents the clean signal, $n(t)$ is the additive noise, which is uncorrelated with the speech signal and unknown. Different techniques have been proposed based on voice activity detection based noise estimation, minimum statistics noise estimation, histogram and quantile based methods, and estimation of the posteriori and a priori signal-to-noise ratio [4]. In Ephraim and Cohen [5], various approaches to speech enhancement based on noise estimation and spectral subtraction are discussed.

Apart from the stationary background noise, another important source of degradation is caused by reverberation produced in acoustic environment. The speech signal acquired in a reverberant room can be modeled as convolution of the speech signal with the room impulse response,

$$y(t) = x(t) * h(t) \tag{2}$$

where $y(t)$ is the degraded speech signal, $x(t)$ represents the clean signal, $h(t)$ is the impulse response of the room. The impulse response depends upon the distance between the speaker and the microphone, and room conditions, such as movement of people in the room, clapping, opening or closing doors, etc. Thus extracting robust features which can handle various room impulse responses is a complex and challenging task. A variant of spectral subtraction has been proposed in [6] to enhance speech degraded by reverberation.

In general to improve robustness of the noisy speech, processing can be performed at signal, feature or model level. Speech enhancement techniques aim at improving the quality of speech signal captured through single microphone or microphone array [7,8]. Robust acoustic features attempt to represent parameters less sensitive to noise by modifying the extracted features. Common techniques include cepstral mean normalization (CMN) and cepstral mean subtraction and variance normalization (CMSVN) and relative spectral (RASTA) filtering [2,9]. Model adaptation approach modify the acoustic model parameters to fit better with the observed speech features [7,10].

Performance of the human auditory system is more adept at noisy speech recognition. Auditory modeling, which simulates some properties of the human auditory system have been applied to speech recognition system to enhance its robustness. The information coded in auditory spike trains and the information transfer processing principles found in the auditory pathway are used in [11,12]. The neural synchrony is used for creating noise-robust representations of speech [12]. The model parameters are fine-tuned to conform to the population discharge patterns in the auditory nerve which are then used to derive estimates of the spectrum on a frame-by-frame basis. This was extremely effective in noise and improved performance of the ASR dramatically. Various auditory processing based approaches were proposed to improve robustness [13,14,15] and in particular, the works described in [12,16] were focused to address the additive noise problem. Further, in [17] a model of auditory perception (PEMO) developed by Dau et al. [15] is used as a front end for ASR, which performed better than the standard Mel-frequency based cepstral coefficients (MFCC) for an isolated word recognition task. Principles and models relating to auditory processing, which attempt to model human hearing to some extent have been applied for speech recognition in [9,18].

The important aspect in a speech recognition system is to have abstract representation of highly redundant speech signal, which is achieved by frequency analysis. The cochlea and hair cells of the inner ear perform spectrum analysis to extract relevant features. The models for auditory spectrum analysis are based on filterbank design, which are usually characterized by non-uniform frequency resolution and non-uniform bandwidth on linear scale. Examples include popular speech analysis techniques, namely Mel frequency cepstrum and perceptual linear prediction which try to emulate human auditory perception. Other important processing is based upon Gammatone filter bank, which is designed to model human cochlear filtering and is shown to provide robustness in adverse noise conditions for speech recognition tasks [16,19]. In [16], gammatone based auditory front-end exhibited robust performance compared to traditional front-ends based on MFCC, PLP and standard ETSI frontend. For large vocabulary speech recognition tasks, the performance of these features have been competitive with standard features like MFCC and PLP [19]. Another important psychoacoustic property is modulation spectrum of speech, which is important for speech intelligibility. The relative prominence of slow temporal modulations is different at various frequencies, similar to perceptual ability of human auditory system. Particularly, most of the useful linguistic information is in the modulation frequency components from the range between 2 and 16 Hz, with dominant component at around 4 Hz [20,21,18]. Modulation spectrum based features computed over longer windows have been effective in measuring speech intelligibility in noisy environments and speech detection [22,23,24].

In this work, an alternate approach based on psychoacoustic properties combining gammatone filtering and modulation spectrum of speech, to preserve both *quality* and *intelligibility* for feature extraction is presented. Gammatone frequency resolution reduces the ASR system sensitivity to environmental reverberant signal attributes and improve the speech signal characteristics. Further, long-term modulation preserves the linguistic information in the speech signal, improving the accuracy of the system. The features derived from the combination are used to provide robustness, particularly in the context of mismatch between training and testing reverberant environments. The studied features are shown to be reliable and robust to the effects of the hands-free recordings in the reverberant meeting room. The effectiveness of the proposed features is demonstrated with experiments which use real-time reverberant speech acquired through four different microphones. For comparison purposes the recognition results obtained using conventional features are tested, and usage of the proposed features proved to be efficient.

The paper is organized as follows: Section 2 gives an overview of the auditory inspired features, including gammatone filter bank processing and modulation spectrum processing. Section 3 describes the methodology for feature extraction. Section 4 presents database description, experiments and results. Section 5 discusses the results. Finally, Section 6 concludes the paper.

2 Feature Description

In this section, a brief introduction and general overview of auditory features based on gammatone filter bank and modulation spectrum is presented.

2.1 Gammatone Filter Bank

The gammatone filter was first conceptualized by Flanagan as a model of the basilar membrane displacement in the human ear [25]. Johannesma used it to approximate responses recorded from the cochlear nucleus in the cat [26]. de Boer and de Jongh used a gammatone function to model impulse responses from auditory nerve fiber recordings, which have been estimated using a linear reverse-correlation technique [27]. Patterson et al. showed that the gammatone filter also delineates psychoacoustically determined auditory filters in humans [28].

Gammatone filters are linear approximation of physiologically motivated processing performed by the cochlea[29], comprise series of bandpass filters, whose impulse response is defined by:

$$g(t) = at^{n-1}cos(2\pi f_c t + \phi)e^{-2\pi bt} \tag{3}$$

where n is the order of the filter, b is the bandwidth of the filter, a is the amplitude, f_c is the filter center frequency and ϕ is the phase.

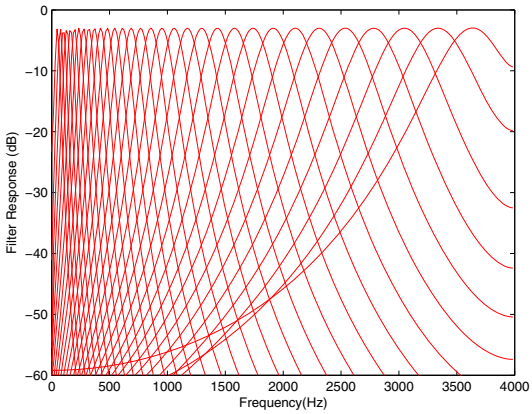


Fig. 1. Frequency response for the 32-channel gammatone filterbank

The filter center frequencies and bandwidths are derived from the filter's Equivalent Rectangular Bandwidth (ERB) as detailed in [29]. In [30], Glasberg and Moore relate center frequency and the ERB of an auditory filter as

$$ERB(f_c) = 24.7(\frac{4.37f_c}{1000} + 1) \tag{4}$$

The filter output of the m^{th} gammatone filter , X_m can be expressed by

$$X_m(t) = x(t) * h_m(t) \tag{5}$$

where $h_m(t)$ is the impulse response of the filter.

The frequency response of the 32-channel gammatone filterbank is as shown in Fig. 1.

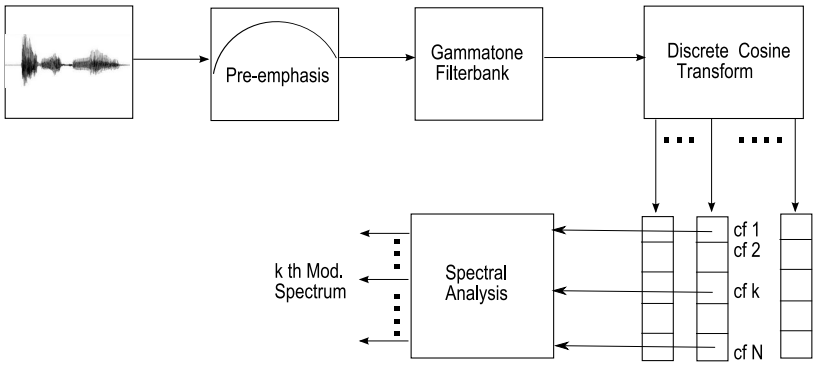


Fig. 2. Processing stages of the gammatone modulation spectral feature

2.2 Modulation Spectrum

The temporal evolution of speech spectral parameters, which describe slow variation in energy represent important information associated with phonetic segments [31]. The low-frequency modulations encode information pertaining to syllables, by virtue of variation in the modulation pattern across the acoustic spectrum. Dudley showed that essential information in speech is embedded in modulation patterns lower than 25 Hz distributed over a few as 10 discrete spectral channels [32]. Further, studies by Drullman et al. confirmed the importance of amplitude modulation frequencies on speech intelligibility, particularly modulation frequencies below 16Hz contributing to speech intelligibility [33]. Houtgast and Steeneeken demonstrated that modulation frequencies between 2 and 10 Hz can be used as an objective measure of speech intelligibility, for assessing quality of speech over wide range of acoustic environments [22].

The long-term modulations examine the slow temporal evolution of the speech energy with time windows in the range of 160 - 800 ms, contrary to the conventional short-term modulations studied with time windows of 10 -30 ms which capture rapid changes of the speech signals. Generally, the modulation spectrum is computed as following: speech signal $X(k)$ is segmented into frames by a window function $w(k,t)$, where t is frame number. Short-time Fourier transform of the windowed speech signal $X(t,f)$ is calculated as

$$Y(t,f) = \sum_{i=-\infty}^{\infty} X(f-i)W(i,t) \tag{6}$$

The modulation spectrum $Y_m(f,g)$ is obtained by applying Fourier transform on the running spectra, obtained by taking absolute values $|Y(t,f)|$ at each frequency, expressed as

$$Y_m(f,g) = FT[|Y(t,f)|]|_{t=1,\ldots T} \tag{7}$$

where T is the total number of frames and g is the modulation frequency. The relative prominence of slow temporal modulations is different at various frequencies, similar to

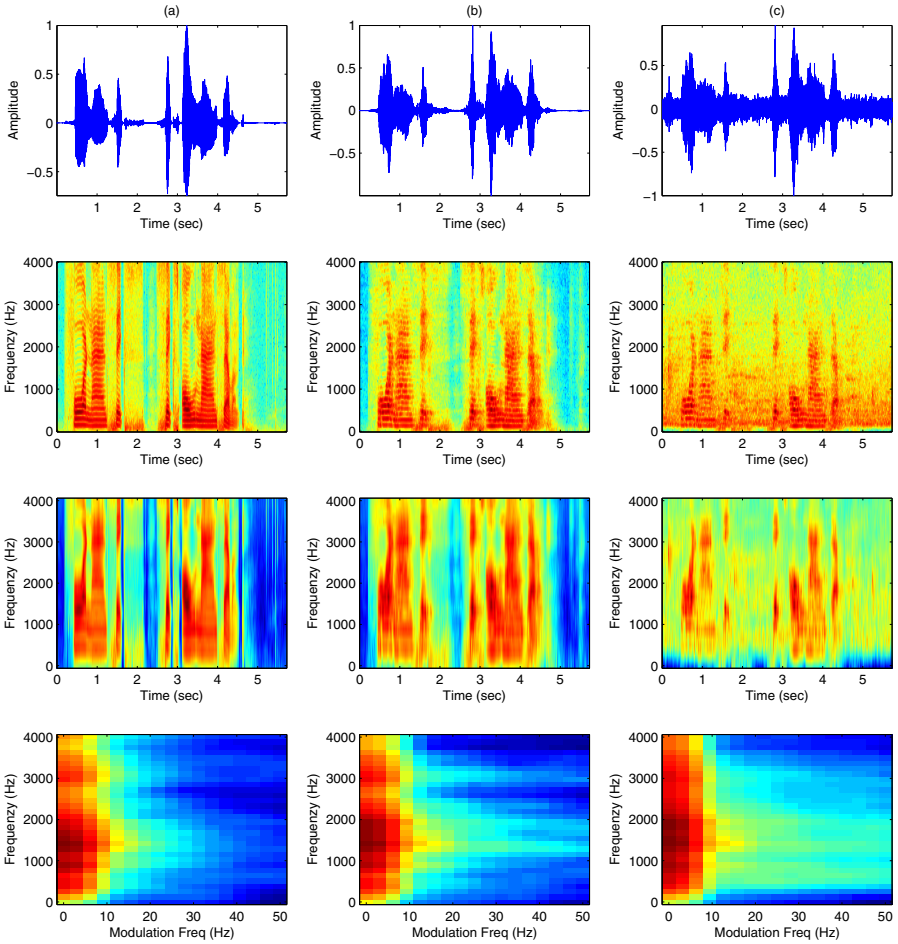


Fig. 3. Waveform, spectrogram, gammatonegram, and modulation spectrum density plots for the (a)clean, (b)reverberant and (c)additive noise corrupted speech

perceptual ability of human auditory system. Most of the useful linguistic information is in the modulation frequency components from the range between 2 and 16 Hz, with dominant component at around 4 Hz [33,21]. In [21], it has been shown that for noisy environments, the components of the modulation spectrum below 2 Hz and above 10 Hz are less important for speech intelligibility, particularly the band below 1 Hz contains mostly information about the environment. Therefore the recognition performance can be improved by suppressing this band in the feature extraction.

The comparative waveforms, spectrograms, gammatonegrams and modulation spectrum density plots of the clean and noisy versions corrupted with convolutive and additive noises of the same speech utterance are as shown in Fig. 3. The example is from Aurora 5 and as following:

(a) clean sentence "4966o97" from TI-DIGITS
(b) sentence in reverberant environment (living room, T60 appr. 0.5s)
(c) sentence in reverberant environment (living room, T60 appr. 0.5s) + additive noise (interior noise at 10dB).

From modulation spectrum density plots, some of the important characteristics of the modulation spectrum can be observed. The important information of speech is concentrated in the area from 2 Hz and 16 Hz, particularly 2 Hz and 4 Hz contain crucial information related to the variation of phonemes.

3 Methodology

The block schematic for the gammatone modulation spectrum based feature extraction technique is shown in Fig. 5. The speech signal first undergoes pre-emphasis, which flatten the frequency characteristics of the speech signal. The signal is then processed by a gammatone filterbank which uses 32 frequency channels equally spaced on the equivalent ERB scale as shown in Fig. 1. The impulse responses of the gammatone filterbank are similar to the impulse responses of the auditory system found in physiological measurements [27]. The filterbank is linear and does not consider nonlinear effects such as level-dependent upward spread of masking and combination tones. The computationally effective gammatone filter bank implementation as described in [34] is used. The gammatone filter bank transform is computed over L ms and the segment is shifted by n ms. The log magnitude resulting coefficients are then decorrelated by applying a discrete cosine transform (DCT). The computations are made over all the incoming signal, resulting in a sequence of energy magnitudes for each band sampled at $1/n$ Hz. Then, frame by frame analysis is performed and a N-dimensional parameter is obtained for each frame. The modulation spectrum of each coefficient which is defined as the Fourier transform of its temporal evolution is computed. In each band, the modulations of the signal are analyzed by computing FFT over the P ms Hamming window and the segment is shifted by p ms. The energies for the frequencies between the 2 - 16 Hz, which represent the important components for the speech signal are computed.

For example, if the given signal $x(t)$ is sampled at 8 kHz, a first-order high pass pre-emphasis filter is applied and short segments of speech are extracted with a 25 window. The window is shifted by 10 ms which corresponds to a frame rate of 100 Hz. Each speech frame is then processed by a 32-channel gammatone filterbank. The 32 logarithmic gammatone spectral values are transformed to the cepstral domain by means of a DCT. Thirteen cepstral coefficients $C0$ to $C12$ are calculated. The modulation spectrum of each coefficient, (sampled at 100Hz) is calculated with a 160 ms window, shifted by 10 ms. Thirteen coefficients $C13$ to $C26$ which are first-order derivatives are further extracted. The features are named gammatone filterbank modulation cepstral (GFMC) features.

The same processing is also performed by replacing gammatone filterbank with Mel filterbank in the Figure 5 resulting in Mel-frequency modulation cepstral (MFMC) features. The performance of these features in comparison to GFMC features are discussed in Section 4.

4 Experiments and Results

To evaluate the performance, a full HTK based recognition system is used. The HMM-based recognizer architecture specified for use with the Aurora 5 database is used [35]. The training data is downsampled version of clean TIDIGITS at a sampling frequency of 8 kHz, with 8623 utterances. There are eleven whole word HMMs each with 16 states and with each state having four Gaussian mixtures. The sil model has three states and each state has four mixtures.

4.1 Convolutive Noise

The experiments are conducted on a subset of the Aurora-5 corpus - meeting recorder digits. The data comprise real recordings in a meeting room, recorded in a hands-free mode at the International Computer Science Institute in Berkeley. The dataset consists of 2400 utterances from 24 speakers, with 7800 digits in total. The speech was captured with four different microphones, placed at the middle of the table in the meeting room. The recordings contain only a small amount of additive noise, but have the effects of hands-free recording in the reverberant room. There are four different versions of all utterances recorded with four different microphones, with recording levels kept low.

Table 1 shows the results in % word accuracies for meeting recording digits recorded with four different microphones, labeled as 6, 7, E and F. The average performance of four microphones for different features is shown at the last column of the table. ETSI-2 correspond to the standard advanced front-end as described in [35]. PLP and MFCC are the standard 39-dimensional Perceptual linear prediction and Mel frequency features along with their delta and acceleration derivatives. MFMC indicate Mel Frequency Modulation Spectral based Cepstral (MFMC) features where the first thirteen features are extracted in a traditional way, and the rest are the modulation features (13) and their derivatives (13) derived as discussed in Section 3, except for Gammatone filterbank being replaced with Mel filterbank. The GFCC features are extracted in a similar way as reported in Section 3 with $C0$ to $C12$ being the corresponding cepstral coefficients. GFMC indicate Gammatone Frequency Modulation Spectral based Cepstral (GFMC) features derived in a same way as GFCC but appended with modulation spectral features corresponding to $C13$ to $C26$ and their corresponding derivatives as discussed in Section 3.

Table 1. Word recognition accuracies (%) for different feature extraction techniques on four different microphones

Channel	6	7	E	F	Average
ETSI-2	64.3	47.6	58.1	62.7	58.1
PLP	73.8	63.8	68.1	71.4	69.2
MFCC	75.8	64.7	67.3	75.9	70.9
MFMC	75.6	61.0	70.8	77.9	71.3
GFCC	86.0	79.0	78.3	84.2	81.9
GFMC	87.8	82.7	82.2	86.9	84.9

From Table 1, it is evident that the advanced ETSI front-end has highest error rates compared to the MFCC and PLP. This demonstrates that for reverberant environments the advanced ETSI front-end is not effective as compared to its performance in the presence of additive background noise. It can be inferred that the techniques applicable for additive background noise removal are not suitable to handle reverberant conditions. The MFMC features have better performance than MFCC, which in turn had better performance than PLP. It can also be seen that the GFCC features were effective, performing better than any of the baseline systems (ETSI-2, PLP, MFCC). This is consistent with the earlier studies which have shown that gammatone based features exhibit robust performance compared to MFCC, PLP features and ETSI frontend [16,19].

It can also be observed that the performance of GFMC is the best among all the baselines and features compared, and consistent across all the channels. However, the combination of Mel filtering and modulation spectral features is not as beneficial as gammatone filtering with modulation spectral features. This clearly demonstrates the efficiency of this combination of these features in reverberation conditions.

4.2 Additive Noise

Further, to test the efficiency for practical conditions which contain additive noises along with reverberation effects, experiments are conducted on hands free office and hands free livingroom simulated data with clean and 15 dB, 10 dB, 5 dB, 0 dB SNR additive noise corrupted signals. The data is from Aurora-5 database, where condition is simulated as combination of additive noise and reverberation[35]. Aurora-5 covers all effects of noises as they occur in realistic application scenarios. In this experiments, hands free speech input in a office and in a living room is considered. The reverberation time for the office room was randomly varied in the range of 0.3 to 0.4 s and for the living room was in the range of 0.4 to 0.5 s. In Table 2, HFOffice, HFLroom, - represents hands free office, hands free living room and no additive noise respectively.

Table 2. Word recognition accuracies (%) for clean, hands free office and hands free living room conditions

Feature	HFOffice					HFLroom				
	-	15dB	10dB	5dB	0dB	-	15dB	10dB	5dB	0dB
PLP	88.6	65.1	40.4	21.1	6.7	74.3	46.9	27.7	14.2	3.7
MFCC	90.1	61.6	41.5	18.0	4.0	75.8	40.2	23.8	8.6	1.9
MFMC	94.5	68.1	40.5	18.5	8.7	85.2	52	28.2	13.5	7.8
GFCC	89.1	65.6	39.5	18.4	8.3	73.8	48.9	29.1	13.7	6.2
GFMC	92.2	73.3	44.8	20.7	10.1	78.6	57.4	34.1	15.6	8.2

From Table 2, it can be observed that for all the features the performance degrades significantly in additive noise compared to no additive noise case. Also, it can be seen that GFCC has better performance than MFCC and PLP. It can also be observed that MFMC has better performance than GFCC, showing that the combination of modulation features in mel domain is beneficial for this task. It can be observed that GFMC has better performance than GFCC, MFCC, MFMC and PLP indicating efficiency of this features in additive noise and reverberant conditions.

4.3 CHiME Challenge

The task of Computational Hearing in Multisource Environments (CHiME) challenge is to separate the speech and recognise the commands being spoken using systems that have been trained on noise-free commands and room noise recordings [36]. The CHiME background noise is recorded separately from the target speech. The target speech is subsequently artificially added but in a manner that closely simulates the effect of the speech being present in the room. This controls the target speech SNR, target talker location, talker characteristics etc. For the background noise, a domestic environment was considered, such as would be encountered in a home automation application. It provides rich mix of sound sources, some of which may be easy to model (e.g. a washing machine that remains in a fixed position and runs a predictable program) and some which are not (e.g. children running around while talking, screaming and laughing).

The ASR task is speaker dependent and employs a small but phonetically confusable vocabulary. The recordings are from a genuine living room (in a house with two small children) measured over a period of several weeks and the SNRs employed in the challenge range between 18 dB down to -6 dB.

Figure 4 shows comparison of GFMC features and MFCC before separation of the speech signals. It can be observed that GFMC is much effective than MFCC for all ranges of SNRs, particularly at very low SNRs.

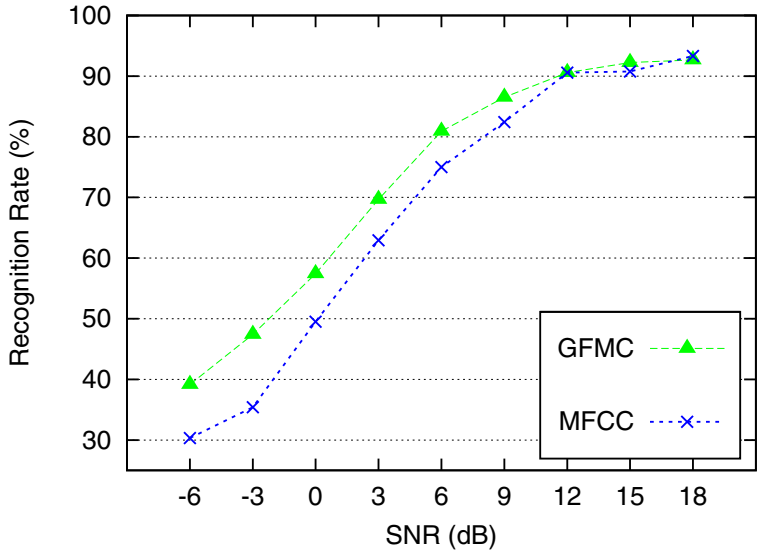


Fig. 4. Comparison of the proposed GFMC features and MFCC before separation of the speech signals

Figure 5 shows comparison of GFMC features and MFCC after separation of the speech signals which is achieved through semi-blind source separation (SBSS). The BSS algorithm is a modified Recursively Regularized Independent Component Analysis [37] according to a semi-blind structure as in [38], where the used prior is the mixing

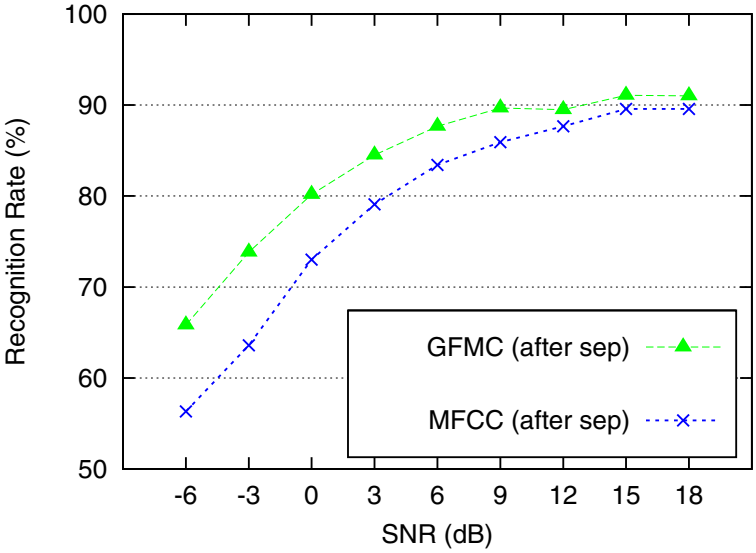


Fig. 5. Comparison of the proposed GFMC features and MFCC after separation of the speech signals achieved through semi-blind source separation

parameters of the target source (estimated beforehand). It can be observed that GFMC is much efficient than MFCC for all ranges of SNRs, with significant performance at very low SNRs.

5 Discussion

The results from both Table 1 and 2 and Figures 4 and 5 indicate that the gammatone frequency resolution was effective in reducing system sensitivity to reverberation and additive noise, and improved the speech signal characteristics. It can also be observed from Table 1, that the combination of gammatone filtering with modulation spectral features is beneficial than the combination of Mel filtering and modulation spectral features. The emphasis on slow temporal changes in the spectral structure of long-term modulations preserved the required speech intelligibility information in the signal which further improved the accuracy of the system. Thus, by extracting features that model human hearing to some extent mimicking the processing performed by cochlea, particularly emulating cochlea frequency resolution was beneficial for speech feature enhancement.

6 Conclusions

The paper has presented auditory inspired modulation spectral features for improving ASR performance in presence of room reverberation. The proposed features were derived from features based on emulating the processing performed by cochlea to improve the robustness, specifically gammatone frequency filtering and long-term modulations of the speech signal. The features were evaluated on Aurora-5 database, meeting

recorder digit task and living room and office room simulated data corrupted with different levels of additive noises. Results were compared with standard ETSI advanced front-end and conventional features. The results show that the proposed features perform consistently better both in terms of robustness and reliability. The work presented results in both additive noise and reverberant scenario where the speech signal was corrupted with 15 dB, 10dB, 5 dB and 0dB SNR noise, simulated with hands-free office and hands-free living room conditions. The work also presented performance of these features on CHiME challenge before and after separation of acoustic sources. The results are promising, performing better than the conventional features, indicating the efficiency of this features in practical scenarios.

Our study raised number of issues, including study of auditory inspired techniques for improvement of standard additive noise removal techniques to deal with reverberation condition. The gammatone filter implemented in this work is linear and does not consider nonlinear effects such as level-dependent upward spread of masking and combination tones. For the future, we like to investigate these issues to efficiently deal with real world noisy speech, and evaluate these features on large vocabulary tasks.

References

1. Kellermann, W.: Some current challenges in multichannel acoustic signal processing. The Journal of the Acoustical Society of America 120, 3177–3178 (2006)
2. Droppo, J., Acero, A.: Environmental Robustness. In: Handbook of Speech Processing, pp. 653–679. Springer, Heidelberg (2008)
3. Maganti, H.K., Member, S., Gatica-perez, D., Mccowan, I.: Speech enhancement and recognition in meetings with an audio-visual sensor array. In: IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne, EPFL (2006)
4. Woelfel, J., McDonough, J.: Distant Speech Recognition, 1st edn. John Wiley (2009)
5. Ephraim, Y., Cohen, I.: Recent Advances in Speech Enhancement. CRC Press (2006)
6. Habets, E.A.P.: Single-channel speech dereverberation based on spectral subtraction. In: PRORISC, Veldhoven, The Netherlands, pp. 250–254 (2004)
7. Omologo, M., Svaizer, P., Matassoni, M.: Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Communication 25, 75–95 (1998)
8. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing 9, 504–512 (2001)
9. Hermansky, H., Morgan, N.: Rasta processing of speech. IEEE Transactions on Speech and Audio Processing 2, 578–589 (1994)
10. Gales, M., Young, S.: A fast and flexible implementation of parallel model combination. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1995, vol. 1, pp. 133–136 (1995)
11. Holmberg, M., Gelbart, D., Ramacher, U., Hemmert, W.: Automatic Speech Recognition with Neural Spike Trains. In: INTERSPEECH (2005)
12. Deng, L., Sheikhzadeh, H.: Use of Temporal Codes Computed From a Cochlear Model for Speech Recognition. Psychology Press (2006)
13. Ghitza, O.: Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. Journal of Phonetics (1988)
14. Seneff, S.: A joint synchrony/mean-rate model of auditory speech processing. Journal of Phonetics 16, 55–76 (1988)

15. Dau, T., Pueschel, D., Kohlrausch, A.: A quantitative model of the effective signal processing in the auditory system. The Journal of the Acoustical Society of America 99, 3615–3622 (1996)
16. Flynn, R., Jones, E.: A comparative study of auditory-based front-ends for robust speech recognition using the aurora 2 database. In: Irish Signals and Systems Conference, 2006, pp. 111–116. IET (2006)
17. Kleinschmidt, M., Tchorz, J., Kollmeier, B.: Combining speech enhancement and auditory feature extraction for robust speech recognition. Speech Commun. 34, 75–91 (2000)
18. Hermansky, H.: Auditory modeling in automatic recognition of speech. ECSAP (1996)
19. Schluter, R., Bezrukov, L., Wagner, H., Ney, H.: Gammatone features and feature combination for large vocabulary speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, pp. IV-649–IV-652 (2007)
20. Drullman, R., Festen, J.M., Plomp, R.: Effect of reducing slow temporal modulations on speech reception. The Journal of the Acoustical Society of America 95, 2670–2680 (1994)
21. Kanedera, N., Arai, T., Hermansky, H., Pavel, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28, 43–55 (1999)
22. Houtgast, T., Steeneken, H.J.M., Plomp, R.: Predicting speech intelligibility in rooms from the modulation transfer function. Acustica 46, 60–72 (1980)
23. Kingsbury, B.: Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments. PhD thesis, Michigan State University (1998)
24. Maganti, H.K., Motlicek, P., Gatica-Perez, D.: Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP (2007)
25. Flanagan, J.L.: Models for approximating basilar membrane displacement. Journal of the Acoustical Society of America 32 (1960)
26. Johannesma, P.I.: The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: Symposium on Hearing Theory (Institute for Perception Research), Eindhoven, Holland, pp. 58–69 (1972)
27. Boer, E.D.: On the principle of specific coding. Journal of Dynamic Systems, Measurement, and Control 95, 265–273 (1973)
28. Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P.: An efficient auditory filterbank based on the gammatone function. In: Meeting of the IOC Speech Group on Auditory Modelling at RSRE (1987)
29. Slaney, M.: An efficient implementation of the patterson holdsworth auditory filterbank. Technical report, Apple Computers, Perception Group (1993)
30. Glasberg, B.R., Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data. Hearing Research 47, 103–138 (1990)
31. Greenberg, S.: On the origins of speech intelligibility in the real world. In: ESCA Workshop on Robust Speech Recognition for Unkown Communication Channels, pp. 23–32 (1997)
32. Dudley, H.: Remarking speech. The Journal of the Acoustical Society of America 11, 169–177 (1939)
33. Drullman, R., Festen, J.M., Plomp, R.: Effect of temporal envelope smearing on speech reception. Journal of The Acoustical Society of America 95 (1994)
34. Ellis, D.: Gammatone-like spectrograms (2010), `http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram`
35. Hirsch, H.: Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments (2007), `http://aurora.hsnr.de/aurora-5/reports.html`

36. Christensen, H., Baker, J., Ma, N., Green, P.: The chime corpus: a resource and a challenge for computational hearing in multisource environments. In: Interspeech 2010 (2010)
37. Nesta, F., Wada, T., Juang, B.H.: Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation. IEEE Transactions on Audio, Speech, and Language Processing 19, 583–599 (2011)
38. Nesta, F., Svaizer, P., Omologo, M.: Convolutive bss of short mixtures by ica recursively regularized across frequencies. IEEE Transactions on Audio, Speech, and Language Processing 19, 624–639 (2011)

A Robust and Efficient Spatio-Temporal Feature Selection for Interpretation of EEG Single Trials

Yehudit Meir-Hasson[1], Andrey Zhdanov[1],
Talma Hendler[2], and Nathan Intrator[1]

[1] Balvatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel
yehudit.hasson@gmail.com
[2] The Functional Brain Imaging Unit, Wohl Institute for Advanced Imaging
Tel Aviv Sourasky Medical Center, Tel-Aviv, Israel

Abstract. Interpretation of brain states from EEG single trials, multiple electrodes and time points, is addressed. A computationally efficient and robust framework for spatio-temporal feature selection is introduced. The framework is generic and can be applied to different classification tasks. Here, it is applied to a visual task of distinguishing between faces and houses. The framework includes training of regularized logistic regression classifier with cross-validation and the usage of a wrapper approach to find the optimal model. It was compared with two other methods for selection of time points. The spatial-temporal information of brain activity obtained using this framework, can give an indication to correlated activity of regions in the brain (spatial) as well as temporal activity correlations between and within EEG electrodes. This spatial-temporal analysis can render a far more holistic interpretability for visual perception mechanism without any a priori bias on certain time periods or scalp locations.

Keywords: EEG, BCI, Regularization, Spatio-temporal analysis.

1 Introduction

A major challenge in neuroscience is inferring how momentary mental states are mapped into a particular pattern of brain activity. Inference, which is based on EEG single-trial (i.e. short segment of the EEG) has practical implementations for brain computer interface (BCI) applications. BCI applications are designated for people suffering from physical disabilities, helping them to communicate with an electronic device through real time interpretation of their brain signals ([1]-[4]).

A common way to analyze EEG single-trials is through classification (for review, see [5]). One of the main challenges is the amount of data needed to properly describe the different states; It increases exponentially with the dimensionality of the data and is termed the *curse of dimensionality* [6].

To reduce the dimension of the data, many feature selection methods have been developed for identifying and choosing an optimal subset of features from the data. Often, researchers focus on few electrodes based on algorithms for channel selection that pick the most promising channels for classification. Muller et al. [7] utilized Spatial Pattern Analysis and PCA for channel selection and compared it to a set of four electrodes

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 219–232, 2013.
© Springer-Verlag Berlin Heidelberg 2013

chosen based on prior knowledge. Palaniappan et al. [8] and Schroder et al. [9] found appropriate channels via a Genetic Algorithm; Lal et al. [10] used Recursive Feature Elimination and Zero-Norm Optimization to reduce the number of electrodes from 39 to 12. Tomika and Muller [11] reduced the dimension of the data by down-sampling the signals. Another way to alleviate the curse of dimensionality is via regularization methods, which stabilize the solutions by introducing prior knowledge or by restricting the solution ([12]-[13]). Cross-validation can be used to find the optimized model and its regularization parameter ([11], [14]-[16]).

As data is becoming more readily available, it is more desirable to let the data guide the choice of an optimal model (determine relevant electrodes and time points) while minimizing a-priori assumptions. For this, a two-dimensional representation of the spatio-temporal predictive information of the brain activity is highly desirable, especially for development of new paradigms, for which the neural correlates may not be known in advance [17].

Modern data-driven analyses, such as microstate segmentation ([18]-[19]), have been developed and used to study the spatio-temporal activity in the brain. Microstate segmentation uses the spatial distribution of the ERP which involves averaging over multiple trials of similar brain activity (for review, see [17]). Such a predictive map lacks the correlated activity between electrodes, which is lost in the traditional ERP approach. Moreover, the ability to assess the trial-to-trial variability in event-related potential experiments can provide new insights into brain functioning which may be ignored during ERP averaging.

Tomioka and Muller [11] suggested an EEG single trials spatio-temporal interpretation that was based on three different regularizers. The regularizers were used to reveal different and complementary aspects of the localization of the discriminative information. (The channel selection regularizer was used for spatially localizing the discriminative information, the temporal-basis selection regularizer localized the discriminative information in the temporal domain and the DS regularizer provided a small number of pairs of spatial and temporal filters that showed both spatial and temporal localization of the discriminative information in a compact manner). The regularizers were applied on a block diagonal data matrix which was composed of first order changes (short segment of filtered EEG signal with C channels and T sampled time-points) and second order changes (the covariance matrix of a short segment of band-pass filtered EEG). Their method requires an estimation of a large-parameter model, which may be prone to overfitting. Furthermore, reducing the sampling rate may ignore important properties of the signal, which are visible in the EEG high temporal resolution. The use of different regularizers [11] may be problematic as it may produce contrasting interpretations with no clear ability to determine which of them is more accurate.

In this work, we follow the framework introduced by [16] and present an efficient and robust computational model for brain state interpretation from EEG single trials. Our approach is based on the use of regularization techniques to optimize classifier coefficients. We further demonstrate how to identify the most relevant time points and electrodes that might be most pertinent in contributing to differentiation between the mental states investigated.

Our approach employs a two-step classification scheme; First, potential time points are selected (i.e. the electrode activity in these time points is found to best separate between the two mental states). Second, the selected time points are fused to obtain a holistic network related to the paradigm. This paper extends previous work [20] and includes a comparison to time-points-selection via ERP method and to the sequential fusion of time points method. This two-step framework leads to an effective feature selection using a small number of parameters, dozens compared with thousands in [11], while maintaining a high temporal resolution.

The proposed framework is generic and can be applied to different brain imaging modalities and classification tasks. In this study we applied it to a visual task of distinction between faces and houses.

2 Materials and Methods

2.1 Experiment Setup

Four subjects (SUBJ1-SUBJ4, 4 females, two left handed, aged 23-28), participated in this experiment. All subjects gave informed consent to participate in the study, which was approved by the ethics committee of the Tel Aviv Sourasky Medical Center. Subjects were presented with images from two different categories-faces and houses. The images of faces were taken from the [21] and [22] databases and include fearful or neutral facial expression. The experiment included 4 sessions, each of 138 epochs 2-seconds-long. During each epoch, the subject was presented with one image of a fearful face, neutral face, house or blank (32, 32, 64 and 10 epochs respectively). To achieve visual field segregation, participants were explicitly instructed to ignore the pictures and to concentrate on a fixation dot at the center of the screen. Throughout the experiment, participants were asked to report the color change of the central fixation dot.

2.2 EEG Data Acquisition

Continuous EEG data was recorded simultaneously with fMRI acquisition. In this study, we are focusing on the EEG data and have set aside the combined fMRI data for further research. Good signal-to-noise ratio of the EEG data in the combined approach was previously shown at our lab ([23]-[24]).

We used a 32-channel BrainCap electrode cap with sintered Ag/AgCl ring electrodes (30 EEG channels, 1 ECG channel and 1 EOG cannel, Falk Minow Services, Herrsching-Breitbrunn, Germany) and a MR-compatible, 32-channel, battery-operated amplifier (Brain Products, GmBH, Germany). The electrodes were positioned according to the 10/20 system. The reference electrode was between Fz and Cz [25]. The signal was amplified, and sampled at 5000 Hz using the Brain Vision Recorder software (Brain Products). The EEG data was transmitted from the scanner room via an optical fiber to a PC in the control room. The exact timing of stimulus onset and MRI scanner gradient switching was transmitted to the EEG amplifier and recorded together with the EEG signal.

2.3 EEG Analysis

EEG data analysis was performed with EEGLAB 6.01 software package (Schwartz Center for Computational Neuroscience, University of California, San Diego), MATLAB software and FMRIB plug-in for EEGLAB, provided by the University of Oxford Centre for Functional MRI of the Brain (FMRIB). Pre-processing of the EEG data included the following steps: MR gradient artifacts removal and Cardio-ballistic artifacts removal using a FASTR algorithm implemented in FMRIB plug-in for EEGLAB ([23]-[24]). For computational efficiency, the EEG signals were down-sampled to 250 Hz and eye blinking artifacts were removed using ICA [26]. The data was then filtered with a 0.5-45 Hz band-pass filter and segmented into epochs starting 100 ms before the stimulus onset and ending 600 after the stimulus onset. Baseline correction was performed using the 100ms of pre-stimulus activity. In this manner for each subject, we obtained several dozens of epochs, each containing 32 (number of channels) x 175 (number of time sampling points in the segmented interval) values. Each epoch was associated with a class label *face* or *house* according to the stimulus which was presented.

3 Brain State Modeling

This section introduces the proposed brain state modeling approach for EEG single trials spatio-temporal analysis. Figure 1 demonstrates the flowchart of the ensemble method.

The essence of the modeling approach is creating a parametric family of classifiers and seeking an optimal member of this family by model selection techniques. The parameter that forms the collection of classifiers controls the bias/variance tradeoff (i.e. regularization parameter), thus a classifier with optimal bias/variance is chosen [27]. Each member of the family attempts to predict the mental states of the brain by finding the coefficients of the model which mostly differentiate the EEG data into two mental states. The selection of the optimal member is done based on classifier mental states prediction.

3.1 Model Estimation

Cross-validation is used for choosing the best model and estimating its predictive accuracy. This method is computationally expensive but is especially important when the number of samples is small. Cross-validation is applied twice: first for dividing the original data into train and test sets. We search for the optimal model on the train sets and record its accuracy on the test sets. For this we used m-k-fold cross validation, where k is the number of unique test sets, and m is the number of times, this process is repeated. Second, an additional inner n-fold cross-validation procedure is applied for selecting the optimal model on the training sets, where n is the number of averaged cross-validation iterations.

In the first cross validation procedure, the original data is partitioned into k disjoint sets. A single dataset is retained as the test data for testing the model, and the remaining $k-1$ disjoint datasets are used as training data. The cross-validation process is then

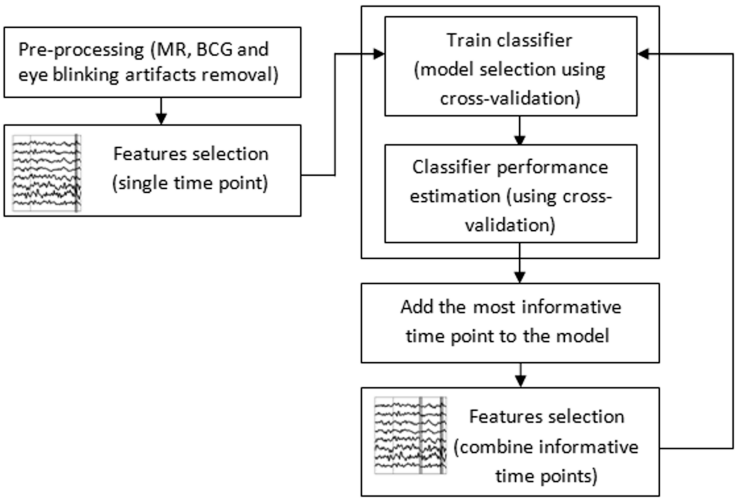


Fig. 1. Brain state modelling flow chart

repeated k times, with each of the k sets used exactly once as the test data. We repeat this process m times. The training sets are used for choosing the best model and the test sets are used to check its predictive accuracy. The predictive accuracy of the model is defined as the number of wrongly predicted samples divided by the overall number of samples.

The second cross-validation operation is used for choosing the optimal model. The training dataset is randomly splitted, n times, into 80-20% training and validation sets respectively. The classifier runs on the training set with different values of the regularization parameter (within the range of interest) and selects the one that yields the best results (i.e. bring mean square error, MSE, to minimum) (see Fig. 2).

The range of regularization values of interest is determined using the singular values, which are obtained from SVD decomposition of the processed data matrix (used for training and testing). The range is bounded between the minimal and the maximal singular values. For computational efficiency, the actual regularization values in that range are distributed uniformly on the logarithmic scale (i.e. the ratio of the two successive samples is constant).

3.2 Regularized Logistic Regression

The proposed regularized brain state interpretation can be used with a variety of linear and nonlinear classifiers. The, logistic regression model is the appropriate one for a binary classification task. It is also optimal in terms of simplicity, interpretability of its coefficients and speed ([28]-[29]).

A useful variable is the odds ratio, which is defined as the ratio of the probability that an event occurs to the probability that it fails. The logit (log odds) of the logistic regression model is given by the following equations, where w_i are the model's coefficients:

$$g(x) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_n x_n \tag{1}$$

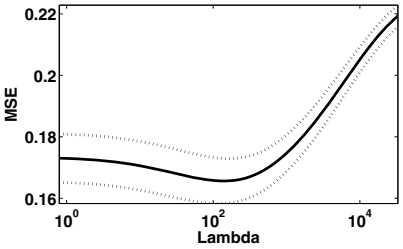


Fig. 2. MSE received on the validation set at the best time point versus the log of the regularization parameter. The lambda that minimizes the average error across iterations is chosen to be the optimal regularization parameter for the model.

$$P(Y = 1|x) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x))}} \tag{2}$$

$$logodds = log\left(\frac{\pi(x)}{1 - \pi(x)}\right) \tag{3}$$

The coefficients are often estimated via the Maximum Likelihood Estimation (MLE) method, which seeks to maximize the log likelihood over the entire observed data:

$$l(w) = \sum_{i=1}^{n} logP(Y = y_i|x_i) \tag{4}$$

The log likelihood value represents how likely the dependent variable can be predicted from the observed values of the independent variables. Maximization of the above expression can be done in various ways, most popular being the Newton-Raphson (NR) algorithm.

The regularized version of the logistic regression algorithm seeks to find the weights (w) which maximizes the equation:

$$l^{\lambda}(w) = l(w) - \frac{\lambda}{2} w^T w \tag{5}$$

We use the Matlab-based MVPA toolbox [30], which implements regularized logistic regression following notes from [31].

3.3 Feature Selection

As mentioned before, one of the main challenges while working with EEG signals is the high data dimensionality. In this case, feature selection is important for reducing the dimensionality of the input signal, removing noise, improving learning performance, speeding up the learning process and improving predictive accuracy. Feature selection has been extensively researched by the machine learning/pattern recognition community over the years (for review in the context of BCI, see [5]).

A two step feature selection algorithm is implemented. First, potential time points are selected, which are found to best separate between the two mental states. Second, the selected time points are fused to obtain a holistic network related to the paradigm. In the first step of the algorithm, the selection of the time points is done in two different ways: using the ERP approach and using a predictive approach based on a single trial performance evaluation. In the ERP approach the averaged EEG signal over trials are constructed (separatly for each condition). Significant time points, that have the highest difference between the average signals are chosen. In the predictive approach, 32 electrodes are selected from a single time point as an input for the classifier in the same manner as in [16]; A set of T trials labeled data samples is obtained, each represented by NxM signal matrix, where N is the number of channels and M is the number of time sampling points in the segmented interval. For each time point (from M), a feature vector that contains the EEG data of the entire electrodes in this time point, is created. (This reduces the dimension of the data from 32x175 to 32). Then, a family of classifiers is constructed with different regularization parameters and applied on the different time points. The model that achieved the minimum MSE on the validation set, over the entire time points, is chosen.

After selecting the model, the predictive accuracy of each time point is evaluated using the test sets, by applying the best model on each time point and averaging the results. The outcome of this stage is a ranking of the entire time points according to the performance of the model (Fig. 3). The coefficients of the regression equation at the time point where minimal prediction error is achieved indicate the contribution of activity in different electrodes in this time point towards the prediction. This can be interpreted as the strength of activity in electrodes, which best contribute to the mental states separation.

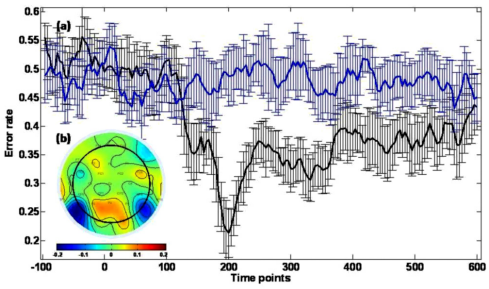


Fig. 3. (a) Predictive accuracy of each time point, on the testing set. The black line depicts the average error rate over the cross-validation iterations and the blue line represents control results obtained using the same algorithm on data with randomly scrambled target labels. It can be seen that the best prediction is achieved around 200ms after the stimulus onset (N170). (b) The coefficients of the regression equation in the best time point. The coefficients indicate the most contributing electrodes in this time point; Blue color indicates strong negative effect of faces compared to houses.

The formulation presented so far indicates the most predictive time point and the configuration of electrodes at that time point. This spatial coding, where the prediction depends on a configuration of electrodes activity at a single time point, may not be the optimal code used by the brain in interpreting the stimuli. Therefore it is possible that a temporal or spatio/temporal coding is more appropriate. The presented model can address this question, although the computational problem involved becomes too big for a single computer to handle, but using a computer grid of several hundred personal computers, the model can be extended in this direction [32].

We sort the local minima in the prediction graph to find different distinct temporal locations with prediction error minimum. The sorting is done in an increasing order (starting from the most predictable time point to the least predictable time point). Then a collection of models is applied, each using an increasing amount of information, where new time points (electrode information) are added into the model. In each such input data configuration we perform the full cross-validation estimation to estimate optimal regularization and prediction error.

The selection of the ensemble time points is done in two different ways; Primary selection using sequential assembling (i.e. the sorted time points are added to the model one after the other without additional search) and secondary selection using a wrapper algorithm [33], which is a technique for selecting an optimal subset of features from a large search space. In the wrapper algorithm, the features are assessed by their predictive power and are added to the subset one by one according to their contribution to the overall prediction. In both methods the ten most predictable time points were included in the process.

The outcome of the classifier for a different number of time points is compared and the ideal number of time points which had significantly lower error prediction is chosen (Fig. 4). Increasing the input vector adds electrode activity data, but also adds free parameters to the model leading to higher chance of overfitting the training data. We thus search for the ideal number of time points which balances between the two effects. Figure 5 depicts the best time points found for one subject and the electrodes activity in these time points contributing towards mental state discrimination.

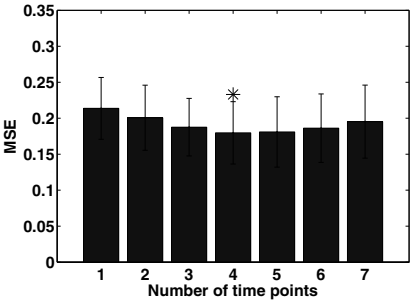


Fig. 4. Prediction error vs. number of time points. For this subject the optimal is 4, namely there was a significant prediction improvement up to that point ($*p < 0.05$).

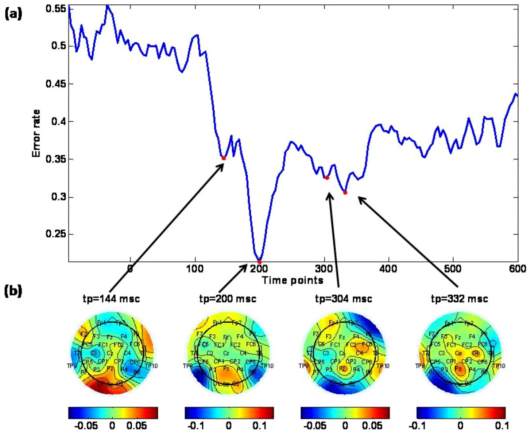


Fig. 5. (a) The ideal number of time points chosen as input for the classifier. (b) The regression coefficients received in those time points.

4 Results

4.1 Spatio-Temporal Analysis

Many studies have shown that pictures of faces elicit a much larger ERP of negative polarity than other object categories. This component peaks at occipital-temporal electrode sites at about 170 ms following stimulus onset [34]. The larger response of the N170 complex to faces is an undisputed observation among researchers in the field of face processing (Fig. 6).

We reinforce this result using EEG single trial classification (Fig. 7). For all subjects the best prediction achieved around 200 ms after the stimulus onset and the electrodes that contribute to the maximum separation between the mental states investigated are located in the occipital area. The coefficients obtained on single trial training correlated to the ERP of the corresponding electrodes. Negative coefficients indicate the ERP for faces is lower than the ERP for houses. In addition, both occipital electrodes (P7 and P8) are correlated in that time point.

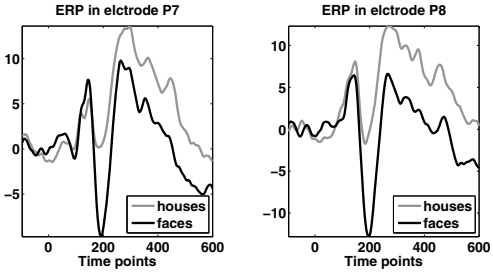


Fig. 6. ERP in electrodes P7 and P8

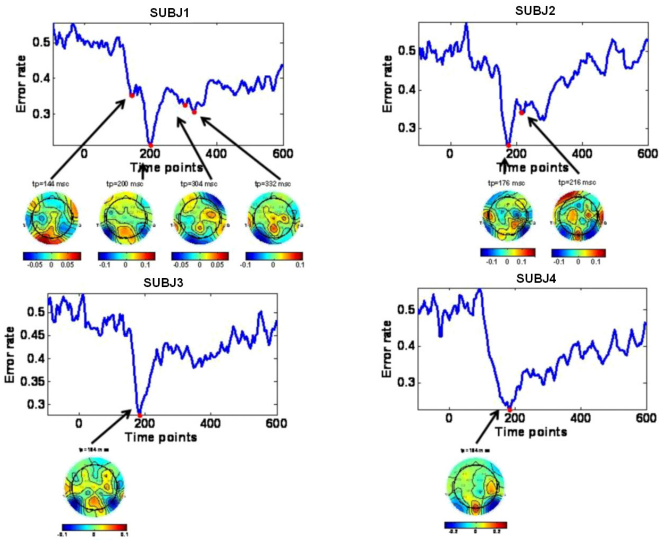


Fig. 7. Best time points found and the coefficients in these time points for different subjects in the houses and faces experiment. As expected, for the entire subjects the best prediction is achieved around 200ms after the stimulus onset and the most activated electrodes are in the occipital-temporal area.

The resulting spatial-temporal weight matrix provides a summary representation which is easily interpretable. A result of the dimensionality reduction, which is performed during the pre-processing stage, where relevant time points and electrodes are chosen, leads to simpler computational model training. This is in contrast to reducing the dimensionality via reducing the sampling rate [11]. The lowest error rates achieved for each subject using a single time point are summarized in Fig. 8 (the prediction error with the optimal number of time points is lower). The results were compared to the control experimental results, which were obtained using the same algorithm on randomly scrambled labels. The difference between the mean error estimates is significant for all subjects ($P < 0.05$).

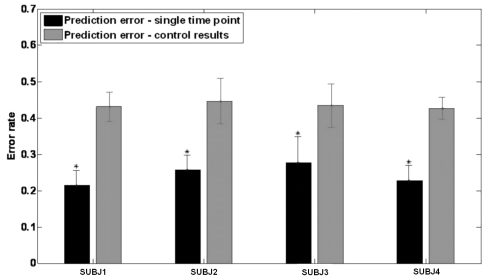


Fig. 8. Classification error rate for all 4 subjects. The classification error is compared to control results obtained using the same algorithm on randomly scrambled labels. The difference between the mean error estimates is significant for all subjects ($P < 0.05$).

4.2 Comparison with Existing Methods

As described in 3.3, our algorithm employed a two-step classification. The first step included selection of time points that were used to predict the two mental states. The time points were selected using two criteria: the ERP approach, which is based on the EEG signal average over trials and a predictive approach based on a single trial prediction performance. Table 1 demonstrates the first 10 time points selected by each method, ordered by their corresponding selection criteria. Following the time-points selection, the wrapper algorithm, which was described in 3.3 was employed to select the ensemble of time points with the highest prediction power. The time points selected by the wrapper, are marked in gray. The results indicate that the optimal time points found by the wrapper were found earlier in the prediction approach than in the ERP approach. However, the performance is similar since points which are not useful are later eliminated due to a comprehensive search of the wrapper.

Two methods for selection of the ensemble time points were compared: sequential selection and a wrapper algorithm. Figure 9 depicts the prediction error for increasing number of time points using the different approaches for time points selection and time point assembling; The results suggest that the contribution of the wrapper to eliminate time-points, which do not contribute to ensemble performance, is significant ($P < 0.05$).

Table 1. The first 10 time points found by the ERP and the predictive approaches. The time points which were selected by the wrapper are marked in gray.

ERP points (ms)	Prediction points(ms)
200	200
380	332
456	352
236	304
328	144
308	456
524	160
140	496
184	568
500	636

4.3 The Impact of Regularization

The amount of data needed to properly describe the different mental states increases exponentially with the dimensionality of the feature vectors. As the amount of training data is small compared to the size of the feature vectors, the classifier is likely to overfit to the training data and thus producing a model which does not uncover the true brain state discrimination. The only way to avoid this and still get a reliable brain state interpretation is a robust training with a regularizer which has to be carefully picked. To demonstrate the effect of non optimal regularization selection, we applied the same algorithm, with and without a regularization parameter on feature vector of size 96 (three best time points). As it can be seen in Fig. 10, the classification error with regularization is significantly lower ($P < 0.05$).

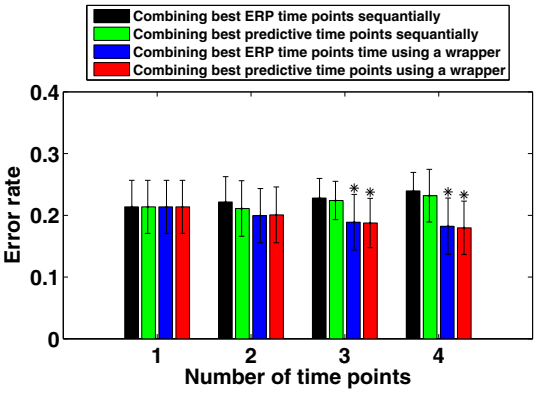


Fig. 9. Classification error rate vs. number of time points for different time-points selection and time-point assembling. It is evident that the contribution of the wrapper is significant ($P < 0.05$).

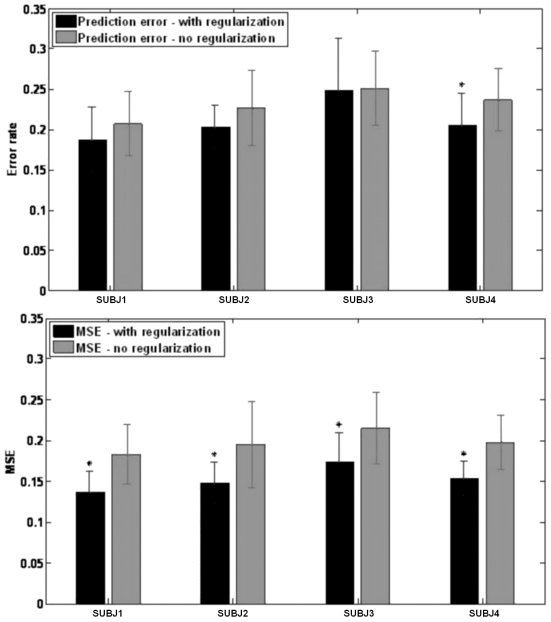


Fig. 10. Prediction error (Error rate and MSE) received for 3 best time points chosen, with and without regularization. These figures show that results with regularization are significantly better ($^*P < 0.05$).

5 Conclusions

This paper proposed a robust and efficient framework for brain state interpretation using EEG single trials. This framework is based on extensive feature selection using a regularized logistic regression classifier and can be used for spatial-temporal analysis of the

EEG data as well as other 2 or 3 dimensional brain imaging modalities. The proposed framework was compared to time-points selection via ERP method and to the sequential fusion of time-points. The results indicate that the wrapper approach offers a significant improvement over sequential selection. This spatial-temporal analysis can render a far more holistic interpretability without any a priori information on certain optimal time points or electrode locations. It can thus indicate whether the coding related to the brain state discrimination task is spatial, temporal or joint, and can indicate the network of information propagation (at high temporal resolution) following the stimuli. This method, which can also be applied to a Time/Frequency representation of the signal, can reveal the different frequency bands at which brain state discrimination is optimal.

Acknowledgements. This research was supported by the Israel Science Foundation Converging Technologies Program.

References

1. Wolpaw, J., Birbaumer, N., McFarland, D., Pfurtscheller, G., Vaughan, T.: Brain-Computer Interfaces for Communication and Control. Clin. Neurophy. 113, 767–791 (2002)
2. Allison, B., Wolpaw, E., Wolpaw, J.: Brain Computer Interface Systems: Progress and Prospects. Expert Review of Medical Devices 4, 463–474 (2007)
3. Dornhege, G., del R. Millan, J., Hinterberger, T., McFarland, D., Muller, K.R.: Towards Brain-Computer Interfacing. MIT Press (2007)
4. Blankertz, B., Dornhege, G., Krauledat, M., Muller, K.R., Curio, G.: The noninvasive Berlin brain-computer interface: fast acquisition of effective performance in untrained subjects. NeuroImage 37, 539–550 (2007)
5. Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., Arnaldi, B.: A Review of Classification Algorithms for Eeg-Based Brain-Computer Interfaces. J. Neural Eng. 4, R1–R13 (2007)
6. Bellman, R.: Adaptive Control Processes. Princeton University Press, Princeton (1961)
7. Muller, T., Ball, T., Kristeva-Feige, R., Mergner, T., Timmer, J.: Selecting Relevant Electrode Positions for Classification Tasks Based on the Electro-Encephalogram. Med. Biol. Eng. Comput. 38, 62–67 (2000)
8. Palaniappan, R., Raveendran, P., Omatu, S.: VEP Optimal Channel Selection Using Genetic Algorithm for Neural Network Classification of Alcoholics. IEEE Trans. Neural Netw. 13, 486–491 (2002)
9. Schroder, M., Bogdan, M., Rosenstiel, W., Hinterberger, T., Birbaumer, N.: Automated EEG Feature Selection for Brain Computer Interfaces. In: Proc. 1st Intern. IEEE-EMBS Conf. on Neural Eng., Capri Island, Italy (2003)
10. Lal, T., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., Scholkopf, B.: Support Vector Channel Selection in BCI. IEEE Trans. Biomed. Eng. 51, 1003–1010 (2004)
11. Tomioka, R., Muller, K.R.: A regularized discriminative framework for EEG analysis with application to brain-computer interface. Neuroimage 49, 415–432 (2010)
12. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. 22, 4–37 (2000)
13. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Recognition, 2nd edn. Wiley-Interscience, New York (2001)
14. Christoforou, C., Sajda, P., Parra, L.C.: Second order bilinear discriminant analysis for single trial EEG analysis. Advances in Neural Info. Proc. Sys. 20, 313–320 (2008)

15. Tomioka, R., Aihara, K., Muller, K.R.: Logistic regression for single trial eeg classification. In: Scholkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, vol. 19, pp. 1377–1384 (2007)
16. Zhdanov, A., Hendler, T., Ungerleider, L., Intrator, N.: Inferring Functional Brain States Using Temporal Evolution of Regularized Classifiers. Comput. Intell. Neurosci., Article ID. 52609 (2007)
17. Murray, M., Brunet, M., Brunet, D., Michel, C.: Topographic ERP analyses: step-by-step tutorial review. Brain Topography 20, 249–269 (2008)
18. Lehmann, D., Skrandies, W.: Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroenceph. Clin. Neurophysiol. 48, 609–621 (1980)
19. Lehmann, D., Ozaki, H., Pal, I.: EEG alpha map series: brain microstates by space-oriented adaptive segmentation. Electroenceph. Clin. Neurophysiol. 67, 271–288 (1987)
20. Hasson-Meir, Y., Zhdanov, A., Hendler, T., Intrator, N.: Inference of Brain Mental States from Spatio-Temporal Analysis of EEG Single Trials. In: Proc. of Biosignals (2011)
21. Ekman, P., Friesen, W.: Pictures of Facial Affect. Consulting Psychologists Press, Palo Alto (1976)
22. Lundqvist, D., Flykt, A., Ohman, A.: The Karolinska Directed Emotional Faces (KDEF). Dept. of Neurosci., Karolinska Hospital, Stockholm, UK (1998)
23. Sadeh, B., Zhdanov, A., Podlipsky, I., Hendler, T., Yovel, G.: The Validity of the Face-Selective ERP N170 Component During Simultaneous Recording with Functional MRI. Neuroimage 42, 778–786 (2008)
24. Ben-Simon, E., Podlipsky, I., Arieli, A., Zhdanov, A., Hendler, T.: Never Resting Brain: Simultaneous Representation of Two Alpha Related Processes in Humans. PloS ONE 3, e3984 (2008)
25. Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A.: Electroencephalographic signatures of attentional and cogntive default modes in spontaneous brain activity fluctuations at rest. Proc. of the National Academy of Sciences, U.S.A. 100, 11053–11058 (2003)
26. Delorme, A., Makeig, S., Sejnowski, T.: Automatic Artifact Rejection for EEG Data Using High-Order Statistics and Independent Component Analysis. In: Proc. of the 3th Intern. ICA Conf. (2001)
27. Geman, S., Bienenstock, E.: Neural Networks and The Bias / Variance Dilemma. Neural Computation 4, 1–58 (1992)
28. Hosmer, D., Lemeshow, S.: Applied Logistic Regression, pp. 118–124. John Wiley, New York (1989)
29. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer Series in Statistics (2001)
30. Detre, G., Polyn, S., Moore, C., Natu, V., Singer, B., Cohen, J., Haxby, J., Norman, K.: The Multi-Voxel Pattern Analysis (MVPA) Toolbox. In: The Annual Meeting of the Organization for Human Brain Mapping, Florence, Italy (2006), `http://www.csbmb.princeton.edu/mvpa`
31. Minka, T.: A Comparison of Numerical Optimizers for Logistic Regression. technical report, Dept. of Statistics, Carnegie Mellon Univ. (2003)
32. `http://www.cs.wisc.edu/condor/manual/v7.0.1`
33. Kohavi, R., John, G.: Wrappers for Feature Subset Selection. Artificial Intelligence 97, 273–324 (1997)
34. Bentin, S., Allison, T., Puce, A., Perez, E., McCarthy, G.: Electrophysiological Studies of Faces Perception in Humans. J. Cognitive Neurosci. 8, 551–565 (1996)

Time Series Clustering Algorithm for Two-Modes Cyclic Biosignals

Neuza Nunes[1], Tiago Araújo[1,2], and Hugo Gamboa[1,2]

[1] Physics Department, FCT-UNL, 2829-516 Caparica, Portugal
[2] PLUX – Wireless Biosignals, Av. 5 de Outubro, nr. 70, 6th floor, 1050-059 Lisbon, Portugal
{neuzanunes,taraujo87,hgamboa}@gmail.com

Abstract. In this study, an automatic algorithm which computes a *meanwave* is introduced. The *meanwave* is produced by averaging all cycles of a cyclic signal, sample by sample. With that information, the signal's morphology is captured and the similarity among its cycles is measured. A k-means clustering procedure is used to distinguish different modes in a cyclic signal, using the distance metric computed with the *meanwave* information. The algorithm produced is signal-independent, and therefore can be applied to any cyclic signal with no major changes in the fundamental frequency. To test the effectiveness of the proposed method, we've acquired several biosignals in context tasks performed by the subjects with two distinct modes in each. The algorithm successfully separates the two modes with 99.3% of efficiency. The fact that this approach doesn't require any prior information and its preliminary good performance makes it a powerful tool for biosignals analysis and classification.

Keywords: Biosignals, Waves, *Meanwave*, k-Means, Clustering, Algorithm, Signal-processing.

1 Introduction

Biomedical signal analysis is a method of the utmost importance for visualization and interpretation in biology and medicine, as the manipulation and processing of data provide the researcher with vital information on the condition of the subject or the status of the experiment. An awareness of the power of computer-aided techniques, coupled with a continuing need to obtain more information from biomedical signals, has led to a growing application of signal processing techniques to be applied in medicine, sports and research [1].

Human-activity tracking techniques focus on direct observation of people and their behavior. This could be done, as an example, with cameras [2], accelerometers to track human motion [3], or contact switches to compute facial expressions with the electromyography patterns [4,5]. When watching a biosignal's evolution it's important to understand the changes in its morphology over time. Abrupt changes often contain critically important information, and hence, the problem of discovering time points where changes occur has received much attention on statistics and data mining [6].

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 233–245, 2013.
© Springer-Verlag Berlin Heidelberg 2013

In this work, we've acquired several cyclic biosignals – such as accelerometry (ACC), electromyography (EMG) and blood volume pressure (BVP) signals – from subjects performing context tasks with two different modes. It was developed an algorithm capable to distinguish the two different modes in the same continuous signal. The developed algorithm follows an unsupervised learning approach, as it doesn't require any prior information [7], and a k-means algorithm is used to cluster the signals due to its efficiency and effectiveness [8]. As a clustering method, our algorithm is signal-independent as it doesn't use specific information about the input signals. Although our algorithm is signal-independent, the signals used must be cyclic, and in this study we only used signals with two different modes and a small variation of fundamental frequency between those modes.

Warren Liao [9] presents a survey on time series data clustering, exposing past researches on the subject. He organizes the studies in three groups: whether they work directly with the raw data, indirectly with features extracted or indirectly with models built from the raw data. In this study we work directly with the data, but in our approach we use a continuous signal with different modes in it - the previous studies use mostly various separated signals with only one mode or activity in each. A more resemble approach, as the clustering is based on the similarity of wave shapes in a single time series data, is the work of Dr. Rodrigo Quiroga [10] with spike sorting. However, as the neuron activity is not periodic, the spikes are detected with a threshold method and the clustering procedure uses features extracted from those parts of the signal.

In this paper we present the concept of "*meanwave*". Having a set of signal's cycles, the *meanwave* is a wave constructed by the mean value for each time-sample of those cycles. Following the same principle, we define a *deviation wave* by computing the standard deviation error value for each time-sample instead of the mean value. Our algorithm separates each signal's cycles and computes a *meanwave*, with center on a chosen notable point of each cycle. With that information the signal's morphology is captured. Our clustering algorithm uses a distance metric, gathered from the information of the signal's cycles, to separate the two modes of the entire signal. As our *meanwave* approach effectively captures the morphology of a signal, it can be useful in several areas, as a clustering basis or just for signal analysis.

In the following section the signal acquisition methodology is presented. In section 3 we expose the signal processing detailing all algorithms steps. Finally in sections 4 and 5 we detail and discuss our results and algorithm performance, concluding the work.

2 Methods

2.1 Acquisition System and Sensors

To acquire the biosignals necessary for this study, we used a surface EMG sensor (*emgPLUX*), a triaxial accelerometer (*xyzPLUX*), and a finger BVP sensor (b*vpPLUX*). A wireless signal acquisition system (bioPLUX research) was used for the signal's analog to digital conversion and bluetooth transmission to the computer.

This system has 12 bit ADC and a sampling frequency of 1000 Hz [11]. In the acquisitions with the triaxial accelerometers, only the axis with inferior-superior direction was connected to the bioPLUX.

2.2 Data Acquisition and Data Format

Several tasks were designed and executed in order to acquire signals with two distinct modes. We conceived a synthetic digital signal and collected signals from four different activities scenarios with the accelerometer sensor, and one for each EMG and BVP sensors.

Synthetic Signal

To test our algorithm, a synthetic cycle was created using a low-pass filtered random walk (of 100 samples), with a moving average smoothing window of 10% of signal's length, and multiplying it by a *hanning* window. That cycle was repeated 30 times for the first mode, so all the cycles were identical. After a small break on the signal, the cycle was repeated 20 more times, but with an identical small change of 40 samples in all waves, creating a second mode. These two modes construct the synthetic wave represented in Figure 1.

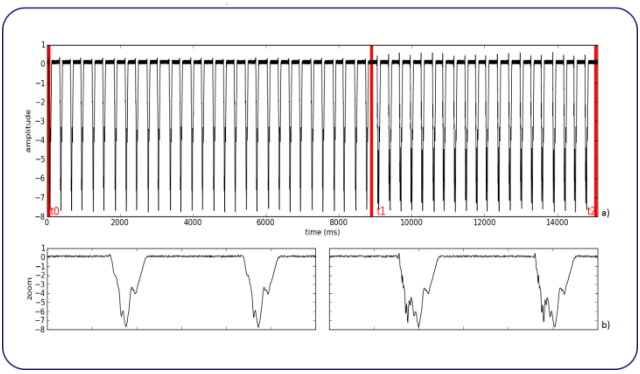


Fig. 1. a): Synthetic signal with identical waves from *t0* to *t1* and from *t1* to *t2*; b): corresponding zoomed waves

Walking and Running

With an accelerometer located at the right hip and oriented so the y axis of the accelerometer (the only connected to the bioPLUX) was pointing upward, the subjects performed a task of walking and running non-stop, on a large circle drawn on the floor. The subjects walked for about 1 minute at a slow speed, then spent 1 minute running, and ended with 1 minute walking again. The signal acquired is represented in figure 2.

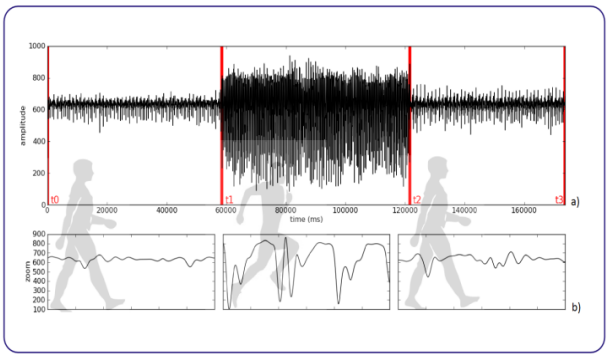


Fig. 2. a): Acceleration signal of a subject walking (*t0* to *t1* and *t2* to *t3*) and running (*t1* to *t2*); b): corresponding zoomed waves

Running and Jumping

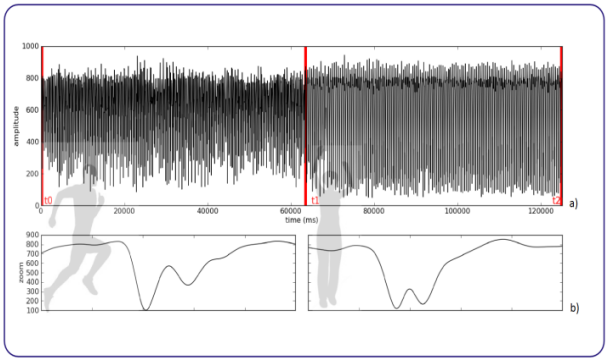


Fig. 3. a): Acceleration signal of a subject running (*t0* to *t1*) and jumping (*t1* to *t2*); b): corresponding zoomed waves

With an accelerometer located at the right hip and oriented so the y axis of the accelerometer was pointing upward, the subjects performed a task of running non-stop, on a large circle drawn on the floor, and jumping also continuously but at the same place. The subjects spent 1 minute running, followed by 1 minute jumping. The signal acquired is represented in figure 3.

Jumping with and without Impulsion

In this task, the following procedure was executed: 14 seconds of "normal" jumping (small jumps without a big impulsion), 24 seconds of jumping with some boost and 7 seconds of normal jumping again. The subjects used an accelerometer located at the right hip and oriented so the y axis of the accelerometer was pointing upward. The signal acquired is represented in figure 4.

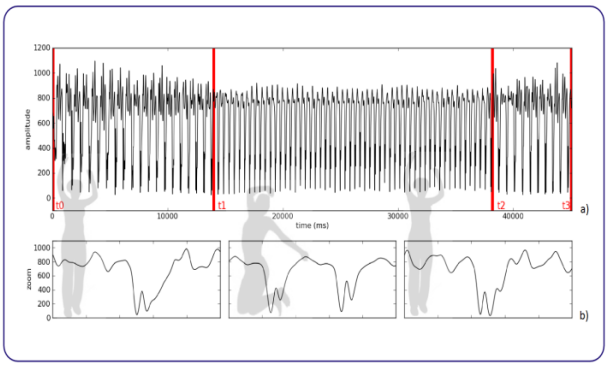


Fig. 4. a): Acceleration signal of a subject performing normal jumps (*t0* to *t1* and *t2* to *t3*) and jumps with boost (*t1* to *t2*); b): corresponding zoomed waves

Skiing with V1 and V2 Technique

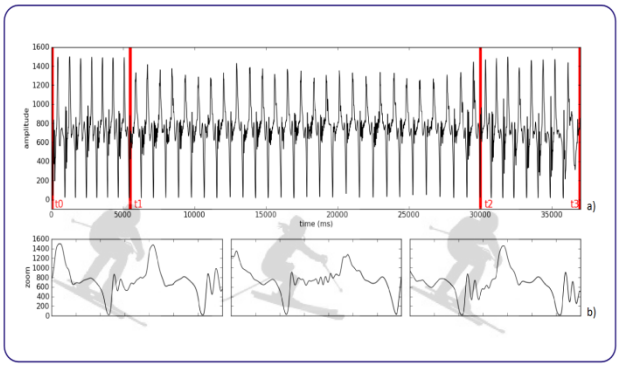


Fig. 5. a): Acceleration signal of a subject skiing with a technique called V2 (*t0* to *t1* and *t2* to *t3*) and skiing with a technique called V1 (*t1* to *t2*); b): corresponding zoomed waves

This acquisition was made during a cross-country skiing study, in which the subject had an accelerometer attached to the ski pole, bellow the handgrip. Figure 5 illustrates the acceleration signal of that accelerometer.[1]

In the 37 seconds of the signal the subject performed two different techniques, called V1 and V2. V1 skate is an asymmetrical uphill technique involving one poling action over every second leg stroke. V2 skate is used for moderate uphill slopes and on level terrain, involving one poling action for each leg stroke [12]. The first 7 cycles of the signal (about 5 seconds) were produced through a V2 technique, the next 27 cycles (about 25 seconds) used a V1 technique and the final 8 cycles the technique was V2 again.

[1] This signal was acquired by Håvard Myklebust and Jostein Hallén, researchers at NIH, Norway [13].

Elevation and Squat of the Legs

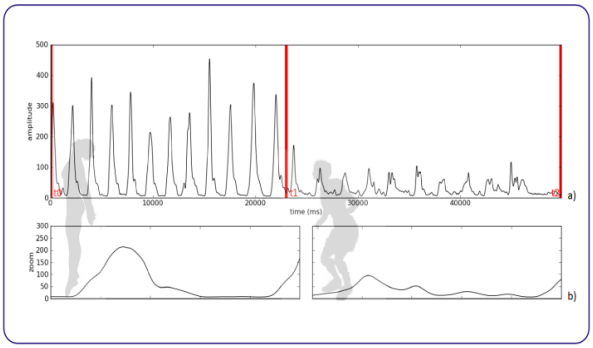


Fig. 6. a): EMG signal of the gastrocnemius muscle's contraction through the elevation (*t0* to *t1*) and squat (*t1* to *t2*) of the inferior members; b): corresponding zoomed waves

In this task, the subject was standing straight with both feet completely on the ground and was asked to perform 12 elevations of the legs - getting on the tiptoes and back with both feet completely on the ground - followed by 11 squats - bending the knees and back standing straight - (Figure 6). The EMG data were collected using bipolar electrodes at the *gastrocnemius* muscles of the right leg.

At Rest and After Exercise BVP

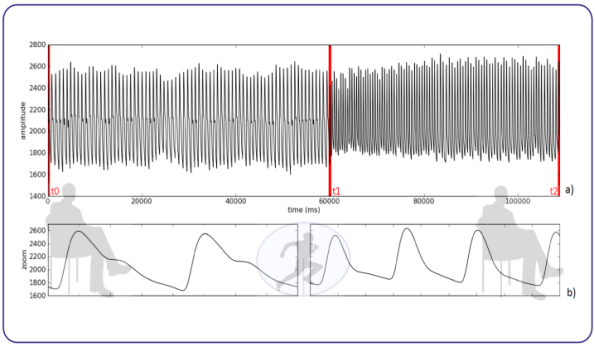


Fig. 7. a): BVP signal with the subject at rest (*t0* to *t1*) and after exercise (*t1* to *t2*); b): corresponding zoomed waves

The subjects were instrumented with a BVP sensor on the fourth finger of the left hand and were sitting with the left forearm resting on a platform. One acquisition was made with the subjects at rest and then the subjects were asked to perform intensive exercise which was not collected to avoid undesirable movement artifacts. Immediately after the exercise, another BVP signal was acquired with the subject sitting again but tired. For the purpose of this study, both signals (at rest and after exercise) were used in the same file, cutting a part of each signal and concatenating them offline. The resulting signal is represented in figure 7.

All the signals referenced above are available at OpenSignals website [14].

3 Signal Processing

The collected data was processed offline using Python with the numpy [15] and scipy [16] packages.

Signal processing algorithms were developed for an automatic detection of a *meanwave* representative of the signal's behavior and the k-means algorithm was used to cluster the signals. The main idea of this algorithm is to define a loop with k centroids far away from each other, taking each point belonging to a given dataset and associate it to the nearest centroid. Repeating the loop, the centroids position will change because they are re-calculated as barycenters of the clustering result, and after several iterations the position will stabilize, achieving the final clusters [17].

Figure 8 describes the method used to process the signals. All biosignals were submitted to a signal-specific pre-processing phase and then to a generic signal-independent phase (composed with an automatic *meanwave* and clustering procedure) which was applied to all the signals of this study.

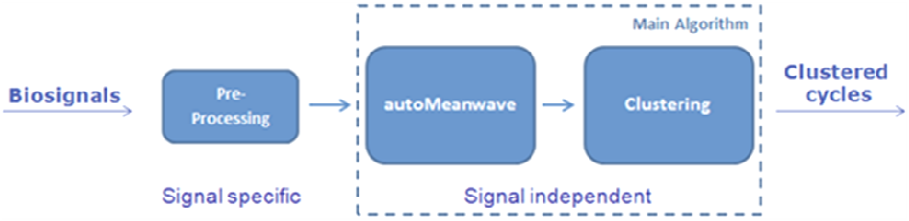


Fig. 8. Signal's processing schematics

For the pre-processing phase, the acceleration and BVP signals were low-pass filtered using a smoothing filter with a moving average window of 50 points. Random noise with 1/5 of original amplitude was added to the synthetic signal. The EMG signal was centered at y axis zero, by subtracting its mean value, and then rectified. Then we applied the smoothing filter with a moving average window of 300 points, to get the EMG envelope signal.

In the next sections the generic processing procedure will be described.

3.1 Automatic *Meanwave*

`Algorithm 1: autoMeanwave`

```
Input:      signal, sampling frequency, trigger mode.
Output:     fundamental frequency, window size, events,
            meanwave, deviation wave.
```

The `autoMeanwave` algorithm is the base function to identify the individual waves. After running this algorithm we will have the *meanwave* computed with the individual wave's information. This algorithm receives, by input, a signal, its sampling frequency and a trigger mode - this one can be omitted and the algorithm will use the maximum point as default.

As we're working with cyclic signals, the first step of the automatic *meanwave* algorithm is the detection of the signal's fundamental frequency. For that we create the `fundamentalFrequency` algorithm (algorithm 2). With the result we compute the window size value and randomly selected a part of the signal with the same number of samples as the window size. With that signal's part and the original signal itself, we run the `sumvolve` algorithm (algorithm 3) to get the signal events, which are a series of points that we consider as the center of each cycle. After this we have all the information necessary to compute the *meanwave*, which we do in the `computeMeanwave` algorithm (algorithm 4).

Following we will describe minutely the sub-algorithms referenced above.

Algorithm 2: fundamentalFrequency
Input: signal, sampling frequency.
Output: fundamental frequency.

There are many ways of computing the fundamental frequency (f_0) and there isn't any ultimate method to estimate it as a procedure that returns good results for one type of signal can perform poorly for others [18].

For the purpose of this study, the estimation of the f_0 was based on the extraction of the first signal's harmonic. So, the first step of this function was to smooth the result of the signal's fast fourier transform with a moving average window of 5% of the signal's length. We assumed the frequency value of the first big peak located at the smoothed FFT signal as the f_0 of the original signal.

With the fundamental frequency value we could compute the sampling size of a signal's cycle. We call that value "window size", with a 20% margin:

$$\text{winsize} = (f_s/f_0)*1.2 \tag{1}$$

Being f_S the sampling frequency and f_0 the fundamental frequency. We open the window 20% to use more samples than a cycle.

Although there are more robust methods to determine the fundamental frequency of a signal, this approach is adequate for our study, as the purpose was to have a close idea of the size of a cycle. We actually use more than one exact cycle as we open the window calculated with the fundamental frequency by 20%. Notice that further on we use a correlation function to detect meaningful events on a cycle, so the fundamental frequency is just used as a preliminary estimation to support other algorithms.

Algorithm 3: sumvolve
Input: two signals.
Output: distance values.

This algorithm works as a correlation function. Sliding the smaller window part of the signal (given by argument) through the original signal, one sample at a time, this algorithm compares the distance of the two overlapped waves. The expression used to compute the distance value for each sample is exposed in equation (2).

$$distance_i = \frac{\sum_{j=1}^{N} \left| sig_{[i:i+N]j} - window_j \right|}{N} \tag{2}$$

With *i* ranging from 1 to number of signal's samples minus the window size. The result of this algorithm is a signal composed with distance values. That distance signal shows the difference between each overlapped cycle and the window selected at the first place. After that, all the minimum peaks of the resulted correlation signal were found. Those peaks were assumed as the cycle's events.

Algorithm 4: computeMeanwave
Input: signal, events, window size.
Output: *meanwave* and *deviation wave*.

With the events and the window size, the signals can be separated into periods that are assumed as the signal's cycles:

$$\text{cycle} = \text{signal}[\text{event} - \text{winsize}/2 : \text{event} + \text{winsize}/2] \quad (3)$$

This way, based on all cycles, we could compute the mean value to each cycle sample, and compose a *meanwave*. The *deviation wave* is computed with the same principle, calculating the standard deviation error instead of the mean value. For a better visualization of the results, we computed a deviation area with the *deviation wave* obtained. For that, we added and subtracted the *deviation wave* to the *meanwave*, getting an upper and lower wave, to graphically represent the deviation area. This is shown in the results section.

After the procedure described above, a final adjustment was made: the rearrangement of the cycle's events positions. A trigger position is computed to rearrange the cycle's events. The trigger position is a notable point of the previously traced *meanwave*. The chosen trigger mode for this study was the *meanwave*'s minimum point; other possibilities were designed – we could use the maximum (of the signal or the derivative signal), or the zero crossings, for example.

With this trigger position we recalculate the peak events, or cutting points, used in the `computeMeanwave` algorithm:

$$\text{events} = \text{events} + \text{trigger} - \text{winsize}/2 \quad (4)$$

With the events variable recalculated, we used the `computeMeanWave` function again, so the cycles had the center on a notable point visually more recognizable in the *meanwave*.

3.2 Clustering

Algorithm 5: distanceMatrix
Input: signal, events, window size.
Output: *matrix with wave-to-wave distances*.

For the clustering procedure we developed a function that receives the signal to cluster, the window size and cutting events produced with the `autoMeanwave` algorithm.

We go through all the cutting events, selecting for each a part of the signal with center at that event and a number of samples to both sizes equal to the window size. Then, that cut is compared with each of the others (with the center in the other cutting events and the same window size), using the mean square error distance formula:

$$distance = \sqrt{\frac{\sum_{i=1}^{l}(cycle1_i - cycle2_i)^2}{l}} \tag{5}$$

With **cycle$_1$** and **cycle$_2$** being the parts of the signal selected before.

With the similarity values for each wave, we built a matrix of distances. Figure 9 presents two matrices of distances, obtained with the *imshow* command. Figure 9 a) shows the matrix of the synthetic waves distances and figure 9 b) the matrix for the skiing task distances. As we can see, the synthetic matrix is almost ideal, as all the waves of each mode are equal – the distance between waves is a minimum or a maximum. In the skiing matrix, there's a greater variation of distances, as the cycles are not exactly the same. However, it's visible the similarity between the cycles of the same technique (7 cycles V2, 27 cycles V1 and 8 cycles V2).

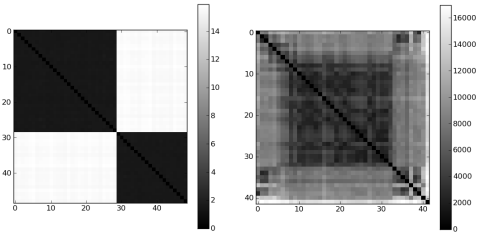


Fig. 9. Matrix of distances produced for the synthetic waves (a)) and the skiing task (b))

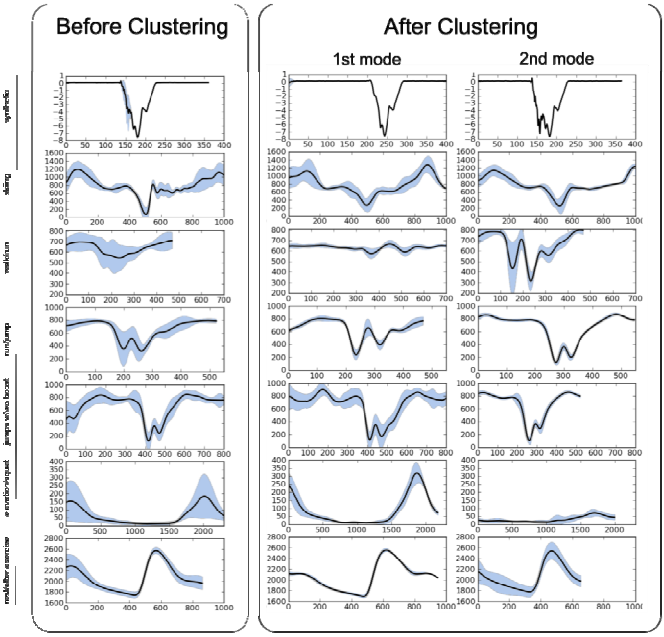


Fig. 10. Resulting *meanwave*s for all tasks, before and after the clustering procedure

To cluster the signal we used a k-means method. This algorithm received the matrices created with the `distanceMatrix` algorithm and the number of clusters expected in the data, returning the clusters and distances to the clusters.

4 Results and Discussion

Figure 10 shows the graphics of the resulting *meanwave*s (line) and deviation area (filling) after running the algorithms referenced above. At the left the graphics represent the initial *meanwave*s traced, before the clustering procedure. At the right we have the *meanwaves* representative of the signal parts that were divided according to the resultant clustering codebook. It's visible that the *meanwave*s at the left gather information about the signal's behavior, even if there are some changes in its shape or frequency. After the clustering procedure there are some predictable variations in the resultant *meanwave*s. We notice an overall reduction of deviation area after the clustering procedure and also a reshaping of the *meanwave*.

After running the clustering procedure, we gathered the results for each task performed. These results are exposed in table 1.

It is important to note that some cycles weren't classified, and that occurred because sometimes the borders of the signal didn't have a full cycle - the `distanceMatrix` algorithm (algorithm 5) cannot be used to compare a short cycle with the regular ones. Therefore, those cycles have been rejected for lack of pattern quality, and won't be taken into account.

Table 1. Clustering Results

Task	*Number of Cycles*	*Cycles correctly clustered*	*Errors*	*Misses*
Synthetic	50	49	0	1
Walk and run	343	342	1	0
Run and jump	296	295	1	0
Jumps	85	84	1	0
Skiing	42	41	0	1
Elevation and squat	23	23	0	0
BVP rest and after exercise	165	159	4	2
All	1004	993	7	4

In the "walk and run" activity there were some extra classification points. The cycles were correctly clustered (with only 1 error encountered), but in the "walking" mode there were some extra points between those cycles that were also counted. That occurs due to a relatively large variation in the fundamental frequency from the walking to the running activity - despite one activity has all cycles well defined by the window size variable, the other has less than one cycle per window size. This condition shows a limitation of our algorithm: it doesn't allow big changes in the frequency domain for the different modes presented on the signal.

Only 7 errors resulted from this algorithm, and 2 of those errors were in transition periods – where the cycle wave is still reshaping to form the other activity and the distance value to the *meanwave* or to the clusters mean values is bigger than anywhere else on the signal. This occurred in the jumps and in the walk and run activities.

Given the results we can affirm that our clustering algorithm based on the *meanwave* information only returned 7 errors out of 1000 cycles with pattern quality, and therefore we achieved 99.3% of efficiency.

5 Conclusions

In this paper we presented a new algorithm for biosignal's processing and time series clustering. Physiological and activity data were acquired and a synthetic signal was produced to test the effectiveness of the created algorithm. The cyclic data acquired had two different modes, distinguished by the cycle's wave-shapes, disposed continuously along the signal. Our algorithm proves to be a good method to cluster single signals with different modes in it, as it doesn't need prior information about the signal's behavior or to previously learn the wave-shapes of the different modes.

The proposed algorithm represents an advance in the biosignals processing field, as it has an effective detection of signal variations, tracing different patterns for distinct clusters, whether it's an activity, synthetic or physiological signal, without the need of the user's intervention to previously analyze the signals.

6 Future Work

In the continuity of our research, we are now working with models built from the raw data. A method has been widely referenced in literature to cluster time series data: clustering using the distances between the data's cepstral coefficients. We're using this approach along with the one presented in this paper to cluster acquired data and compare the efficiency of both algorithms.

In future works, we intend to repeat this procedure to a wide range of subjects doing the same task, perform a noise immunity test and also study the response of this algorithm to signals with more than two modes. A preliminary test with a few signals with three and four modes has already been done, and the results achieved were highly satisfactory. It's also our intention to automatically perceive the number of clusters present in the signals and input that estimation into the k-means algorithm.

We intend to introduce an automatically perception of the cycles which are too distant from the cluster and assign those cycles to a new "rejection class". This will reduce the number of errors due to a strange cycle, in particular the mode's transition cycles.

The local detection of the fundamental frequency is also a future goal, as we intend to realize when there's a major variation of fundamental frequency and make our algorithms adapt its behavior according to that variation.

Finally, we have the intention to create a multimodal algorithm, which can receive and process more than one signal at the same time and with the same treatment. This could be useful, for example, if we want to use the three axis of a triaxial accelerometer, a BVP with an electrocardiography (ECG) signal, or to conciliate the information of synchronous acquired signals.

Acknowledgements. The authors would like to thank PLUX – Wireless Biosignals for providing the acquisition system and sensors necessary to this investigation. We also like to thank NIH, the Norwegian School of Sports and Science, in the person of

Håvard Myklebust and Jostein Hallén, for acquiring and allowing us to work with the Skiing signal used in this study. We acknowledge Rui Martins and José Medeiros for their help and advices on the BVP acquisition procedure.

References

1. Theis, F., Meyer-Base, A.: Biomedical Signal Analysis: Contemporary Methods and Applications. The MIT Press (2010)
2. Ben-Arie, J., Wang, Z., Pandit, P., Rajaram, S.: Human Activity Recognition Using Multidimensional Indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(8), 1091–1104 (2002)
3. Baek, J., Lee, G., Park, W., Yun, B.: Accelerometer Signal Processing for User Activity Detection. Knowledge-Based Intelligent Information & Engineering Systems 3215, 610–617 (2004)
4. Smith, J., Fishkin, K., Jiang, B., Mamishev, A., Philipose, M., Rea, A., Roy, S., Sundara-Rajan, K.: RFID-Based Techniques for Human-Activity Detection. Communications of the ACM 48(9), 39–44 (2005)
5. Fridlund, A., Schwartz, G., Fowler, S.: Pattern Recognition of Self-Reported Emotional State from Multiple-Site Facial EMG Activity During Affective Imagery. Society for Psychophysiological Research 21(6), 622–637 (2007)
6. Basseville, M., Nikiforov, I.: Detection of Abrupt Changes: Theory and Applications. Prentice-Hall Inc. (1993)
7. Ghahramani, Z.: Unsupervised learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) Machine Learning 2003. LNCS (LNAI), vol. 3176, pp. 72–112. Springer, Heidelberg (2004)
8. Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D., Steinberg, D.: Top 10 algorithms in data mining. Knowledge Information Systems 14(1), 1–37 (2007)
9. Liao, W.: Clustering of time series data – a survey. Pattern Recognition 38, 1857–1874 (2005)
10. Quiroga, R.: Spike sorting. Scholarpedia 2(12), 3583 (2007)
11. PLUX – Wireless Biosignals, bioPLUX Research Manual - internal report (2010)
12. Andersson, E., Supej, M., Sandbakk, Ø., Sperlich, B., Stöggl, T., Holmberg, H.: Analysis of sprint cross-country skiing using a differential global navigation satellite system. European Journal of Applied Physiology 110(3), 585–595 (2010)
13. Myklebust, H., Nunes, N., Hallén, J., Gamboa, H.: Morphological analysis of acceleration signals in cross-country skiing - information extraction and technique transitions detection. In: Procedings of the 4th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2011), Rome, Italy (2011)
14. OpenSignals, `http://www.opensignals.net`
15. Oliphant, T.: Guide to Numpy. Tregol Publishing (2006)
16. Oliphant, T.: SciPy Tutorial, `http://www.scipy.org/`
17. Martins, D., Mattos, M., Simões, P., Cechinel, C., Bettiol, J., Barbosa, A.: Aplicação do Algoritmo K-Means em Dados de Prevalência da Asma e Rinite em Escolares. In: XI Congresso Brasileiro de Informática em Saúde (2008)
18. Gerhard, D.: Pitch extraction and fundamental frequency: History and current techniques. Technical Report (2003)

Non-contact Pulse Wave Velocity Assessment Using Optical Methods

T. Pereira[1], M. Cabeleira[1], P. Matos[1], E. Borges[1], V. Almeida[1], H.C. Pereira[1,2], J. Cardoso[1], and C. Correia[1]

[1] Instrumentation Center, Physics Department, University of Coimbra, Coimbra, Portugal
[2] ISA- Intelligent Sensing Anywhere, Coimbra, Portugal
taniapereira10@gmail.com, correia@berta.fis.uc.pt

Abstract. The clinical relevance of pulse wave velocity (PWV), as an indicator of cardiac risk associated to arterial stiffness, has gained clinical relevance over the last years. Optic sensors are an attractive instrumental solution for this type of measurement due to their truly non-contact operation capability, which has the potential of an interference free measurement. The nature of the optically originated signals, however, poses new challenges to the designer, either at the probe design level as at the signal processing required to extract the timing information that yields PWV. In this work we describe the construction of two prototype optical probes and discuss their evaluation using three algorithms for pulse transit time (PTT) evaluation. Results, obtained in a dedicated test bench, that is also described, demonstrate the possibility of measuring pulse transit times as short as 1ms with less than 1% error.

Keywords: Optical probes, Photodiode, Waveform distension, Pulse transit time, Pulse wave velocity.

1 Introduction

Pulse wave velocity (PWV) is defined as the velocity at which the pressure waves, generated by the systolic contraction of the heart, propagate along the arterial tree. PWV is a measure of regional arterial stiffness of the arterial territory between the two measurement sites. This parameter is related to the elastic modulus (E) of the arterial wall (which represents the intrinsic wall stiffness), and the arterial geometry (thickness: h) and blood density (ρ). The first relationship was formulated by Moens and Korteweg and expresses:

$$PWV = \sqrt{\frac{Eh}{d\rho}} \tag{1}$$

Later on, Bramwell and Hill described (1) the association in terms of distensibility (D), which is determined by the blood vessel's compliance (C), the former relation can be expressed:

$$PWV = \frac{1}{\sqrt{\rho DC}} \tag{2}$$

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 246–257, 2013.
© Springer-Verlag Berlin Heidelberg 2013

From the expression, we can deduce that higher PWV corresponds to lower vessel distensibility and compliance and therefore to higher arterial stiffness.

The pulse waves travel through the arteries at a speed of 4 to 10 meters per second depending on the vessel (PWV increases with the distance from the heart), and the elastic condition of the arterial wall, which is affected by a variety of factors in health and disease [1].

The most common technique to assess non-invasively PWV is based on the acquisition of pulse waves generated by the systolic ejection at two distinct locations, separated by a distance *d*, and determination the time delay, or pulse transit time, due to the pulse wave propagation along the arterial tree [2]. The PWV parameter is then simply calculated as the linear ratio between *d* and the PTT.

Many different pulse waves have been used to assess pulse wave velocity, such as pressure wave, distension wave or flow wave. The gold standard in PWV assessment uses pressure waves measured by pressure sensors [3]. These sensors need to exert pressure in the blood vessel this will distort the waveform, and may lead to inaccurate measurements. Another drawback of this method is the fact that the predicted PWV is relative to a large extension of the arterial tree and therefore is the conjunction of different local PWVs.

Other studies describe ultrasonic probes that predict PWV using Doppler Effect and modified ecography probes [4], but the PWV measurements were unreliable.

Recently [5] [6], described alternative approaches for estimating carotid artery pressures with an ultrasound system. Calibrated diameter distension waveforms were compared to the more common approach based on pressure waves, proving to be a valid alternative to local pressure assessment at the carotid artery.

All the previous techniques are minimally invasive, but the probe has to be in contact with the patient's tissues at the artery site. This contact, as stated above, can distort the signal integrity and thus rise the interest in exploring true non-contact technique.

The propagation of pressure waves in arterial vessels generates distensions in the vessel's walls. These distensions can be optically measured in peripheral arteries like the carotid that, as they run very close to the surface impart a visible distension. This distension, as it modulates the reflection characteristics of the skin, can be used to generating an optical signal correlated with the passing pressure wave.

The probes developed in this work, gather the light generated by LED illumination and reflected by the skin, using two photodiodes placed 3 cm apart, all assemble in a single probe. PWV is assessed by measuring the time delay between the signals of the two photo-sensors using different algorithms that are also discussed.

2 Technologies

Two distinct types of silicon optical sensors – planar and avalanche photodiode (APD) – are used in this work, each one requiring a particular electronic circuitry. Results, however, are derived by the same signal processing algorithms.

Each probe incorporates two identical optical sensors placed 3 cm apart and signal conditioning electronics based on a transconductance amplifier and low-pass filter.

The APD probe includes the high voltage biasing circuitry (250V) necessary to guarantee the avalanche effect. Illumination is provided by local, high brightness, 635 nm light-emitting diodes (LEDs).

A photodiode (PD) is a type of photo-detector with the ability of converting light into either current or voltage, according to the *modus operandi*. One decided to use a planar, rectangular-shaped photodiode, its dimensions being 10.2x5.1mm. This is silicon solderable photodiode feature low cost, high reliability and a linear short circuit current over a wide range of illumination.

Analogously to the conventional photodiodes, APDs operate from the electron-hole pairs created by the absorption of incident photons. The high reverse bias voltage of APDs, however, originates a strong internal electric field, which accelerates the electrons through the silicon crystal lattice and produces secondary electrons by impact ionization. This avalanche effect is responsible for a gain factor up to several hundred.

APDs are operated with a relatively high reverse voltage and will typically require 200 to 300 volts of reverse bias. Under these conditions, gains of around 50 will result from the avalanche effect, providing a larger signal from small variations of light reflected from the skin and will, at least theoretically, improve the signal-to-noise ratio (SNR).

On the other hand, since the sensitive area of this sensor is very small (1 mm^2), the accuracy of the estimations increases. In fact, comparatively to the planar photodiode, in which the detection of light takes place over a much larger area, this sensor can measure an almost punctual section of the skin, thus decreasing the error associated to the detection solid angle.

The two prototype probes, on which we support this work, incorporate an APD from Advanced Photonics (SD 012-70-62-541) and a planar type from Silonex (SLCD-61N3) respectively.

3 Test Setup

The test setup was designed to assess the two main parameters of in PWV measurements: linearity and time resolution.

Their assessment was carried out in a test setup where illumination is provided by two LEDs whose light intensities reproduce the same signal with a variable time delay between them, as shown in Figure 1.

Two arbitrary waveform generators, *Agilent 33220A* (AWG1 and AWG2), are synchronously triggered by an external signal. The waveform generators have been previously loaded with the same typical cardiac waveforms and the mutual delay is selected in order to simulate different pulse transit times (Figure 2). These signals must be added to a small offset of the order of the magnitude of the forward voltage drop of the LED, so that the resulting light intensity is linearly modulated by the LED signal current. A 16-bit resolution data acquisition system (National Instruments, USB6210) samples the signals at a 20 kHz rate, adequate for PTTs as low as 100 μs and stores them for off-line analysis using Matlab™.

In the test setup, the probe is placed in front of a test device, see Figure 1, which holds the two modulated LEDs and provides light isolation to prevent crosstalk.

During the tests, the LEDs of the probe itself are deactivated and all light comes from the LEDs in the test device.

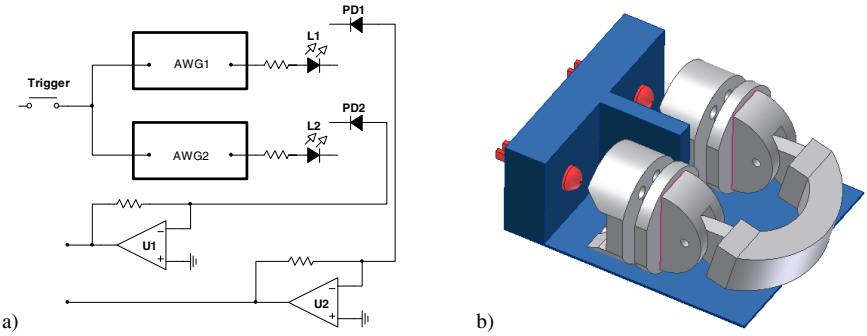


Fig. 1. Test setup; a) light modulation and detection circuit b) during test the probe is held to the blue part

Figure 2 shows a typical set of signals generated and detected by the circuit of Figure 1.

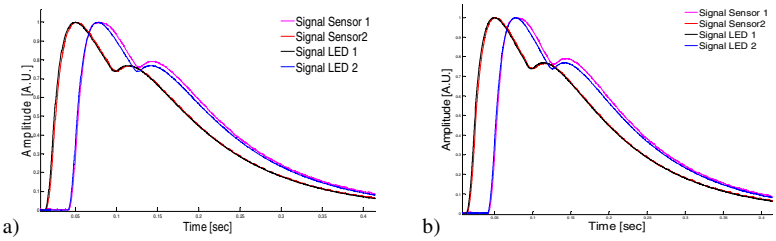


Fig. 2. Excitation and detector responses for a) Planar Photodiode b) Avalanche Photodiode

To assess the operational limits of our probes and algorithms, we designed three different tests. In the first one, signals with frequency similar to the normal heart rate but with delays within the interesting PTT range are fed to the system to investigate the integral linearity error. This test was performed at a constant frequency of 1.5 Hz and time delays varying from 1ms to 100ms, corresponding to PWVs in a 30m/s to 0.3m/s interval. This range of values includes the normal PTT range of values in humans.

In the second test we assess the robustness of the algorithms to noise. To do this, we add white noise of amplitudes ranging from 1% to 50% of the signal amplitude in 0.02% steps, to the isolated pair of pulses. For each noise level, 1000 samples produced in order to obtain reasonable statistics. The resulting PTT distribution is then studied.

The third test was intended to validate our algorithm's operability under a wide range of frequencies (simulating different Heart rates) with a time lag far greater than the maximum PTT seen in humans. It consisted of varying the output frequency (1 Hz to 200 Hz) of the cardiac pulses keeping the time lag between the two signals at 1.1ms.

4 Signal Processing

Three different algorithms for extracting the time delay from the detector's signals are considered. They are referred to as foot-to-foot, cross-correlation and phase spectra. Their basis derives from the homonymous mathematical functions.

The accuracy of the results delivered by the algorithms discussed in this section is compared with the reference delay selected at the waveform generators.

In the foot-to-foot method and in spite of more complex methods [7], a simple detection of the time lag between the start of the upstroke of the two consecutive pulses is carried out. This is possible due to the well behaved nature and low noise levels of our signals. A different situation occurs in signals collected from a patient, mostly due to baseline drift.

The cross-correlation method is based on the well known property of the peak of cross-correlogram that allows delays to be calculated by subtracting the peak time position from the pulse length [8]. Two different correlation functions are used: one that belongs to the Matlab™ core (*Xcorr*) and another one that generates the cross-correlation making direct use of the cross-correlation theorem (*Fcorr*).

The third method uses data in the phase spectra of the signals. In this method, we first identify the exact frequency of the signal's harmonics, using the amplitude spectra, and then, extract the corresponding phase angles from the phase spectra.

The phase angle, θ, is related with angular frequency of the phase spectrum, ω and with the time delay, t, according to:

$$\theta = \omega \cdot t \,. \tag{3}$$

On its turn, the time delay is computed from the phase angles of the same harmonic in the phase spectra of each signal, θ_1 and θ_2:

$$t = \frac{(\theta_1 - \theta_2)}{\omega} \tag{4}$$

Despite the fact that, theoretically, the time delay can be determined at any harmonic of the complete spectrum, the practice, however, differs, given their affectation by noise. Nevertheless, by performing the filtering at the detector amplifier level, one is able to obtain a lower error, as long as the best harmonics (that is, with the highest SNR possible) are selected. For the circuits used in this study, one checked best performances when the time delay was computed at the 2nd harmonic in the APD case and in the 4th one for the PP circuits.

5 Results

This section is dedicated to the discussion of results obtained with the two probes using the previously mentioned algorithms.

5.1 Integral Linearity

By definition, integral linearity is the maximum deviation of the results from the reference straight line, expressed as a percentage of the maximum. We explore delays in the 1 to 100 ms interval. Results are shown in Figure 3.

A higher number of points are taken close to the origin since this is the interesting range of values in human PTT studies using the optical probes.

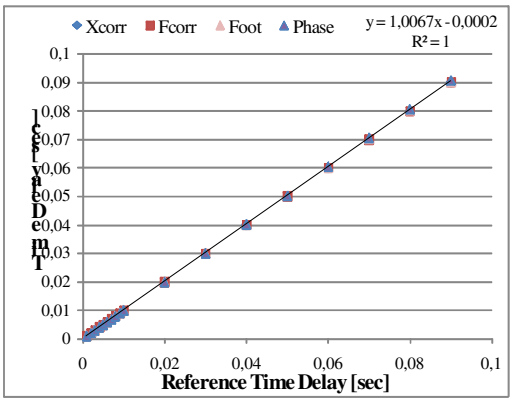


Fig. 3. Reference Delay versus measured delay for the PP probe. The APD curve practically coincides with this one.

For both probes, all the algorithms produce highly linear (better than 1%) results as well as low error agreement with the reference time delay.

5.2 PTT Error

Error plots, expressed as a percentage of the corresponding reference value, are shown in Figures 4 and 5. We discuss the main differences between the PP and the APD probes.

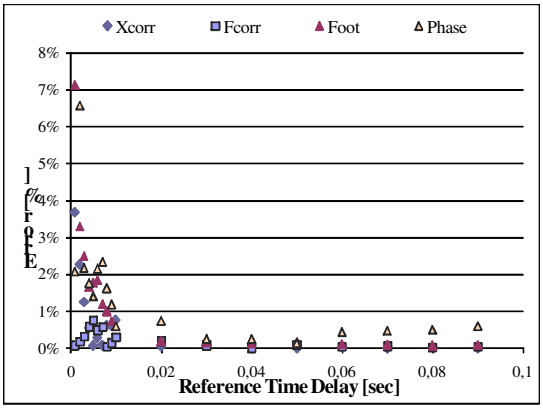


Fig. 4. Relative errors by algorithm, for the PP probe

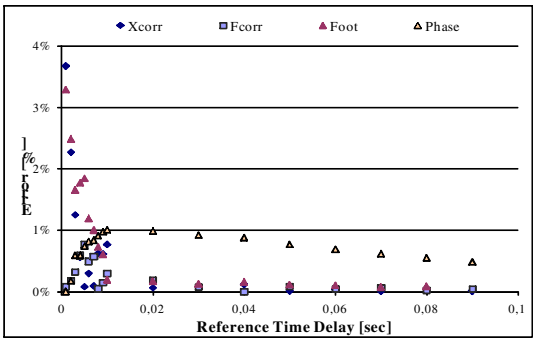


Fig. 5. Relative errors by algorithm, for the APD probe

While the PP probe exhibits lower than 8% error, the APD one never exceeds the 4% limit.

Cross-correlation (*Fcorr* version) can be identified as the best performing algorithm with a relative error never exceeding 1% in any probe.

In the APD probe, the phase angle detection method also yields very good (lower than 1%) error, but poor performance for the PP probe, mainly in the small time lag region.

As expected all the algorithms performed almost perfectly for higher than 10ms time delays.

5.3 Noise Tolerance

Robustness of the algorithms to noise is assessed by adding normal distribution noise to the photodiode readings and studying the resulting effect on the algorithm output.

This test was performed just for the correlation and phase methods. It was not used in foot-to-foot detection, because, as long as added noise is of the order of magnitude of the threshold used to detect the upstroke, the upstroke will not be detected at all.

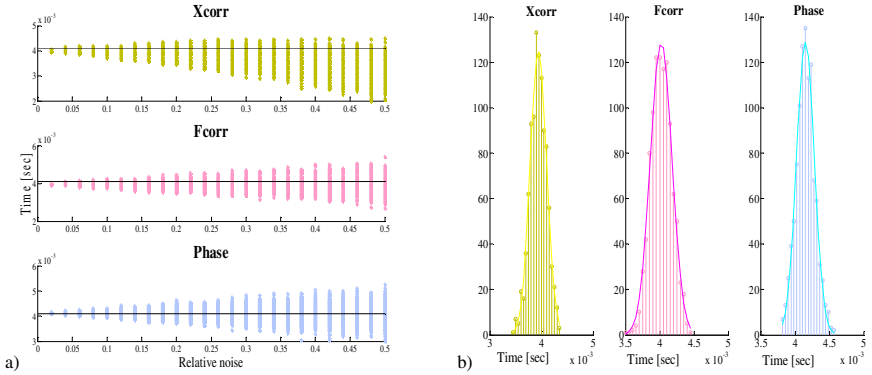


Fig. 6. Dispersion introduced by noise in the PP probe a). PTT dispersion plots for each algorithm, for a relative noise level of 0.14, the gaussian fittings stress the normal nature of the distributions b).

Data collected by the PP and APD probes was submitted to this test using the following procedure: for each noise level, the algorithm under test was run 1000 times, with an independent noise vector affecting every run.

In total, 25 relative noise levels, from zero to 0.5 of peak amplitude, were explored. Figures 6b) and 7 show the dispersion introduced by noise for a reference delay of 4.1ms. The resulting PTT values, taken as the mean value of each distribution, are plotted in Figure 7.

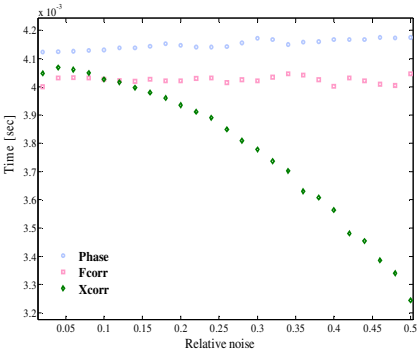


Fig. 7. Mean of distribution vs. relative noise for a 4.1 ms reference delay in the PP probe

While these figures concern the PP probe, Figures 8 and 9 represent the same study for the APD probe. As mentioned before, noise is expressed as a fraction of the peak amplitude of the signal.

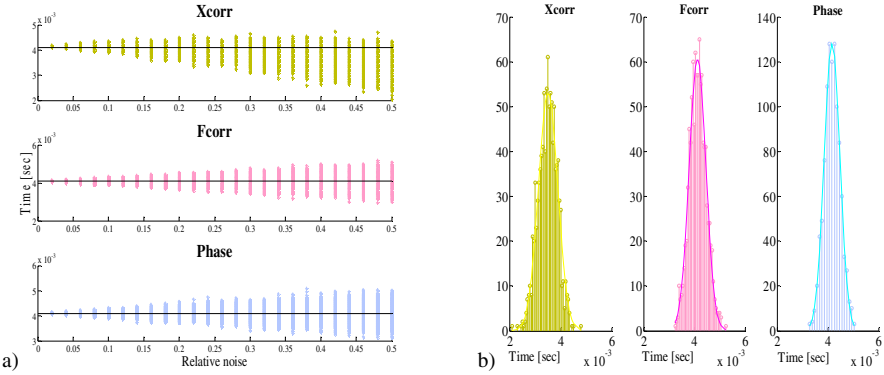


Fig. 8. Dispersion introduced by noise for the APD probe a).PTT dispersion plots for each algorithm, for a relative noise level of 0.14. the gaussian fitting stresses the normal nature of the distribution b).

Not surprisingly, the dispersion introduced by adding noise is also gaussian with variance proportional to the noise level (Figures 6a and 8a).

However, different robustness to noise is exhibited by each of the three tested algorithms, with the phase and *Fcorr* methods showing the lowest errors when subject to high levels of noise.

It is also clear that the *Xcorr* based algorithm is not robust to noise and, under high noise conditions it shows a strong tendency to under-evaluate PTT, as shown in Figures 7 and 9.

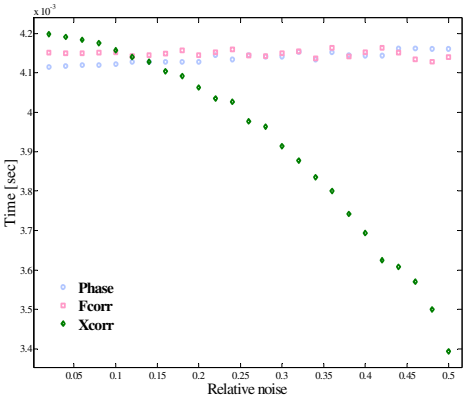


Fig. 9. Mean of distribution vs. relative noise for a 4.1 ms reference delay in the APD probe

The phase method exhibits the higher levels of robustness since its median remains constant for high noise levels and, in addition, the corresponding distribution shows the lower variance. The large offset yielded by this algorithm in the PP probe (but not in the APD probe) is rather puzzling and is probably associated to the particular shape of PP signals which, very much unlike the APD, are conditioned by the large equivalent capacity of the photosensor.

Another clarifying way to look at the overall performance of probes and algorithms is shown in Figure 10 where the probabilities of the algorithm returning a PTT value with less than 5% and less than 10% error are plotted against noise. Results, expressed as a percentage, are derived from 1000 runs per curve.

Data in Figure 10 confirms the superior robustness of the phase method.

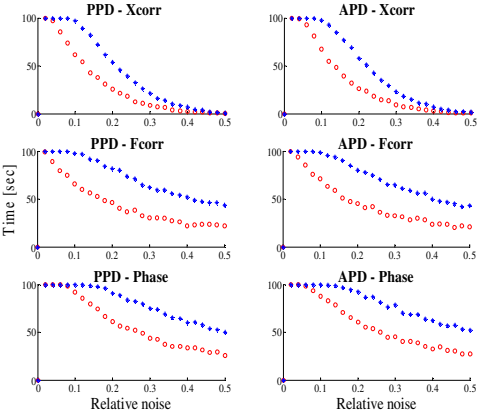


Fig. 10. Measurements with less than 5% error (blue dots) and less than 10% error (red circles) vs. relative noise

As can be stated, all algorithms can deliver 100% measurements within the specied error threshold, up to a certain noise level, where the curves show a turning point and start decaying towards zero. The phase algorithm not only shows a higher turning point but also decays much slowly as noise increases, denoting extra robustness to noise.

5.4 Algorithm Robustness

A final test was carried out in order to study the effect of different heart rates on the performance of the algorithms. In fact, all the data mentioned so far was acquired at a rate of 1 pulse per second, thus, any conclusive notes might not be valid for other acquisition rates. Accordingly, the referred test was performed for signal repetition rates varying from 1 to 200 Hz, without artificial noise added to the readings and for a known constant time delay.

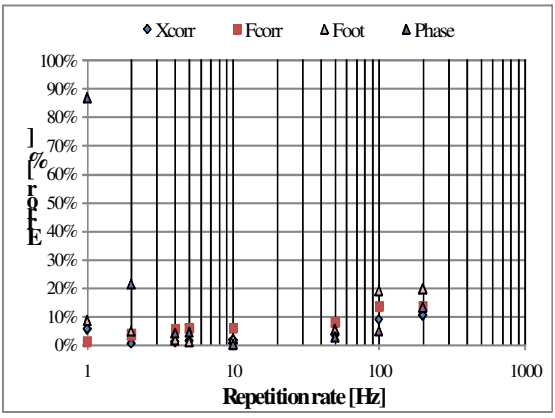


Fig. 11. Plot of the relative errors for each algorithm for a range of frequencies of signal, for the PP probe

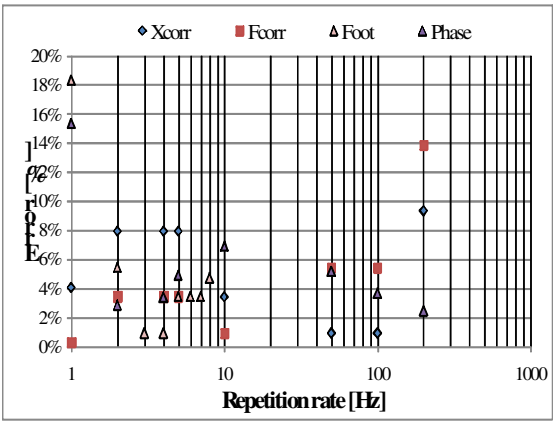


Fig. 12. Plot of the relative errors for each algorithm for a range of frequencies of signal, for the APD probe

The value used for the time delay, 1.1 ms, was selected by mere convenience. At this point it's not unimportant to remark that the AWG2 (Figure 1) can define the time delay as an angle, the *delay angle*, with a precision of a tenth of a degree; on the other side, for the specific used set of repetition rates, the value of 1.1 ms yields feasible values for delay angles that, otherwise, could not be loaded by the equipment.

In conclusion, as Figures 11 and 12 reveal, the APD probe performs superiorly (note that the vertical scales of the figures are different). It is also noticeable that the *Fcorr* and the phase algorithms produce the best results if the entire range of repetition rates is considered.

6 Conclusions

Two optical probes specifically designed to measure PTT have been developed and tested along with three different signal processing algorithms.

Tests show that although both probes are capable of measuring PTT accurately, the APD based one is more precise and accurate.

All three tested algorithms can measure PTT with an error below 8%. Nevertheless, just the one designated by *Fcorr* exhibits the capability of measuring PTT with an error bellow 1%, for the complete range of delays. The phase method shows the higher levels of robustness to noise.

When the signal repetition rate spans over a large range of values, the *Fcorr* algorithm can deliver PTTs with the lowest errors.

The natural follow-up of this work will be start acquiring pulse data in humans. Figure 13 shows a preliminary acquisition in human using the APD probe. The shape of the pulses is very clear, not to much affected by noise and allows the anticipation of good results.

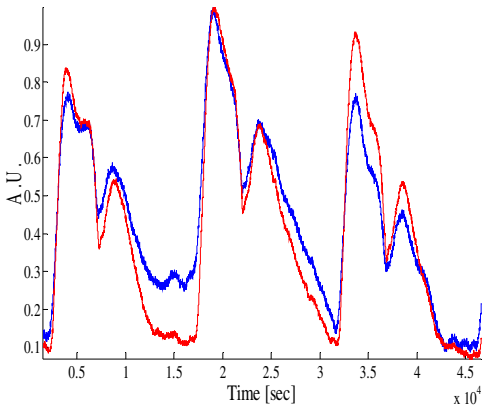


Fig. 13. Preliminary results of the APD probe acquiring data in humans

In the future, we expect to test new light sources with different wavelengths, such as infrared, with greater ability to penetrate in tissues and wavelengths with peak

sensitivity closer to each probe. For PPD the Maximum Sensitivity Wavelength is 930nm, and for the APD is the 830nm.

The ergonomics of the probe is very important factor, in this regard, new probes are being designed, seeking more user-friendly and stable platforms. For the PPD probe we test smaller sensors, with reduced area for decreasing the error associated to the detection solid angle.

Acknowledgements. The authors acknowledge the support from Fundação para a Ciência e Tecnologia (FCT) for funding (PTDC/SAU-BEB/100650/2008). Project developed under the initiative of QREN, funding by UE/FEDER, through COMPETE - Programa Operacional Factores de Competitividade.

References

1. Bramwell, J., Hill, A.: The velocity of the pulse wave in man. Proc. Roy. Soc. Lond. [Biol.] 93, 298–306 (1922)
2. Rajzer, M., et al.: Comparison of aortic pulse wave velocity measured by three techniques: Complior, Spymocor and Arteriograph. J. Hypertens 26, 2001–2007 (2008)
3. Laurent, S., et al.: Expert consensus document on arterial stiffness: methodological issues and clinical applications. Eur. Heart. J. 27, 2588–2605 (2006)
4. Xu, M.: Local Measurement of the Pulse Wave Velocity using Doppler Ultrasound. Massachusetts Institute of Tecnology (2002)
5. Kips, J., et al.: The use of diameter distension waveforms as an alternative for tonometric pressure to assess carotid blood pressure. Physiol. Meas. 31(2010), 543–553 (2010)
6. Vermeersch, S.J., et al.: Determining carotid artery pressure from scaled diameter waveforms: comparison and validation of calibration techniques in 2026 subjects. Physiol. Meas. 29, 1267–1280 (2008)
7. Kazanavicius, E., Gircys, R.: Mathematical Methods for determining the foot point of the Arterial Pulse Wave and Evaluation of Proposed Methods. Information Technology and Control (2005) ISSN 1392 – 124X
8. Azaria, M., Hertz, D.: Time delay estimation by generalized cross correlation methods. IEEE Trans. Acoustics, Speech, and Signal Processing ASSP-32, 280–285 (1984)

Fractal-Based Brain State Recognition from EEG in Human Computer Interaction

Olga Sourina, Qiang Wang, Yisi Liu, and Minh Khoa Nguyen

Nanyang Technological University, Nanyang Ave, Singapore
{eosourina,wang0586,liuy0053,raymondkhoa}@ntu.edu.sg

Abstract. Real-time brain states recognition from Electroencephalogram (EEG) could add a new dimension in an immersive human-computer interaction. As EEG signal is considered to have a fractal nature, we proposed and developed a general fractal based spatio-temporal approach to brain states recognition including the concentration level, stress level, and emotion recognition. Our hypothesis is that changes of fractal dimension values of EEG over time correspond to the brain states changes. Overall brain state recognition algorithms were proposed and described. Fractal dimension values were calculated by the implemented Higuchi and Box-counting methods. Real-time subject-dependent classification algorithms based on threshold FD values calculated during a short training session were proposed and implemented. Based on the proposed real-time algorithms, neurofeedback games for concentration and stress management training such as "Brain Chi", "Dancing Robot", "Escape", and "Apples", and emotion-enabled applications such as emotion-enabled avatar, music therapy, and emotion-based search were designed and implemented.

Keywords: BCI, Neurofeedback, Emotion recognition, Fractal dimension, EEG, Real-time applications.

1 Introduction

An immersive human interaction with computer systems should use all human senses such as visual, audio, tactile, odor, taste, etc. to make virtual or even non-virtual experience more real. In the case of human computer systems, human receives information from the system using eyes, ears, skin, nose, etc. Then, the information is processed by the corresponding lobes of the brain. After that, the user could make a decision to enter information into the computer system. It could be done according the implemented human-computer system interface scenarios. To make the human-computer interfaces more seamlessness the information could be entered involuntary as well. It could be entered into the computer by cameras, sensors based tracking systems, and by biofeedback sensors. Then, the information could be processed by the corresponding algorithms depending on the system application. In this paper, we study a novel dimension of human-computer interfaces that is based on real-time EEG recordings and its recognition. Electroencephalogram (EEG) is a non-invasive technique recording the electrical potential over the scalp which is produced by the activities of brain cortex, and reflects the state of the brain [1]. EEG-based technology is widely used in serious

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 258–272, 2013.
© Springer-Verlag Berlin Heidelberg 2013

games design since more wireless headsets that meet consumer criteria for wearability, price, portability and ease-of-use came to the market. EEG techniques give us an easy and portable way to monitor brain activities by using suitable signal processing and classification methods and algorithms. Originally, such technologies were mostly used in different medical applications, Brain Computer Interfaces (BCI), and neurofeedback games. The algorithms adopted in such applications are mainly based on power spectrum analysis, which may not be fully revealing the nonlinear complexity of the brain activities.

We proposed new algorithms of brain state recognition and innovative integrated methods and tools for implementation of the EEG-based user immersion and interaction. Algorithms of the "inner" brain state quantification including emotion recognition would advance research on the human computer interaction bringing up the proposed novel quantification methods and algorithms as new research tools in medical applications, entertainment, and even novel digital art methodology applications, and allowing us an integration of the brain state quantification algorithms in the human computer interfaces. It would lead to the implementation of the applications such as EEG-based serious games including neurofeedback games, emotion-enabled personalized search on the Web, experimental art animation, personalized avatars seamlessness communicating with virtual objects, other avatars, or even social robots working with elderly people, etc.

In this paper, we describe a novel spatio-temporal fractal-based approach to brain state recognition particularly to recognition of the concentration levels and emotion recognition. The method is general, and it could be used as a basis for the development of other brain states recognition algorithms as well. Our approach consists from two parts: interactive tools allowing dynamic analysis of EEG signals amplitudes and other parameters distribution over the brain lobes, and fractal dimension algorithms with sliding window that could estimate complexity of the signal by calculating fractal dimensions values or fractal dimension spectrum changing over time. The algorithms of concentration level recognition and emotion recognition could use just one fractal feature per channel that allows us to implement real-time EEG-enabled applications with acceptable accuracy. The accuracy could be improved by combination of fractal dimensions and by applying the generalized fractal spectrum. We implemented the real-time applications such as blobby 3D mapping, color mapping, concentration and stress management training games, emotion-enabled avatar, emotion-enabled search on the Web, music therapy, etc. that are described in this paper.

In Section 2.1, we describe neurofeedback systems and its applications. In Section 2.2, emotion recognition algorithms are reviewed. Then, we describe the proposed general spatio-temporal fractal-based approach to the brain states recognition. In Section 3.1, 3D Mapping of EEG signal with blobby model is presented. In Section 3.2, fractal dimension algorithms applied for extraction of fractal dimension features of EEG signal with sliding window are elaborated. An implementation of the overall algorithm that includes fractal feature extraction and classification is given in Section 4. Real-time EEG-enabled applications including EEG-based concentration and stress management training games, emotion-based music search, emotion-enabled music therapy are described in Section 5.

2 Related Work

2.1 Neurofeedback Systems

EEG-based systems using the brain signals could create a new channel of interaction of humans with computers or other devices. There are two types of real-time EEG-based systems: Brain Computer Interfaces (BCI) and neurofeedback systems. BCI systems are mostly used for disabled persons to communicate with computers or devices. Neurofeedback is a technique that allows the user voluntarily change his/her brain state based on the visual or audio feedback from the system corresponding to the recognition from the user EEG state of the brain. The user by doing some exercises recommended by the doctor or just by playing the serious game with the neurofeedback learns how to improve his/her brain plasticity. Neurofeedback could recover some psychological disorders or just help to improve some skills of concentration, meditation, etc. EEG based neurofeedback systems have significant treating effects on Attention Deficit Hyperactivity Disorder (ADHD) patients using sensorimotor power and low Beta power training [2], and theta-beta training and SCP training [3]. For Autistic Spectrum Disorders (ASD) patients, the EEG power ratio training [4] and theta power training [5] could be used. And for Substance Use Disorders (SUD) patient, alpha and theta power training has good effects [6]. Neurofeedback systems could also be applied for special brain function improvement, e.g. to enhance the focused attention process [7], mental rotation ability [8], and memory work [9]. As EEG signals could be used to characterize the subject's concentration level [10,11], the neurofeedback system can be used for concentration enhancement [12,13,14,15].

Although an efficiency of EEG linear features application were proved in clinical treatments, the nonlinear methods, e.g. entropy analysis and fractal dimension analysis, became popular in EEG processing due to the nonlinearity of the EEG signals. The hypothesis is that a non-linear fractal dimension approach allows quantify brain states corresponding to the concentration levels, pain levels, etc. In work [16,17], two well-known algorithms such as Box-counting [18] and Higuchi [19] were applied for the concentration level recognition in neurofeedback games, and the efficiency was proved. Recently, efforts have been made on the development of EEG-based real-time applications in multimedia communication, rehabilitation games, interaction in virtual environments, etc. In this paper, we describe new EEG-enabled applications that employ the proposed concentration level recognition and emotion recognition algorithms.

2.2 Emotion Recognition Algorithms

Emotion recognition from EEG could reveal the "inner" feeling of the user, and then, it could be used in a therapy or to create an emotion-enabled avatar of the user or other real-time applications. Emotion recognition algorithms consist from two parts: feature extraction and classification. For real-time applications, an objective is to develop fast algorithms recognizing more emotions with fewer electrodes used. Currently, mostly off-line recognition algorithms were proposed which are shown in the Table 1 and Table 2. EEG-based emotion recognition algorithms could be divided into two groups: a subject-dependent (as shown in Table 1) and a subject-independent one (as shown

in Table 2). In the Table 1 and 2, the algorithms are compared by feature extraction and classification algorithms used, by emotion types recognized, and by the algorithms accuracy. The algorithms are also differed by the number of the electrodes used in the emotion recognition. In the Table 1 and 2, in works [20,21,22,23,24], from 2 to 4 electrodes were used. All other works employed more than 32 electrodes to collect EEG data.

Table 1. Subject-dependent emotion recognition algorithms

Author	Feature and Classification	Emotion	Result
Ishino and Hagiwara, 2003 [20]	**Feature** FFT; Wavelet transform; Variance, mean **Classification** Neural Network	Joy; sad; angry; relaxed	Joy:54.5% Anger: 67.7% Sorrow: 59% Relaxation:62.9%
Zhang and Lee, 2009 [24]	**Feature** PCA **Classification** Linear Kernel SVM; RBF Kernel SVM	Negative and positive	73%
Chanel et al., 2006 [25]	**Feature** 6 frequency bands from different locations **Classification** Naive Bayes; Fisher Discriminant Analysis	3 degree of arousal	58%
Chanel et al., 2009 [26]	**Feature** Short Time Fourier Transform; Mutual Information **Classification** Discriminant Analysis; SVM; Relevance Vector Machine	Positive/ arousal; neutral/ calm; negative/ arousal	63%
Lin et al., 2009 [27]	**Feature** ASM 12 **Classification** SVM	Joy; anger; sadness; pleasure	90.72%
Schaaff and Schultz, 2009 [28]	**Feature** peak alpha frequency, alpha power, and cross-correlation features **Classification** SVM	Pleasant; neutral; unpleasant	66.7%

Table 2. Subject-independent emotion recognition algorithms

Author	Feature and Classification	Emotion	Result
Khalili and Moradi, 2009 [29]	**Feature** Statistical feature combined with correlation dimension **Classification** Quadratic Discriminant Analysis	Positive/ arousal; calm; negative/ arousal	76.66%
Takahashi, 2004 [23]	**Feature** Statistical feature **Classification** SVM; Neural Networks	Joy; anger; sadness; fear; realization	41.7% for five emotions; 66.7% for three emotions
Petrantonakis and Hadjileontiadis, 2010 [21]	**Feature** Statistical feature; wavelet based features; higher order crossings **Classification** SVM; QDA; KNN; Mahalanobis Distance	Happy; surprised; angry; fear; disgust; sad	62.3% for single channel case; 83.33% for combined channel case

Although, in the past few years, the field of EEG-based emotion recognition was receiving more and more attention, it is still a relatively new area. There are the following limitations: most of the works are off-line implementation of the algorithms; the number of electrodes used is usually large; the types of emotions recognized are limited, etc. In [30], we proposed a real-time algorithm only using 3 channels in total. Fractal dimension algorithms were applied to compute fractal based features, and the real time EEG-based emotion recognition algorithm was implemented with predefined thresholds based on the training session analysis. In our work, by recognizing arousal and valence level with an accuracy of 84.9% and 90% respectively, the satisfied, pleasant, happy, frustrated, sad, and fear emotions were differentiated. Since the discrete emotions can be mapped to the 2D emotional model, and fractal dimension values can be mapped to 2D emotional model as well, even more emotions defined in the 2D model could be distinguished with the implemented algorithm.

3 Spatio-Temporal Approach

The spatio-temporal approach combines two methods: a spatio-temporal analysis and fractal based analysis. The spatio-temporal analysis includes real-time 3D mapping of EEG signal amplitude or other parameters, for example, fractal dimension values, with blobby model defined by implicit functions. Set-theoretic operations could be applied over the moving blobby shapes to isolate activities common for the signal during the

time interval, as well as those that are unique one. The proposed fractal based method allows us to estimate the signal complexity changing over time and then, recognize the brain state.

3.1 3D Mapping of EEG

We proposed a novel method of EEG analysis based on 3D mapping of EEG data. We employed a concept of a dynamic 3D volumetric "blobby" shape to visualize the EEG signal changes over time. The blobby-like objects were firstly introduced in [31,32]. A time-dependent "blobby" object is defined using implicit functions that allow us to propose and implement set-theoretic operations over the time-dependent shapes for further analysis.

This object is defined using so-called FRep representation proposed in [33] and extended to spatio-temporal model in [34,35] where it was described by the following formula:

$$f(x,y,z,t) = \sum_{i=1}^{M} a_i e^{-r_i \cdot b_i(t)} - g \geqslant 0 \,. \qquad r_i = \sqrt{(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2} \,. \tag{1}$$

where a is a scale factor, b is an exponent scale factor changing over time, g is a threshold value, and m is the number of electrodes used.

At any given point (x,y,z,t), function f can take negative, positive or zero values. The point is considered on the surface of the object if the function value is zero, inside the object if the function value is positive, and outside the object otherwise. The shape changes through time due to the variable values of the exponent factor b according to the signal. Its size and appearance visually reflect the brain activity. For a better visual impression, the blobby shape is superimposed on a 3D head model. Besides just a visual comparison, we proposed to apply set-theoretic ("Boolean") operations to the moving shapes to isolate activities common for them per time point, as well as those that are unique for either one. Furthermore, the group set-theoretic operations applied to the individual time frames of the moving shape allow us to isolate idle parts of the brain as well as to estimate an average level of the brain activity. The proposed operations could be applied over one or/and over two datasets. On one data set, we can do intersection of all shapes to show the constant activity on the time interval, and union of all shapes to show the overall maximum activity. It could be done in real time. On two data sets, we could apply an intersection to show common activity, a union to show overall maximum activity, and subtraction to show activities which are characteristic of one set.

3.2 Fractal-Based Approach

Fractal dimension (FD) is a measurement of complexity and irregularity of the object based on an entropy analysis. Entropy is a measure of the disorder in physical systems, or an amount of information that may be gained by observations of the disordered systems. A common practice to distinguish among possible classes of time series is to

determine their so-called correlation dimension. The correlation dimension, however, belongs to an infinite family of fractal dimensions [36]. Hence, there is a hope that the use of the whole family of fractal dimensions may be advantageous in comparison to using only some of these dimensions. The concept of generalized entropy of a probability distribution was introduced by Alfred Renyi [37]. Based on the moments of order q of the probability p_i,Renyi obtained the following expression for entropy

$$S_q = \frac{1}{q-1} \log \sum_{i=1}^{N} p_i^q \, . \tag{2}$$

where q is not necessarily an integer and log denotes $\log_2$. Note that for $q \to 1$, Eq. 2 yields the well-known entropy of a discrete probability distribution [38]:

$$S_1 = -\sum_{i=1}^{N} p_i \log p_i \, . \tag{3}$$

There are various methods to calculate fractal dimensions. In works [39,34], the generalized Renyi approach based on Renyi entropy and calculation of the whole spectra of fractal dimensions to quantify brain states were studied. In work [40], another generalization of fractal dimension based on correlation dimension was proposed and described. As EEG signal is not mono-fractal, calculation of fractal spectrum of EEG with sliding window could improve EEG signals classification accuracy. In our real-time applications, we apply two well-known Higuchi [19] and Box-counting [18] algorithms to calculate the fractal dimensions. Both of the algorithms were evaluated using mono-fractal Brownian and Weierstrass functions where the theoretical FD values are known [17]. The algorithms were used in the proposed algorithm of FD feature extraction with sliding window for concentration level recognition and emotion recognition. Higuchi algorithm gave a better accuracy as FD values were closer to the theoretical FD ones, and it was applied in the proposed emotion recognition algorithm. Both Higuchi and box-counting algorithms were used in the concentration level recognition algorithm to calculate two FD values to improve the algorithm accuracy.

The Box-counting and Higuchi algorithms are described as follows.

Box-counting Method. Fractal dimension D_B is defined in Box counting method [18] as

$$D_B = \lim_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon)} \, . \tag{4}$$

Equivalently,

$$D_B = -\lim_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{\ln(\varepsilon)} \, . \tag{5}$$

where $N(\varepsilon)$ is the number of boxes of length ε which cover the whole data set.

Our implementation of the above box counting formula is based on [41]. A time series data are covered by a grid of boxes of length ε. $N(\varepsilon)$ is the number of boxes

that the curve intersects with the grid. Different values of ε result in different number of boxes and hence different values of $N(\varepsilon)$. Therefore, the curve of $\ln N(\varepsilon)$ is plotted versus $\ln(\varepsilon)$ and D_B is calculated as the slope of that curve multiplied by -1.

Higuchi Method. The following implementation is based on work [19]. Suppose we want to calculate fractal dimension for a time series $x(1), x(2), \ldots, x(n)$.

step 1: Choose one value of k.

step 2: Construct the sub-series X_k^m from the time series as following

$$x(m), x(m+k), \ldots, x(m+[\frac{n-m}{k}]\cdot k)\ . \tag{6}$$

where $m = 1, 2, \ldots, k$ and [ ] denotes Gaussian notation which rounds a number in the brackets to its largest integer which is equal to or smaller than itself, m is the initial time and k is the interval time. For example, when $k = 3$ and $n = 100$ we have 3 sub-series as follows:

$$\begin{gathered} X_3^1 : x(1), x(4), x(7), \ldots, x(100) \\ X_3^2 : x(2), x(5), x(8), \ldots, x(98) \\ X_3^3 : x(3), x(6), x(9), \ldots, x(99) \end{gathered}$$

Then, every length of each sub-series X_k^m is calculated. Length $L_m(k)$ of X_k^m is equal to

$$\frac{\left\{ (\sum_{i=1}^{[\frac{n-m}{k}]} |x(m+ik) - x(m+(i-1)\cdot k)|) \frac{n-1}{[\frac{n-m}{k}]\cdot k} \right\}}{k}\ . \tag{7}$$

step 3: Calculate the average length $L(k)$ of all $L_m(k)$.

step 4: Repeat step 1 to 3 for several values of k.

step 5: Slope of the curve of $\ln(L(k))$ versus $\ln(k)$ is approximated. Fractal dimension value is the slope multiplied by -1.

4 Implementation

The proposed real-time system diagram is shown in Fig. 1. The user receives stimuli from the computer system such as visual, audio, etc. Then, the corresponding brain states changing over time are recognized by the proposed algorithms from his/her EEG. The EEG data is acquired by the EEG device. An overall recognition algorithm used in the real-time applications consists from the following steps: data sampling and pre-processing including data filtering, feature extraction, and subject-dependent machine learning algorithm. Then, the command to the feedback system is formed based on the recognition results.

4.1 Pre-processing

The collected data are filtered by a 2-42 Hz bandpass filter since major waves of EEG lie in this band [42].

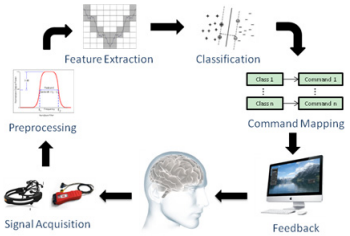


Fig. 1. Diagram for non-invasive BCI system

4.2 Features Extraction

The next step after the data pre-processing is the features extraction. We apply a sliding window to calculate FD-based feature vector per window per channel. The FD-based feature vector is defined as follows:

$$FV_{FD} = [FD_1, FD_2, \cdots, FD_n] \,. \tag{8}$$

where n is the number of FD values used in the feature vector.

Then, if more than one channel is used to recognize the brain states, the feature vector FV_{FD}^{C} is defined as follows:

$$FV_{FD}^{C} = [FV_{FD}^{1}, FV_{FD}^{2}, \cdots, FV_{FD}^{m}] \,. \tag{9}$$

where m is number of channels.

In the proposed concentration level recognition algorithm, we have two FD values per channel: n equals to 2 in the Eq. 8, and m equals to 1 in the Eq. 9 [43]. In the real-time emotion recognition algorithm, there is one FD value per channel and 3 channels are used: n equals to 1 in the Eq. 8, and m equals to 3 in the Eq. 9 [30].

4.3 Classification Algorithms

Currently, we implemented real-time subject-dependent classification algorithms based on the threshold FD values that were calculated during short training sessions. Note that off-line processing with SVM classifier of the EEG labeled with emotions and concentration levels gave us similar accuracy as the real time implementation algorithms that used the thresholds.

In the concentration recognition algorithm, an adaptive threshold was implemented to enhance the user performance. If it is hard for the user to concentrate during the predefined time, the threshold is automatically reduced to make the concentration process easier for the user. On contrary, if he/she has attained a certain concentration level for a certain amount of time, the threshold is automatically increased to raise the difficulty level of the training.

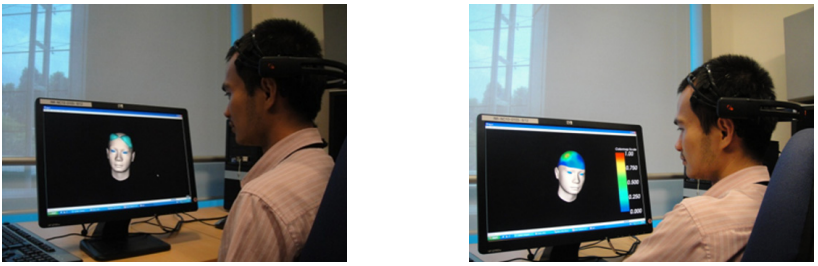
5 Applications

EEG data are collected by Emotiv device with 14 electrodes located at AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 standardized by the American

Electroencephalographic Society [44]. The sampling rate is 128Hz. To be able to use any EEG device a program reading raw EEG signals is needed to be implemented. Currently, our applications could also work with Pet 2 and Mindset 24. All electrodes could be active in the system. The steps of the overall algorithm of the real-time application are as follows. First, raw data are read from the EEG device, filtered with band pass filter 2-42 Hz, and entered to the corresponding brain state recognition algorithm. Then, the results of the recognition are fed to the developed game, web site, or any other real-time application.

5.1 3D Mapping of EEG

We proposed a spatio-temporal approach to the EEG analysis. 3D blobby-based EEG mapping was implemented for offline processing. To monitor and analyze the subject/user brain state in real time, a system "VisBrain" was implemented. Signal amplitude values are visualized with blobby model, color or "pins". In Fig. 2a, the real-time spatio-temporal visualization of EEG signals is shown with the 3D blobby mapping. The blobby model allows assessing a spatio-temporal pattern of the subject/user EEGs corresponding to the different brain states. The user could visually access his/her own brain state pattern corresponding to different mental tasks. In Fig. 2b, color mapping of the user's EEG parameters is shown.



(a) (b)

Fig. 2. Real-time system "VisBrain"for the spatio-temporal analysis based on (a) blobby model; (b) color map

5.2 EEG-Based Serious Games

The EEG-based serious game design includes two parts: signal processing algorithms and a 2D/3D or virtual reality game part. Raw EEG signals collected by the device from the user brain are filtered and analyzed by signal processing algorithms in real-time, and the resulting values are interpreted in the game as an additional game control using just the "brain power". A therapeutic effect of such games could consist from the combination of a distraction effect of the game and an effect from the learning by the user/patient how to control the game by changing voluntary his/her brain state. For example, the user could learn how to improve his/her concentration level. We developed

the games for concentration training named “Brain Chi”, “Dancing Robot”, “Explosion”, and the game for stress management training named “Pipe”, and “Apples”. They are simple single-player games implemented with the game engines SDL, Panda3D, and Adobe Flash CS4. The recognized relaxation/ concentration/ stress level values from EEG could be interpreted in the games as any visual/audio effects or even as a behavior change of the game characters.

In the “Brain Chi” game, the relaxation/ concentration level of the user is associated with the radius of a “growing/shrinking” ball. It allows the “little boy” character to fight enemies by “growing” the ball. In the “Dancing Robot” game, the relaxation/concentration level is associated with the “robot” character behavior. When the concentration level of the user increases, the robot character starts to move faster. If the user is fully relaxed, the robot stops dancing. In our implementation, the concentration and relaxation levels could be easily associated either with concentration training or relaxation training depending on the therapeutic purpose of the game. A change of the quantified level of the user concentration level is interpreted as a “faster/slower” movement of the “robot”. In Fig. 3a, the screenshot of the EEG-enabled “Explosion” game for concentration training is shown. The level of the user concentration is presented with the green bar. The time of achieving the highest concentration level is calculated. The highest concentration level would result in the 3D wall explosion.

The “Pipe” game is implemented more as a traditional neurofeedback game. The bars on the screen show the level of the user stress and the number of the points gained. In the “Pipe” game, water flows faster when the player’s stress level increases, and hence it makes the game playing more difficult. In Fig. 3b, the “Apples” game is shown. The user moves the trolley to catch the falling apples using the keyboard. If the user is stressed the apples fall slower, and it takes more time to complete the game. We also did preliminary study how to use the EEG-enabled serious games for pain management and have got some promising results.

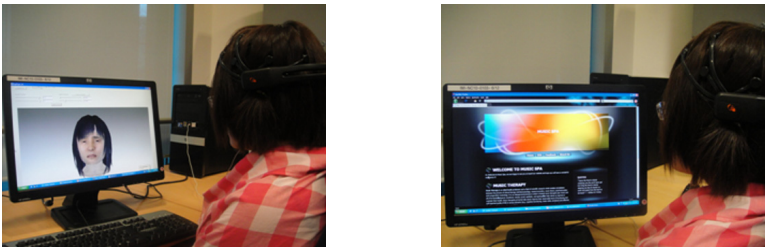
5.3 Emotion-Based Digital Experience

From our previous work [30], we have the following results: 1) the arousal level of emotions could be detected by the FD values computed from FC6 channel; 2) the valence level of emotions could be detected by the difference of FD values computed from AF3 and F4 channels; 3) the change of the FD values responsible for the arousal level recognition could be mapped to the arousal level axis in the 2D emotion model. As a result, we proposed and implemented a real-time fractal-based emotion recognition algorithm. It is possible to recognize in real time any discrete emotions that could be defined with the 2-dimensional emotion model. Satisfied, pleasant, happy, frustrated, sad, and fear emotions were recognized. In our algorithm, only 3 channels are used. We implemented three emotion-enabled real-time applications. First, we implemented an application with the EEG-enabled avatar [30]. The music stimuli were used for emotion induction as it was proposed in [45]. We used an avatar available with Haptek development package for our application [46]. Haptek Activex control allows the users to make their own 3D avatars by using their own photos and provides functions and commands to change facial expressions of 3D avatars. We defined six emotions by changing the parameters controlling the facial muscles of the Haptek emotion avatar. Those emotions



(a) (b)

Fig. 3. (a) EEG-enabled "Explosion" game for the concentration training; (b) EEG-enabled "Apples" game for the stress management training



(a) (b)

Fig. 4. Emotion-enabled applications: a) "Sad" emotion is recognized and visualized on the user 3D avatar; b) The music therapy Web site

are: fear, frustrated, sad, happy, pleasant and satisfied. In the application, emotions of the user are recognized from EEG and visualized in real-time on the user's avatar with Haptek system. In Fig. 4a, the user was listening to music pieces as external stimuli to evoke certain emotions, and the algorithm has recognized "sad" emotion that was visualized on the user's avatar. The avatar emotions are changed according to the emotions that the user is feeling during the music listening. Second, we implemented an EEG-enabled music therapy Web site. Usually, a therapist is needed to be present during music therapy sessions; however, with the help of EEG-based emotion recognition, the music therapy process could be adjusted automatically. The user/patient's emotion is recognized from EEG, and the corresponding music piece according to the music therapy plan is played to the user. In Fig. 4b, the "inner" user's emotion is recognized from the EEG signal in real-time. Then, if the music therapy for the pain management is selected, a song evoking "happy" emotion is played to the user to distract his/her attention from the pain he/she is suffering. A music therapy algorithm was proposed and implemented allowing automatically switch the music pieces if one could not evoke the targeted emotion during the planned time. Thirdly, an EEG-enabled music player was

implemented. It is an entertainment application. The music could be played according to the emotional states of the subject recognized from the EEG in real-time. For example, if the subject is feeling happy, then the happy music or opposite emotion music would be played depending on the choice of the user.

6 Conclusions

In this paper, we proposed and described a novel general spatio-temporal fractal based approach to recognize different brain states such as concentration and stress levels, and human emotions. We expect such approach could be used for other brain states recognition as well. We are going to study on "central pain"level feeling recognition. Using fractal dimension features and the machine learning algorithm allows us to implement real-time brain state recognition algorithms and its applications with acceptable accuracy that is improved by the short subject-based training sessions. The accuracy of the proposed algorithms could be also improved by using generalized fractal dimension spectrum of EEG. We expect further use of the proposed approach in different real-time applications. It could be also used in validation of the hypotheses: emotion induction could change pain level in the patients; the positive emotions could improve human performance, etc. We also work on the improvement of the real-time filtering of artifacts of different origin. The work described in the paper is a part of the project EmoDEx presented in [47].

Acknowledgements. This project is supported by the grant NRF2008 IDM-IDM004-020 "Emotion-based personalized digital media experience in Co-Spaces" of National Research Fund of Singapore.

References

1. Nunez, P.L., Srinivasan, R.: Electric Fields of the Brain. Oxford University Press (2006)
2. Fuchs, T., Birbaumer, N., Lutzenberger, W., Gruzelier, J.H., Kaiser, J.: Neurofeedback treatment for attention-deficit/hyperactivity disorder in children: A comparison with methylphenidate. Applied Psychophysiology Biofeedback 28, 1–12 (2003)
3. Gevensleben, H., Holl, B., Albrecht, B., Schlamp, D., Kratz, O., Studer, P., Wangler, S., Rothenberger, A., Moll, G.H., Heinrich, H.: Distinct eeg effects related to neurofeedback training in children with adhd: A randomized controlled trial. International Journal of Psychophysiology 74, 149–157 (2009)
4. Thompson, L., Thompson, M., Reid, A.: Neurofeedback outcomes in clients with asperger's syndrome. Applied Psychophysiology Biofeedback 35, 63–81 (2010)
5. Kouijzer, M.E.J., van Schie, H.T., de Moor, J.M.H., Gerrits, B.J.L., Buitelaar, J.K.: Neurofeedback treatment in autism. preliminary findings in behavioral, cognitive, and neurophysiological functioning. Research in Autism Spectrum Disorders 4, 386–399 (2010)
6. Saxby, E., Peniston, E.G.: Alpha-theta brainwave neurofeedback training: An effective treatment for male and female alcoholics with depressive symptoms. Journal of Clinical Psychology 51, 685–693 (1995)
7. Vernon, D., Egner, T., Cooper, N., Compton, T., Neilands, C., Sheri, A., Gruzelier, J.: The effect of training distinct neurofeedback protocols on aspects of cognitive performance. International Journal of Psychophysiology 47, 75–85 (2003)

8. Hanslmayr, S., Sauseng, P., Doppelmayr, M., Schabus, M., Klimesch, W.: Increasing individual upper alpha power by neurofeedback improves cognitive performance in human subjects. Applied Psychophysiology Biofeedback 30, 1–10 (2005)
9. Heinrich, H., Gevensleben, H., Strehl, U.: Annotation: Neurofeedback - train your brain to train behaviour. Journal of Child Psychology and Psychiatry and Allied Disciplines 48, 3–16 (2007)
10. Davidson, P.R., Jones, R.D., Peiris, M.T.R.: Eeg-based lapse detection with high temporal resolution. IEEE Transactions on Biomedical Engineering 54, 832–839 (2007)
11. Lin, C.T., Wu, R.C., Jung, T.P., Liang, S.F., Huang, T.Y.: Estimating driving performance based on eeg spectrum analysis. Eurasip Journal on Applied Signal Processing, 3165–3174 (2005)
12. Huang, R.S., Jung, T.P., Makeig, S.: Multi-scale eeg brain dynamics during sustained attention tasks. In: Proc. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. IV1173–IV1176 (2007)
13. Lutsyuk, N.V., Eismont, E.V., Pavlenko, V.B.: Modulation of attention in healthy children using a course of eeg-feedback sessions. Neurophysiology 38, 389–395 (2006)
14. Pop-Jordanov, J., Pop-Jordanova, N.: Neurophysical substrates of arousal and attention. Cognitive Processing 10, 71–79 (2009)
15. Schier, M.A.: Changes in eeg alpha power during simulated driving: A demonstration. International Journal of Psychophysiology 37, 155–162 (2000)
16. Wang, Q., Sourina, O., Nguyen, M.K.: Fractal dimension based algorithm for neurofeedback games. In: Proc. CGI 2010, p. SP25 (2010)
17. Wang, Q., Sourina, O., Nguyen, M.K.: EEG-based "serious" games design for medical applications. In: Proc. 2010 Int. Conf. on Cyberworlds, Singapore, pp. 270–276 (2010)
18. Block, A., Von Bloh, W., Schellnhuber, H.J.: Efficient box-counting determination of generalized fractal dimensions. Physical Review A 42, 1869–1874 (1990)
19. Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Physica D: Nonlinear Phenomena 31, 277–283 (1988)
20. Ishino, K., Hagiwara, M.: A feeling estimation system using a simple electroencephalograph. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 5, pp. 4204–4209 (2003)
21. Petrantonakis, P.C., Hadjileontiadis, L.J.: Emotion recognition from EEG using higher order crossings. IEEE Transactions on Information Technology in Biomedicine 14, 186–197 (2010)
22. Schaaff, K.: EEG-based emotion recognition. PhD thesis, Universitat Karlsruhe, TH (2008)
23. Takahashi, K.: Remarks on emotion recognition from multi-modal bio-potential signals. In: IEEE International Conference on Industrial Technology, IEEE ICIT 2004, vol. 3, pp. 1138–1143 (2004)
24. Zhang, Q., Lee, M.: Analysis of positive and negative emotions in natural scene using brain activity and gist. Neurocomputing 72, 1302–1306 (2009)
25. Chanel, G., Kronegg, J., Grandjean, D., Pun, T.: Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals (2006)
26. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. International Journal of Human Computer Studies 67, 607–627 (2009)
27. Lin, Y.P., Wang, C.H., Wu, T.L., Jeng, S.K., Chen, J.H.: EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In: Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, pp. 489–492 (2009)
28. Schaaff, K., Schultz, T.: Towards emotion recognition from electroencephalographic signals. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 1–6 (2009)

29. Khalili, Z., Moradi, M.H.: Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of eeg. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1571–1575 (2009)
30. Liu, Y., Sourina, O., Nguyen, M.K.: Real-time EEG-based human emotion recognition and visualization. In: Proc. 2010 Int. Conf. on Cyberworlds, Singapore, pp. 262–269 (2010)
31. Blinn, J.F.: A generalization of algebraic surface drawing. SIGGRAPH Comput. Graph. 16, 273 (1982)
32. Wyvill, G., McPheeters, C., Wyvill, B.: Data structure for soft objects. The Visual Computer 2, 227–234 (1986)
33. Pasko, A., Adzhiev, V., Sourin, A., Savchenko, V.: Function representation in geometric modeling: concepts, implementation and applications. The Visual Computer 11 (1995)
34. Kulish, V., Sourin, A., Sourina, O.: Analysis and visualization of human electroencephalograms seen as fractal time series. Journal of Mechanics in Medicine and Biology 26, 175–188 (2006)
35. Sourina, O., Sourin, A., Kulish, V.: EEG Data Driven Animation and Its Application. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2009. LNCS, vol. 5496, pp. 380–388. Springer, Heidelberg (2009)
36. Hentschel, H.G.E., Procaccia, I.: The infinite number of generalized dimensions of fractals and strange attractors. Physica D: Nonlinear Phenomena 8, 435–444 (1983)
37. Renyi, A.: On a new axiomatic theory of probability. Acta Mathematica Academiae Scientiarum Hungaricae 6, 285–335 (1955)
38. Shannon, C.: The mathematical theory of communication. M.D. Computing 14 (1997)
39. Kulish, V., Sourin, A., Sourina, O.: Human electroencephalograms seen as fractal time series: Mathematical analysis and visualization. Computers in Biology and Medicine 36, 291–302 (2006)
40. Pawelzik, K., Schuster, H.G.: Generalized dimensions and entropies from a measured time series. Physical Review A 35, 481–484 (1987)
41. Phothisonothai, M., Nakagawa, M.: Fractal-based eeg data analysis of body parts movement imagery tasks. Journal of Physiological Sciences 57, 217–226 (2007)
42. Sanei, S., Chambers, J.: EEG signal processing. John Wiley & Sons, Chichester (2007)
43. Wang, Q., Sourina, O., Nguyen, M.K.: Fractal dimension based neurofeedback. Visual Computer 27, 299–309 (2011)
44. AES: American electroencephalographic society guidelines for standard electrode position nomenclature. Journal of Clinical Neurophysiology 8, 200–202 (1991)
45. Sourina, O., Kulish, V.V., Sourin, A.: Novel tools for quantification of brain responses to music stimuli. In: Proc. of 13th International Conference on Biomedical Engineering, ICBME 2008 (2008)
46. Haptek: Haptek (2010)
47. IDM-Project: Emotion-based personalized digital media experience in co-spaces (2008)

Understanding Cerebral Activations during the Observation of Marketing Stimuli: A Neuroelectrical Perspective

Giovanni Vecchiato[1,2], Laura Astolfi[2,3], Fabrizio De Vico Fallani[1,2], Jlenia Toppi[2,3], Fabio Aloise[2], Anton Giulio Maglione[4], Febo Cincotti[2], Donatella Mattia[2], and Fabio Babiloni[1,2]

[1] Dept. Physiology and Pharmacology, University of Rome "Sapienza", 00185, Rome, Italy
[2] IRCCS Fondazione Santa Lucia, via Ardeatina 306, 00179, Rome, Italy
[3] Dept. of Computer Science and Systems, University of Rome "Sapienza" via Ariosto 25, Rome, Italy
[4] Dept. of Anatomy, Histology, Forensic Medicine and Orthopedics, University "Sapienza", Rome, Italy
Giovanni.Vecchiato@uniroma1.it

Abstract. This paper aims to be a survey of recent experiments performed in the Neuromarketing field. Our purpose is to illustrate results obtained by employing the popular tools of investigation well known in the international neuroelectrical community such as the MEG, High Resolution EEG techniques and steady-state visually evoked potentials. By means of temporal and frequency patterns of cortical activations we intend to show how the neuroscientific community is nowadays sensible to the needs of companies and, at the same time, how the same tools are able to retrieve hidden information about the demands of consumers. These instruments could be of help both in pre- and post-design stage of a product, or a service, that a marketer is going to promote.

Keywords: Neuromarketing, MEG, High resolution EEG, SSVEP, Functional connectivity.

1 Introduction

For a long time in the neuroscience the experimental psychology has been the main method to investigate the human being. In that field, they measure the execution time of the experimental subject as well as his/her reaction time to particular stimuli. In particular, the subject can answer, move the finger, do free mental associations and the experimenter will measure some variables correlated with the internal processes of the subject. Of course, the access to the internal state of the subject is restricted, since it is observed by means of his/her behavioural answers. The experimental context plays an important role in these kind of experiments, since the subjects are required to answer in a short time in order to cause errors. The measure of the amount and frequency of such errors is important as an indirect measure of the internal processes of the subject. These behavioural techniques can be applied with a low cost to a large

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 273–281, 2013.
© Springer-Verlag Berlin Heidelberg 2013

amount of people. The brain imaging techniques have the capability to show images of the cerebral activity during the execution of a particular experimental task. The most popular method is the functional Magnetic Resonance Image (fMRI), which returns a sequence of images of the cerebral activity by means of the measure of the cerebral blood flow. Although such as images are "static", i.e. they are related to around ten seconds activity, they have a high special resolution that no other neuroimaging method can offer. Nowadays, fMRI scanners are employed in the neuromarketing field and in literature there exists some scientific studies showing the activation of particular cerebral areas during the tasting of a couple of popular drinks such as Coca-Cola and Pepsi [1]. It must be noted that the design of the fMRI studies rely on highlighting cortical areas which differ between the experimental task and the control one. The issue is connected with the data interpretation which can be given by the activation of particular cerebral areas during the experimental task proposed. In order to solve this problem there is the need to generate a proper experimental design in order to remove these kind of confounding factors. As stated above, the brain imaging allows us to observe cerebral areas activating during a particular task but it does not explain why, neither in which way the information is processed within the experimental subject. In this way we are not accessing the cerebral answer of the subject by his/her behave but directly by means of the activity of his/her neurons. The role of the brain imaging in the neuromarketing field is to refine, on the basis of a selected sample of people, some theories and hypotheses the neuropsychological research draw from the same experimentation on a larger number of subjects.

In the following paragraphs we intend to give a survey about the neuromarketing research performed in the neuroelectrical community. Our purpose is to show the results obtained by different group of research, their congruence and similarity leading towards a shared model which is able to extract information about the memorization and the pleasantness perceive by subjects while watching TV advertisements. In particular, this paper will focus on the results achieved by employing the magnetoencephalography (MEG) and electroencephalography (EEG) technique along with a discussion about patterns of cortical functional connectivity and indexes derived from the graph theory.

Finally, we describe the experimental findings obtained by a steady-state visually evoked potential analysis.

2 Temporal Patterns of Cortical Activity

In this context, the research team of Sven Braeutigam [2-4] has employed the MEG in order to study the temporal relationship of cerebral areas involved in consumers' choices when they have to make decisions among different items within a laboratory.

In this study they wanted to analyse the cerebral behaviour by distinguishing male subjects from females during a simulated shopping.

Cerebral activations induced by choices to make reflect the level of familiarity or the preference that a particular experimental subject has with the presented products. These factors can be considered by taking into account the relationship between the current choice of a product on the shelf and the relative frequency of choice and usage of that product in the past.

In particular, the main observation came to light from these studies presents the consumer's choice like a complex sequence of cerebral activations that greatly differ according to the consumer's sex and to the probability of choice. From a behavioural point of view, choices with a high probability were faster than those less predictable. This can be interpreted by supposing that in the case of more difficult choices the cortical activities are more complex than those simple to make. As illustrated in Fig. 1, they distinguished two distinct cerebral paths. The first one is referred to predictable choices, i.e. associated to products that the experimental subject already used in the past or said to prefer; the second one is related to unpredictable choices, i.e. associated to unfamiliar products to the subject. In the experiment, a first stage in the decisional process has been individuated around 100 ms after the stimulus onset with an activity located in the occipital cortex. At this stage of decision (W) the subject compare the product to choose with the list of products seen before, by involving the working memory.

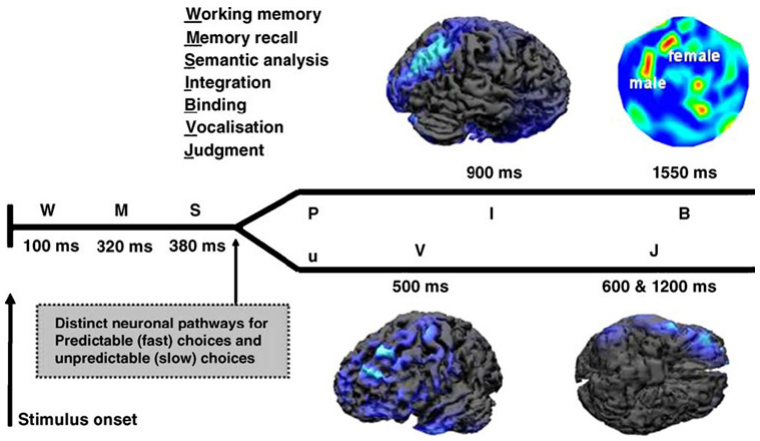


Fig. 1. Cortical activations associated to the decisions of an experimental subject. Predictable choices are the ones related to familiar items which has been often bought or used in the past. Cortical maps present the brain areas activated during the different decision stages in the frequency range of [30, 40] Hz. Modified by Braeutigam et al., 2005.

The sequence of cortical activations observed in the experimental subjects continues with two neuronal stages partially correlated (M, S) which can be observed between 280 and 400 ms after the beginning of the decisional process. In this period the selective attention of the subject is oriented towards images of products to identify, classify and compare with those stored in the memory related to the preferred products and brands. This memory can involve the past experience to have bought the particular item or to have watched the commercial of the specific brand. The cerebral activation differs in this time interval between men and women. In particular, female subjects showed a stronger activation with respect to the males in the left parieto-occipital lobule of the brain while males presented a stronger cortical activity in the right temporal lobe. These differences connected with the sex of the

subjects characterize both the stage of choice of the product and its discrimination. This observation suggest that, at this temporal stage, women tend to employ a strategy based on the knowledge of the product to buy, while men tend to act according to a spatial memory strategy [5].

After 500 ms from the beginning of the decisional process, two patterns of cortical activations can be identified according to the predictability of the choices adopted by the subjects. In particular, as to the predictable choices, we can observe a strong activation in the right parietal areas around 900 ms after the beginning of the experiment (I). In later time latencies, predictable choices of products recall strong MEG oscillations in the frequency band between [30, 40] Hz in the left prefrontal cortex (B). Parietal cortex receive inputs from many cortical areas since it is involved in the spatial integration of sensorial information. Differences in the cortical activity between men and women can strengthen the hypothesis of two different groups of strategies. On the contrary, unpredictable choices generate a strong activation in the right inferior frontal cortex (V), at a latency of around 500 ms, and in the left orbitofrontal cortex (J) between 600 and 1200 ms after the stimulus presentation. In the case V, the cortical patterns are consistent with the activity in the Broca's area, which is involved in the spoken language, which is also active during the observation of videoclips. Hence, the cortical activity at this latency may indicate a tendency to vocalize brands, as a part of strategy which helps in the decision when it is difficult. The activity in the orbitofrontal cortex (J) can be explained by stating that during an unpredictable choice we have to evaluate the outcome in terms of convenience. Overall, these results explain a complex neuronal network which is active during a simple decisional process connected to the purchase of a product. The generation of a choice is considered as an information processing which can be highly influenced, sensible to the complexity of the decision to make and to the rush in which the decision is made and many other factors.

A strong involvement of parietal areas during the observation of the TV commercials with an affective and cognitive content was also noted in a previous study, performed by using sophisticated MEG recordings [6]. In this study, cognitive frames elicited a stronger activity in the parietal areas and superior prefrontal cortex while the observation of the affective ones is correlated with the activation of the orbitofrontal and retrosplenial cortex, amygdala and brainstem. The magneto field tomography (MFT) results showed an increasing activity during the observation of cognitive stimuli rather than affective commercials in parietal and superior prefrontal areas known to be associated with executive control of working memory and maintenance of highly processed representation of complex stimuli [7]. Although the affect related activations are more variable across subjects, these findings are consistent with previous PET and fMRI studies [8-10] showing that stimuli with affective content modulates activity in the orbitofrontal and retrosplenial cortex, amygdala and brainstem.

3 Frequency Patterns of Cortical Activity and Connectivity

The research in the neuromarketing field performed in the recent years showed result suggesting that the cortical activity elicited during the observation of the TV

commercials that were forgotten (FRG) is different from the cortical activity observed in subjects that remembered the same TV commercials (RMB). In fact, the principal areas of statistical differences in power spectra between such conditions are located almost bilaterally in the prefrontal Broadmann Areas (BAs) 8, 9 as well as in the parietal BAs 7, as shown in Fig. 2.

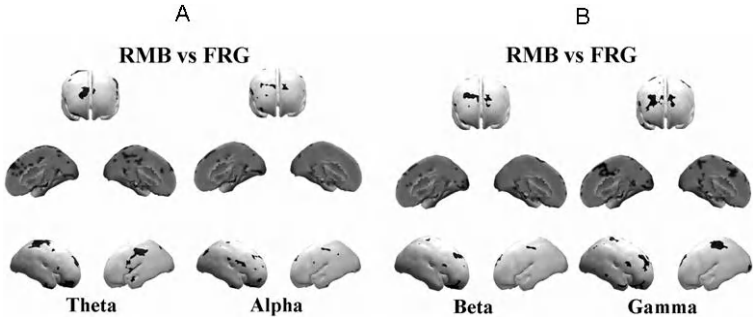


Fig. 2. Figure presents the results of statistical comparisons of RMB and FRG groups in the theta and alpha frequency ranges (panel A), and in the beta and gamma bands (panel B). In particular, the picture is composed by three rows: the first one shows the brain from a frontal perspective, the second one is related to a medial- sagittal perspective while the third one presents statistically significant images associated to the left and right lateral vision. Modified by Astolfi et al., 2008.

The spectral amplitude in the RMB condition was always higher than the power spectra in the FRG conditions over the BAs 8, 9 and 7 [11]. A statistical increase of EEG spectral power in the prefrontal and parietal areas for the RMB dataset compared with the FRG one is in agreement with the suggested role of these regions during the transfer of sensory percepts from short-term memory to long-term memory storage. Although in this study the differences in the cortical power spectra between the RMB and FRG conditions are relatively insensitive to the particular frequency bands considered, there are experimental evidences showing that there is a stronger engagement of the left frontal areas in all the subjects analyzed during the observation of the TV commercials that were remembered [12].

In particular, the analysis of the statistical cortical maps in the condition RMB vs FRG suggested that the left frontal hemisphere was highly active during the RMB condition, especially in the theta and gamma band. These results are in agreement with different observations on the RMB condition performed in literature by studying different experimental paradigm [7,13]. Taken together, the results indicated the cortical activity in the theta band on the left frontal areas was increased during the memorization of commercials, and it is also increased during the observation of commercials that were judged pleasant by subjects. These results are in agreement with the role that has been advocated for the left pre and frontal regions during the transfer of sensory percepts from the short-term memory toward the long-term memory storage by the HERA model [14,15].

In fact, in such model the left hemisphere plays a key role during the encoding phase of information from the short term memory to the long term memory, whereas the right hemisphere plays a role in the retrieval of such information.

It must be noted, however, that the role of the right cortices in storing images has been also recognized for many years in neuroscience [2,3,16,17,11]. It is worthy of note that the subjects were unaware of the kind of questions that the researcher asked them after the viewing of the documentary. Hence, the cortical areas elicited by this study are likely to be involved just in the process of the memorization of the pictorial material, owing to an increase of attention during the observation of the TV commercial. In addition, there was no particular set of commercials remembered that was common to all the subjects. The results also suggest an active involvement of the Anterior Cingulate Cortex (ACC) and the Cingulate Motor Area (CMA) as sources of links to all the other cortical areas during the observation of the TV commercials remembered after ten days. In this case the increased activity related to an increase of the outflow of PDC links from ACC and CMA towards other cortical regions could be taken as a sign of increased 'emotive' attention to the stories proposed by different TV commercials that significantly aid successive memorization.

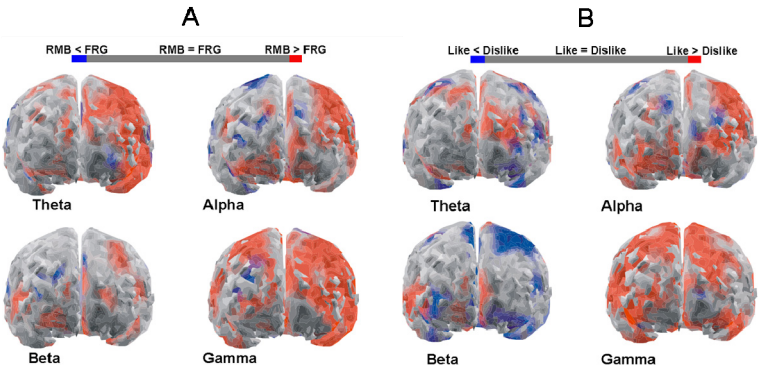


Fig. 3. Figure presents four cortical z-score maps, in the four frequency bands employed. Colour bar represents cortical areas in which increased statistically significant activity occurs in the RMB group when compared to the FRG group in red, while blue is used otherwise ($p < 0.05$ Bonferroni corrected). Grey colour is used to map cortical areas where there are no significant differences between the cortical activity in the RMB and FRG groups (panel A). Panel B refers to the statistical comparison LIKE vs DISLIKE with the same conventions of panel A. Modified by Vecchiato et al., 2010.

The EEG spectral and cortical network analyses performed in these study also suggest a key role of the parietal areas as targets of the incoming information flow from all the other cortical areas. Functional networks in the frequency domain were also estimated by evaluating the global- and local-efficiency indexes derived from the graph theory, employed as a measure of the level of communication in the networks [18]. The changes of these indexes could be related both to memory coding activity as well as to increase/decrease of attentive state of the subjects. As to the RMB condition, the functional network in the beta and gamma band state a significant non-homogeneous allocation of the involved information flows and a consequent

reduction of the efficiency in the overall communication between the network nodes. In the beta and gamma frequency bands, the respective reduction of global-efficiency, as well as the reduction of local-efficiency for the alpha band of the cortical network communication could represent a predictive measure for the accurate recall of the commercials that will be remembered.

A contrast of the activity elicited by observing pleasant (LIKE dataset) and unpleasant (DISLIKE dataset) audiovisual content has been performed in a previous study [12]. The result of this experiment shows that the activity of the brain is greater in the LIKE condition than in the DISLIKE except that in beta band, being the activity in the LIKE condition for the gamma band rather symmetrical. The results here obtained for the LIKE condition are also congruent with other observations performed with EEG in a group of 20 subjects during the observation of pictures from the International Affective Picture System (IAPS, [19]). Such observations indicated an increase of the EEG activity in the theta band for the anterior areas of the left hemisphere.

4 Steady-State Visually Evoked Potentials

These results are also supported by findings obtained from the group of Richard Silberstein which measured the steady-state visually evoked potential (SSVEP) by means of the steady-state probe topography (SSPT), which is a particular version of the EEG technology [20-22]. In this study [21], they collected the cerebral activity from thirty five women that were subjected to the exposition of eighteen minutes documentary in which 12 US TV commercials were inserted within. Seven days after the recording, the participants were asked to recall the viewed advertisements from a series of frames taken from the same commercials. They found out that images corresponding to a minima of the posterior frontal latency were more likely to be recognized than images associated to a SSVEP latency maxima. Moreover, they showed a significant correlation between the recognition performance and SSVEP latency measured at electrode sites located in the left posterior frontal site suggesting that this kind of result can be employed in order to asses the strength of long-term memory encoding for the audiovisual stimuli they proposed.

5 Conclusions

In recent years, Neuromarketing has gained always more interest and attention in both the scientific community and mass media. Findings obtained so far show that results from these studies can be of help for many areas of marketing. For instance, marketers could exploit neuroimaging tools in order to achieve hidden information about products and services to advertise that is impossible to acquire. This information could be employed both during the design process of an item and during its commercial campaign. In fact, one could think to adopt these neuroelectrical tools to test different versions of the same object, evaluate the cerebral results and use them according to the requirements of the company and at the same time facing with the

consumers' need. In addition, marketers could use an ad pre-test in order to create a TV commercial which is as closer as possible to the demand of the same. In this case we can distinguish two different point of view of the powerful and innovative tool: from one hand, product manufacturers could use cerebral information in order to force people to buy and consume products that they don't want and neither need; from the other hand, we hope that neuromarketing will be of help in design objects and the environment following the pleasures of each of us and for identifying new and exciting products that people want and find useful.

References

1. McClure, S.M., Li, J., Tomlin, D., Cypert, K.S., Montague, L.M., Montague, P.R.: Neural correlates of behavioral preference for culturally familiar drinks. Neuron 44(2), 379–387 (2004)
2. Braeutigam, S.: Neuroeconomics–from neural systems to economic behaviour. Brain Research Bulletin 67(5), 355–360 (2005)
3. Braeutigam, S., Rose, S.P., Swithenby, S.J., Ambler, T.: The distributed neuronal systems supporting choice-making in real-life situations: differences between men and women when choosing groceries detected using magnetoencephalography. The European Journal of Neuroscience 20(1), 293–302 (2004)
4. Braeutigam, S., Stins, J.F., Rose, S.P., Swithenby, S.J., Ambler, T.: Magnetoencephalographic signals identify stages in real-life decision processes. Neural Plasticity 8(4), 241–254 (2001)
5. Kimura, D.: Sex, sexual orientation and sex hormones influence human cognitive function. Current Opinion in Neurobiology 6(2), 259–263 (1996)
6. Ioannides, A.A., Liu, L., Theofilou, D., Dammers, J., Burne, T., Ambler, T., Rose, S.: Real time processing of affective and cognitive stimuli in the human brain extracted from MEG signals. Brain Topography 13(1), 11–19 (2000)
7. Summerfield, C., Mangels, J.A.: Coherent theta-band EEG activity predicts item-context binding during encoding. NeuroImage 24(3), 692–703 (2005)
8. Cahill, L., Haier, R.J., Fallon, J., Alkire, M.T., Tang, C., Keator, D., Wu, J., McGaugh, J.L.: Amygdala activity at encoding correlated with long-term, free recall of emotional information. Proceedings of the National Academy of Sciences of the United States of America 93(15), 8016–8021 (1996)
9. Maddock, R.J.: The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain. Trends in Neurosciences 22(7), 310–316 (1999)
10. Grabenhorst, F., Rolls, E.T., Parris, B.A.: From affective value to decision-making in the prefrontal cortex. The European Journal of Neuroscience 28(9), 1930–1939 (2008)
11. Astolfi, L., De Vico Fallani, F., Cincotti, F., Mattia, D., Bianchi, L., Marciani, M.G., Salinari, S., Colosimo, A., Tocci, A., Soranzo, R., Babiloni, F.: Neural basis for brain responses to TV commercials: a high-resolution EEG study. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society 16(6), 522–531 (2008)
12. Vecchiato, G., Astolfi, L., De Vico Fallani, F., Cincotti, F., Mattia, D., Salinari, S., Soranzo, R., Babiloni, F.: Changes in brain activity during the observation of TV commercials by using EEG, GSR and HR measurements. Brain Topography 23(2), 165–179 (2010)

13. Werkle-Bergner, M., Muller, V., Li, S.C., Lindenberger, U.: Cortical EEG correlates of successful memory encoding: implications for lifespan comparisons. Neuroscience and Biobehavioral Reviews 30(6), 839–854 (2006)
14. Tulving, E., Kapur, S., Craik, F.I., Moscovitch, M., Houle, S.: Hemispheric encoding/retrieval asymmetry in episodic memory: positron emission tomography findings. Proceedings of the National Academy of Sciences of the United States of America 91(6), 2016–2020 (1994)
15. Habib, R., Nyberg, L., Tulving, E.: Hemispheric asymmetries of memory: the HERA model revisited. Trends in Cognitive Sciences 7(6), 241–245 (2003)
16. Babiloni, F., Babiloni, C., Locche, L., Cincotti, F., Rossini, P.M., Carducci, F.: High-resolution electro-encephalogram: source estimates of Laplacian-transformed somatosensory-evoked potentials using a realistic subject head model constructed from magnetic resonance images. Medical & Biological Engineering & Computing 38(5), 512–519 (2000)
17. Astolfi, L., Vecchiato, G., De Vico Fallani, F., Salinari, S., Cincotti, F., Aloise, F., Mattia, D., Marciani, M.G., Bianchi, L., Soranzo, R., Babiloni, F.: The track of brain activity during the observation of TV commercials with the high-resolution EEG technology. Computational Intelligence and Neuroscience 652078 (2009)
18. De Vico Fallani, F., Astolfi, L., Cincotti, F., Mattia, D., Marciani, M.G., Gao, S., Salinari, S., Soranzo, R., Colosimo, A., Babiloni, F.: Structure of the cortical networks during successful memory encoding in TV commercials. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology 119(10), 2231–2237 (2008)
19. Aftanas, L.I., Reva, N.V., Varlamov, A.A., Pavlov, S.V., Makhnev, V.P.: Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. Neuroscience and Behavioral Physiology 34(8), 859–867 (2004)
20. Silberstein, R.B., Schier, M.A., Pipingas, A., Ciorciari, J., Wood, S.R., Simpson, D.G.: Steady-state visually evoked potential topography associated with a visual vigilance task. Brain Topography 3(2), 337–347 (1990)
21. Silberstein, R.B., Harris, P.G., Nield, G.A., Pipingas, A.: Frontal steady-state potential changes predict long-term recognition memory performance. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology 39(1), 79–85 (2000)
22. Silberstein: Steady State Visually Evoked Potentials, Brain Resonances and Cognitive Processes. In: Neocortical Dynamics and Human EEG Rhythms. Oxford University Press (1995)

PART IV

Health Informatics

Semantic-Based Conformance Checking of Computer Interpretable Medical Guidelines

M.A. Grando[1], M.H. Schonenberg[2,⋆], and W. van der Aalst[2,⋆]

[1] School of Informatics, Edinburgh University, Edinburgh, U.K.
[2] Department of Mathematics and Computer Science
Technische Universiteit Eindhoven, Eindhoven, The Netherlands

Abstract. Medical recommendations are generally expressed in natural language and therefore their ambiguity can lead to miss interpretations and medical errors. In this paper we propose an approach to 1) disambiguate medical recommendations by specifying them in a declarative language and 2) check the conformance of Computer Interpretable Guidelines (CIGs) with respect to the declarative specification of the medical recommendations. Our approach is based on semantic process mining techniques. Furthermore, we explain two ways to further exploit the information provided by the semantic conformance checker. To increase the accuracy of the model checker we suggest medical scenarios that were not considered for the enactment of the CIG and could show a violation of the medical constraints. Moreover, we discover scenarios which were not covered by the CIG but were considered for the medical recommendations.

1 Introduction

Medical guidelines are used to disseminate the consensus reached on best medical practices. They are paper-based and expressed in natural language, therefore they can be ambiguously interpreted if the medical evidence provided in the guideline is not carefully considered. For some guidelines a general practice is to provide workflow diagrams (networks of tasks and states) that model the care paths explained in the natural language description. In [15] the declarative language CIGDec has been proposed to precisely specify medical recommendations. The CIGDec specification language is supported by DECLARE [18], a constraint-based Workflow Management System. DECLARE offers a graphical notation and a precise formal semantics in Linear Temporal Logic (LTL).

Unfortunately, the translation of medical recommendations into Computer Interpretable Guidelines (CIGs) is not automatic and therefore not necessarily error free. However, some languages for the specification of CIGs, like Asbru [20], have been provided with formal techniques based on theorem proving and model checking for verifying the satisfaction of properties [1] [21] . Moreover, the area of business process mining has contributed with techniques to analyze processes based on their execution history (event logs). Process mining can amongst others be applied to automatically

⋆ The authors would like to thank the LOIS initiative at TU/e and the NWO project "MinAdept" for their support.

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 285–300, 2013.
© Springer-Verlag Berlin Heidelberg 2013

derive process models from the event logs, to check the conformance of the actual execution with a presumed model of the process and to check properties on the execution traces in the log.

More recently more accurate and robust process mining techniques, known as semantic conformance checking, were introduced in [3] [5] to analyze processes by adding semantics to the event logs. Here our aim is to explore the use of semantic conformance checking for verifying if a CIG satisfies/violates the recommendations from the medical guideline on which it is based.

We propose a novel strategy for a further analysis and interpretation of the results generated during semantic conformance checking. The strategy is based on 1) generating the classes of unexplored scenarios that could expose untested cases of violation of medical recommendations with the DECLARE model of medical recommendations, and 2) reusing the scenarios from (1) exposing modeling decisions that reduce the flexibility of the CIG with respect to the medical recommendations on which it is based.

In Section 2 we take as starting point two natural language recommendations from the chronic cough guideline from [9] and we disambiguate them by considering the medical evidence provided in the guideline. Then in Section 3 we explain how to specify the chosen recommendations in the DECLARE framework. While the analysis presented in sections 2 and 3 can not be automatized, the methodology presented in sections 4 and 5 has been implemented in the DECLARE tool and the ProM framework (www.processmining.org). In Section 4 we explain how to use the ProM framework for checking semantic conformance of a CIG with respect to the DECLARE constraints from Section 3. The CIG used in Section 4 is specified using the PRO*forma* [6] language and it has been taken from the Open Clinical repository [16] where other CIGs for the same medical recommendations are available. The PRO*forma*-based CIG has been selected for pragmatic reasons and the explained technique is generic and independent of the language and decision-support system used for the specification of the CIG. In Section 5 we explain some techniques for allowing further analysis of the results obtained from the DECLARE tool and the ProM framework after performing the semantic conformance checking. After mentioning related approaches in Section 6 we finish the conclusions in Section 7.

2 Chronic Cough Guideline Recommendations

In this section we explain our disambiguation of the natural language medical recommendations from the chronic cough guideline [9].

According to [9] if a patient has a cough which last at least 3 weeks the cough is considered chronic. The chronic cough guideline distinguishes 2 different patient classes for which different diagnostic treatments are prescribed in order to discover the most likely cause of cough and treat it. Here we only consider the case of immunocompetent adult patients.

Through out this paper we will consider the following medical recommendations from the chronic cough guideline for inmunocompetent adult patients:

R1) “chest radiographs should be ordered before any therapy is prescribed in nearly all patients with chronic cough. Chest radiographs do not have to be routinely

obtained before beginning treatment for presumed PNDS [post nasal drip syndrome] in young nonsmoker, or in pregnant women, or before observing the result of discontinuation of an ACEI [angiotensin-converting enzyme inhibitor]."

R2) "When the chest X-ray is normal, PNDS, Asthma, and GERD [Gastroesophageal reflux disease] are the likely causes of chronic cough."

We start by explaining our interpretation of recommendations R1) and R2):

R1)(a) **Pregnant Patient or Young Non Smoker with Presumed PNDS.** In the case of pregnant women there is medical evidence of grade II-2 that the *x-ray expose the embryo to radiation*. Evidence of grade II-2 is obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group. Medical evidence of grade II-2 is ranked bellow the evidence of type I (obtained from at least one randomized controlled trial), and bellow the evidence of type II (obtained from well-designed controlled non-randomized trials). This recommendation is *critical* and provided with a high medical evidence, therefore it is also *mandatory* and it should be satisfied in every CIG that models the chronic cough guideline from [9].

In the case of young non smoker with presumed PNDS there is medical evidence of grade II-2 that the probability of PNDS/Asthma/GERD is higher than the average population, therefore it is more *cost-effective* and less time consuming to skip *Chest X-ray*. This recommendation is *not critical* but provided with a high medical evidence and therefore should be *mandatorily enforced*.

R1) (b) **Patients for whom Recommendation R1)(a) does not Apply (Not Pregnant and not Young Non Smokers with Presumed PNDS).** Therefore for this class of patients obtaining a *Chest X-ray* is strongly recommended based on evidence of grade II-2, promoting the values of maximizing *likelihood of diagnosis* and maximizing *cost-effectiveness* because the X-ray may contain results that can aid in making a correct diagnosis. This recommendation is *not critical* but is provided with a high medical evidence and therefore should be *mandatorily enforced*.

R2) For the same reasons the opposite recommendations apply to the treatment of PNDS/Asthma/GERD with the same grade II-2. For pregnant women and for young non-smokers with presumed PNDS this plan is recommended without need to perform first a *Chest X-ray*. Quoting [9] " PNDS either singly or in combination with other conditions, is the most common cause of chronic cough, followed by asthma and GERD" (grade II-2), therefore the multi-treatment is recommended to any patient to maximize the *likelihood of diagnosis*. The treatment of PNDS/Asthma/GERD could be prescribed in the following way:

1) Sequentially treating PNDS, followed by treating asthma and finally treating GERD;

2) Sequentially treating two conditions while concurrently treating the third condition (6 possible combinations);

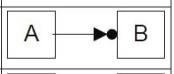
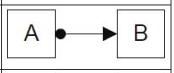
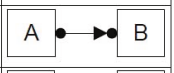
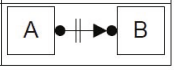
3) Concurrently treating PNDS, asthma and GERD: quoting [9] "Properly chosen empiric therapy for PNDS should start to yield a favorable response within 1 week; for asthma within 1 week; for GERD within 1 week to 3 months" (grade III). Evidence of grade III has less strength than evidence of type II-2 and type II-3 (obtained from multiple time series with or without the interventions). From this and the previous quote from the chronic cough guideline that indicates that PNDS is the most common cause

of chronic cough we can deduce that it is discovered first that PNDS is the reason for chronic cough then all the other treatment can be dropped. The intention of this recommendation is to minimize the *cost of the treatment*.

3 DECLARE Model

In Section 2 we presented our interpretation of recommendations R1) and R2) from the chronic cough guideline [9] for immunocompetent adults. Based on our interpretation we can express those recommendations in the constraint-based Workflow Management System DECLARE [18]. A DECLARE model is defined by a set of constraint templates, each template having a name, a graphical representation and an LTL expression (as shown in Table 1). The provided graphical representation allow the DECLARE users to define very intuitive models which are supported by a very formal and precise semantic in LTL. Here we show how to define a DECLARE model to restrict the way medical actions should be performed in order to comply with the chronic cough medical recommendations explained in Section 2.

Table 1. Relevant CIGDec constraints

Name	LTL expression	Graphical
init(A)	$((A,t_s) \vee (A,t_x))W(A,t_c)$	
existence(A)	$\Diamond(A,t_c)$	
exactly1(A)	$existence(c,A) \wedge !(\Diamond((A,t_c) \wedge \bigcirc(existence(A))))$	
absence2(A)	$!(\Diamond((A,t_c) \wedge \bigcirc(existence(A))))$	
precedence(A,B)	$(!((B,t_s) \vee (B,t_c) \vee (B,t_x)))W(A,t_c)$	
response(A,B)	$\Box((A,t_c) \Rightarrow \Diamond(B,t_c))$	
succession(A,B)	$response(A,B) \wedge precedence(A,B)$	
not succession(A,B)	$not\ response(A,B) \wedge not\ precedence(A,B)$	

In DECLARE constraints can be *mandatory* or *optional*. If a mandatory constraint is violated the DECLARE tool triggers a warning and it explains the violation. Graphically, mandatory constraints are depicted as solid lines and optional constraint as dashed lines.

Data attributes can be specified and associated to relevant tasks, e.g. the X-ray task reads the pregnancy status of a patient. While executing a task, its data attributes can be read or written, as specified for that task at design-time.

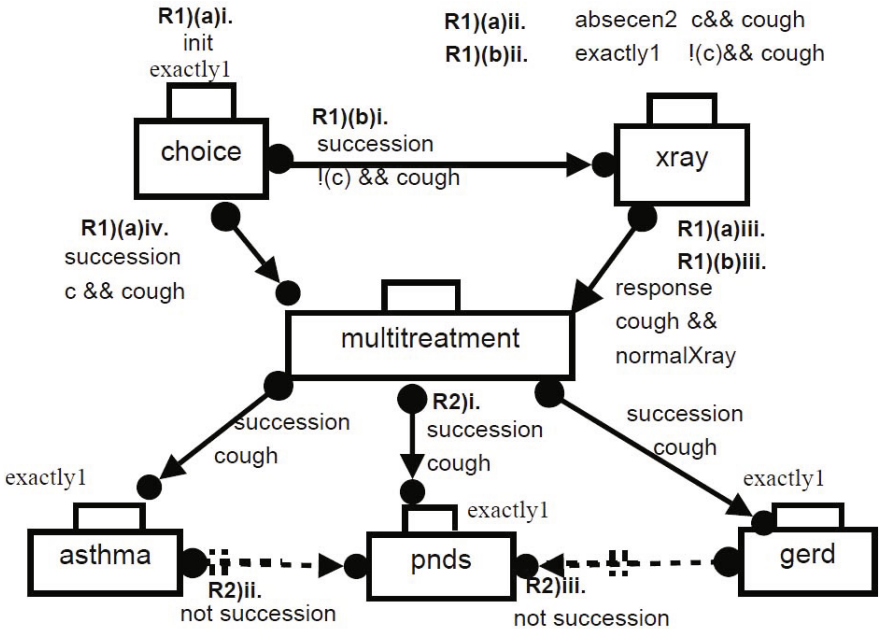


Fig. 1. DECLARE model for the considered recommendations from the chronic cough guideline

Constraints can be *conditional*. A boolean expression can be associated to a constraint, such that if the condition is true then the constraint should be satisfied. While the condition is false the constraint is inactive. The boolean expression can be defined in terms of data attributes. For example an X-ray should be performed only if the patient is not pregnant. Graphically we represent conditions between brackets.

The DECLARE specifications explain bellow are summarized in the first three rows of Table 2.

Before we describe how to model the recommendations from the chronic cough guideline (Section 2) in DECLARE, we need to explain the following general recommendations for the chronic cough guideline, which should always be satisfied:

RG1) The *choice* of ordering a chest X-ray is made once (constraint exactly1 over task *choice*);

RG2) Performing the treatment of asthma, GERD or PNDS more than once will not change the diagnosis therefore the constraint exactly1 is associated to the tasks *multitreatment*, *asthma*, *GERD* and *PNDS*.

Next, we explain how to model the recommendations from the chronic cough guideline in DECLARE:

R1) (a) **Pregnant patient or young smoker with presumed PNDS:** for patients in this class who have chronic cough $(c \;\&\&\; cough)$ it is recommended to first optionally performing a chest X-ray and then if the cough persists treat for PNDS/Asthma/GERD:

i. decide first if a *X-ray* has to be taken (init constraint over task *choice*);
ii. the task *X-ray* becomes optional, depending on the decision taken in R1)(a)i. (conditional absence2 constraint in task *xray*);

iii. start the treatment for PNDS/Asthma/GERD if after the *X-ray* the patient has persistent cough and the result of the *X-ray* is normal $(cough \;\&\&\; normalXray)$ (conditional response relation between tasks *xray* and *multitreatment*);
iv. decide not to perform the *Chest X-ray* and immediately start the treatment of PNDS/Asthma/GERD (conditional succession between tasks *choice* and *multitreatment*). In this case the enactment of *xray* task is not possible anymore because it would violate the precedence constraint between tasks *choice* and *xray*.

R1)(b) **Patients for whom Recommendation R1)(a) does not Apply.** For patients in this class who have chronic cough $(!(c) \;\&\&\; cough)$ a *Chest X-ray* should be taken and then in case of normal X-ray irritants should be avoided. If after avoiding irritants the cough persist then the PNDS/Asthma/GERD treatment should be started:

i. a chest X-ray should be taken (conditional succession constraint between tasks *choice* and *xray*).
ii. exactly one chest X-ray has to be done (conditional exactly1 constraint on task *xray*);
iii. similarly to R1) (a) iii. if the result of the X-ray is normal and the patient still has a cough $(cough \;\&\&\; normalXray)$ then the multi-treatment of PNDS/Asthma/GER D has to be started (conditional response constraint between tasks *xray* and *multitreatment*).

R2) Instead of specifying all possible combinations of treatments, for PNDS, asthma and GERD where the treatment of PNDS should not preceded by the other treatments, it is enough to define the preferred scenarios by conditional constraints, that hold in case condition *cough* is true:

i. a mandatory relation of conditional succession between the task *multitreatment* and each of the treatments;
ii. an optional negative relationship of succession between *asthma* treatment and *PND S* treatment, so before treating *PNDS* the patient cannot be treated for *asthma* and after treating for *asthma* the patient cannot be treated for *PNDS*.
iii. similarly to R2)ii. define an optional negative succession constraint between *GERD* and *PNDS*.

Here we suggest simple questions that can help to determine the level of compliance of a medical recommendation based on its attached values:

1) Which properties are related to the patient's safeness and therefore are critical to verify?

For the case of pregnant women the recommendation R1)(a) of not performing a chest X-ray has associated the value of minimizing the *risk of damaging the embryo*. This value is critical and therefore this property should be satisfied in every CIG that models the chronic cough guideline from [9].

2) Which properties are not critical but are provided with a high medical evidence and therefore should be mandatorily enforced?

Table 2. Summary of the DECLARE constraints from Figure 1 and the results obtained after checking their conformance in the Tallis CIG for chronic cough

Medical recommendation	**DECLARE constraint**	**Natural Language**	**Semantic checking**
RG1) exactly1(choice)	Non critical mandatory	Only once it is possible to decide if a X-ray is performed	+
RG2) exactly1(asthma)	Non critical mandatory	The diagnosis of asthma as cause of cough is evaluated only once	+/-
RG2) exactly1(GERD)	Non critical mandatory	The diagnosis of gerd as cause of cough is evaluated only once	+/-
RG2) exactly1(PNDS)	Non critical mandatory	The diagnosis of PNDS as cause of cough is evaluated only once	+/-
RG2) exactly1(multitreatment)	Non critical mandatory	The diagnosis by multi-treatment is evaluated only once	+
R1)(a)i. init(choice)	Non critical mandatory	Initially for any patient decide if X-ray should be made	+
R1) (a)ii. absence2(xray)	Critical mandatory	For pregnant women or young non smoker with presumed PNDS X-ray is optional	+
R1) (a)iii., R1) (b) iii. response(xray,multitreatment)	Non critical mandatory	After X-ray multi-treatment should be started	+
R1) (a)iv. succ(choice, multitreatment)	Non critical mandatory	For pregnant women or young non smoker X-ray can be skipped	+
R1)(b)i. succ(choice,xray)	Non critical mandatory	When recommendation R1)(a) does not apply, an X-ray should be made	+
R1) (b)ii. exactly1(xray)	Non critical mandatory	When recommendation R1)(a) does not apply, exactly one X-ray should be taken	+
R 2)i. succ(mutitreatment, asthma) succ(mutitreatment, PNDS) succ(mutitreatment, GERD)	Non critical mandatory	The muti-treatment consists on the treatment of asthma, GERD and PNDS	+/-
R 2) ii. not succ(asthma, PNDS) not succ(GERD, PNDS)	Non critical optional	PNDS can not be preceded by Asthma or GERD	+

For instance the following recommendations supported by medical evidence of grade II-2 are critical:

a) performing multi-treatment for pregnant women in recommendation R1) (a); b) performing a chest X-ray and a multi-treatment for asthma/GERD/PNDS for the case of young non smoker with presumed PNDS in recommendation R1) (a) and c) performing a chest X-ray and a multi-treatment to patients for whom the recommendation R1) (a) does not apply.

3) Which properties are related to the efficient use of resources (time, money, medical staff, etcetera) and can mainly suggest optional improvements in the quality of medical treatment?

This is the case of the recommendation R2) (medical evidence of grade III) that suggests that the treatment of PNDS should not be preceded by the treatment of Asthma and GERD in order to minimize the *cost of the treatment*.

4 Semantic Model Checking

The methodology proposed here to check the conformance in a CIG of the DECLARE specification of medical recommendations is illustrated by Figure 2 and explained in detail below.

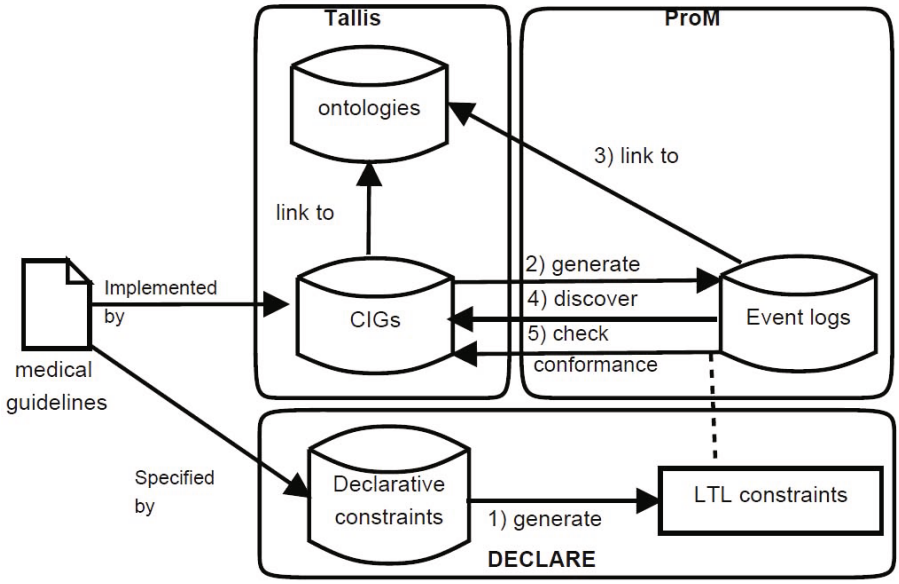


Fig. 2. Steps required by the methodology proposed here: 1) Generate LTL constraints from the DECLARE model, 2) Generate logs by enacting the CIG in the Tallis engine, 3) Link ontologies, 4) Discover the model mined from the event logs using ProM, 5) Check conformance using ProM

1) Generate LTL properties from the DECLARE model: the DECLARE tool automatically generates the LTL properties from the constraint model of the medical recommendations explained in Section 3.

2) Enact the CIG to generate event logs: in Mor Peleg et al. work [17] the developers of Asbru [20], GLIF, GUIDE, EON and PRO*forma* [6] languages were asked to specify CIGs for similar recommendations to the ones we explained in Section 3. But the developers of the CIGs did not have any access to the chronic cough medical guideline from [9]. The recommendations on which the study [17] was based on and the repository of the resulting CIGs is available at the Open Clinical repository [16]. None of the CIGs from the repository have been used in any real medical environment. Therefore, we decided to explain the methodology with the PRO*forma*-based CIG from the Open Clinical repository which has been enacted using the Tallis engine

(http://www.cossac.org/tallis). But the strategy explained here requires the execution history (event logs) of the CIG and therefore it is independent of the language used for the specification of the CIG. An event log contains the executions of one or more processes. To construct such log is the required that each event in the log (e.g. an X-ray) can be mapped to a single case or process instance (e.g. a patient treated for cough) and that each process instance can be mapped to a single process (e.g. the process for treating chronical cough). The more information is available in the log, the better the quality of the results and the larger the number of questions that can be answered. For instance, event time stamps can be used to do performance analysis. Similarly, every process instance has zero or more tasks. Every task or audit trail entry must have at least a name and an event type. The event type determines the state of the tasks. Timestamps can be used for capturing timing information and to analyze performance related aspect. The originator element records the person/system that performs the task. Because the Tallis CIG specified for the chronic cough recommendations has not been used in a real medical environment, the event logs have to be generated considering generic patient cases. According to the DECLARE constraints from Section 3 six different patient medical conditions are significant: (1) if the cough is persistent, (2) if the result of the X-ray is normal, (3) if the patient is pregnant, (4) if the patient is less than 18 years old, (5) if the patient is a smoker, (6) if PNDS is presumed. But the enactment of the Tallis CIG showed that the developers of the CIG also considered two additional medical conditions: (7) if the patient has ACE-related cough and (8) if the cough is productive. Therefore firstly generic patient cases were created based on all the possible combinations of the eight mentioned medical conditions. Secondly the generic patient cases were used to enact the Tallis CIG. This is a standard practice in software engineering when system's correctness needs to be tested before the system's release. Finally the events generated during the CIG enactment were recorded as event logs using the MXML (Mining eXtensible Markup Language) format. The schema for the MXML format is available at www.processmining.org.

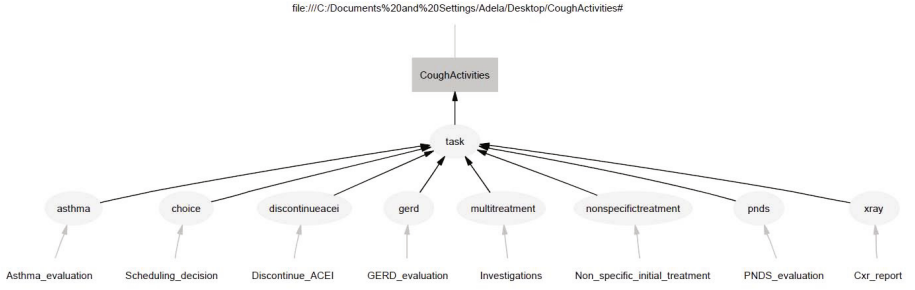


Fig. 3. Ontology of activities, where the leafs are the annotated Tallis activities

3) Link ontologies by semantically annotating the event logs generated by the CIG: this requires mapping the concepts used in the CIG with the concepts from the DECLARE specification. For instance as Figure 3 shows the tasks *Scheduling_decision* and *Cxr_report* from the Tallis implementation presented here can be mapped into the semantically equivalent concepts *choice*, *xray* from the DECLARE model from Section 3.

Only after mapping concepts it is possible to verify if the DECLARE constraints are satisfied in the mined model. For example the DECLARE constraint that specifies that task *choice* has to be enacted exactly once is verified in the model mined from the Tallis implementation as the constraint that the Tallis task *Scheduling_decision* has to be enacted exactly once.

The same DECLARE model used to perform the conformance checking of the Tallis implementation can be reused to check the conformance of any other implementation from the Open Clinical repository on which Mor Peleg et al. work [17] is based on. The only requirement is that previously the corresponding mapping between the DECLARE model and the new implementation has been done.

In particular from the DECLARE model from Section 3 we can create ontologies for data, activities, event types, process instances and originators (actors who enact activities). We have called the ontologies of data and activities *CoughData* and *CoughActivities* respectively. Figure 3 shows the graphical representation of the ontology *CoughActivites*, where the leafs correspond to the annotated activity instances from the Tallis CIG. To perform the ontology linking the MXML file generated in 2) is annotated with concepts from these created ontologies. The resulting semantic annotation is expressed in the SA-MXML (Semantically Annotated MXML) format used by the ProM framework. The SA-MXML format is available at www.processmining.org and it is an extension of the MXML formant where all elements (except for audit trail entries and time stamps) have an optional extra attribute that links to a list of concepts in the ontologies. For instance in the SA-MXML log shown in Figure 4 for the process instance with identifier "Case0100a" the variables *young* and *old* from the DECLARE ontology *CoughData* are linked by an attribute to the variables *younger* and *older* from the data ontology in the Tallis CIG.

4) Discover the PRO*forma* model from the semantically annotated event logs: using the alpha ++ algorithm that is provided as a plug-in of the ProM framework it is possible to extract (mine) the PRO*forma* model based on the dependency relations that could be inferred among the activities in the Tallis event logs from the SA-MXML file generated in 3). Figure 5 shows the resulting mined PRO*forma* process for generated event logs for the 64 combinations of medical conditions.

5) Perform semantic conformance checking of the discovered PRO*forma* model: using the semantic LTL checker plug-in from the ProM framework it is possible to perform semantic conformance checking of the PRO*forma* semantically annotated model discovered in 4) and the DECLARE model from Section 3.

In Section 3 we explained that it is possible to differentiate in DECLARE between medical recommendations that are mandatory or optional. Only if a DECLARE constraint is mandatory and it is not satisfied in all the event logs generated from the CIG, a warning should be given during conformance checking. Otherwise if the DECLARE constraint is optional and not satisfied in all the generated event logs it is considered that the CIG complies with the recommendation though it is not optimal. The level of satisfaction of a property can be seen as the percentage of event logs that satisfy the property. This information is provided by the semantic LTL checker as two lists: the first list corresponds to the event logs that satisfy the property and the second list contains the event logs that do not satisfy that property.

```
        <Source program="Rapid Synthesizer" >
                <Data>
                        <Attribute name="program" >Rapid Synthesizer</Attribute>
                </Data>
        </Source>
        <Process id="log2" description=""
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/ProcessOntology.wsml#Proce
ss">
<ProcessInstance id="Case0100a" description=""
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/ProcessInstanceOntology.ws
ml#ProcessInstance" >
<Data>
<Attribute name="Pregnant"
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughData.wsml#Pregnant">
no  </Attribute>
<Attribute name="Smoker"
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughData.wsml#Smoker"> no
</Attribute>
<Attribute name="Younger"
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughData.wsml#Young"> yes
</Attribute>
<Attribute name="Older"
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughData.wsml#Old"> no
</Attribute>
<Attribute name="PressumedPNDS"
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughData.wsml#PressumedPN
DS"> no </Attribute>
</Data>
<AuditTrailEntry>
        <WorkflowModelElement
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughActivities.wsml#choic
e">Scheduling_decision</WorkflowModelElement>
                                <EventType
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/Events_Ontology.wsml#Compl
ete" >Complete</EventType>
                                <Timestamp>2010-04-17T09:00:02.968+00:00</Timestamp>
                                <Originator>UNDEFINED</Originator>
</AuditTrailEntry>
<AuditTrailEntry >
        <WorkflowModelElement
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughActivities.wsml#nonsp
ecifictreatment">Non_specific_initial_treatment</WorkflowModelElement>
                                <EventType
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/Events_Ontology.wsml#Compl
ete" >Complete</EventType>
                                <Timestamp>2010-04-17T09:00:06.968+00:00</Timestamp>
                                <Originator>UNDEFINED</Originator>
</AuditTrailEntry>
        <AuditTrailEntry >
<WorkflowModelElement
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/CoughActivities.wsml#xray"
>Cxr_report</WorkflowModelElement>
                                <EventType
modelReference="file:///C:/Documents%20and%20Settings/Adela/Desktop/Events_Ontology.wsml#Compl
ete" >Complete</EventType>
```

Fig. 4. SA-MXML file of annotated event logs from the Tallis CIG for chronic cough

In the last row of Table 2 we show the results of performing semantic conformance checking as explained in Section 4 over the Tallis CIG. A property can be unsatisfied (-), fully satisfied (+) or partially satisfied (+/-).

For instance the recommendations R1)(a)ii. and R1)(a)vi. for the case of pregnant women are related to the decision of performing a chest X-ray and they have associated the value of minimizing the *risk of damaging the embryo*. This value is critical and therefore was modeled in Section 4 as mandatory DECLARE constraints. We have shown that all the event logs that we generated with the Tallis CIG satisfy these properties.

We did not find any property that is fully unsatisfied by the generated event logs. For fully unsatisfied properties the semantic LTL model checker returns an empty list of event logs satisfying the property.

For the properties that we showed to be partially satisfied by the Tallis CIG some of the generated traces of the event logs are in the first list generated by the semantic LTL checker (traces satisfying the property) and some others are in the second list (traces

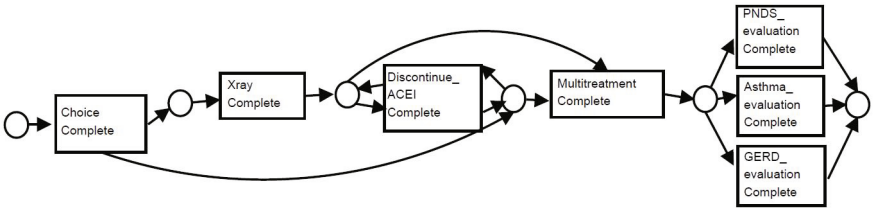


Fig. 5. Model of the PRO*forma*-based CIG discovered by the ProM framework from the Tallis event logs

not satisfying the property). For example this was the case of the non critical mandatory chronic cough medical recommendations R2)i. which specifies that the diagnosis by multi-treatment consists on evaluating for Asthma, GERD and PNDS. Therefore a notification should be given in order to inform that this recommendation has been violated though it was mandatory. From the partial satisfaction of this property and from the observation of the event logs contained in each list computed by the semantic LTL model checker we can infer that the PRO*forma*-based CIG was designed to ask the user to choose to evaluate only one of the mentioned possible causes of chronic cough.

5 Scenario-Based Analysis

Semantic conformance checking allows to automatically show that certain properties are not satisfied by a CIG. Here we propose a novel methodology to allow further automatic analysis over a CIG once the semantic conformance checking was performed as explained in Section 4. Our methodology takes advantage of: 1) the DECLARE specification of the medical recommendation on which the semantic conformance checking is based on, and 2) results computed by the semantic LTL checker enacted in the ProM framework.

The methodology is based on the scenarios generated by the DECLARE model which can be used to 1) suggest event logs corresponding to unexplored critical behaviors of the CIGs, and 2) analyze if the CIG is more or less restricted than the medical recommendations that were used as starting point.

DECLARE models can be used to suggest significant testing scenarios because they are provided with algorithms to generate: the finite automata that corresponds to the scenarios that do not violate any constraint, and the finite automata that represents all the possible scenarios that violate same of the model constraints (see [19]).

The information provided by the automaton could be used to:

1) *Suggest classes of scenarios that comply with/violate the medical recommendations.* For instance these scenarios can be used to generate test cases to verify properties that are satisfied because no event log could be use to check its satisfaction.

2)*Provide classes of scenarios that are generated by the DECLARE model and therefore comply with the medical recommendations but can not be generated by the CIG.* The generated scenarios can be used to verify if the CIG is more restricted than the medical recommendations on which it is based. For instance we can construct the automata

that generates all the scenarios which do not violate the constraints defined by the DECLARE model from Figure 1. According to this automata there is a direct transition between the activities *choice* and *multitreatment* for the class of patients corresponding to pregnant women or non-smokers with presumed PNDS. This means that it is possible to chose to start the muti-treatment of asthma/PNDS/GERD without need to perform before a chest X-ray. So in the mined PRO*forma* model (Figure 5) there should be a direct transition between the semantically equivalent activities *Scheduling_decision* and *Investigations*. But this is not the case, the enactment of activity *Scheduling_decision* is only possible if the activity *Cxr_report* has been enacted before. From this analysis we deduce that the mined PRO*forma* model is more restricted that the medical recommendations from the DECLARE model: for the class of pregnant women or non-smokers with presumed PNDS the mined PRO*forma* model obliges to perform a X-ray, instead of offering it as an option.

6 Related Work

In this paper, we focussed on the problem of checking a posteriori whether a CIG satisfies a set of medical recommendations specified in a declarative formal notation called DECLARE. Another a posteriori approach is presented in [10] where the DECLARE constraints satisfied by a mined process can be discovered. In [13] Mans et al. have addressed this problem a priori by first specifying the medical recommendations as a Coloured Petri Net (CPN) and then mapping the CPN into CIGs specified in different workflow-based language. It seems that using this mechanism it should be possible to obtain CIGs that are closer to the behavior described by the medical recommendations modeled by the CPN, but there is no claim in [13] that the obtained CIG preserves all the behavior from the CPN used as starting point.

The idea of checking properties in medical applications is not totally new. In [21] a theorem proving approach was proposed to verify Asbru guidelines by mapping then to KIV, and interactive theorem prover for higher order logic. In [7] a model checking approach was proposed to verify GLARE guidelines by translating them to the Promela language of the model checker SPIN. The properties were formulated as LTL formulas. Also in [1] a model checking approach was proposed to verify Asbru guidelines using the SMV checker. The properties were formalized as ACTL formulas.

By theorem proving and model checking it is possible to guaranty the absence of errors, though these strategies are highly costly because they require to provide a mapping of the language on which the CIG is implement into a language which allows theorem proving or model checking. In contrast an advantage of the semantic conformance checking based approach presented here is that it only requires the process history and therefore it can be applied over any CIG independently of the language used for its implementation. This feature is very important, considering that currently there are numerous incompatible languages available for the specification of medical guidelines [17]. Nevertheless, using the LTL constraints that can be generated by DECLARE also conventional model checking could be used without needing event logs.

The idea of using semantic process mining techniques to analyze healthcare processes is not new. In [11] [12] Mans et al. extract relevant event logs from real applications running in hospitals' information systems and they analyze these logs using the

ProM framework. Their results show that process mining can be use to provide new insights that facilitate the improvement of existing careflows. Their work differs from us on the fact that they do not use semantic conformance checking to verify if the medical recommendations on which the care process is based are satisfied/violated.

7 Conclusions

In this paper we proposed a technique for checking the compliance of CIGs with respect to medical recommendations specified using a precise declarative language. The technique is based on the execution history of the CIGs and therefore can be applied at post-execution time. While the semantic-based process mining technique described here is generic, we have explained it with the Tallis CIG from the Open Clinical repository [16].

There are various reasons why we think that in a near future it is possible that the Health Informatics community could benefit from semantic conformance-based approaches as the one we present here. First of all, the use of specification languages as DECLARE open the possibility that medical recommendations become available, besides its paper version, as formal models based on standard medical ontologies like Systematized Nomenclature of Medicine (SNOMED) or the Unified Medical Language System (UMLS). The use of standard ontologies could simplify the reusability of the methodology explained here. Recently, we have applied the techniques introduced in this paper, explained in Section 4, over the the GLIF CIG for chronic cough from the same Open Clinical repository [16] from where the PROforma CIG used here was extracted. Reusing the explained techniques was possible because a suitable ontology mapping between the terms from the GLIF CIG and the DECLARE ontology could be defined. The results presented here for the PRO*forma*-based CIG and our further analysis of the GLIF chronic cough CIG show that it is possible to check the conformance of CIGs defined in different languages with respect to the same DECLARE specification of medical recommendations. The reusability of the DECLARE specification is possible if an ontology mapping between the terms used in the DECLARE model and the concepts used for an arbitrary CIG can be defined.

Secondly, the use of PAIS [4] which automatically collect event logs generated by running business applications is growing and could be adopted in a near future in clinical environments. PAIS have been successfully used before for analyzing non-trivial careflow processes, where the patient's real data has been previously anonymised. For instance in [12] they have used PAIS to analyze a group of 267 gynecological oncology patients treated in 2005 and 2006 in the AMC hospital in Amsterdam, covering diagnostic and treatment activities. Because the Tallis CIG that we have use here as case study has not been tested in a real medical environment we have manually generated the event logs. But by means of PAIS the event logs could be automatically generated from real medical applications based on the Tallis CIG from the Open Clinical repository.

In a near the future we are also interested on considering the incorporation of the scenario-based information provided by the DECLARE models for the run-time prevention and detection of violations to medical recommendations and policies. Taking in to account that existing medical recommendations are periodically revised/updated and

new medical recommendations can be added, this is an important problem to consider. Some formal frameworks like [8] are available in the field of health informatics for specifying exception managers to detect and recover from undesirable states happening during the enactment of CIGs. But so far no strategy has been proposed to automatically suggest the scenarios or states that violate the medical policies and recommendations. As shown in Section 5, from DECLARE models it is possible to automatically generate the automaton that describes the scenarios that violate the model constraints.

References

1. Bäumler, S., Balser, M., Dunets, A., Reif, W., Schmitt, J.: Verification of Medical Guidelines by Model Checking – A Case Study. In: Valmari, A. (ed.) SPIN 2006. LNCS, vol. 3925, pp. 219–233. Springer, Heidelberg (2006)
2. Black, E., Glasspool, D., Grando, M.A., Patkar, V., Fox, J.: Goal-Based Decisions for Dynamic Planning. In: AIME Conference, pp. 96–100 (2009)
3. Casati, F., Shan, M.C.: Semantic Analysis of Business Process Executions. In: Jensen, C.S., Jeffery, K., Pokorný, J., Šaltenis, S., Bertino, E., Böhm, K., Jarke, M. (eds.) EDBT 2002. LNCS, vol. 2287, pp. 287–296. Springer, Heidelberg (2002)
4. Dumas, M., van der Aalst, W.M.P., ter Hofstede, A.H.: Process-Aware Information systems: Bridging People and Software Through Process Technology. Wiley (2005)
5. El Kharbili, M., Stein, S.: Policy-Based Semantic Compliance Checking for Business Process Management. In: MobIS Workshops, pp. 178–192 (2008)
6. Fox, J., Johns, N., Lyons, C., Rahmanzadeh, A., Thomson, R., Wilson, P.: PROforma: a general technology for clinical decision support systems. Computer Methods and Programs in Biomedicine 54(1-2), 59–67 (1997)
7. Giordano, L., Terenziani, P., Bottrighi, A., Montani, S., Donzella, L.: Model checking for clinical guidelines: an agent-based approach. AMIA, 289–293 (2006)
8. Grando, A., Peleg, M., Glasspool, D.: A goal-oriented framework for specifying clinical guidelines and handling medical errors. J. Biomed. Inform. 43(2), 287–299 (2010)
9. Irwin, R.S., Boulet, L.S., Cloutier, M.M., et al.: Managing Cough as a Defense Mechanism and as a Symptom, A Consensus Panel Report of the American College of Chest Physicians. Chest 114(2), 133–181 (1998)
10. Maggi, F., Mooij, A., van der Aalst, W.P.M.: User-guided discovery of declarative process models. In: IEEE Symposium Series in Computational Intelligence 2011 - CIDM - 2011 IEEE Symposium on Computational Intelligence and Data Mining (2011)
11. Mans, R.S., Schonenberg, H., Leonardi, G., Panzarasa, S., Cavallini, A., Quaglini, S., van der Aalst, W.M.P.: Process Mining Techniques: an Application to Stroke Care. MIE, 573–578 (2008)
12. Mans, R.S., Schonenberg, H., Song, M., van der Aalst, W.M.P., Bakker, P.J.M.: Application of Process Mining in Healthcare - A Case Study in a Dutch Hospital, BIOSTEC Conference (Selected Papers), pp. 425–438 (2008)
13. Mans, R.S., van der Aalst, W.M.P., Russell, N.C., Bakker, P.J.M., Moleman, A.J., Lassen, K.B., Jørgensen, J.B.: From Requirements via Colored Workflow Nets to an Implementation in Several Workflow Systems. In: Jensen, K., Billington, J., Koutny, M. (eds.) ToPNoC III. LNCS, vol. 5800, pp. 25–49. Springer, Heidelberg (2009)
14. McCartney, J.J.: Values Based Decision making in Helathcare: Introduction. Journal HEC Forum 17(1), 1–5 (2005)
15. Mulyar, N., Pesic, M., van der Aalst, W.M.P., Peleg, M.: Declarative and Procedural Approaches for Modelling Clinical Guidelines: Addressing Flexibility Issues. In: Informal Proceedings of the International Workshop ProHealth 2007, pp. 17–28 (2007)

16. Open Clinical repository, http://www.openclinical.org (March 2011)
17. Peleg, M., Tu, S.W., Bury, J., Ciccarese, P., Fox, J., et al.: Comparing Computer-Interpretable Guideline Models: A Case-Study Approach. J. Am. Med. Inform. Assoc. 10(1), 52–68 (2003)
18. Pesic, M., van der Aalst, W.M.P.: A Declarative Approach for Flexible Business Processes Management. In: Eder, J., Dustdar, S. (eds.) BPM Workshops 2006. LNCS, vol. 4103, pp. 169–180. Springer, Heidelberg (2006)
19. Pesic, M.: Constraint-based workflow management systems: Shifting controls to users. Beta Research School for Operations Management and Logistics, Eindhoven (2008)
20. Shahar, Y., Miksch, S., Johnson, P.: The Asgaard Project: A Task-Specific Framework for the Application and Critiquing of Time-Oriented Clinical Guidelines. Artificial Intelligence in Medicine 14(1-2), 29–51 (1998)
21. ten Teije, A., Marcos, M., Blaser, M., et al.: Improving medical protocols by formal methods. AI in Medicine (2006)

REST-Style Architecture and the Development of Mobile Health Care Solutions

François Andry, Lin Wan, and Daren Nicholson

Axolotl Corp., 160 West Santa Clara Street, Suite 1000, San Jose, CA 95113, U.S.A.
{fandry,lwan,dnicholson}@axolotl.com

Abstract. The growing demand for mobile applications in the health care industry is changing the way software solutions are created, deployed, tested and maintained. The REST-Style architecture is particularly well suitable to quickly integrate legacy systems and offer new highly scalable and secure APIs that can be used as gateways to health care data and services. This paper describes the motivations and technical solutions chosen to build a REST API that can be easily integrated with mobile applications. As a result, care givers can have better access to patient health records from local, regional or state wide Health Information Exchanges (HIE).

Keywords: Mobile application, Health records, RHIO, Lab results, CCD, HL7, REST API, Web services, JAX-RS, Security, Software development process.

1 Introduction

In the ambulatory health care environment, providers spend the majority of their time in an examination room with patients. Although some clinics have installed personal computers in the exam room for use at the point of care, many physician practices have yet to do so or have no such intention. Reasons for not installing PCs in the exam room include (among others) lack of space, security concerns, and cost. Often, clinics have PCs installed outside of the exam room to be used for encounter documentation or health history research (i.e., reviewing the patient's health records).

This physical setup is often satisfactory for providers to complete their documentation needs. Providers often scratch rough notes on paper during an encounter, then dictate or type their notes after the visit has ended. The absence of computers in the exam room, however, is a disadvantage for research activities. Frequently, after listening to the patient's verbal health history, a provider wishes to read past records. If those records are in an electronic format, it is optimal to access those records at the point of care (i.e., in the exam room) [10]. Thus, computer devices that are smaller and more mobile than a PC (e.g., smart phones, PDAs, tablets) would be the optimal hardware choice to access these electronic records [4], [9]. Given that many physicians carry smart phones, such mobile devices would be the ultimate tools to look up patient records [6], [11].

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 301–311, 2013.
© Springer-Verlag Berlin Heidelberg 2013

1.1 APIs for Mobile Applications

Over the years, our group has developed advanced Clinical Networking™ solutions for hospitals, RHIOs and state-wide health information exchanges (HIE). We have created a backend infrastructure accessible for integration as software-as-a-service (SaaS), most of them relying on Simple Object Access Protocol (SOAP) APIs. One of the issues we have encountered when creating an architecture for mobile applications, was the type of API to use between backend servers and the new mobile clients.

In this paper, we explain the challenges [7] faced when creating a new high performance API accessible by mobile health care applications for physicians and medical personnel. We describe the technical choices that we have made to simplify and to speed up the development, management, extension and maintainability of the features involved in such applications and the associated service layer.

2 VHR Mobile Application

The initial mobile application that we built on top of the Elysium Virtual Health Record (VHR) infrastructure only displays lab results. Other types of data (vital signs, radiology and transcribed reports, medications, etc.), which are currently accessible via a web-based application, will be offered at a later time. The usage scenario for the mobile application is as follows:

1. Physician logs in using her/his credentials.
2. After successfully logging in, the physician is presented with a list of patients who have consented to allow the physican to view their patient records. In addition, there is also a search box that will allow the physician to manually search for patients.
3. The physician chooses the patient of interest on the patient list (or search results).
4. Once the patient is selected, the physician sees a list of available lab results for that patient.
5. The physician selects a lab result of interest by tapping it. This opens the lab result and displays all of its details in a new screen, with the option of going back to the previous screen to retrieve another lab result.

2.1 Security Concerns

To minimize security risks and to comply with HIPAA [3] security regulations, we chose not to store any patient, login or password data on the mobile device, thereby preventing non-authorized users from gaining any access to a patient's personal health information (PHI).

Access to the patients' health records are provided after the user (typically a doctor) enters his or her login and password on the client application. These are sent to the server via an application programming interface (API). After authentication is successful, the server sends back a security token that the client must reuse for each subsequent request to the server API. Initially all users (physicians) have the same access control to the data as long as the patients have given consent to specific physician practices. As a result, the authorization function is very simple.

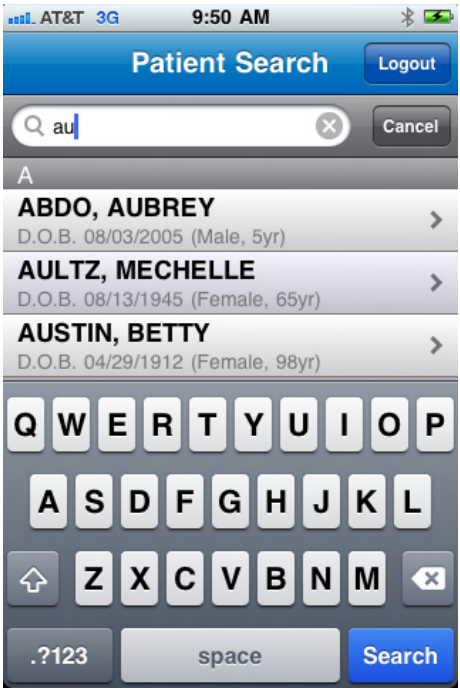


Fig. 1. Patient search (iPhone application)

2.2 Client Platforms

The initial platforms chosen to deploy our mobile applications are the iPhone and iPad. Additional platforms (Google Android OS, RIM BlackBerry OS, Windows Mobile OS) are planned for future releases.

The process of building, deploying and maintaining stand-alone applications for mobile platforms is different from building browser-based web applications. Each platform has its own design, development and deployment process and tools.

Since the development of client applications on different mobile platforms requires more time than creating web applications for a handful of browsers, it is important to minimize the complexity of the integration with the back-end services and to try to decouple the development and maintenance of the client- and server-side components.

3 Backend Infrastructure

For over 15 years, Axolotl (now part of Ingenix) has developed and offered web-based applications to access patient records. Until recently, these web-application components were either tightly coupled with the backend data sources or were accessing data through Simple Object Access Protocol (SOAP) based web services. In health care, SOAP is often the top choice when it comes to web services. The advantages of SOAP are:

- Type checking (via WSDL files)
- Availability of development tools
- Suitability for securely transferring large numbers of data.

Even though it was tempting for us to reuse the existing SOAP APIs to leverage the existing infrastructure and avoid new development on the server side, we decided to create a more appropriate architecture that would be easier to integrate and extend as more types of data are added to the mobile health care applications.

4 VHR REST API

The Representational State Transfer (REST) architecture [1], [2] is an alternative to SOAP and offers certain characteristics that were relevant to our use case:

- Lightweight and easy to build
- Extensible
- Scalable
- Easy to debug (human readable results).

SOAP based Web Services offer no separation of concern, since the supported programming model does not separate network centric operations from local operations [5]. On the other hand with REST APIs, the separation of concern is clear. All resources are accessible using the same protocol (HTTP). Also with SOAP, interoperability problems can occur when native types are present in the interface, whereas REST is simpler as it completely constrains the set of operations [8].

4.1 API Design and Development

With REST we are able to dynamically build unique URLs to represent remote health records objects as needed. The mobile application sends HTTP requests over Secure Sockets Layer (SSL) to obtain a JavaScript Object Notation (JSON) of patient demographic info (e.g. DOB, name, etc) or health records. JSON was chosen because, as a compact data format, it offers better performance compared to the complexity of an XML representation. In addition to this, requests coming back from the REST API are compressed using GZIP, which further improves the performance between the server and the client.

The API has been built using JBOSS JAX-RS RestEasy library, which leverages Java annotations, making the definition of the resource mapping easy to do in a very declarative manner. We use the Jackson JSON/Java library to marshal and unmarshal java objects into JSON format. The integration of the API layer with the backend data source, as well as other services such as configuration, is done using the Java Spring framework.

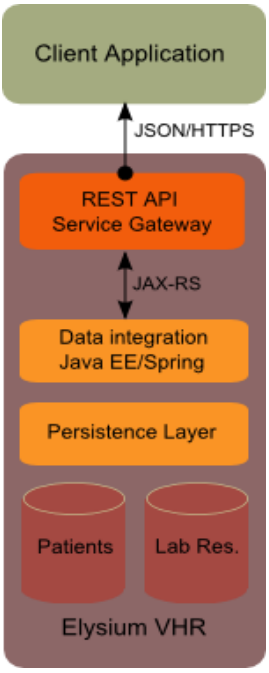


Fig. 2. Integrated Mobile/REST API architecture

To define the API dependencies, modules, libraries and to build the web application archive (war), we use Apache Maven 2. Throughout the development cycle, we have used a continuous integration process with unit tests (JUnit).

The overall project has been managed through an agile/SCRUM process, with iterative, incremental development sprints, each lasting three weeks.

4.2 API Operations

With REST, the identification of the resources (user, patient, lab results, etc.) is straightforward. There is no need for the client to create complex request envelopes to query the server.

The initial features of the mobile client were to access the health record in a read-only mode; no data is created or modified. As a result, most of the operations use GET HTTP methods, with some restrictions for certain operations, where parts of the patient's protected health information (PHI) may be sent to the server. To be compliant with HIPAA [3] security regulations, a POST method is used and the sensitive information (e.g., a query on a patient name) is passed in the body of the request as a JSON object over an SSL connection.

Table 1. Health Record API operations subset

POST /users/tokens	authentication for a user
GET /users/{user-id}/patients	retrieve patients
GET or POST /users/{user-id}/patients/{patient-id}	profile of a specific patient
GET /users/{user-id}/patients/{patient-id}/lab-results	lab-results of a patient

All request and response headers have a content type "application/json," which means that complex queries and responses are in the form of JSON arrays or JSON objects. In addition to this, the requests coming from the client need to provide the security token provided initially by the server as a cookie in the request header.

Because the REST API is stateless, queries for large amounts of data which require more than one request - have to provide a limit (maximum) for the number of resource items and an offset where to start the query to avoid returning the same set of data. These are defined as string parameters or posted queries.

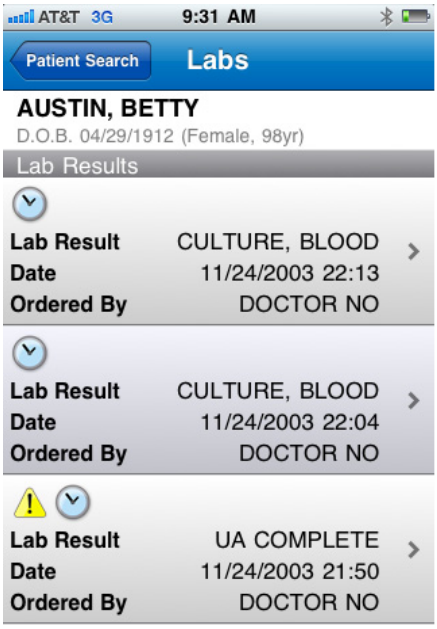


Fig. 3. VHR patients's lab results summary (iPhone)

Example of a JSON query object for patients:

```
{"query":
  {"family":"...",
   "max":N,
   "offset":M}
}
```

Example of a JSON response object for patients:

```
{"patients":
  {"list":[list of patients],
   "count":X,
   "offset":Y,
   "remain":Z}
}
```

Heath Record data can be very complex. For example in our data source, data is often stored as HL7v2 fields. The REST API uses an HL7v2 to HL7v3 transformation library to access Java canonical representation of the observations and unmarshal these plain Java objects (POJO) into an HL7v3 JSON representation of the observation.

JSON response object for lab results:

```
{"lab-results":
  {"list":
        [{"lab-result":
             {"entry":"...",
              "facility":"...",
              "normalcy":"...",
              "orderedBy":"...",
              "status":"...",
              "subject":"...",
              "urgency":"..."}},
         {"lab-result":{...}},...],
 "count":{count},
 "offset":{offset},
 "remain":{remain}}
```

HL7 V3 Lab result entry (JSON format):

```
{"entry":
  {"organizer":
        {"code":{"displayName":"..."},
         "components":[{"component":...},...],
         "notes":[...]}}}
```

HL7 V3 Lab result component (JSON format):

```
{"component":
  {"observation":
       {"code":{"displayName":"..."},
        "effectiveTime":{"value":<ISO-8601>},
        "value":...,
        "interpretationCode":{"code":"..."},
        "referenceRange":{"observationRange":{...}},
        "notes":[...]}}}
```



Fig. 4. VHR patients's lab results details (iPhone)

4.3 REST API Extension

Using JSON as an exchange format minimizes the amount of data that is transferred between the server and the client. Also, parsing JSON rather than a more complex format such as XML is faster and easier to perform on a mobile device.

However, JSON marshalling/un-marshalling from and to Java classes has to be implemented for each clinical object (e.g. allergies, encounters, immunizations, procedures). This is appropriate when an existing XML marshalling mechanism can be extended, directly adding JSON as an alternative format for the service layer that the REST API consumes. In this case, the REST API is very easy to create and offers a fast performance since the clinical objects are marshalled only one time closer to the source of the documents.

Most of the time, health record feeds and services are not provided in JSON format but most likely as XML structures encapsulated in SOAP envelopes or directly obtained from low level programming API.

In our case, we were able to integrate with a Java based API that is producing XML Continuity of Care Documents (CCD) sections. We used this new source of clinical documents to extend our VHR REST API with a more generic operation that would generate CCD sections directly in XML. Since the XML marshalling was already provided by the low level Java API, we were able to avoid creating the intermediate Java representation in the REST API and re-marshall the clinical objects into JSON.

On the other hand, one drawback of the new operation is that the consumer of the REST API (e.g. mobile devices) will have to receive and parse XML instead of a more compact and simpler JSON structure.

Table 2. API extension for CCD sections

POST /users/{user-id}/patients/{patient-id}/CCD	CCD from a patient

In this case, the request header for the request to query the CDD is still of type application/json, but the response is a CCD in XML format.

Request body for the new CCD POST operation (JSON format):

```
{"ccd-query":
  {"fromDate":<ISO-8601>},
   "toDate":<ISO-8601>},
   "sections":[
        {"section":
            {"fromDate":<ISO-8601>,
             "toDate":<ISO-8601>},
             "id":<templateIdCCD>}},...]}}
```

The CCD query specifies the date range for the CCD and the sections to retrieve using CCD standard template Ids (e.g Advance Directives = 2.16.840.1.113883.10.20.1.1).

Date and section selectors are optional and applied as follow:

- The date range at the ccd-query level is used as a selector for all the sections that are not specified.
- The date range at the section level is used as a selector for this specific section and overwrite the ccd-query date range selector.
- If no date range is specified, all the documents will be added to the CCD for all the sections specified.
- If no date is specified, all the documents will be added to the CCD for all the sections specified.
- If the ccd-query is empty (no date and no section), the resulting CCD will contain all available documents for all sections for the selected patient.

5 Testing and Integration

The REST API is very easy to test using HTTP requests from simple command tools such as cURL or a web browser. Early in the project we were able to create and deploy, in a few days, a full working prototype of the API that was producing mocked-up resources so that mobile clients were able to retrieve sample data in the same fashion as the final API production version.

During development we used a very small footprint embedded java web server (TJWS) that was running JUnit integration tests every time we were compiling the project. As a result, the REST API was always operational and ready to be deployed as a war file to the staging server when necessary.

On the client side, the very simple host, port and base URL configuration is necessary to switch between a staging and production server, which makes development, testing and deploying extremely easy.

The first version of our mobile health record will be available in early 2011 as a pilot program for a small group of physicians participating in the New York Rochester Regional Health Information Organization (Rochester RHIO). It has already been tested by the technical services of the New York Rochester RHIO and we have incorporated changes in the iPhone interface and REST API format to support the variability of real world lab results (which sometimes can be either incomplete or contain badly formatted HL7 data sets).

Initially only the records of patients who have provided consent will be accessible by the mobile application. This represented, in summer 2010 a total of 290,000 patients from more than 140 practices.

6 Conclusions

In this paper, we have presented the advantages of using a REST architecture for designing a lightweight and scalable API that is extremely easy to build, integrate, test, extend and maintain. We were able to create a working API prototype in a matter of days and a full functioning set of sophisticated health record web services accessible by our mobile client application in few weeks.

In the next few months, we will be able to gather user feedback from a large group of users (physicians and nurses) and statistics about usage that will be very valuable to improve the performance of our health record REST API and the user experience of the mobile application. The auditing of server logs for the REST API and the use of data mining tools will also be valuable to see performance bottlenecks and usage trends. In parallel, we are extending our REST API to additional operations such as CCD operations and additional authentication mechanisms to offer our VHR REST API as a useful alternative to existing SOAP web services. As a consequence, this architecture is not only suitable to be accessed by mobile applications, but can also be part of other types of loosely coupled solutions such as mash-up and portal applications.

References

1. Fielding, R.: Architectural Styles and the Design of Network-based Software Architectures. Doctoral dissertation, University of California, Irvine (2000)
2. Fielding, R., Taylor, R.: Principled Design of the Modern Web Architecture. In: TOIT 2002, ACM Transactions on Internet Technology, pp. 115–150 (2002)

3. HIPAA, The Health Insurance Portability and Accountability Act of (HIPAA) Privacy and Security Rules (1996), http://www.hhs.gov/ocr/privacy
4. Kumar, et al.: ELMR: Lightweight Mobile Health Records: A Study of Access Control. In: SIGMOD 2009: Proceedings of the 35th SIGMOD International Conference on Management of Data (2009)
5. Landre, E., Wesenberg, H.: Rest versus soap: as architectural style for web services. In: 5th International OOPSLA Workshop on SOA & Web Services Best Practices (2007)
6. Luo, J., Wesenberg, H.: Mobile computing in healthcare: the dreams and wishes of clinicians. In: HealthNet 2008: Proceedings of the 2nd International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments (2008)
7. Pabllo, C., Soto, R., Campos, J.: Rest Mobile Medication Administration System: Application and Architecture. In: EATIS 2008: Proceedings of the 2008 Euro American Conference on Telematics and Information Systems (2008)
8. Pautasso, C., Zimmermann, O., Leymann, F.: RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision. In: WWW 2008: 17th International World Wide Web Conference, Beijing, China (2008)
9. Sammon, et al.: MACCS: Enabling Communications for Mobile Workers within Healthcare Environments. In: Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2006 (2006)
10. Shiffman, et al.: Pen-Based, Mobile Decision Support in Healthcare. SIGBIO Newsletter 19(2) (1999)
11. Watson, M., Wesenberg, H.: Mobile Healthcare Applications: A Study of Access Control. In: Proceedings of the 2006 International Conference on Privacy, Security and Trust (2006)

Biomedical Literature Retrieval Based on Patient Information

Ana Jimenez-Castellanos[1], Izaskun Fernandez[2], David Perez-Rey[1], Elisa Viejo[3], Francisco Javier Díez[2], Xabier García de Kortazar[2], Miguel Garcia-Remesal[1], Victor Maojo[1], Antonio Cobo[4,3], and Francisco del Pozo[3,4]

[1] Dept Inteligencia Artificial, Facultad de Informática
Universidad Politécnica de Madrid, Campus de Montegancedo S/N
28660 Boadilla del Monte, Madrid, Spain
[2] Tekniker-IK4, Av Otaola, 20, 20600 Eibar, Guipuzcoa, Spain
[3] Centre for Biomedical Technology, Technical University of Madrid
Campus de Montegancedo - UPM Autopista M-40, km 38. Pozuelo de Alarcón
28223 Madrid, Spain
[4] Biomedical Research Networking Center in Bioengineering
Biomaterials and Nanomedicine, Madrid, Spain
{ajimenez,dperez,mgarcia,vmaojo}@infomed.dia.fi.upm.es,
{ifernandez,fdiez,xkortazar}@tekniker.es,
{evdiego,acobo,fpozo}@gbt.tfo.upm.es

Abstract. Information and Communication Technologies has led to a biomedical data explosion. A proportional growth has been produced regarding the amount of scientific literature, but information retrieval methods did not follow the same pattern. By using specialized clinical search engines such as PubMed, Medscape and Cochrane, biomedical publications has became instantly available for clinical users. However, additional parameters, such as user context, are not taken into account yet. Initial queries still retrieve too many results without a relevance-based ranking. The objective of this work was to develop a new method to enhance scientific literature searches from various sources, by including patient information in the retrieval process. Two pathologies have been used to test the proposed method: diabetes and arterial hypertension. Results obtained suggest the suitability of the approach, highlighting the publications related to patient characteristics.

Keywords: Electronic health record, Search engines, Literature retrieval, Integration, Federated search.

1 Introduction

Physicians attending daily a high number of patients hamper their availability to keep up to date with the latest research news in their field. They frequently lack the time to locate relevant information related to the patient Electronic Health Record (EHR). Physicians need specific information rather than large amounts of information, however the amount of data generated nowadays and stored both, within EHR and research literature, is overwhelming. The more data the more challenging is to find relevant information (for

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 312–323, 2013.
© Springer-Verlag Berlin Heidelberg 2013

research and training of the physician) and to integrate it with current resources. Search engines and other Web 2.0 technologies such as RSS feeds facilitate these tasks, but advanced methods are required to retrieve research information related to, not only a pathology, but also a specific demographic group.

Scientific research papers regarding clinical practice deals with population groups rather than specific patients. Physicians must generalize patient data in order to find patient-related clinical literature. They should analyse the EHR of the patient to include relevant keywords in the query that filter the results from several information sources. After that, the physician reviews the scientific articles retrieved and selects the relevant information for the patient, discarding the majority of the results. This is particularly interesting for patients who do not respond to therapies or treatments, who need special attention to avoid further complications. These kind of patients are those for whom physicians need more specific research information about their disease.

In this work we present a new method, implemented by the authors, aiming to provide an advanced method to retrieve biomedical literature based on EHRs. Diabetes and arterial hypertension are the two use cases that have been tested, within the framework of *Tratamiento 2.0*, a research and development project aiming to create a generic middle ware platform that serves as the basis for the development of management services and intelligent application for treatment for patients, especially chronic [1].

2 Background

Patient information inclusion in search queries for biomedical literature, requires techniques and technologies with a high research activity. Those used to store the required information, such as Electronic Health Records and Decision Support Systems, and those used to extract information, such as Natural Language Processing and Federated Searching described in this section.

Electronic Health Records (EHR), defined as electronic objects that contain data, evaluations and information of any kind, on the status and the clinical course of one patient through the care process [2], have been a fundamental advance in clinical practice. Electronic information can be automatically analysed and there exist research efforts to generate knowledge from EHRs such as [3], where the data generated from care delivery and captured in the EHR systems, is used for being analysed to discover and then inform about best practice. In [4], how to transfer knowledge from a medical record written in a free text form into a structured format represented by the EHR is analysed. Finally, [5] is a study aiming to understand user needs as captured by their search queries in an EHR system.

EHR are the essential infrastructure to other software elements in clinical practice such as Decision Support Systems, a kind of information systems that supports decision-making activities [6]. Specifically, Clinical Decision Support Systems (CDSS) are software programs that assist the physicians in the decision making process [7][8]. Some methods to integrate CDSS and EHR are: [9] and [10], systems dealing with CDSS adaptation, to facilitate the integration of individual patients preferences and characteristics and improve decision making.

To retrieve further clinical information, the EBM [11] recommends physicians to formulate clinical questions in terms of the problem/population, intervention, comparison, and outcome for searching clinical reports efficiently. These elements comprise the PICO frame: P represents problem/population; I interventions information; C the comparison; and finally O the outcome. The construction of this kind of query is not an easy task as it is shown in [12]. For instance, defining P requires an exhaustive reading of the patient EHR and the selection of the most relevant characteristics for the current context.

Information extraction dealing with free text instead of structured data is treated with Natural Language Processing (NLP) techniques, that have been applied to many tasks of Biomedicine such as bio-entity recognition, protein/gene normalization, interaction, extraction and many others. As there are many fields where NLP techniques can be used, various NLP tools are available. There are domain specific tools resolving an specific task like GoAnnotatorTool [13], but also we can find more generic and flexible NLP tools, like Freeling [14] or GATE [15] which can be used for many tasks and domains. The fundamental problem of NLP analysis is the fact, that a particular meaning may be expressed using different synonymous expressions. Lexical repositories recording meaning and lexical forms relations are frequently used in NLP techniques for interpreting texts. MeSH [16] or SNOMED [17] are two well known lexical repositories in the medicine domain which can be exploited for scientific article interpretation.

Finally, heterogeneous data sources stored in different locations require distributed or federated methods to extract information. The federated search paradigm was thus created, evolving in response to the vast number of information resources. As defined in [18], federated searching consists of firstly transforming a query and broadcasting it to a group of different sources; merging the results obtained from the different asked sources; presenting it in a unified format without duplications; and providing the option of sorting the result set. Search engines like Sphinx [19] help in the development of this kind of systems, indexing the information and providing a fast information retrieval.

These techniques have been used in projects such as Parallel IE for bio-medical text mining and indexation at Merck kGaA, Darmstadt and Medline Analysis [20] at Institute for Medical Informatics and Biometry, University of Rostock, Germany for extracting causal functional relations on MedLine abstracts using GATE. There are also tools like MeSHMap [21] and MedMeSH Summarizer [22] that exploits MeSH ontology for document indexation and summarization, respectively.

3 Patient-Based Literature Retrieval and Integration

The approach proposed in this paper aims to provide an improved method to search biomedical literature based on patient data. Data used to identify the relevant information for literature retrieval is collected from two sources: sensors on the patients home described elsewhere [1] and data contained in the EHR of the patient. Both sources are stored in the platform of the project and then feed the services explained below.

Fig. 1 presents the two main phases performed to retrieve relevant publications. Firstly, the system should receive both data sources and locate relevant characteristics of the patient and then looking for and integrating the most relevant publications based on these characteristics.

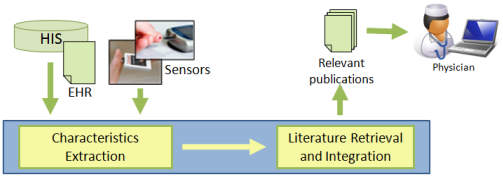


Fig. 1. Patient-based literature retrieval architecture

To automatically query current biomedical literature search engines for relevant publications based on EHRs and sensor info, a web services architecture was proposed. Three main web services were identified to cover the required functionality:

- *patientCharacterictics*, receiving a set of parameters from the EHR and sensors to identify the relevant characteristics.
- *federatedSearch*, receiving the relevant characteristics, generates and launches queries against different data sources.
- *resultAggregation*, that collect, integrates and present the results to the user.

These web services are explained within the next subsections, 3.1, 3.2 and 3.3 respectively.

3.1 Characteristics Extraction

The *patientCharacteristics* web service receives a set of parameters in XML format from the platform of the project. This platform stores data contained in the patient's EHR in addition to other data resulting from the monitoring of the patient at home e.g. certain data is needed in real time (like glucose levels before and after eating). This was implemented using sensors within the patient environment. Fig. 2 shows the structure of the web service which extracts the characteristics of a patient.

Two separated phases can be identified: (i) parameter processing and (ii) curation. Within the first phase, parameters are processed to extract the relevant characteristics. Afterwards, during the curation phase, an expert may evaluate the characteristics extracted to filter those relevant for the publication search.

Parameters from EHR or sensors are integrated into a single XML parameter list:

```
<patient_parameters>
   <pathology>diabetes</pathology>
   <parameters_list>
      <parameter>
         <name>age</name>
         <value>62</value>
      </parameter>
      ...
   </parameters_list>
</patient_parameters>
```

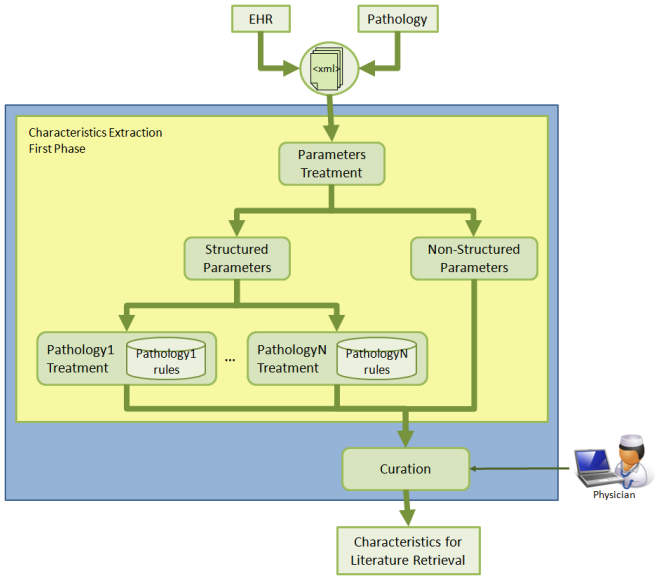


Fig. 2. Characteristics Extraction

For each parameter its name and its value are required. The *patientCharacteristics* web service extracts the significant characteristics focused on a given pathology.

The set of parameters is not closed, the service can receive any type of parameters, i.e. that the application will analyze only the parameters that can be treated according to certain rules, discarding those that are not recognized.

After receiving the parameters, the web service separates among structured and non-structured parameters. Structured data are those which have a numeric or enumerated value (E.g.: the numeric value of the patient's age, or the values *male* or *female* identifying the sex of the patient). Not structured data are free text and they are directly addressed to the curation phase as characteristics of the patient.

Depending on the pathology, structured parameters are selected for its relevance. For each disease there is a knowledge base of rules that is responsible for processing related parameters. Thus, when the characteristics of a patient have to be extracted based on a new disease, only the corresponding rules are required. The knowledge base of each pathology is built based on standards of CDSSs.

Relevant data for the pathology is analysed following rules contained within the knowledge bases, which are associated with the parameters. Those rules deal with patient stratification. Most of the rules are constructed based on expert approved thresholds. Depending on parameters values, characteristics are extracted and mapped to one or several MeSH terms.

All parameters are evaluated to extract significant characteristics related to one patient. In the curation phase, the physician (using a simple interface) may choose characteristics to be used to build the search query. Relevant characteristics selected by the physicians will be sent to the next web service for retrieving the corresponding literature.

3.2 Query Generation

Search federation intends to aggregate information units from heterogeneous sources, presenting the results on a unified way, without duplicates and ordered by relevance. This architecture enables a high scalability and stability, since adding new sources is only a matter of new connectors configuration, and system remains working even if some sources are down.

So the *federatedSearch* web service gets as input the XML characteristics generated by *patientCharacteristics* service as context and a query defined by the expert, which expresses the searching topic. In order to represent efficiently those characteristics and the free query, the service exploits MeSH ontology combined with NLP tools, trying to find corresponding MeSH terms both, for the characteristics and on the input free text. The MeSH lexicon as a flexible Gazetteer on GATE tool has been integrated in the service. In this way MeSH terms on any text are tagged, first parsing it with the morphological module of GATE and then identifying MeSH terms with the gazetteers module.

The characteristics are extracted from the patient EHR, so it can be considered that the selected MeSH terms define the problem (P) of the PICO query. Intervention (I) and outcome (O) will be completed with the identified and after selected terms on the expert query.

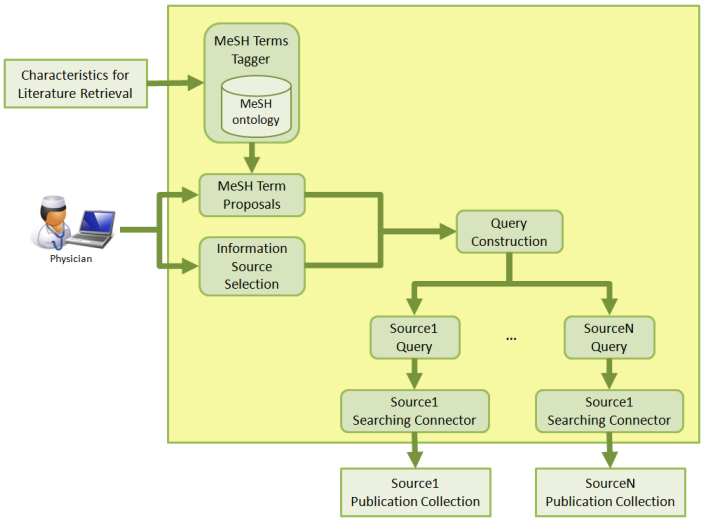


Fig. 3. Query Generation and Information Retrieval

In order to get the federated search, the web service queries different information sources. Each registered expert has associated a collection of information sources. For each new query the expert selects a set of those sources to search. Each information source encloses its own query format. Before searching, it is necessary to transform the

PICO query to the each source query format. For the transformation process, regardless of the source, the system applies always the same methodology: taking PICO query as input, it applies an xslt based transformation. The xslt object is a source dependent element where the source specific query syntax is considered for the final query representation. Together with query representations, the service launches different queries, getting one result set from each source in XML format. This procedure is graphically represented at Fig. 3.

3.3 Integration and Presentation

The publication sets obtained from the previous web service are the input of the *resultAggregation* service. The aim of this service is to aggregate all the result sets presenting to the expert the publications as a unique collection. The final collection should not have duplicated items and must be sorted by relevance regardless of the source. The web service processes the publications, tagging MeSH terms and storing both the item itself and the tagged MeSH terms in a MySQL database in order to index the entire collection with Sphinx and improve the efficiency of the sorting search.

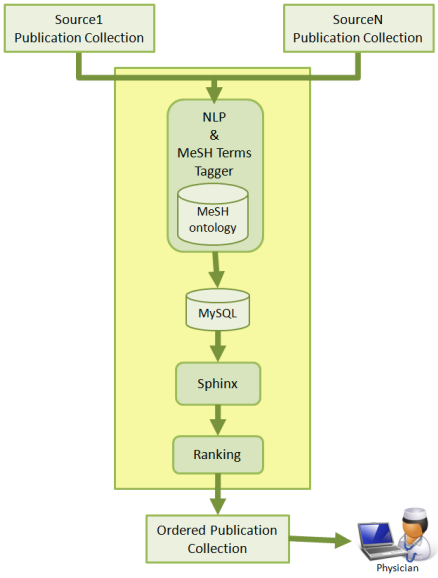


Fig. 4. Information Aggregation

Once the collection is indexed, the service ranks the publications based on the initial query using Sphinx. The result is a sorted collection based on the relevance of each publication respect the initial query as it is shown in Fig. 4.

4 Use Cases: Diabetes and Arterial Hypertension

Two use cases were developed to test the proposed method: (i) diabetes and (ii) arterial hypertension. A knowledge base to analyze the parameters related with them has been implemented to test the system. Information needed to evaluate each parameter received by the *patientCharacteristics* web service is stored in different knowledge (one for each pathology). Structured parameters were selected by experts within the project using a focus group methodology. These parameters are shown in table. 1.

Table 1. Recognized structured data

Common parameters	Diabetes specific	Arterial Hypertension
age	diabetic_familiar	hypertensive_familiar
race	neuropathy	total_cholesterol
sex	basal_glycemia	triglycerides
body_mass_index	gestational_diabetes	systolic_pressure
	retinopathy	diastolic_pressure
	nephropathy	HDL
	cardiopathy	abdominal_wall_size
	glucose_intolerance	smoker
	glycated_hemoglobin	cLDL
	diabetic_foot	

Depending on the pathology, structured parameters are selected by its relevance. Due to the characteristics of the diseases, there are some parameters that are common to both and others are specific for each pathology (table. 1).

Finally, for each structured parameter, the Web Service applies several rules, extracting the related characteristic. For example, treatment of the glycated hemoglobin (HbA1c) and related characteristics are:

- HbA1c < 6.5 : "normal HbA1c"
- HbA1c < 7 : "high HbA1c"
- HbA1c < 8 : "high HbA1c AND glycemic control"
- HbA1c $>= 8$: "high HbA1c AND intensive glycemic control"

Once all the characteristics are extracted went through the curation phase, the system constructs a searching query with the pathology, the relevant characteristics previously gathered and the MeSH terms identified with the experts input query (with the searching topic).

With queries close to the mentioned pathologies, the system searches on expert defined sources using their corresponding connectors to obtain the most relevant publications and evidences. Nowadays, sources covered by the system are PubMed, Cochrane and MedScape, where the most relevant publications about diabetes and hypertension are published.

Finally, the results of each source are integrated, indexed and presented to the expert as a unique collection as shown in the following section.

5 Results

Characteristics are extracted according to the pathology and the parameters introduced. From this output, the expert will select the most relevant features to be launched to the search.

A preliminary test set of ten patient data and fifty queries were used to check the functionality of the system. The retrieved characteristics were correct for all the cases. The results correctly identified the 90% of the relevant papers, according to assessment of experts of the project.

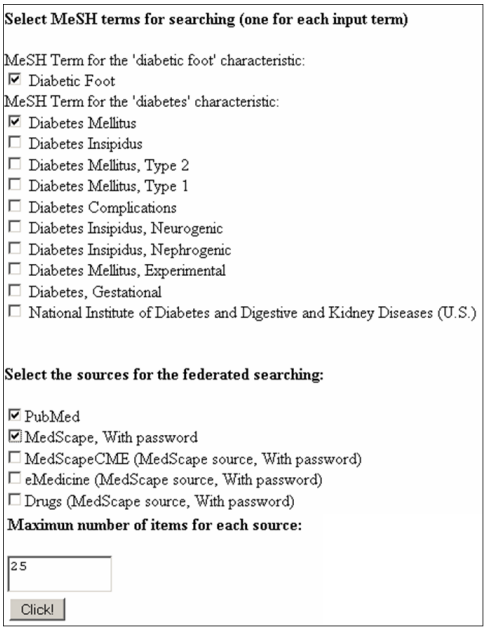


Fig. 5. Interface to select the parameters and sources of the search

As it is shown in Fig. 5, the expert can choose a more specific term for the search, i.e. optimizing the query that will be launched against the selected sources. The expert can also specify the number of results obtained from the query.

With the optimized query, the system accesses the different sources and aggregates the results, filtering those fulfilling the initial requirements and finally presenting them ordered by relevance, as it is shown on Fig. 6.

Two filters were applied to the parameters in order to extract the most relevant literature referred to a patient: (i) a transformation of the parameters into Mesh terms and (ii) an optimization of the characteristics, using the MeSH ontology and the medical knowledge of the expert. Through the final query and the federated search the expert will retrieve biomedical literature based on patient data.

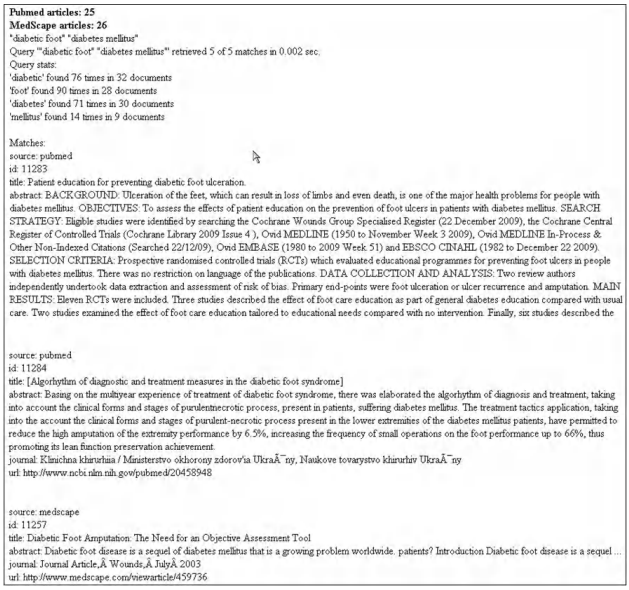


Fig. 6. Aggregated Results

6 Conclusions and Future Lines

In this work, a new method for search clinical literature has been proposed and implemented within the framework of a research project for the personalization of treatments and therapeutic strategies. Promising results of the system suggest that EHR-based literature retrieval may facilitate the work of physicians obtaining biomedical patient related research publications. Since there is not any gold-standard corpus defined for evaluating the system empirically, we have done a qualitative evaluation of it testing with real patients and experts from the Hospital Universitario de Valencia and most of the clinicians were satisfied with the usability of the system for their daily practice.

As work in progress, extensions of services described within this paper are being implemented: further non-structured data processing, clustering techniques to group the results, open of generic connectors and source weighting for final ranking.

First, the service will use Natural Language Processing techniques to process non-structured data of the patient for characteristic extraction.

Clustering techniques are being used to present results not only as a sorted collection but also grouped by their semantic similarity, providing a more intuitive representation for the navigation.

Regarding information sources, connector designing process is being generalized to facilitate new source integration within the system. This service will add a weighting functionality for sources, to adjust the final publication relevance not only based on the initial query but also considering the relevance of the source.

The system presented in this work has proven the suitability of a patient-based literature retrieval approach. And we are confident that current extensions will improve the

location of relevant publications for physicians, facilitating more relevant results when searching at main biomedical literature search engines.

Acknowledgements. The present work has been funded by the Spanish Ministry of Industry, Tourism and Trade in the framework of the "Avanza I+D" programme under the project *Tratamiento 2.0* (project number TSI-020301-2008-15). We would like to express our gratitude to the *Hospital Universitario de Valencia* for their collaboration in the election of the parameters used in this work.

References

1. Tratamiento 2.0, `http://www.tratamiento20.com`
2. Healthcare Information and Management Systems Society: Consensus definition of an Electronic Health Record, `http://www.himss.org/ASP/topics_ehr.asp`
3. Weaver, C.A., Warren, J.J., Delaney, C., et al.: Bedside, Classroom and Bench: Collaborative Strategies to Generate Evidence-Based Knowledge for Nursing Practice. Int. J. Med. Inform. 74, 989–999 (2005)
4. Antolik, J.: Automatic Annotation of Medical Records. Stud. Health Technol. Inform. 116, 817–822 (2005)
5. Natarajan, K., Stein, D., Jain, S., et al.: An Analysis of Clinical Queries in an Electronic Health Record Search Utility. Int. J. Med. Inform. (2010)
6. Keen, P.G.W., Morton, M.S.S.: Decision support systems: An organizational perspective. Addison Wesley Publishing Company (1978)
7. Kawamoto, K., Houlihan, C.A., Balas, E.A., et al.: Improving Clinical Practice using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success. BMJ 330, 765 (2005)
8. Sim, I., Gorman, P., Greenes, R.A., et al.: Clinical Decision Support Systems for the Practice of Evidence-Based Medicine. J. Am. Med. Inform. Assoc. 8, 527–534 (2001)
9. van der Weijden, T., Legare, F., Boivin, A., et al.: How to Integrate Individual Patient Values and Preferences in Clinical Practice Guidelines? A Research Protocol. Implement Sci. 5, 10 (2010)
10. Seidling, H.M., Schmitt, S.P., Bruckner, T., et al.: Patient-Specific Electronic Decision Support Reduces Prescription of Excessive Doses. Qual. Saf. Health. Care (2010)
11. Evidence-Based Medicine Working Group: Evidence-Based Medicine. A New Approach to Teaching the Practice of Medicine. JAMA 268, 2420–2425 (1992)
12. Huang, X., Lin, J., Demner-Fushman, D.: Evaluation of PICO as a Knowledge Representation for Clinical Questions. In: AMIA. Annu. Symp. Proc., pp. 359–363 (2006)
13. Couto, F.M., Silva, M.J., Lee, V., et al.: GOAnnotator: Linking Protein GO Annotations to Evidence Text. J. Biomed. Discov. Collab. 1, 19 (2006)
14. Atserias, J., Casas, B., Comelles, E., et al.: FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 48–55 (2006)
15. Cunningham, H.: GATE, a General Architecture for Text Engineering. Computers and the Humanities 36, 223–254 (2002)
16. Díaz-Galiano, M.C., García-Cumbreras, M.Á., Martín-Valdivia, M.T., Montejo-Ráez, A., Ureña-López, L.A.: Integrating MeSH Ontology to Improve Medical Information Retrieval. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 601–606. Springer, Heidelberg (2008)

17. Spackman, K.A., Reynoso, G.: Examining SNOMED from the Perspective of Formal Ontological Principles: Some Preliminary Analysis and Observations. KR-MED., 72–80 (2004)
18. Jacso, P.: Thoughts about Federated Searching. Information Today 21, 17 (2004)
19. Lee, K.F.: Automatic speech recognition: The development of the SPHINX system. Kluwer Academic Pub. (1989)
20. Karopka, T., Scheel, T., Bansemer, S., et al.: Automatic Construction of Gene Relation Networks using Text Mining and Gene Expression Data. Med. Inform. Internet Med. 29, 169–183 (2004)
21. Srinivasan, P.: MeSHmap: A Text Mining Tool for MEDLINE. In: Proc. AMIA. Symp., pp. 642–646 (2001)
22. Kankar, P., Adak, S., Sarkar, A., et al.: MedMeSH Summarizer: Text Mining for Gene Clusters, pp. 11–13 (2002)

Evaluating Information Security Effectiveness with Health Professionals

Robin Krens[1], Marco Spruit[2], and Nathalie Urbanus[1]

[1] University Medical Center Utrecht, Heidelberglaan 100
Utrecht, The Netherlands
rkrens@umcutrecht.nl
[2] Department of Information and Computing Science, Utrecht University
Padualaan 8, Utrecht, The Netherlands
m.r.spruit@cs.uu.nl

Abstract. This paper outlines an alternative view on the information security discipline. We argue that information security is, in general, viewed from a technological and means-end oriented perspective. Our approach can be seen as an initial attempt to approach information security in a broader, more holistic, sense. For this purpose, we approach information security from a health professional's perspective. An instrument, The Information Security Employee's Evaluation (ISEE), is presented to evaluate and discuss information security with health professionals. The ISEE instrument consists of seven dimensions: priority, responsibility, incident handling, functionality, communication, supervision and training and education. The ISEE instrument can be used to better understand health professional's perception, needs and problems when dealing with information security in practice. Following the design science approach, the ISEE instrument was validated within a focus group of security experts and pilot tested as workshops across five hospital departments in two medical centers. Although the ISEE instrument has by no means the comprehensiveness of existing security standards, we do argue that the instrument can provide valuable insights for both practitioners and research communities.

Keywords: Information security, Evaluation, Human perspective.

1 Introduction

Information security is involved with guaranteeing the availability, integrity and confidentiality of information [18]. In health care, correct and in-time medical information is needed to provide high quality care. Unavailable or unreliable information can have serious consequences for patients, such as incorrect or delayed treatment. Also, since this type of information is uttermost sensitive, protecting the patient's privacy is another major security objective. From a health professional's point of view, information security aspects concern issues such as in-time access to medical information during consultation, fast recovery during system downtime and assurance of data integrity.

1.1 Narrow and Holistic Definition

There is no clear definition of information security in the health care domain: this alone would be sufficient for a research on its own. We see, however, two types of definitions

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 324–334, 2013.
© Springer-Verlag Berlin Heidelberg 2013

most frequently. One is a narrow definition, while the is more of a holistic definition. The narrow definition could be best defined as the protection of information against unauthorized disclosure, modification and withholding [6]. One could argue that this view is almost synonymous with computer security. Identity and Access management is, for example, a discussed theme by authors who use this type of definition. Referring to the aspects of information security, in this type of definition there seems to be a tendency towards the confidentiality aspect, overshadowing the other two: availability and integrity. Barber notices this tendency and states that "the issues of integrity and availability will probably deserve more attention than the issues of confidentiality as medical information systems became more inter-twined with clinical practice" [2].

In contrast, a holistic definition of information security could be defined in terms of quality of care or safety of patients. If for example, information is not available at a given time in a health care environment, would this just be solely an issue for the safety and quality of care, or is this an issue for information security as well? To give another example, if information in an electronic health record is altered unauthorized, would this just be solely an issue for information security, or has this an impact on safety and quality of health care as well? We argue that it involves both domains and thereby use a holistic definition: information security can not be seen as a distinct discipline.

1.2 Effectiveness of Information Security?

Practitioners are most interested in the effectiveness of information security in an organization. They question (1) whether or not security controls (i.e. continuity management) have a positive effect and (2) whether more (or less) security control are needed.

Roughly, methods to test the effectiveness of security could be divided in:

1. **Compliance with Standards, Checklists and Legislation**
   The general idea is that, as long as an organization is compliant with a certain standard (i.e. ISO/IEC 27002 [9]), security is considered sufficient. Of course, compliance is a useful method to test if certain processes of information security are implemented. An objection, however, is clearly stated by Siponen [17]:

   > "[T]he fact that an organization has in place a certain security process or security activity prescribed by the information security standards does not imply the ultimate goal of this process (or security activity) is therefore achieved. It does not mean the organization's systems are secured according to this objective."

   So in this view, for example, a process conducting end-user training, no matter the quality, would be concerned as effective security.

2. **Usage of Performance Indicators**
   Clearly, an indicator has more focus on the quality of a process. The COBIT control objectives framework, for example, describes a few indicators for security, such as the number of incidents damaging the organization's reputation with the public and the number of violations in segregation of duties. The discussion, however, arises

if a complex subject as information security, can be systematically mapped down by a few indicators.

3. **Perception by the End-user**
   This view focuses on the perceived effectiveness of security by end-users. User-related issues and needs are central in this view. Examples of user-related type issues are problems with retrieving data when needed and lack of knowledge on security procedures. Recent research subscribes the need for a more social approach to information security [3] [16]. Part of this approach is to enlighten on the human and cultural elements of information security [21] [7] [1] . Another part is, since the increase in vulnerabilities and complexity to health information systems nowadays, to involve health professionals actively within the domain of information security.

1.3 Outline of Paper

This paper takes the later approach and focuses on all aspects of medical information security seen from a health professional's point of view. The aim of the research is to build an instrument to evaluate and discuss information security with health professionals. The developed instrument, named ISEE (Information Security Employee's Evaluation), can be used to better understand user's perception, needs and problems. The paper is structured as follows. After this introduction, the second section reviews related work. The third section describes the research approach and the development of the ISEE instrument. Subsequently, the fourth section describes the use of ISEE. The fifth and final section discusses contributions, limitations and future research for this study.

2 Related Work

It is widely recognized that information security is much more than technology. Williams [21] states that information security is not a technical problem but mostly a human one. Williams identifies poor implementation of security controls, lack of relevant knowledge and inconsistencies between principles and practice as key issues. Williams also states that a trusting hospital environment undermines the need for proper supervision. In a culture of trust, confidence in medical practice staff is high, resulting in little scrutiny of Internet usage, no policy on changing passwords and unmonitored access to clinical records. Fernando and Dawson [4] show similar findings: poor quality training and the hospital environment are constraints on effective information security. Additionally, they argue that wrongly implemented security controls can result in workarounds such as the sharing of passwords or the usage of written clinical notes in case of systems downtime. Security controls often take time from patient care (i.e. logging out of a system). Health professionals are skeptic about such controls that form a constraint on their daily work and that could, in the worse case, harm the patient. In a complex environment were sensitive information is routinely recorded, spread and used it is a challenge to guarantee the availability, confidentiality and integrity of information. As indicated in the introduction, most evaluation approaches of information security are technical and

risk based. Our aim is to evaluate information security with health professionals and for this purpose we desire a different type of evaluation. In the discipline of information security such a comprehensive type of evaluation does not exist yet. Most existing instruments are prescriptive (i.e. how should end-user perform?) and focus strongly on the confidentiality aspect. We, therefore, adapt an instrument from the health care domain. The instrument, named the Manchester Patient Safety Framework [19] is used to discuss the physical safety of patients with health professionals. The following subsection gives a short overview of this Patient Safety evaluation instrument.

2.1 The MaPSaF Instrument

The Manchester Patient Safety Framework (MaPSaF) is an instrument to help health care teams assess the safety of patients. Assessment with the instrument is usually carried out in workshops, led by a facilitator from a health care department. A workshop starts by letting each health professional individually rate dimensions of the patient safety instrument. Dimensions of this instrument are, for example, staff education and the investigation of patient safety incidents. The dimensions are given a score according to a maturity framework (we discuss this framework in detail in the next section). If, for example, a nurse thinks that staff education is lacking to 'safely' perform her daily job, she can fill out a low score.

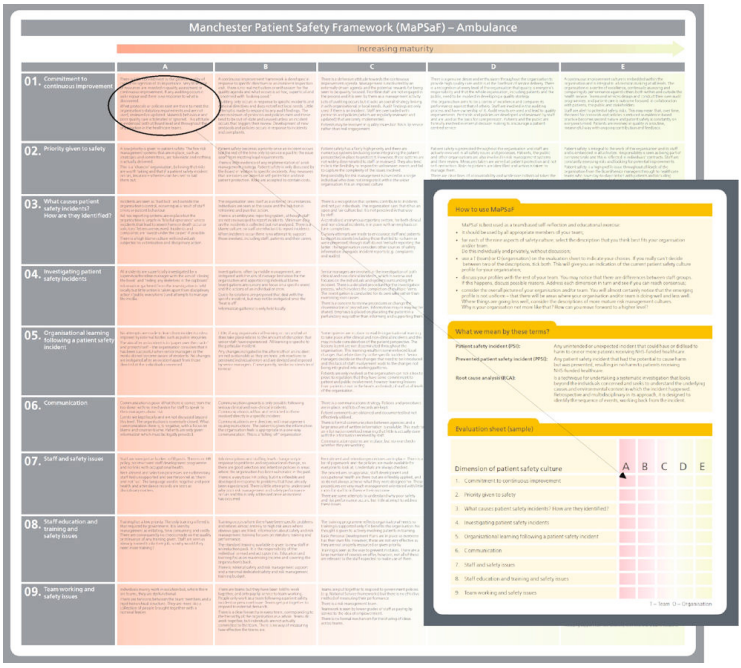


Fig. 1. Evaluation sheet and manual

A sample evaluation form and manual is given in Fig. 1. The evaluation form is the compact version of the instrument, the manual contains an explanation of each dimension and textual illustrations for each level that a user can rate.

The Maturity Framework. MaPSAF is based on a maturity framework. This framework was originally developed by Westrum [15], and was later extended by Reason [15] and Parker and Hudson [13]. The framework identifies five levels an organization can have: pathologic, reactive, bureaucratic, proactive and generative. Pathologic is defined as a situation where safety practices are the barest industry minimum. There is no top level commitment to the pursuit of safety goals. Reactive is a attitude where changes are implemented after incidents or problems occur. Bureaucratic is a situation where a lot is formalized, on paper, but practically a lot is failing. In contrast, proactive and generative form complete opposite and the positive view of this situation. Table 1 shows the MaPSaF safety framework.

Table 1. MAPSAF safety levels

Pathologic	Why waste our time on safety?
Reactive	We act when we have an incident.
Bureaucratic	We have systems in place to manage risks.
Proactive	We are alert on safety related risks.
Generative	Safety is a part of everything we do.

The Dimensions. Health professionals can, as already shortly explained, rate safety dimensions according to the maturity framework. Table 2 shows a sample of a dimension combined with this framework. Each level contains a description with examples on what stereotypical department is like at that particular maturity.

Table 2. A dimension combined with the framework. Illustrative description are given for each level.

Dimension	Pathologic	Reactive	Bureaucratic	Proactive	Generative
Staff education	Training has a low priority. The only training offered is that required by government.	Training occurs where there have been specific problems and relates almost entirely to high risk areas where obvious gaps are filled.	The training program reflects organizational needs so training is supported only if it benefits the organization.	There is an attempt to identify the training needs of the organization, and of individuals, and to match them up.	Individuals are empowered and motivated to undertake their own training needs analysis and negotiate their own training program.

Workshop Set-up. The instrument is typically used practically in workshops. A workshop involves two facilitators and around six to ten participants. MaPSaF encourages to involve a "crosscut" of a department, involving physicians, nurses, management and supportive staff. This will stimulate discussion from different perspectives. Workshops were conducted by a scheme. The following list shows the sequence of steps:

1. Individual evaluation: participants fill out the evaluation individually.
2. Work in pairs: participants discuss their perceptions with other participants. They are encouraged to explain their ratings and exchange anecdotes and personal experiences.
3. Group discussion: general discussion about strength, weaknesses and differences in perceptions.
4. Action planning: the creation of an action plan for weak safety issues.

We have adapted MaPSaF for the purpose of evaluating information security. The next section describes what methods we followed to translate MaPSaF to information security.

3 Development of Information Security Employee's Evaluation (ISEE)

To adapt MaPSaF for evaluating medical information security, we used the design science approach [8]. The design science approach consists of two main steps, namely (1) the development and (2) the validation of the instrument.

3.1 Literature Review

The security dimensions were initially based on a literature review. At first we identified over 30 user-related (i.e. lack of knowledge, poor security implementation, unusable security controls, workarounds) security issues. Since it was not feasible to include each of these issues individually in this type of evaluation, we decided to increase the level of abstraction. We provide a list of seven dimensions: priority, responsibility, incident handling, functionality of security, communication, supervision and training and education. Table 3 provides a general overview of these dimensions. The table also shows how we mapped various security issues from our literature study to a dimension. Though security involves many more topics than discussed within the evaluation (i.e. network control or protection against viruses), our perspective is restricted to those security issues that were relevant to health professionals.

3.2 Expert Review

Consequently, the dimensions were reviewed and approved by security experts from the health care discipline. The first review round was performed in a focus group with nine security experts. A focus group was chosen to stimulate discussion between experts. Focus groups encourage people to talk to one another, ask questions, exchange anecdotes and comment on each others' experiences and points of views. By these means, focus groups are considered to have high face validity [14]. The group consisted of security auditors, security officers and IT executives. The data collected during this session was used to develop descriptions for each of the seven dimensions of information security, at each of the five levels. A second interview round was performed with these experts to revalidate these descriptions. Table 4 shows the constructed and validated instrument (a bit abbreviated due to paper size constraints).

Table 3. Information Security Employee's Evaluation (ISEE) dimensions

Dimension	Information security issues
Priority	Lack of time [4] [11] [21], cost [21], the hospital environment [4], conflicting demands [7] and productivity [4]
Handling of incidents	Lack of incident reporting and handling [11] and response [12]
Responsibility	Attitude and ignorance [21] [7], lack of awareness and responsibility [11] [12], skepticism [4], data fragmentation [4] and underestimation of threats [11]
Functionality	Usability [4] [5], Workarounds [4], poor implementation [21], inadequate systems [7] and security design [12]
Communication	Communication [11], communication and feedback [10] and inconsistent policies and communication [7]
Supervision	Audit and supervision [4], trust [21], ethics [12], reward, punishment and hiring practices [10]
Training and education	Training shortcomings [4] [10], lack of knowledge [21], capability and education [21], [11]

4 Use of the ISEE

How is ISEE perceived in a practical context? For this purpose we conducted a pilot study at five hospital departments in two medical centers in the Netherlands. The goal of the workshops was to test if ISEE is applicable in a practical setting. For each workshop the face validity and utility of the instrument were investigated. We asked the participants if they found the evaluation useful and if they thought the scope of information security was covered. Feasibility concerned boundary conditions such as the amount of time. Since the instrument is not a pure measurement instrument validation was kept qualitative. The original instrument, MaPSaF, was also constructed in this nature. We will discuss one of these workshops in more detail below.

4.1 One Workshop Highlighted

One of the workshops was held at a Radiotherapy department. There was a total of seven participants. See Table 5 for an overview of participants and Table 6 for the descriptive statistics over the scores. The lowest scoring dimensions were *Supervision* and *Training and education*. The highest scoring dimensions were *Responsibility* and *Handling of incidents*. Most standard deviations of the dimensions indicate an acceptable distribution of responses. Handling of incidents and supervision show the highest variance. Management of the radiotherapy department was very positive on handling of security incidents, which explains the variance. The range of scores on supervision

Table 4. The ISEE instrument with abbreviated examples

Dimension	1: Pathologic	2: Reactive	3: Bureaucratic	4: Proactive	5: Generative
Priority: how important is security (availability, integrity and confidentiality of patient information?	Risks are not recognized	After incidents there is an increase in priority	Now and then plans are made for improvements	Plans are made and evaluated	Employees are involved, security is a management cycle
Incident handling: is the importance of reporting incidents (system failure, confidentiality breaches, unsafe systems) recognized?	It is not clear how and where incidents should be reported	Incidents are handled unstructured and on ad-hoc basis	There is a formal reporting systems, but is not fully implemented	Incidents are handled swiftly.	Trend analysis takes place to prevent incidents for future happenings
Responsibility: who or what is responsible for medical information security?	Information security is not my responsibility	Security is something management does	Security is about defining roles and responsibilities	Security is everybody's concern.	Employees know how to enhance security
Functionality: do systems support security in daily working routines?	Functionality comes with the systems	Temporary solutions are constructed	Needed system security functionality is planned	Systems work correctly and new improvements are considered	Systems fully support the process of care!
Communication: how is the communication about medical information security?	There is no possibility to discuss concerns	Communication is one way	Communication is paper work	Communication is a two-way process	Employees are aware and have a questioning attitude
Supervision: is the correct usage of medical information examined?	Incorrect usage has no consequences	Sanction are taken by severe shortcomings	Most of procedures are in place	Evaluation of behavior is done on periodic basis	Management and employees are widely involved on this topic
Training and education: do health care professionals know how to act?	Employees should not be bothered with security	Training is done if it is an absolute necessity	Training is highlighted, but not enforced	Employees are encouraged to participate	Training is part of the day-to-day job

Table 5. An overview of the participants and associated scores of one of the workshops

Department and participants	P	I	R	F	C	S	T
Radiotherapy							
Manager Quality Assurance	3	4	3	3	4	4	1
Manager Department	3	5	3,5	2,5	2	3	2
Physician	4	3,5	3	2	3	1	2
Head of Laboratory	3,5	3,5	3	3	2	3	2,5
Laboratory worker	2	3	4	2	4	2	2
Laboratory worker	3	2	2,5	3,5	2,5	3	2
Front Office / secretary	4	3	4	3	3	2	2
P=priority, I=incident handling R=responsibility, F=functionality, C=communication, S=supervision, T=training and education 1=pathologic, 2=reactive, 3=bureaucratic, 4=proactive, 5=generative							

is also broad. Management was also more positive towards this dimension then direct health care workers. The difference in perception brought to light that access control mechanism were not fully implemented.

Most participants prioritize the availability and sharing of information. This may have consequences on the confidentiality aspects. Most participants agreed that more awareness on confidentiality of patient information is desirable. Some even came up

Table 6. Workshop at a radiotherapy department (statistics are based on seven participants)

INSTR. Dimensions	Mean (1-5)	Std	Range (1-5)	Floor (x)	Ceiling (x)
Priority	3,21	0,69	2-4	0	0
Handling of incidents	3,42	0,93	2-5	0	1
Responsibility	3,29	0,57	2,5-4	0	0
Functionality	2,71	0,57	2-3,5	0	0
Communication	2,93	0,84	2-4	0	0
Supervision	2,57	0,98	1-4	1	0
Training and education	1,93	0,45	1-2,5	1	0

with a proposal, such as introducing privacy concerns to new employees or to make 'confidentiality and electronic medical records' a recurrent theme. The discussion also brought forward that many security issues (such as automatic logging out of systems) can be easily implemented. However, a reactive attitude causes that this does not happen. As one participants stated: "things should go wrong, before something eventually happens".

Furthermore, a variety of contemporary issues were discussed. Amongst these were:

- Unavailability of patient's status information.
- Slow security incident handling according to some health care workers.
- Poor integration with another system which made it impossible to write down medical information.
- The transition towards electronic medical records systems made that information was scattered (partly digital and partly on paper).

Based on these differences between experiences regarding the supervision and functionality dimensions the department created an action plan. Participants considered the workshop useful to discuss information security. Most participants encouraged the multidisciplinary setup to discuss different perceptions on information security.

5 Contributions, Limitations and Future Research

Health professionals are the first in line to experience disturbances with the availability and integrity of medical information. Furthermore, concerning the confidentiality of information, they play an important role in the protection of such information. Based on the MaPSaF instrument, that discusses the safety of patients, we constructed the Information Security Employee's Evaluation (ISEE) to evaluate information security with health professionals.

Overall, the pilot study showed that ISEE is useful to:

- Discuss medical information security within a hospital department.
- Identify and discuss weak and strong points.
- Discuss different perceptions on information security between employees.

A workshop can best be held at one single department (i.e. an outpatient clinic or nursing department). At one workshop two departments participated. Some security issues that were problematic at one department (availability of electronic nursing records) were never heard of at the other department. It was interesting to see such differences between departments. It is, however, hard to discuss and identify single points for improvements with such diverse groups. We, therefore, recommend using the evaluation within a single department.

The multidisciplinary set-up of participants highlighted various perceptions on information security. For instance, at one workshop management indicated that they had a very positive view on incident handling. Further discussion however, showed that staff had no idea how to report problems, and even when they did, they were not pleased with the department's solving skills. At another workshop the multidisciplinary set-up even took care of some quick fixes: A physician indicated that during night shift, magnetic resonance information about patients was not available. An employee of the IT supportive staff argued that this was an unknown issue, yet provided a quick solution.

Reflecting on all five workshops of the pilot study, we found that the dimensions priority and responsibility show the least amount of variance and range of scores. These dimensions, since they relate to attitude, might suffer social desirability bias. Floor effects occurred most frequently at the dimensions functionality and supervision. A majority of these low scores was explained by the participants. Ceiling scores were only given by management staff. Overall, management gave relatively higher scores than direct health care workers which might indicate a too optimistic view by management.

For future purposes, it might be interesting to further develop the instrument and apply it as a measurement instrument in a survey-format. Dimensions can be further defined with specific characteristics. To give an example, the dimension training and education could be further defined on the issues 'knowledge of privacy legislation', 'knowledge of information security' and 'knowledge on how to use security controls'. Such refinement makes the instrument more applicable for actual measurement within a hospital environment. Further work, then, will be needed to address these characteristics specifically. Also, such a measurement instrument, gives opportunities to examine in greater depth the instrument's psychometric properties including measures of internal consistency, reliability and construct validity.

This research has shown that the ISEE instrument can effectively assist health professionals in their efforts to improve information security within their hospital departments. The ISEE instrument has by no means the comprehensiveness and completeness of existing standards or other security checklists. We do, however, argue that the instrument and the human perspective can provide additional insights. Implementing secure systems does involve health care workers, both in respect of functional security controls as in human characteristics such as awareness, responsibility and knowledge.

References

1. Ashenden, D.: Information security management: A human challenge? Information Security Technical Report 13(4), 195–201 (2008)
2. Barber, B.: Patient data and security: an overview. International Journal of Medical Informatics 49(1), 19–30 (1998)

3. Dhillon, G., Backhouse, J.: Current directions in IS security research: towards socio-organizational perspectives. Information Systems Journal 11(2), 127–154 (2001)
4. Fernando, J.I., Dawson, L.L.: The health information system security threat lifecycle: An informatics theory. International Journal of Medical Informatics 78(12), 815–826 (2009)
5. Ferreira, A., Antunes, L., Chadwick, D., Correia, R.: Grounding information security in healthcare. International Journal of Medical Informatics 79(4), 268–283 (2010)
6. Gollman, D.: Computer Security, 1st edn. John Wiley & Sons (1999)
7. Gaunt, N.: Practical approaches to creating a security culture. International Journal of Medical Informatics 60(2), 151–157 (2000)
8. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly 28(1), 75–105 (2004)
9. International Organization for Standardization. Information technology – security techniques – code of practice for information security management. Technical Report ISO/IEC 27002:2005 (2005)
10. Kraemer, S., Carayon, P.: Computer and information security culture: Findings from two studies. In: Human factors and the Ergonomics Environment, pp. 1483–1487. Human Factors and the Ergonomics Society, Orlando (2005)
11. Nosworthy, J.D.: Implementing information security in the 21st century do you have the balancing factors? Computers & Security 19(4), 337–347 (2000)
12. OECD: Guidelines for the security of information systems and networks: Towards a culture of security. Technical report, Organization for Economic Cooperation and Development, Paris (2002)
13. Parker, D., Hudson, P.T.: HSE: Understanding your culture. Shell International Exploration and Production EP 2001 - 5124 (2001)
14. Pope, C., Mays, N., Kitzinger, J. (eds.): Qualitative research in health care, chapter Focus Groups, 3rd edn., pp. 21–31. Blackwell Publishing, Oxford (2006)
15. Reason, J.: The identification of latent organizational failures in complex systems. In: Wise, J.A., Hopkin, V.D., Stager, P. (eds.) Verification and Identification of Complex Systems: Human Factor Issues, pp. 223–237. Springer, New York (1993)
16. Siponen, M.T.: An analysis of the traditional IS security approaches: implications for research and practice. European Journal of Information Systems 14(3), 305–315 (2005)
17. Siponen, M.T.: Information security standards focus on the existence of process, not its content. Communications of the ACM 49(8), 97–100 (2006)
18. Stamp, M.: Information security: principles and practice, 2nd edn. John Wiley & Sons, Hoboken (2006)
19. University of Manchester and National Patient Safety Agency: Manchester Patient Safety Framework MaPSaF, `http://www.nrls.npsa.nhs.uk`
20. Westrum, R.: Cultures with requisite imagination. In: Wise, J.A., Hopkin, V.D., Stager, P. (eds.) Verification and Validation in Complex Man-machine Systems, pp. 401–416. Springer, New York (1993)
21. Williams, P.A.H.: When trust defies common security sense. Health Informatics Journal 14(3), 211–221 (2008)

Using Social Network Analysis to Study the Knowledge Sharing Patterns of Health Professionals Using Web 2.0 Tools

Samuel Alan Stewart and Syed Sibte Raza Abidi

NICHE Research Group, Dalhousie University
6050 University Ave., Halifax, NS, Canada
sam.stewart@dal.ca
http://www.cs.dal.ca/∼niche

Abstract. Peer communication is a vital component of the knowledge translation process for healthcare practitioners, and emerging web 2.0 tools are providing virtual venues to facilitate this communication. Using social network analysis methods this paper will attempt to explore the communication patterns that emerge out of the Pediatric Pain Mailing List. The analysis will assess the overall health of the communication network, identify users and subjects of interest, and it will isolate potential subgroups that exist within the community. These results will be presented to the user using the VECoN system, developed as part of this project to present the structure of communication networks graphically to the user through the use of social network analysis methods.

Keywords: Social network analysis, Data visualization, Knowledge translation, Web 2.0.

1 Introduction

Experiential healthcare knowledge manifests in a variety of modalities: clinical case studies, problem-based discussions between clinicians, experience-based insights, diagnostic heuristics et cetera. This knowledge accounts for the intrinsic experiential know-how, insights, judgements and problem-solving strategies of healthcare practitioners. Such knowledge is not 'published' as evidence-based, yet it holds vital insights into solving atypical clinical problems. The key issues related to experiential healthcare knowledge are: (a) how to formulate a community of healthcare practitioners; (b) how to explicate and share their experiential healthcare knowledge; and (c) how to put value on experiential healthcare knowledge, especially for clinical decision making, since it is not systematically evaluated in the same manner as evidence-based studies.

In the realm of Web 2.0, the emergence of Medicine/Health 2.0 presents 'virtual' community-driven environments to create and share healthcare knowledge. The key idea is that the community creates and validates experiential knowledge in an organic manner, and applies it to provide feedback to its effectiveness. Web 2.0 based knowledge sharing mediums include online discussion forums, email-based mailing lists, web blogs, et cetera. Through these mediums, healthcare practitioners are able to articulate

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 335–352, 2013.
© Springer-Verlag Berlin Heidelberg 2013

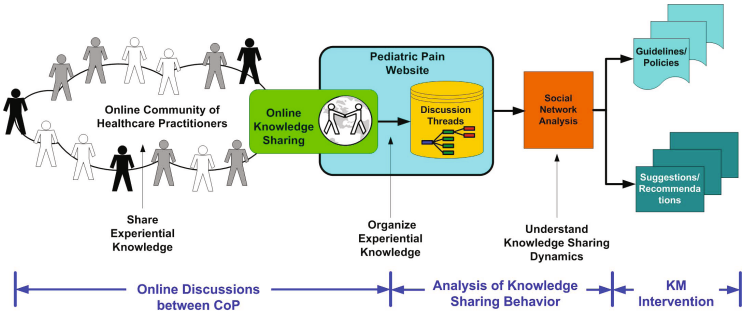


Fig. 1. Depiction of the experiential knowledge sharing framework and how SNA can leverage the knowledge sharing to inform knowledge creation

and share clinical, operational and even psycho-social experiences, along with insights and know-how about particular healthcare topics. The efficacy of this approach is that healthcare practitioners, originating from different backgrounds and expertise levels, can engage, collaborate and share their experiential knowledge for the benefit of the entire community.

Given the virtual nature of the community it is of interest to get insights about the knowledge sharing dynamics (active participants, key contributors of knowledge, topics of interest, influential members, et cetera) of the virtual community, as it helps to put a value on the knowledge created and shared there. We argue that a study of the community's communication patterns and of the knowledge content being shared can provide insights into experiential knowledge sharing dynamics of a specialized healthcare practitioner community. Social Network Analysis (SNA) allows us to analyze the communication networks within a socially-connected community (such as an online healthcare practitioner community) and highlight the key actors, interest groups, sub-networks, content seekers and experts, collaboration opportunities, communication barriers, and other network attributes. SNA focuses on analyzing the attributes that arise out of the structural properties of a social network, rather than the properties of the actors themselves, providing an overview of the community and how people communicate within it.

In order to make the results of the SNA useful to the PPML community at large, a visualization tool has been developed to allow the members to visually navigate the network and explore both the conversations on the mailing list and the social connections that have arisen out of it. VECoN (Visual Exploration of Communication Networks) is a tool that provides the means to visualize the social aspects of the network, along with functions to navigate to individual threads on the mailing list. A beta release of VECoN has been produced as a proof of concept of how we can provide further insights into the different uses of Web 2.0 technologies in healthcare.

2 Experiential Healthcare Knowledge Sharing Framework

From a health knowledge management perspective, an online mailing list or discussion forum provides a collaborative learning environment in which domain experts can

disseminate their wealth of knowledge and experience, and junior health practitioners can learn by leveraging the explicated experiential knowledge. This brings into relief an experiential knowledge sharing framework (figure 1) that allows a community of healthcare practitioners to interact and collaborate to create and share experiential knowledge, while organizing the knowledge in terms of domain specific topics. Each topic is pursed by a group of practitioners who interact through online tools, such as email or discussion forums. The discussion around a topic can be organized in terms of a 'discussion thread': a series of emails/posts by healthcare practitioners around a specific topic. Using this framework analysis of emergent social networks within the community can be performed (where the social network depicts the collaboration/communication dynamics of the community) and analysis of the content of the discussions can be done through the use of intelligent text analytics and text/knowledge creation methods. It is anticipated that subsequent SNA of the knowledge sharing behaviour of the virtual community will provide useful insights into the patterns of knowledge flow amongst healthcare practitioners. The resultant social network will provide an objective measure of the effectiveness of the online knowledge sharing medium to support collaborative learning.

The objective of the SNA is to provide meaningful insight into the flow of communication within the network. This will be provided at both the micro and macro level. Centrality measures provide insight into the roles individuals play within the network, and will help identify content experts, those members that are actively communicating with their peers and facilitating knowledge transfer. Clique and structural equivalence analysis is performed at the macro level, identifying subgroups of actors that are tightly connected. The presence of subgroups may represent particular sub-topics of interest, or groups of clinicians that are not fully communicating with the rest of the network.

The social network analysis and visualization framework is outlined in figure 2. The following sections will outlined the process in detail.

3 Background

3.1 Social Network Analysis

Social Network Analysis (SNA) is the analysis of the relations between actors, i.e. understanding the underlying social structure of a community of practice. SNA utilizes the principles of graph theory from the world of mathematics to represent communication networks in terms of actors (nodes) and ties between actors (edges) [17,8]. Traditional statistical analysis focuses on actors as independent units, and analyzes them in terms of their personal attributes. SNA instead focuses on the structures that emerge out of the relations between actors, and not on the actors themselves.

Important to this project is the notion of one- and two-mode networks. In one-mode networks the nodes in the network are homogeneous, i.e., they all belong to the same class. This is the traditional network layout, in which nodes represent people and ties represent some sort of social construct that connects them: friendship, advice, work, et cetera. Two-mode networks contain two different classes of network, and ties exist only from one mode to another. These most commonly occur when one set of nodes represents people, and the second set represents events, and ties go strictly from one mode

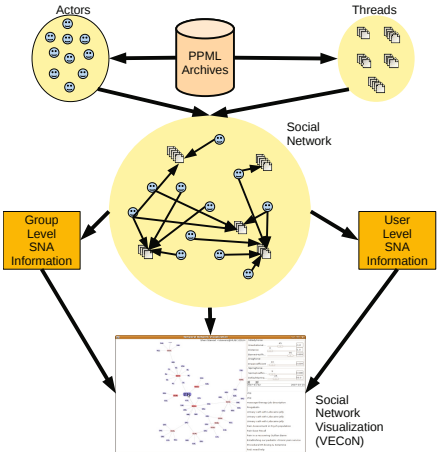


Fig. 2. SNA and visualization process. The network information is extracted from the PPML database, where it is processed to produce the network. The network is used to general social networking information, which is incorporated with the network to produce the VECoN system.

to another (indicating that a person has attended an event). This person-event design is called an *affiliation* network, and is the structure of the VECoN system. The network represented in this project presents healthcare practitioners and discussion topics as two independent classes of nodes, and the ties from a practitioner to a topic indicates that that user has communicated on the connected topic.

Because the majority of SNA methods involve one-mode networks, a common component of two-mode network analysis is to transform the data into two one-mode networks. The first network, referred to as the actor network, is created from the users in the network, and a tie is created between them if they both communicate on the same thread. The second network, referred to as the thread network, is created from the threads, and a tie is created between two threads if the same user communicated on both of them. It is preferable to analyze the two-mode network when possible, as some information is lost in the transformation, but due to the existing body of literature for SNA it is necessary to perform some analysis on the one-mode networks.

This paper will begin by analyzing the network at the micro level, using centrality measures to identify the active members and threads in the network. It will then move on to identifying potential subgroups of actors within the network using blockmodeling.

Centrality. The goal of centrality analysis is to recognize the most important actors in the network; those actors that are at the centre of the action in terms of communication between individuals. Three different centrality statistics are going to be presented; the actors identified through these measures occupy key roles in the network, and will be considered content experts.

Degree centrality is the simplest of the centrality measures, calculating the number of ties one node has to the others. For the two-mode network this is the number of threads each actor communicates on or the number of users communicating on a specific

thread. For the one-mode networks it represents the number of other actors a user has communicated with, or the number of threads one thread is connected to through shared users.

Closeness Centrality extends the idea of degree centrality beyond one step. Closeness centrality considers an actor central to the network if they can reach all the other nodes in the network in as few steps as possible. The calculation of closeness centrality differs little between one and two mode networks. If two actors are one-step away in the one mode network, they are necessarily two steps away in the two-mode network. Likewise, if two actors in the one-mode network share a common partner, then each of them must have shared a thread with that actor in the two-mode network. Though the normalization is different [3] the general ranking is the same. As such only one-mode closeness will be presented.

Betweenness Centrality deems nodes central if they are hubs of information. Where closeness deems a node central if it can quickly reach other nodes, betweenness deems a node central if it is used as a path between other nodes. A node has a high betweenness score if it falls in the shortest path between many other pairs of nodes. High betweenness scores can indicate that a specific user or thread is acting as a bridge of information, requiring others to communicate through that node. Low betweenness scores indicate that knowledge can flow freely through the network without depending on a specific user or thread. As with closeness centrality betweenness centrality does not significantly differ between one and two mode networks [3].

The next step in the SNA is to try and isolate potential subgroups within the network. With a topic as large as pediatric pain there may be evidence of subgroups in which actors are more active around a certain topic of interest. If the network were broken into groups, one would expect a lot of communication within groups, and relatively little communication between groups.

Blockmodeling. A blockmodel is a partitioning of the network into exclusive, non-overlapping groups, such that most communications are within groups rather than between them. This project uses structural equivalence blockmodeling for the 1-mode networks, and generalized blockmodeling for the 2-mode network.

Formally, two nodes are structurally equivalent (SE) if they have the same ties to all other nodes in the network. If two nodes are SE then one can replace the other without interfering with the flow of information. In reality true SE is rare, so approximate SE needs to be measured. There are many different methods used for approximating structural equivalence, but for this project Hamming distance was selected . The Hamming distance between two nodes is the number of ties that would have to have to change in order for the nodes to be SE.

Regardless of which SE measure is used, a SE matrix is developed, which records the SE between all the actors or threads. This matrix is used to group similar nodes using a hierarchical clustering algorithm. The result is a binary tree, or *dendogram*, depicting a hierarchical ranking of similarities, as in figures 6(a) and 6(b). Cutting the tree off at a particular level results in partitions being created from the clusterings. The red blocks in figures 6(a) and 6(b) represent the cutpoint at which the clusters are created. Assigning the actors to these groups creates the blockmodel.

Structural equivalence blockmodeling is sufficient for 1-mode networks, but generalized blockmodeling [4] is the method of choice for two-mode networks. The general idea is to partition the rows and the columns of the matrix into groups such that the groups are as pure as possible (either all connected or all disconnected). Generalized blockmodeling is performed using a local optimization procedure, details of which can be found in [4,18].

3.2 Network Visualization

The visualization of networks is a key component of SNA, and as such there is a rich literature base describing methods of presenting networks visually. Linton Freeman [5] documents the history of social network visualization from a sociological perspective, including theories on node layout (both information-based and algorithmic theories) along with the use of colour, size and shape to encode network information. There are many current tools for analytic network visualization, including `UCINet` [2] and an extension for the `R` statistical language called `statnet` [7].

Previous work on social network visualization has also been directed towards network navigation. Examples include ContactMap [14] for identifying community groups within email contacts, PieSpy [13], which provides a real-time visualization of social networks for Internet Relay Chat (IRC) members, and Vizster [9], a tool for exploring the Friendster (www.friendster.com) social networking site. These tools are all designed for 1-mode networks, for example, the nodes in the Vizster program all represent users of Friendster, and the ties represent friendship links between them. In contrast, this project is visualizing a 2-mode network, where the first class of nodes represent mailing list members and the second represents threads, and the links between a node and a thread indicate that a certain list member has communicated on that thread.

The software being used to implement this project is the `prefuse` toolkit in Java [10]. `Prefuse` was chosen because it provides a full Java library, and previous implementations of `prefuse`, including the Vizster program, have proven successful.

4 VECoN System

4.1 Visualization

The purpose of the visualization is to first provide a tool for visually exploring social networks, and secondly to provide some insight into the underlying social structure of the network. This section will outline the graph-theoretic layout decisions for the network and the visualization tools implemented to help the exploration of the mailing list.

Graph Structure. The network is laid out using a force-directed layout, in which the nodes repel one another and the edges act as "springs" that hold the nodes together. Because of this spring effect, the layout is also sometimes referred to as a "spring embedding" algorithm. `Prefuse` implements the Barnes-Hut algorithm [1] which allows for real time calculation of spring-embedding forces. The algorithm is an iterative process, and following the lead of Vizster, this project chooses to not limit the number of iterations of the algorithm, resulting in a visualization in which the nodes migrate to

their optimal positions but continue to move subtly. The effect is "a living or 'breathing' feel, connoting social energy or playfulness."[9].

Two changes that need to be made to the spring embedding algorithm are i) an adaptation to two-mode networks, and ii) dealing with the problem of components (disconnected sections of the graph). The two-mode issue is addressed by Krempel [12], in which he suggests fixing the second mode and allowing the first mode to vary. To this end, the algorithm will be adjusted slightly: the first step is to disperse the threads evenly around the space, and then allow the actors to move according to the spring embedding algorithm. This method has proven to be effective, but a more formal evaluation of its efficiency is required.

The issue of components is addressed in Kamada and Kawai's seminal work on force directed layouts [11]. The solution is to partition the space according to the number of components in the graph, with each graph being allocated space proportional to its size (number of nodes). The forces are then calculated only on the nodes within the component. This means that components do not effect each other in terms of layout, and avoids the "drift" that is caused when disconnected components continue to push each other away (see figure 72 in [6]).

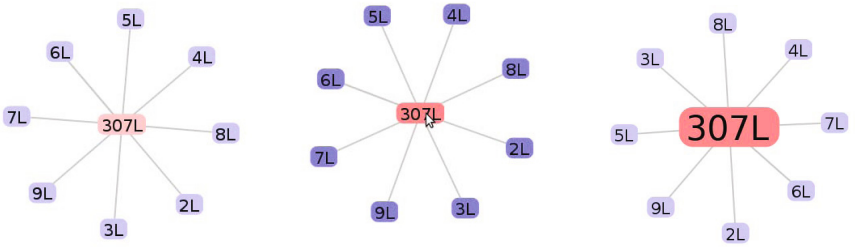
The colouring of the nodes is defined by their mode: blue indicates actors and red indicates threads. Following Vizster's lead, nodes and their neighbours are highlighted when the mouse hovers over them: when hovering over an actor that actor's threads are highlighted, and conversely when hovering over a thread the actors that communicated on that thread are highlighted. The highlighting is done by increasing the saturation of the colour (see figure 3).

Exploration. The visualization provides several different exploration methods, which will each be described in detail. The objective of each of the tools is to provide a different way of exploring the mailing list to retrieve pertinent messages.

Hover Over As mentioned before, the visualization implements a hover-over feature. When the mouse hovers over a node, that node becomes fixed, and it and its neighbours are "highlighted". A node is highlighted by increasing the saturation. Figure 3 demonstrates the difference between the regular and highlighted nodes.

Along with changing the representation of the nodes, the hover-over feature presents additional information about the node/edge in question. For actors it presents the actor's name in the top-right corner of the visualization, and for threads it presents the thread's subject line (see figure 7). The node names were left out to avoid cluttering the visualization: current node labels are numerical identifiers, provided in order to differentiate between them.

On the right side of the visualization is the control bar, where users can manage their search features and control the factors that effect the spring embedding. At the bottom of the control bar is a text box that presents the detailed content of the selected node. For actors a list of the threads they have communicated on is presented, and for threads the conversation is presented. When an edge is selected then that specific message is presented. The objective is to allow the user to quickly browse through the conversations, and upon finding the desired message to explore it in more detail. Future work will pursue connecting the visualization directly to the online discussion forum, but as the forum is not yet developed this feature is not implemented.



(a) Regular. (b) Highlighted. (c) Focus Node.

Fig. 3. Highlighting is achieved by increasing the saturation from 25 to 50. A focus node is activated by clicking it, at which point it doubles in size.

Point and Click. The objective of the point and click option is to allow the user to explore a particular node in more detail. When a node is clicked it doubles in size, and the text associated with it becomes fixed. The user can then adjust the temporal filter to see where that node fits within the filtered data. While the node is in focus (i.e. while its size is doubled) the text stays static, though the rest of the highlighting effects associated with hovering remain. A second click on the node reduces it back to normal size and allows the text to change freely. Figure 3 demonstrates what a focus node looks like.

Temporal Filtering. Adding a temporal filter provides a way to reduce the volume of information being presented. The temporal filtering is accomplished using a horizontal scroll bar, located within the control panel. The user can manipulate the lower-end or upper-end of the bar to adjust the visualization, or can set the bar to a specific window width and slide the window itself. When the mouse is released the graph is re-drawn with only those messages that are within the window presented. The user can also manipulate the dates manually by typing dates into the two date boxes that are below the bar.

When the visualization is adjusted, only those nodes that are connected to each other are presented, reducing clutter by removing obsolete nodes. There is still a potential for isolates in the filtered network, however. If an actor participates in a thread, but that contribution is not made during the window, then the edge will be removed but the actor will remain in the network, and that contribution will still be presented in the hover-over effect. Future work should explore how to incorporate the disconnected user into the force-directed layout, but for now the user is assigned a separate segment of the space.

5 Methods

5.1 Data

This project will use the archives of the Pediatric Pain Mailing List (PPML), provided to the VECoN project by the administrators of the mailing list. There are over 13,000 messages in the PPML archives, dating back to 1991; for this project a sample of the

messages from 2007 and 2008 was used (1096 messages). The sample has been parsed from simple ASCII text files and the messages have been written into a MySQL database. Along with sender, subject and date information about each message, a thread number has been assigned to indicate which messages combine to form conversations. The threads and senders will be used as nodes in the network, with ties between them indicating that a sender has communicated on a particular thread. Currently the PPML administrators are in the process of transferring the archives of the mailing list to an online discussion forum. Once this forum is active the VECoN system will be connected to it.

There are 270 individual users in the network, communicating on 353 threads. Many of these users/threads are isolates in the network, only sending one or two messages that were not responded to. Many SNA methods require a connected network, therefore those users that did not connect with the main body of the network and those threads that did not receive a reply are removed from the analysis. The result is a network with 1051 messages from 232 users in 304 threads.

The data were analyzed with R 2.12.2 [15] using the `statnet` [7] and `blockmodeling` [19] libraries.

Social Network Visualization. Beyond simple network exploration the tool provides functions to explore the SNA dimensions of the PPML. Within the control panel there are buttons to toggle centrality indicators and blockmodeling. When activated the centrality indicators redefine the actor and thread nodes in the network, setting the node size relative to that node's degree. Currently only degree centrality is implemented, but all three centrality measures will eventually be added.

The visualization also attempts to present the blockmodels to the user. For the 1-mode methods, it is left to the user to decide how many blocks to assign to the network, presenting the blocks as "blobs" surrounding the component members. When activated the blobs modify the spring embedding algorithm to increase tensions within the blobs and decrease tension outside them, resulting in tighter groups of block members. Currently the 2-mode clustering has not been implemented in VECoN, due to the challenges of creating clusters of different sizes in real time.

6 Results

6.1 SNA

The SNA will be broken into three sections: a micro-level analysis of the actors and threads individually, then a macro-level analysis of the blockmodels.

Micro Level Actor Analysis. Table 1(a) lists the actors with the highest degrees in the two-mode network, and table 1(b) lists the same for the one-mode network.

With a maximum normalized degree of 0.181 in the two-mode network no single actor is participating in many threads, indicating that conversations are occurring between different users. A normalized degree of 0.368 in the one-mode network demonstrates a slightly higher level of connectivity between users. Though there is not a single user

Table 1. Centrality measures calculated for the actors in the PPML network

(a) Actors with the highest degree centrality in the two-mode network.

Actor	Degree	NormDegree
121	42	0.1810
167	41	0.1767
066	36	0.1552
055	35	0.1509
170	31	0.1336

(b) Actors with the highest degree centrality in the one-mode network.

Actor	Degree	NormDegree
167	85	0.3680
128	66	0.2857
066	65	0.2814
170	63	0.2727
055	58	0.2511

(c) Closeness scores for the Actor network.

Actor	Closeness
167	0.5938
128	0.5607
066	0.5553
055	0.5474
170	0.5461

(d) Betweenness scores for the actor network.

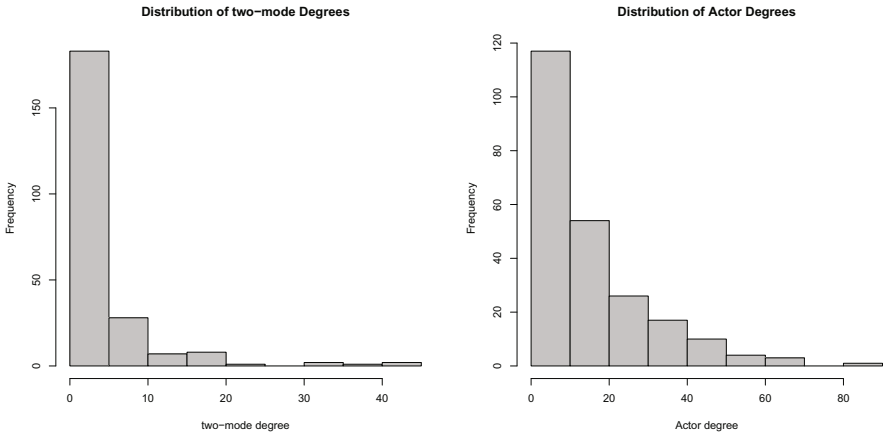
Actor	Betweenness
167	0.119
066	0.088
170	0.070
128	0.070
179	0.061

present in all conversations, the users are well connected to one another. The histograms in figure 4(a) confirm these findings; the majority of actors in both the two-mode network have low degree measures, while they are slightly higher in the 1-mode network.

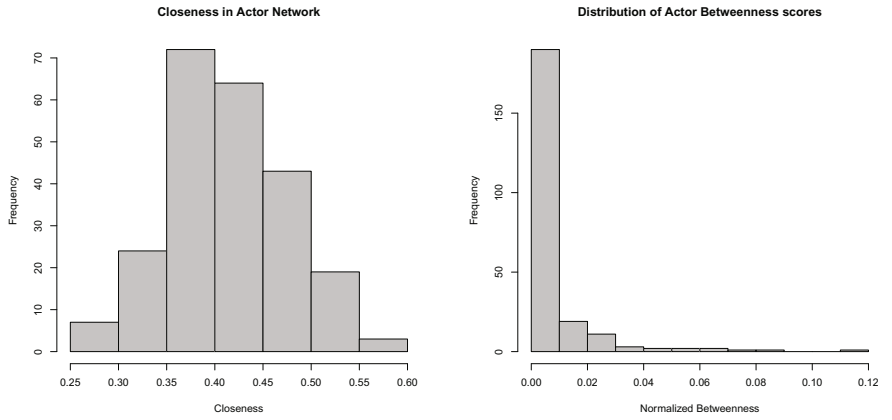
Table 1(c) lists the highest closeness scores in the network. With a max closeness of 0.594 the network seems to have a very high closeness measure. This means that it is easy to get from any node to any other node. These findings are confirmed in the histograms in figure 4(b). This histogram seems to follow a normal distribution, centred at 0.414 with a standard deviation of 0.061, meaning that 97.5% of the members of the network have a closeness centrality greater than 0.292.

Table 1(d) lists the highest betweenness scores from the network. The low maximum betweenness of 0.119 indicates that the network does not depend on any particular actor to facilitate communication. The histogram in figure 4(c) confirms that finding, with the majority of users having low betweenness.

Overall, the centrality scores for the users indicate a healthy and active knowledge sharing network. Low degree centralities mean that there is not a single user, or a set of users, that are required to initiate conversation. High closeness scores mean that it is easy to connect one user to another, either through a thread they have both communicated on, or a short series of "mutual friends"; this is key to facilitating knowledge transfer, allowing users to connect directly to the source of experiential knowledge in as few steps as possible. Low betweenness scores compound this finding by demonstrating that, when trying to connect with someone through a series of users, there are always multiple paths available, without the need to always communicate through a single user. These results combine to describe a knowledge sharing network that has a large number of healthy communicators and an active user base.



(a) Degree Distribution.



(b) Closeness Distribution. (c) Betweenness Distribution.

Fig. 4. Distribution of centrality scores for the actors in the PPML network

Though the centrality results do not demonstrate the presence of any dominant users, it is noteworthy that the same actors appear at the top of all the centrality measures. Table 2 lists the top ranked actors for each of the centrality measures, and the same set actors appear in each list. If further investigation into the content of the mailing list were required, these actors should be the first to be contacted, as they seem to be the most active and most important to the communication network.

Micro Level Thread Analysis. Tables 3(a) and 3(b) list the highest degrees in the two-mode and one-mode networks respectively.

The degree centrality analysis of the 2-mode network reveals that there are not any dominant threads. A maximum normalized degree of 0.076 in the two-mode network indicates that there is not one subject that everyone commented on. The one-mode

Table 2. The ranking of actors provided by each of the three centrality measures

Degree	Degree.2M	Closeness	Betweenness
167	121	167	167
128	167	128	066
066	066	066	170
170	055	055	128
055	170	170	179
056	035	121	121
121	148	042	035
184	179	184	184
020	184	035	020
179	020	056	056
035	128	020	055
042	224	045	266
254	102	077	015
045	146	179	224
077	015	015	096

Table 3. Centrality measures for the threads in the PPML network

(a) Threads with the highest degree centrality in the two-mode network.

thread	Degree	NormDegree
T090	24	0.0764
T091	22	0.0701
T218	19	0.0605
T293	19	0.0605
T247	15	0.0478

(b) Threads with the highest degree centrality in the one-mode network.

Thread	Degree	NormDegree
T037	108	0.3450
T090	105	0.3355
T248	99	0.3163
T170	98	0.3131
T017	94	0.3003

(c) Closeness scores for the threads in the PPML network.

Thread	Closeness
T248	0.5712
T090	0.5670
T037	0.5660
T017	0.5619
T293	0.5511

(d) Betweenness scores for the threads in the PPML network.

Thread	Betweenness
T090	0.0815
T293	0.0639
T248	0.0453
T328	0.0429
T017	0.0384

network is more centralized, with a max normalized degree of 0.345. This means that the most popular thread is only one step removed from 34.5% of the other threads. The histograms in figure 5 indicate that the majority of threads have fewer than 10 comments, but that they are more connected than the actors are.

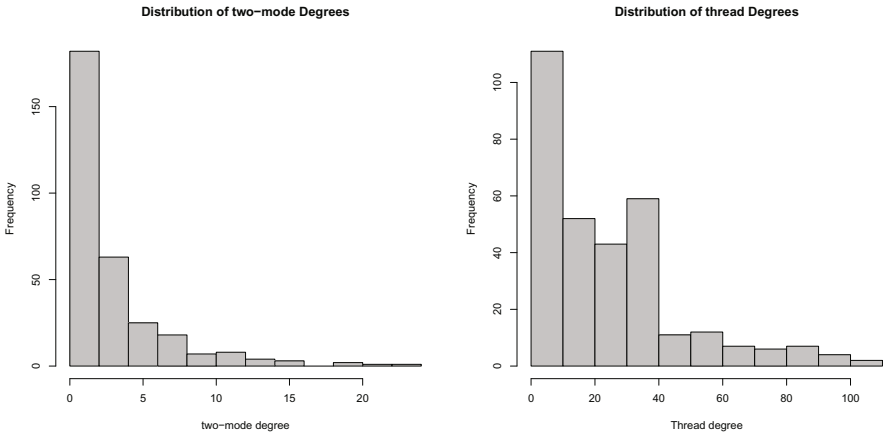
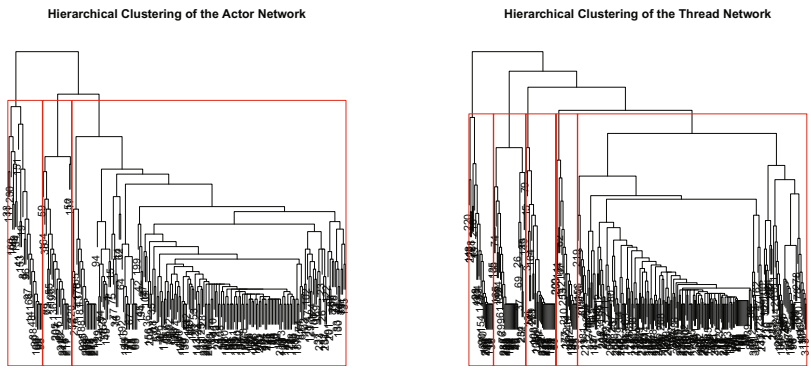


Fig. 5. Distribution of thread degrees for the PPML data. The histograms indicate that threads usually received fewer than 10 comments, but that the threads are highly connected.

Tables 3(c) and 3(d) present the closeness and betweenness scores for the threads, and confirm the results of the actor measures: that the network is tightly connected, without a single thread dominating the flow of knowledge.

The thread centrality measures confirm many of the findings from the actor-level analysis. There is not a single thread that dominates the network, and threads seem to be well connected, without one single thread dominating the flow of information. One noticeable difference is the distribution of one-mode degrees, indicating that the thread network is a slightly more connected network.



(a) 1-Mode Actor Clusters. (b) 1-Mode Thread Clusters.

Fig. 6. Hierarchical clustering of the actors and threads using structural equivalence. The red blocks indicate the points at which the tree is cut to create the blockmodel.

Macro Level Analysis. A blockmodel was developed for each of the 1-mode networks using structural equivalence, as shown in figures 6(a) and 6(b) for actors and threads

respectively. These blocks produce *image matrices*, that report the density of communication in each of the blocks. The image matrices for the actors and threads are in tables 4(a) and 4(b) respectively. For both of the block models, the algorithm identified

Table 4. Image matrices depicting the communication densities between blocks in the 1-mode actor and thread blockmodels

(a) Actor Image Matrix.

	B1	B2	B3
B1(n=188)	0.03493	0.09796	0.07128
B2(n=24)	0.09796	0.67754	0.13958
B3(n=20)	0.07128	0.13958	1.00000

(b) Thread Image Matrix.

	T1	T2	T3	T4	T5
T1(n=213)	0.040	0.039	0.046	0.067	0.047
T2(n=28)	0.039	1.000	0.138	0.084	0.114
T3(n=20)	0.046	0.138	1.000	0.154	0.287
T4(n=23)	0.067	0.084	0.154	1.000	0.187
T5(n=30)	0.047	0.114	0.287	0.187	1.000

a large, disconnected block of nodes, and then smaller, denser blocks. For the actors it identified a fully connected clique in block 3, and a well connected group in block 2, and for the threads it identified four perfectly connected cliques, and then a large group of less connected nodes. Note as well that there is some communication between thread blocks, with densities at or above 0.1.

Moving to generalized blockmodeling can provide much more insight into the network by working directly with the two-mode data. Using the blocks found with the 1-mode method, the generalized block model will permute the blocks until it finds the best partitioning of the network such that the blocks are as pure as possible: densities of either 1 or 0. The resulting image matrix in is table 5(a), along with the block sizes. Once again the algorithm has identified a large, sparse block, and several smaller, more active blocks. The interaction between the blocks is much more interesting. Looking at number of messages (table 5(b)) per block, the second actor block, block 2, has a lot of communication with thread blocks 2 and 5 but little communication with thread-blocks 3 and 4. The third actor block is most active with the fourth thread block, rather than the other three small thread blocks.

Table 5. The communication densities and number of messages within the generalized blocks

(a) Image Matrix.

	1(n=193)	2(n=16)	3(n=23)
T1(n=193)	0.006	0.035	0.004
T2(n=17)	0.017	0.287	0.046
T3(n=45)	0.004	0.000	0.048
T4(n=20)	0.017	0.056	0.139
T5(n=39)	0.009	0.144	0.011

(b) Number of Messages.

	1(n=193)	2(n=16)	3(n=23)
T1(n=193)	217	108	18
T2(n=17)	56	78	18
T3(n=45)	34	0	50
T4(n=20)	64	18	64
T5(n=39)	65	90	10

6.2 Visualization

Figure 7 is a picture of the VECoN system. The left pane contains the network visualization, and the right pane contains the spring-embedding controls, the filter controls and the message pane where the contents of the threads or the participating threads are presented.

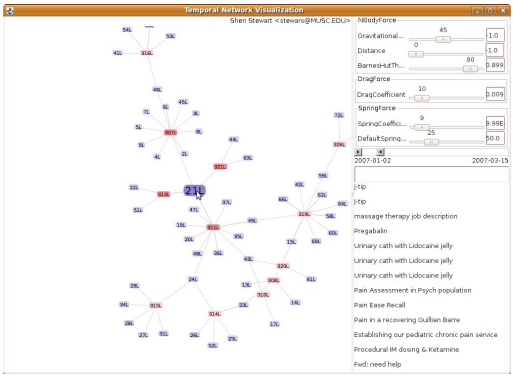


Fig. 7. The PPML visualization. Note that the current network is restricted to messages between 2007-01-02 and 2007-03-21.

7 Discussion

The SNA provided some useful insight into the communication patterns within the network, but further analysis is required to fully flush out what the results mean in the scope of the project.

7.1 Isolates

Formally, an isolate is a node in the network that does not connect to any other node. Expanding the definition to account for the structure of 2-mode networks, isolates in this context are actors or threads that are not connected to the main body of the network. There are 38 users and 39 threads that did not connect to the main body of the network. Reviewing the subject-lines of these threads, many of them were ignored because they were not pertinent discussion subjects: spam/incorrect administrative emails or job and conference announcements. Ignoring those non-pertinent messages there are 17 users who posted legitimate content to the mailing list and got no reply, and 16 threads that received no response.

It is imperative to the health of a mailing list that new users receive responses to their questions. The results of the isolate analysis are positive; evidence that knowledge seeking queries are being answered the majority of the time. Moving forward it is vital that the PPML continues to be an active community by incorporating new members and responding to their queries.

7.2 Centrality

The centrality analysis has indicated that there is not a central user or set of users that control the communications on the network. For each of the four measurements the highest scoring actors do not have disproportionately high normalized values, indicating that there is not one actor that communicates on every thread, or connects disparate groups of actors. Though the low degree centrality scores could be interpreted as members being inactive, a more reasonable interpretation is that there are many messages on the mailing list that are spawning conversations between different actors. This is a very promising result moving forward.

The thread-centrality measures have similar findings. In the 2-mode network there is not a single thread that has received the bulk of the attention, indicating that knowledge is being shared on a wide variety of threads. The 1-mode network has a higher degree than the actor network, indicating a more tightly knit group: it is easier to get from one thread, through it's authors, to another thread in the network. Tightly coupled networks allow knowledge to flow more freely, which is also confirmed by the high closeness and low betweenness scores for the events.

Detecting content experts or knowledge sharing activities strictly using graph-theoretic principles such as centrality analysis poses problems. If someone contributes to the list by asking many questions, without providing answers, they are recognized as central users, but should not be considered content experts. In order to improve the prediction of content experts it is necessary to extract the underlying semantics of the messages being communicated. Previous work, [16], has worked on extracting semantic representations of the messages, and incorporating this information into the SNA could help differentiate the content experts from the "question askers."

7.3 Blockmodeling

The blockmodeling has provided a valuable clustering of the nodes in the network. The 1-mode blockmodels potentially isolated interesting groups of actors/events, separating the active users/threads from the inactive. The one-mode partitions were most valuable, however, in informing the 2-mode clustering, which took the 1-mode clusters and created a generalized blockmodel that split the users into three groups and the threads into five. For each mode a large cluster of less-active users was created, but the interactions between the smaller groups is what is of more interest. There are strong interactions in the second row block and the second column block (the (2,2) block), as well as in the (2,5) block and the (3,4) block. It's also of note that the (2,3) block is completely empty.

The next steps should be to explore these blocks: what is it about these users and these subjects that is causing such a clustering? Surveys of the users should be conducted so that user attributes can be incorporated into the clusters, and the semantic representations of the messages could provide valuable information about the thread clusters.

7.4 Visualization

A beta release of the *VECoN* system has been produced, but it has not yet been tested by the PPML members. Once the mailing list is available as a discussion forum and

links between the visualization and the forum are established the *VECoN* system will be rolled out as a Java applet, and at that point research will be conducted on its utility and on future additions. Future research is being conducted to improve the spring-embedding algorithm and the blockmodeling algorithm, along with adding more centrality measures and a more intuitive control panel.

8 Conclusions

Experiential healthcare knowledge is a vital component of the current healthcare system, and developing new methods to facilitate the sharing of this knowledge is vital to sustaining and improving the medical community. Medicine 2.0 technologies provide online tools for facilitating knowledge sharing, establishing virtual communities of clinicians. Understanding the flow of knowledge in these virtual communities is key to developing new systems, and SNA provides the necessary tools for understanding the flow of communication within the network. It has provided a list of potential content experts within the list, it has recognized several active subgroups, and it has partitioned the network into disparate groups of potentially different clinicians and subjects.

Further research should be directed at better understanding the members of the community. User attributes such as specialty, location, and job description would provide better insight into the structure of the network, and in particular into the structure of the subgroups revealed through blockmodeling. Incorporating the semantic information of the communications themselves would help differentiate between acts of knowledge seeking and knowledge sharing, and improve our overall understanding of the experiential knowledge available in the network.

Though the *VECoN* project is only in its beta stages, preliminary results are promising. The network has been visualized, and SNA tools have been added. Future work will be on adding new SNA tools, improving the visualization, implementing the system online and making the interface more intuitive. The ultimate goal of the *VECoN* system is to provide a novel network exploration tool to help users make new connections within the PPML community, and find new sources of experiential knowledge already available in the network.

Acknowledgements. This work was carried out with partial support from the Global Health Research Initiative (GHRI), a collaborative research funding partnership of the Canadian Institutes of Health Research, the Canadian International Development Agency, Health Canada, the International Development Research Centre, and the Public Health Agency of Canada. The authors would like to acknowledge Dr. Allen Finley for his contribution of data to the project and his ongoing support of this research.

References

1. Barnes, J., Hut, P.: A hierarchical o(n log n) force calculation algorithm. Nature 324 (1986)
2. Borgatti, S., Everett, M., Freeman, L.: Ucinet for windows: Software for social network analysis. Analytic Technologies, Harvard (2002)

3. Borgatti, S.P., Everett, M.G.: Network analysis of 2-mode data. Social Networks 19, 243–269 (1997)
4. Doreian, P., Batagelj, V., Ferligoj, A.: Generalized blockmodeling of two-mode network data. Social Networks 26, 29–53 (2004)
5. Freeman, L.C.: Visualizing social networks. Web (1999), http://www.cmu.edu/joss/content/articles/volume1/Freeman.html
6. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software-Practice and Experience 21, 1129–1164 (1991)
7. Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Morris, M.: statnet: Software Tools for the Statistical Modeling of Network Data. Seattle, WA. Version 2.1 (2003), Project home page at http://statnetproject.org
8. Hanneman, R.A., Riddle, M.: Introduction to social network methods. University of California, Riverside, CA (2005)
9. Heer, J., Boyd, D.: Vizster: Visualizing online social networks. InfoVis (2005)
10. Heer, J., Card, S., Landay., J.: Prefuse: A toolkit for interactive information visualization. In: CHI (2005)
11. Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Information Processing Letters 31, 7–15 (1989)
12. Krempel, L.: Visualizing networks with spring embedder: Two-mode and valued data. In: American Statistical Association, Proceedings of the Section of Statistical Graphics, pp. 36–45 (1999)
13. Mutton, P.: Inferring and visualizing social networks on internet relay chat. InfoVis, 35–43 (2004)
14. Nardi, B., Whittaker, S., Isaacs, E., Creech, M., Johnson, J., Hainsworth, J.: Contactmap: Integrating communication and information through visualizing personal social networks. Communications of the ACM 45(4), 89–95 (2002)
15. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2011) ISBN 3-900051-07-0
16. Stewart, S., Abidi, S., Finley, A.: Pediatric pain management knowledge linkages: mapping experiential knowledge to explicit knowledge. Studies in Health Technology and Informatics 160, 1184–1188 (2010)
17. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
18. Ziberna, A.: Generalized blockmodeling of valued networks. Social Networks 29(1), 105–126 (2007)
19. Ziberna, A.: Blockmodeling: An R package for Generalized and classical blockmodeling of valued networks. R package version 0.1.8 (2010)

A Semantic-Based Platform for Medical Image Storage and Sharing Using the Grid

Daniela Giordano[1], Carmelo Pino[1], Concetto Spampinato[1], Marco Fargetta[2], and Angela Di Stefano[3]

[1] Department of Electrical, Electronics and Computer Engineering
University of Catania, Viale A. Doria, 6, 95125 Catania, Italy
{dgiordan,cpino,cspampin}@dieei.unict.it
[2] Italian Institute for Nuclear Physics (INFN)
Via Santa Sofia 64 95123 Catania, Italy
Marco.Fargetta@ct.infn.it
[3] Institute of Neurological Science, Italian Research Council (CNR)
Via Gaifami, 18. 95127 Catania, Italy
a.distefano@isn.cnr.it

Abstract. Since the introduction of medical imaging techniques such as MRI, X-Ray, SPECT, PET, large amounts of medical images have been produced in the different fields of medicine. Most of this data is usually stored either in paper format (stored in clinical records) or in optical disks (given to the patients) and this information is partially uncorrelated with the clinical history of patients. Furthermore, the image processing carried out by the physicians that perform the exam and used to make the diagnosis is unknown for other researchers that further treat the same image. In this paper we propose a system that allows radiologists to share medical images (and their processing) and the associated metadata both within the same medical institute and with other medical institutes located anywhere in the world by using GRID services for data (LFC) and metadata (AMGA) storage. The system is also provided with a semantic layer for describing the stored images through a novel RDF schema, which integrates existing ontologies and vocabularies such as FOAF, Mesh and GeneOntology with new terms related to the image processing part. This enables the correlation of medical images with other clinical information and makes our system fully compatible with the existing systems compliant to semantic web standards.

1 Introduction

Usually, radiologists carry out examinations, e.g. MRI, X-Ray, SPECT, PET, etc. , without a complete knowledge of the clinical history of the patients. In fact, most of the raw data collected in experiments or clinical trials is usually stored either in different places or in separate paper forms. Therefore there is a real need of integrating heterogeneous data sources in order to have a holistic vision of the patient's health status. Currently, existing methods, e.g. [1], [2] barely address medical data management needs beyond the specific department's boundaries, while it is known that the patient medical folders are wide spread over many medical sites involved in the patient's healthcare. For this

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 353–364, 2013.
© Springer-Verlag Berlin Heidelberg 2013

reason, many efforts have been done in last years to develop an interoperable infrastructure for digital health ([3], [4]) by using semantic web concepts, even if the attention is mainly oriented to metadata sharing. A relevant project in this direction is the e-Child project [4] that develops a healthcare platform for pediatrics and aims at integrating ontologies to homogenize biomedical (from genomic, through cellular, disease, patient and population-related) data. The system proposed in this paper follows, and for some aspects, overcomes the above medical data management's approach because it permits to store images associating them with not directly related metadata (i.e. not included in the DICOM file) by developing an RDF schema thatlinks the most common ontologies used in the medical field, thus making our system interoperable with methods that share data following semantic web (e.g. in [5]).

Given the huge amount of medical data involved in this process and the need of collaboration among researchers our system is provided with a GRID layer. Many GRID systems have been proposed for handling medical data geographically distributed on various medical centers, such as HOPE [6], MediGrid [7] or MAMMOGRID [8], but most of them deal only with data files and do not provide higher level services for manipulating medical data or for handling the associated metadata, and also lack in the integration with other systems.

Therefore, in this paper we propose an approach for sharing and analyzing medical images that extends the system previously proposed in [9]. The system herein proposed handles all the types of medical images, instead of only nuclear medicine images, and it is provided with a novel RDF schema for describing patient data. The key features of the proposed system are:

- Full functionality in cases of network malfunctioning is ensured since data are stored both locally (in the user's computer and in the medical institute server) and globally on GRID;
- It is interoperable with existing sharing methods that use semantic web;
- It is flexible since radiologists can share content choosing when and what to share;
- A RDF schema is proposed to enrich the metadata usually contained in the DICOM files in order to take into account the whole clinical history of patients.
- Data encryption is implemented using SSH, which ensures data integrity over Internet transfers;
- Data privacy is guaranteed by removing sensitive information before any transmission.

The paper is so organized: in section 2 the storage and sharing system is described. In detail, we present how the system works inside a medical institute using a client-server architecture and how it works among different institutes using GRID services. Sect. 3 and Sect. 4 describes, respectively, the high level features for managing the interoperability with other systems, for image analysis and for system querying and the user interface. Finally, concluding remarks are given.

2 Storage and Sharing System Overview

In order to develop a distributed environment for image and information sharing to support the diagnosis, the treatment of patients and for statistical evaluation, the system

is provided with two levels of storage and sharing: the first is locally managed by a client-server architecture, deployed in the medical institute where the radiologists belong to, whereas the second one is on GRID and allows global data sharing, i.e. data may be shared among different institutes using the services offered by the GRID computing. Fig. 1 shows the architecture of the proposed system for the local and global data sharing.

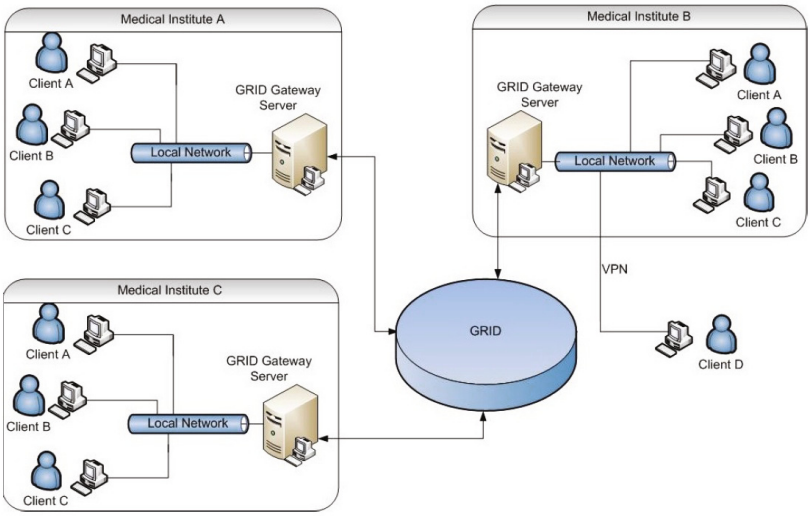


Fig. 1. Architecture for Local and Global Data Storage and Sharing

The typical use case is the following: a radiologist stores the images and the metadata of a performed examination in its own local database (located in his/her computer). Afterwards, the client creates an anonymous version of the data removing all the confidential information so they can be sent to the main server in the respect of privacy issues. Additionally, the client allows users to define the set of metadata he/she wants to share both in GRID and in his/her medical institute.

The data transmission between client and server runs asynchronously in order 1) to make the system robust, in fact in case that, Internet connection is unavailable, data are locally stored and subsequently sent to the main server and to GRID when the connection will be available again and 2) to keep users unaffected by the actual time needed for the data transfer.

The server is provided with repositories where all the metadata produced within the same institute are stored. The communication with GRID is delegated to the server, thus optimizing the bandwidth's use.

2.1 Local Data Storage and Sharing

Inside a medical institute, data are stored and shared using a standard client-server architecture, as shown in fig. 2. The client and the server are connected by a local network or a VPN (Virtual Private Network).

The client contains the user interfaces and implements the logical communication with the GRID infrastructure. It also contains a file repository (for image storage) and a SESAME server [10] (for RDF metadata storage), in order to save patient's data locally. The server also includes a file repository and a SESAME metadata repository for the data produced by all the radiologists within the same institute. Data are sent from the client to the server using File Transfer Protocol (FTP), whereas metadata is transmitted to SESAME server using Simple Object Access Protocol (SOAP) requests by means of a webservice, as shown in fig. 2.

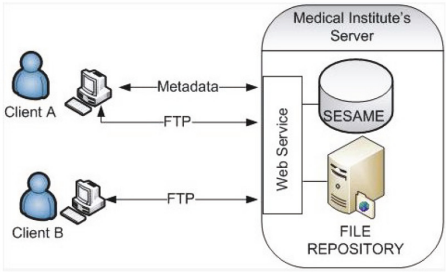


Fig. 2. Local Data Storage and Sharing

By using the client interface, a radiologist can record and manage patients, add information to patient's clinical history (according to the schema shown in the next section), include any relevant documents (textual reports, generic images, DICOM images, etc..), run queries locally or on GRID data, associate the metadata deriving from the queries to the data locally stored and perform statistical analysis on sets of data and virtual data (i.e. coming from the main institute center or from GRID).

2.2 Global Data Sharing on GRID Using LFC and AMGA

Data sharing among different institutes, geographically distributed, is implemented on GRID and it is based on the paradigm to create virtual environments where large amount of data and complex computations can be performed by different communities grouped in *Virtual Organisation (VO)*. Usually, each VO offers several services for GRID participants in order to simplify the data management providing them basic functionality to store and retrieve files.

The access to the GRID is hidden by a middleware. In this work we used the EGEE Grid and the G-Lite middleware[1]. The two main services of the G-lite middleware for data and metadata storage, used in our system, are: the *Logical File Catalogue (LFC)* [11] and the *AMGA Metadata Service* [12]. The LFC allows users to associate a logical name to a file in a hierarchy format like a local file system, hiding the real location of the storage. Moreover, a logical name may refer many replicas, so a user can retrieve the file through its logical name from the nearest location in a transparent way. The AMGA Metadata Service is a special database designed to store metadata associated with files.

[1] EGEE website - `http://www.eu-egee.org/` and gLite Grid middleware website - `http://www.glite.org/`

Therefore, its internal structure reproduces a file system hierarchy where the directories are collections of metadata defined in a custom schema and each file in the directory is an entry containing the values for the metadata. This approach allows to easily map a file name with a set of metadata inside AMGA.

The communication with GRID infrastructure is managed by the medical institute server, which also aims at maintaining aligned the information stored in radiologists' computers and the one stored in GRID. In each institute the main server represents the bridge between clients and GRID, as shown in fig. 3.

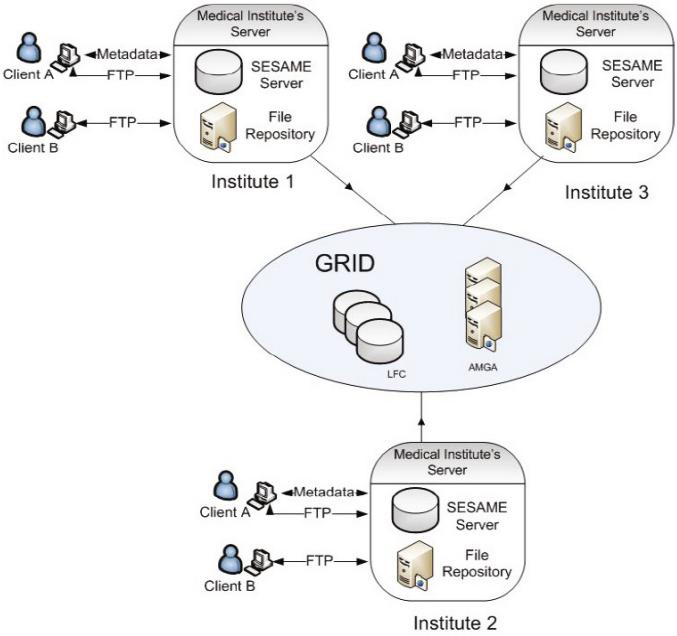


Fig. 3. Global Data Storage and Sharing

The flow diagram of the interaction between a generic medical institute and the GRID infrastructure is shown in fig. 4, when a radiologist requires the storage of a specific image on GRID by using a proper Graphical User Interface (GUI). The steps performed by our system are: 1) the client sends to the server the proxy certificate (needed for the access to GRID) previously created and the identifier of the image to be stored on GRID, 2) the server of the medical institute (for simplicity called proxy server) queries the file repository to check if the file is present or not, if not it asks the client to send the image, then it sets the necessary permissions to read, write for the GRID, 3) the server queries the GRID database using the medical institute Grid ID given by the client in order to see if the image was already stored, if not 4) the server removes sensitive data, sends it to a Grid Storage Element, writes in the LFC catalog an appropriate logical name and writes (uploads if the image was previously stored) the metadata in the AMGA server.

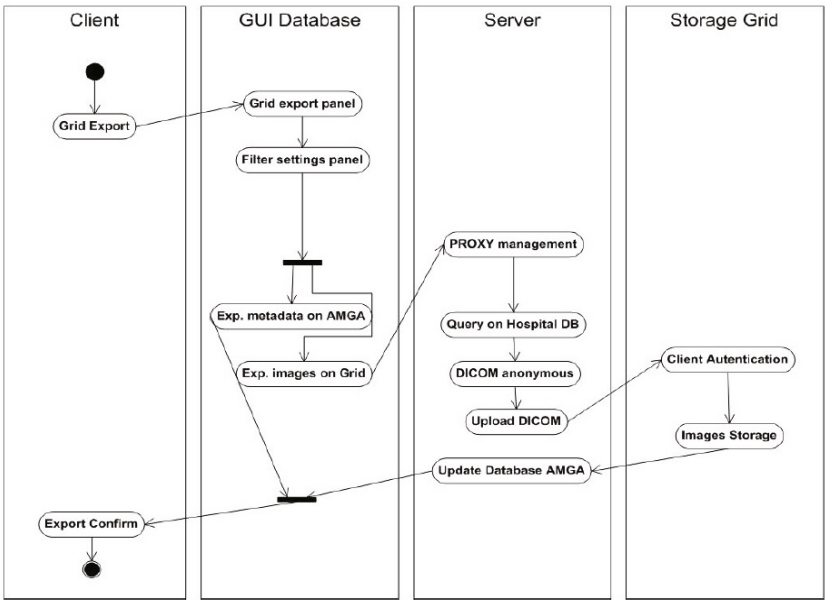


Fig. 4. Interaction with Medical Institute - GRID Infrastructure

3 High Level Features

In order to provide useful information about the stored images and to make them available with the related metadata to the medical imaging community, the system is provided with high level features. More in detail, the system contains three processing levels:

- A semantic layer that enriches patient metadata by an RDF schema. This level guarantees the interoperability with existing semantic-web compliant frameworks;
- An image processing layer that analyzes the stored images. This is an important layer, since sometimes is desirable to share also the results of the image processing. This level performs the image analysis and interacts with the semantic layer for the storage of the processing results in RDF/XML;
- Query Composition for performing complex queries both locally and on GRID. This module allows users to search useful information by processing only the metadata available locally or in GRID.

The interaction between the three levels and the system's architecture is shown in fig. 5.

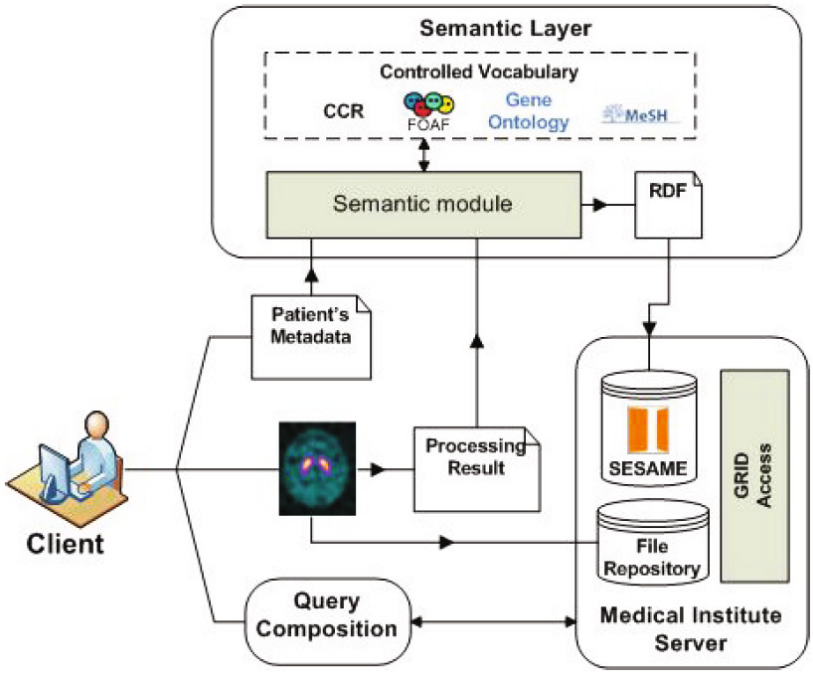


Fig. 5. Interaction between the high level features' systems

3.1 Semantic Layer

Usually medical images (MRI, X-Ray, PET, SPECT) are stored in DICOM format, containing the metadata provided with the standard. These metadata are not sufficient for describing the clinical history of patients. For this reason we enrich the information available in order to give the radiologists the possibility to better figure out a specific disease by developing a model that represents the medical data so that it can be analyzed by semantic tools. In detail, the system stores concepts, specifies relationship types between these concepts using RDF[2] (Resource Description Framework). More in detail, we have developed an RDF schema that includes:

- Personal data by using FOAF ontology[3];
- Genetic information using GeneOntology [13];
- Neurological detailed information by using Mesh [14];
- Image processing information that represents the output of the implemented image processing algorithms and which introduces a new semantic level to the stored metadata.

[2] http://www.w3.org/RDF/
[3] http://www.foaf-project.org/

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xml:base="http://i3s-lab.unict.it/semweb/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:go="http://www.geneontology.org/dtd/go.dtd#"
  xmlns:med="http://i3s-lab.unict.it/semweb/med/ns"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="George_Smith">
    <med:Patient>
      <foaf:Person>
        <foaf:name>George Smith</foaf:name>
        <foaf:firstName>George</foaf:firstName>
        <foaf:surname>Smith</foaf:surname>
        <med:disease rdf:resource="http://i3s-lab.unict.it/semweb/med/IPR003999"/>
        <med:medication rdf:resource="http://i3s-lab.unict.it/semweb/med/Ibuprofen"/>
        <med:imagingData rdf:resource="http://i3s-lab.unict.it/semweb/George_Smith/im"/>
        <med:treatedBy>
          <med:Doctor>
            <foaf:name>John Smith</foaf:name>
            <foaf:firstName>John</foaf:firstName>
            <foaf:surname>Smith</foaf:surname>
          </med:Doctor>
        </med:treatedBy>
      </foaf:Person>
    </med:Patient>
  </rdf:Description>
  <rdf:Description rdf:about="IPR003999">
    <med:goterm>
      <go:dbxref>
        <med:relatedmeshterm>68012772</med:relatedmeshterm>
        <go:database_symbol>interpro</go:database_symbol>
        <go:reference>IPR003999 Staphylococcal toxic shock syndrome toxin</go:reference>
      </go:dbxref>
    </med:goterm>
  </rdf:Description>
  <rdf:Description rdf:about="Ibuprofen">
    <med:medications>
      <med:medication>
        <med:name>Ibuprofen</med:name>
        <med:medicationcategory>NSAID</med:medicationcategory>
        <med:substance>(RS)-2-(4-(2-methylpropyl)phenyl)propanoic acid</med:substance>
      </med:medication>
    </med:medications>
  </rdf:Description>
  <rdf:Description rdf:about="George_Smith/im">
    <med:imagingData>
      <med:MRI>
        <med:bodyPart>Brain</med:bodyPart>
        <med:MRIWeight>T1</med:MRIWeight>
        <med:processingTechnique>Contrast Stretching</med:processingTechnique>
        <med:processingTechnique>Histogramm Equalization</med:processingTechnique>
        <med:report>Negative</med:report>
        <med:processedImage rdf:resource="George_Smith/im/1proc.tiff"/>
        <med:DICOMContainer rdf:resource="George_Smith/DICOM/1.dcm"/>
        <dc:source rdf:resource="George_Smith/im/1.tiff"/>
      </med:MRI>
    </med:imagingData>
  </rdf:Description>
</rdf:RDF>
```

Fig. 6. RDF instance describing a patient data

An instance of the RDF file for a patient is shown in fig. 6. It is notable that this information is inserted by the users, but it can be easily obtained by querying systems that share data using RDF. For instance, personal data in FOAF can be derived from a generic social network or by using a v-card, whereas generic health information can be retrieved (and converted in RDF), for example, from the user's Google Health Account[4] or from other systems that store online health care data such as the one proposed in [15]. Metadata storage has been carried out by using SESAME server [10] so that these information may be available also for other purposes. The sensitive data, such as Name,

[4] `https://www.google.com/health/`

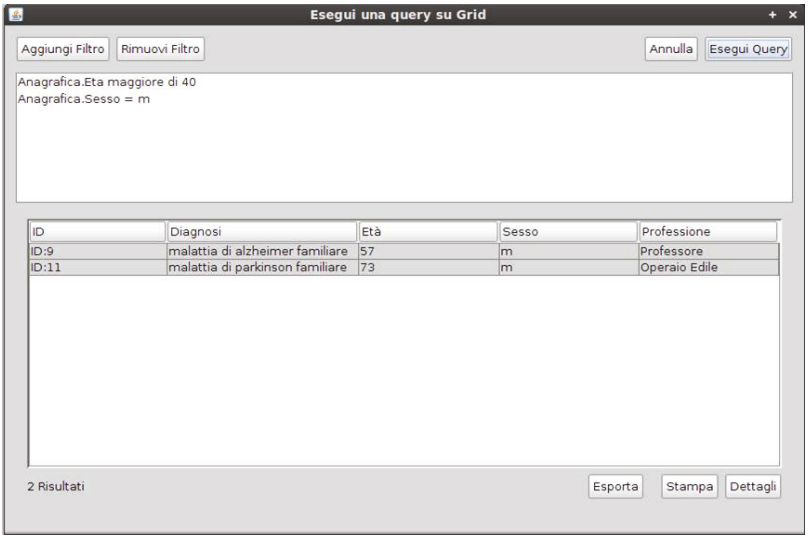


Fig. 7. Graphical User Interface for Automatic Query Composition

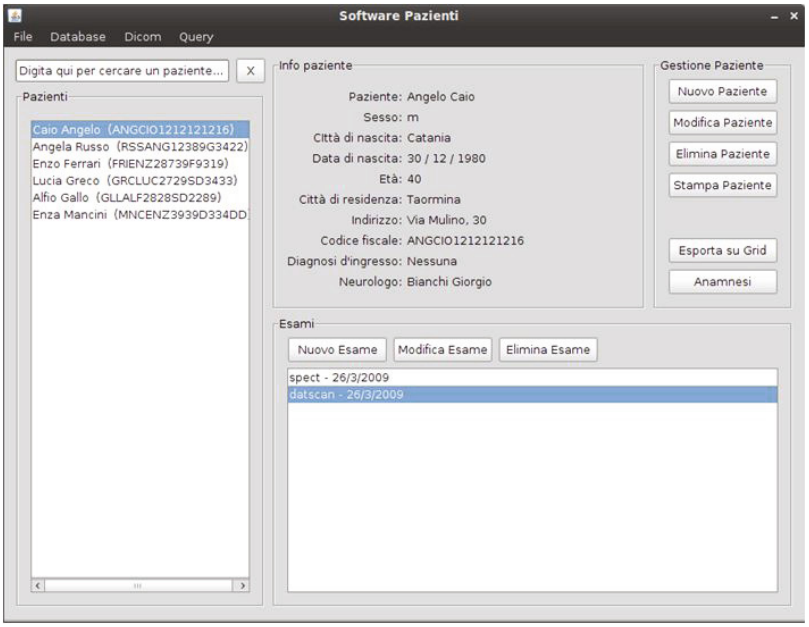


Fig. 8. Main Graphical User Interface

Surname, SSN must be available only for the radiologists who carries out the examination, and are not exported in RDF in order to ensure data privacy.

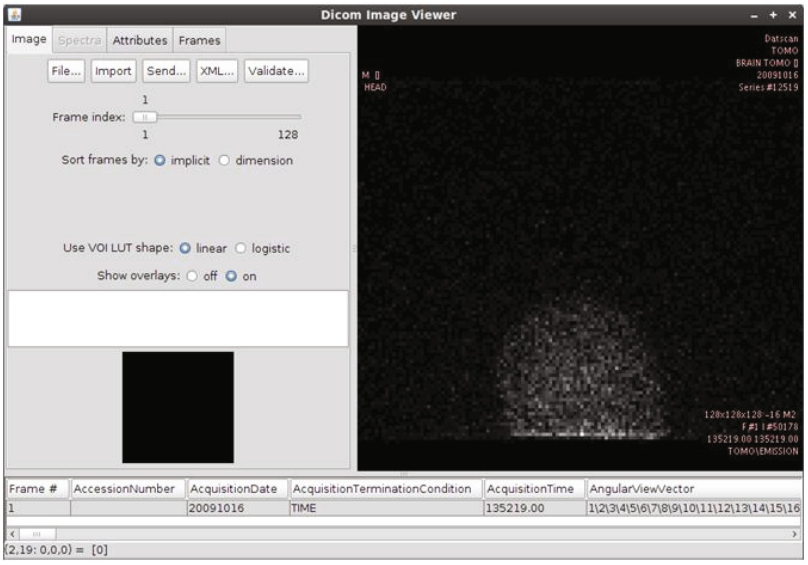


Fig. 9. Graphical User Interface for DICOM File Processing

3.2 Image Processing Layer

This level is provided with a set of processing methods for medical image analysis. The output of this processing is stored according to the semantic layer and is related to the specific processed image. This allows users to also share the results of the processing avoiding to send the original images when it is not required, resulting in less bandwidth usage.

The implemented methods for image analysis, available for SPECT, PET, MRI, X-Rays images, are:

- Measurement of distances, angles and some parameters within the images;
- The contrast absorption curve over time;
- Image Texture and Image Contour Analysis for specific organs;
- Pattern recognition for identifying brain structures.

Example of such algorithms have been proposed by the authors in [16], [17] and [18]. Therefore, when a user performs one of the above methods, the output is treated as metadata and stored in the SESAME server.

3.3 Query Composition

The query composition level aims at building complex queries both locally and on GRID. The queries are performed only on the metadata (stored in the SESAME server and in the AMGA server) since a content based image retrieval module is not present.

This level receives the user's queries (by using a controlled GUI, see fig. 7) and interacts both with the local storage, performing SPARQL queries on the SESAME server, and with the GRID, where queries are performed following the approach proposed in [19].

4 Client Interface

The client interface consists of a multiform Java application, allowing users to manage any type of information concerning the patients. The functionalities available for radiologist are: view information about their patients; search a patient by typing the name; insert, modify or delete a patient; print a patient's report; show the patient exams; insert, modify or delete an exam; management of medical history; view and process DICOM files; export the patient's data on GRID; query the system both locally and globally on GRID. For each functionality a specific GUI has been designed. For example, fig. 8 shows the main interface where a list of patients and a tool to manage patients' data is available. In the bottom area, there is also a section to handle the exams for each patient. Fig. 9, instead, shows the interface for DICOM files and image processing.

5 Concluding Remarks

In this paper we have proposed a Grid-based system for local and global data and metadata storage, sharing and processing. It is based on the RDF standard to enhance the interoperability with existing semantic web compliant applications and the reusability of the stored information. In detail, we have proposed an RDF schema that allows us to integrate heterogeneous data in order to provide radiologists with a holistic view of the patients' data. A future development is to provide the system with a semantic image retrieval layer based on ontologies that will map high level queries to low level image features. The system will be also integrated with available biosignal management systems (such as the proposed in [20]) to allow medical doctors to store and share also biosignals such as EEG, ECG, MEP, etc. Finally, we are currently working on automatically changing the language of the graphical user interfaces and also on integrating a lexical parser for solving automatically the ambiguities and the errors that may occur when data is inserted. A semantic text parsing system is under development for assigning automatically clinical ICD[5] (International Classification of Diseases) codes to patient records written in any language.

References

1. Gabber, E., Fellin, J., Flaster, M., Gu, F., Hillyer, B., Ng, W.T., Özden, B., Shriver, E.A.M.: Starfish: highly-available block storage. In: USENIX Annual Technical Conference, FREENIX Track, pp. 151–163 (2003)
2. Xu, Z., Jiang, H.: Hass: Highly available, scalable and secure distributed data storage systems. In: IEEE International Conference on Computational Science and Engineering, vol. 2, pp. 772–780 (2009)
3. Cheung, K.H., Prud'hommeaux, E., Wang, Y., Stephens, S.: Semantic Web for Health Care and Life Sciences: a review of the state of the art. Brief. Bioinformatics 10, 111–113 (2009)
4. Freund, J.: Health-e-child: an integrated biomedical platform for grid-based paediatric applications. Stud. Health Technol. Inform. 120, 259–270 (2006)
5. Holford, M.E., Rajeevan, H., Zhao, H., Kidd, K.K., Cheung, K.H.: Semantic Web-based integration of cancer pathways and allele frequency data. Cancer Inform. 8, 19–30 (2009)

[5] `http://www.who.int/classifications/icd/en/`

6. Diarena, M., Nowak, S., Boire, J.Y., Bloch, V., Donnarieix, D., Fessy, A., Grenier, B., Irrthum, B., Legre, Y., Maigne, L., Salzemann, J., Thiam, C., Spalinger, N., Verhaeghe, N., de Vlieger, P., Breton, V.: HOPE, an open platform for medical data management on the grid. Stud. Health Technol. Inform. 138, 34–48 (2008)
7. Montagnat, J., Breton, V., Magnin, I.: Partitioning medical image databases for content-based queries on a Grid. Methods Inf. Med. 44, 154–160 (2005)
8. Amendolia, S.R., Estrella, F., Hassan, W., Hauer, T., Manset, D., McClatchey, R., Rogulin, D., Solomonides, T.: MammoGrid: A Service Oriented Architecture Based Medical Grid Application. In: Jin, H., Pan, Y., Xiao, N., Sun, J. (eds.) GCC 2004. LNCS, vol. 3251, pp. 939–942. Springer, Heidelberg (2004)
9. Giordano, D., Pino, C., Spampinato, C., Fargetta, M., Di Stefano, A.: Nuclear medicine image management system for storage and sharing by using grid services and semantic web. In: Proceedings of the Healthinf 2011: International Conference on Health Informatics, Rome, Italy, January 26-29, 2011 (2010)
10. Broekstra, J., Kampman, A., Harmelen, F.V.: Sesame: An architecture for storing and querying rdf data and schema information. In: Semantics for the WWW. MIT Press (2001)
11. Venugopal, S., Buyya, R., Winton, L.: A grid service broker for scheduling distributed data-oriented applications on global grids. In: MGC 2004: Proceedings of the 2nd Workshop on Middleware for Grid Computing, pp. 75–80. ACM, New York (2004)
12. Nuno, N.S.: Distributed metadata with the amga metadata catalog. In: Workshop on Next Generation Distributed Data Management - HPDC-15 (2006)
13. Ashburner, M.: Gene ontology: Tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
14. Soualmia, L., Golbreich, C., Darmoni, S.: Representing the mesh in owl: Towards a semi-automatic migration. In: Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation, pp. 81–87 (2004)
15. Bieliková, M., Moravcík, M.: Modeling the reusable content of adaptive web-based applications using an ontology. Advances in Semantic Media Adaptation and Personalization, 307–327 (2008)
16. Faro, A., Giordano, D., Spampinato, C., Ullo, S., Di Stefano, A.: Basal Ganglia Activity Measurement by Automatic 3-D Striatum Segmentation in SPECT Images. IEEE Transactions on Instrumentation and Measurement 60(10), 3269–3280 (2011)
17. Faro, A., Giordano, D., Spampinato, C., Pennisi, M.: Statistical texture analysis of MRI images to classify patients affected by multiple sclerosis. In: 12th Mediterranean Conference on Medical and Biological Engineering and Computing, MEDICON 2010, Porto Carras, Chalkidiki, Greece. International Proceedings of the IFBME, pp. 236–239. Springer, Heidelberg (2010)
18. Giordano, D., Spampinato, C., Scarciofalo, G., Leonardi, R.: An Automatic System for Skeletal Bone Age Measurement by Robust Processing of Carpal and Epiphysial/Metaphysial Bones. IEEE Transactions on Instrumentation and Measurement 59(10), 2539–2553 (2010)
19. Montagnat, J., Frohner, Á., Jouvenot, D., Pera, C., Kunszt, P., Koblitz, B., Santos, N., Loomis, C., Texier, R., Lingrand, D., Guio, P., Rocha, R.B.D., de Almeida, A.S., Farkas, Z.: A secure grid medical data manager interfaced to the glite middleware. J. Grid Comput. 6, 45–59 (2008)
20. Giordano, D., Kavasidis, I., Spampinato, C., Bella, R., Pennisi, G., Pennisi, M.: An integrated computer-controlled system for assisting researchers in cortical excitability studies by using transcranial magnetic stimulation. Computer Methods and Programs in Biomedicine, December 14 (2011), doi:10.1016/j.cmpb.2011.10.008, ISSN 0169-2607

Securing the Access to Electronic Health Records on Mobile Phones*

Alexandra Dmitrienko[1], Zecir Hadzic[1], Hans Löhr[1],
Ahmad-Reza Sadeghi[2], and Marcel Winandy[1]

[1] Horst Görtz Institute for IT-Security (HGI), Ruhr-University Bochum
Bochum, Germany
{alexandra.dmitrienko,zecir.hadzic,
hans.loehr,marcel.winandy}@trust.rub.de
[2] Center for Advanced Security Research Darmstadt (CASED)
Technische Universität Darmstadt, Darmstadt, Germany
ahmad.sadeghi@trust.cased.de

Abstract. Mobile phones are increasingly used in the e-health domain. In this context, enabling secure access to health records from mobile devices is of particular importance because of the high security and privacy requirements for sensitive medical data. Standard operating systems and software, as they are deployed on current smartphones, cannot protect sensitive data appropriately, even though modern mobile hardware platforms often provide dedicated security features. Current mobile phones are prone to attacks by malicious software, which might gain unauthorized access to sensitive medical data.

In this paper, we present a security architecture for the protection of electronic health records and authentication credentials that are used to access e-health services. Our architecture is derived from a generic solution and tailored specifically to current mobile platforms with hardware security extensions. Authentication data are protected by a trusted wallet (TruWallet), which leverages trusted hardware features of the phone and isolated application environments provided by a secure operating system. A separate application environment is used to provide runtime protection of medical data. Furthermore, we present a prototype implementation of TruWallet on the Nokia N900 mobile phone. In contrast to commodity systems, our architecture enables healthcare professionals to securely access medical data on their mobile devices without the risk of disclosing sensitive information.

1 Introduction

The usage of mobile phones as multi-purpose assistant device in healthcare has been proposed in several application scenarios. Its usefulness is derived from its mobility and flexibility, i.e., today's smartphones offer appropriate computing and storage capacity allowing the realization of various applications that can be used basically from everywhere. For instance, healthcare professionals can use a mobile phone to download and share electronic health records of their patients [9]. In other scenarios, patients use their

* An earlier version of this paper has been published in [11].

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 365–379, 2013.
© Springer-Verlag Berlin Heidelberg 2013

mobile phones to provide personal health data, e.g., taken from additional bio-sensors, to a medical information and diagnosis system [17].

While smartphones are very flexible and cost-efficient computing devices, they generally do not offer sufficient security mechanisms to protect the data they operate on. This is mainly due to the architectural shortcomings of their operating systems, which are derived from the same (security) architecture as desktop operating systems. Typical examples are Google Android [6], Apple iOS [7], Symbian [35], and Windows Mobile [28]. Although, some of them provide more sophisticated security mechanisms than their desktop counterparts, e.g., application-oriented access control in Android [16], they still suffer from fundamental security problems due to their large code base and complexity, lacking of strong isolation of applications (*secure execution*) and insufficient protection of stored data (*secure storage*). Recent attacks on smartphones demonstrate their vulnerability [1,19,39]. But the secure operation of a mobile phone is an important aspect when a user is working with security and privacy-sensitive data such as personal health records on the device.

Especially in healthcare telematics infrastructures, the end-user systems of health professionals have been identified as an insecure and less specified component [34]. Malware on the user's computing platform could steal passwords that are used to access healthcare information systems, manipulate data such as medical prescriptions, or eavesdrop on and copy private data such as personal health records. While the connection of stationary desktop systems to the healthcare telematics may be protected by additional secure hardware network components like, e.g., special firewalls and gateway routers, the situation gets worse when mobile phones are used. Due to their mobility and changing connectivity (wireless LAN or GSM network), mobile phones may usually only use Virtual Private Network (VPN) technology to secure the connection. But the necessary credentials, like user passwords and VPN keys, are not sufficiently protected against malware on the device, and, hence, could be accessed by unauthorized parties.

However, modern smartphone hardware offers advanced security functionality, which are embedded in their processors, but generally not used by the mainstream mobile operating systems. For instance, ARM TrustZone [37] and Texas Instruments M-Shield [8] offer secure boot[1] functionality, secure storage and secure execution environments for security-critical functions, which are isolated based on hardware mechanism from other processes running on the phone.

Contribution. In this paper, we propose a security architecture for accessing e-health services on mobile phones. We present the combination of efficient solutions that current technology can offer on mobile phones for the secure handling of accessing and processing of security-sensitive data such as electronic health records. In particular, we propose (i) a security framework to create a secure runtime environment for medical applications, and (ii) specific tools that protect the authentication of users and their mobile devices to e-health servers.

[1] Secure boot means that a system terminates the boot process in case the integrity check of a component to be loaded fails.

In our security framework, we combine the concept of a security kernel with hardware security features of modern mobile phone processors. On top of this layer, we use isolated execution compartments to separate applications that process medical data (e.g., an EHR viewer) and applications that process non-medical data (e.g., the telephony application or an ordinary web browser).

As a secure authentication tool, we propose a *trusted wallet* that protects the user's login credentials and performs the authentication to e-health (or other) servers on behalf of the user. This tool protects the users from being tricked into entering their credentials in malicious applications or faked web sites, and takes advantage of the underlying security framework to protect the credentials from malicious software potentially running on the phone. We present a new implementation of this wallet for mobile phones based on the Nokia N900 platform.

Compared to commodity mobile phone operating systems, our approach provides a secure environment against software attacks like malware. The usage of security-critical data like patients health records is effectively isolated from other software running on the phone, and secret data like login credentials to healthcare information systems is protected by advanced hardware security features.

In the following, we describe the usage and adversary scenario we consider (Section 2). Then, we present our security architecture (Section 3): first from a generic perspective, which can be used on all platforms, followed by its instantiation on mobile phone platforms. In Section 4, we describe how our architecture can be implemented and we present our Mobile TruWallet prototype. Finally, we conclude in Section 6.

2 Problem Scenario

We consider a scenario in which electronic health records (EHRs) of patients are stored on a local server of a healthcare provider, e.g., in a hospital. Health care professionals, like physicians and nurses, are equipped with mobile computing devices (smartphones) on which they can create, edit, and view EHRs. The EHRs are stored on the e-health server, and the smartphones communicate with the server via wireless network connections. For instance, the access of medical data can be realized with web-based applications, using standard web browser software on mobile devices. Figure 1 depicts the scenario we consider.

Since EHRs are very security-sensitive private data, and in most countries protected under strong privacy laws, unauthorized access to these data must be prevented. An adversary may try to eavesdrop or manipulate the sensitive data. As mentioned before, end-user devices are typically the least specified and least secured devices in healthcare infrastructures. Hence, an adversary would most likely try to attack the mobile phone and its communication connection to the server in order to illegitimately access medical data.

Studies like [40] have analyzed how to secure the data transfer, i.e., via encryption (for confidentiality), digital signatures (for integrity and authenticity), and user authentication (for legitimacy of access). However, the protection of the critical cryptographic keys that are needed for those mechanisms is not addressed appropriately. Hence, an attacker who gains access to these keys can circumvent any other protection mechanism.

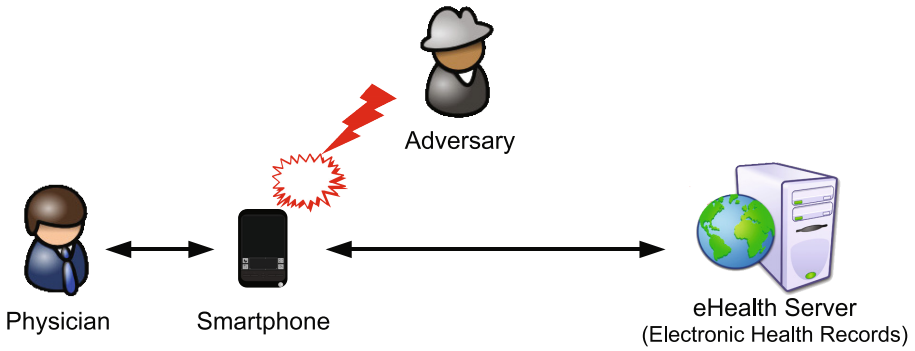


Fig. 1. Use Case and Adversary Model

Therefore, in this paper we concentrate on an adversary model in which the attacker targets the mobile computing device of health care professionals in order to obtain the secret login credentials or keys that are needed to access the EHR server. Once the adversary has access to these credentials, he can download or modify all medical data from the server to which the credentials allow access to. To achieve this goal, the adversary can follow two strategies:

- ***Direct Access.*** The adversary tries to directly access the sensitive data or keys. He could try to manipulate software running on the phone to access the data, or he could steal the device and try to access the data. The former could be achieved by letting the users install malicious software (malware, such as trojan horses) without their notice, e.g., when they browse to a website containing malicious code that exploits a vulnerability of the phone's software to install the malware. Physicians may use their phones also for other tasks and they may want to download additional applications to run them on the phone, which could create the vulnerability for such an attack.
- ***Indirect Access.*** The adversary tries to trick the users to enter their passwords into a faked EHR viewer application. The faked application looks like the real one, but instead of logging into the server, it sends the passwords to the adversary. The faked application could be installed on the phone in the same way as malware described above.

The problem with a commodity mobile phone operating system (OS) is that it cannot provide a sufficient level of protection for the applications or stored credentials. A mobile phone OS that is directly derived from a desktop OS (e.g., Linux or Windows) has limited protection capabilities, i.e., simple process isolation and access control. However, malicious applications can modify or eavesdrop data of other applications since they are running with the same user privileges as other applications.

A more advanced OS, e.g., like SELinux [25], can enforce mandatory access rules, which provide a stronger isolation of different applications from each other. For instance, a text editor could only edit text files, whereas an audio application could not modify text files. The application of such a system in a mobile e-health scenario has been shown earlier [2]. However, SELinux is a very complex system with security policies that are hard to configure correctly even for moderately complicated scenarios.

Moreover, due to a relatively large code base, it is infeasible to perform a comprehensive formal (or even semi-formal) verification of the correctness and security of SELinux. Another example is Android [16], which provides a similar application-oriented access control, i.e., it defines for each application different access rules and privileges — in contrast to user-oriented access control as in normal Linux and Windows, where all programs of one user share the same access rights.

Nevertheless, even advanced mobile phone OS's still suffer from ineffective protection against unauthorized modifications of programs or even modifications of the OS itself. An adversary could install on the user's phone additional (faked) programs or replace existing programs. The user has seldom a chance to notice the modification, and critical data like credentials could be transferred to the adversary.

3 Wallet Architecture

3.1 General Idea

Our security architecture aims to protect against the attacks described above. To counter direct access attacks, our architecture is based on a security kernel that isolates different applications, supports secure boot, and provides secure storage. Hence, authentication data is stored encrypted, and can only be accessed by the legitimate application (TruWallet) when the correct (unmodified) system has been booted.

Our wallet architecture aims to prevent indirect attacks by letting the wallet handle all authentication procedures. During a normal authentication, users do not enter passwords (this is automatically done by the wallet), hence they cannot accidentally disclose them towards a fake application that tries to spoof the look and feel of the legitimate EHR viewer or another application trusted by the user.

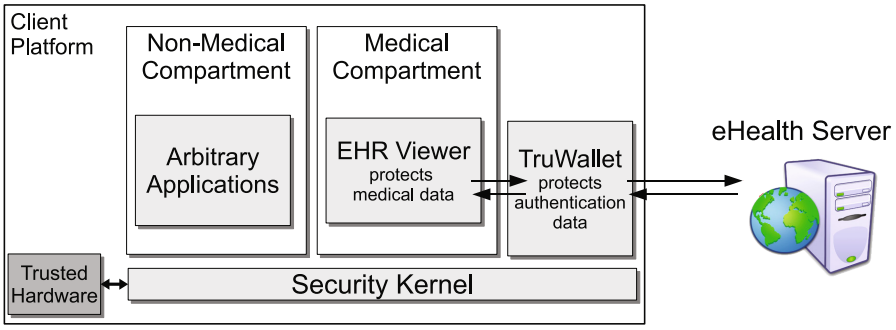


Fig. 2. General idea of TruWallet

Our wallet-based security architecture provides two levels of protection (cf. Figure 2):

1. **Protection of *Authentication Data*.** TruWallet protects the user's credentials (username and password) against unauthorized access. This approach is generic, and can be used also for other scenarios (e.g., web applications like eBay or Amazon).

Indeed, TruWallet can be used simultaneously by different applications, yet it only authenticates each application to the server where it has been registered as legitimate application before.

2. **Protection of *Medical Data*.** An isolated EHR viewer (which can be a special-purpose browser) is used to view EHRs. This viewer cannot be modified because a fixed program image is executed, which is measured by the security kernel by computing a cryptographic hash and compared to a known-good reference value. This ensures that all modifications of the EHR viewer can be detected. In case a browser is used as EHR viewer, this browser is only allowed to contact the EHR server and cannot connect to other sites. For all other websites or services, a separate browser process would be used which is isolated from the EHR viewer.

3.2 System Model

Our system model for TruWallet consists of several parties (see Figure 3): A *user* interacts with a computing platform through a secure graphical user interface *secure GUI*. An *EHR viewer* is used to render content that it gets from the *wallet*, which is acting as a proxy. The wallet obtains the requested content from the server, blinds security-sensitive fields (e.g., password) on the pages presented to the browser, and fills in login credentials when logging into the system. For this, TruWallet has to handle two different SSL sessions: one between wallet and EHR viewer, and one between wallet and server. The secure GUI controls the input/output devices and multiplexes the screen output of the EHR viewer and of the wallet. Moreover, it always indicates the name of the application the user is currently interacting with via a reserved area on the screen, hence providing a *trusted path* between user and application. Moreover, our architecture includes a compartment for non-medical data and applications. This compartment is strictly separated from the EHR viewer and can be used for arbitrary applications.

3.3 Generic Wallet-Based e-Health Architecture

The generic TruWallet architecture is based on a *security kernel*, which is a small trusted software layer, providing *trusted services* and *isolated compartments*. Thus, the security kernel ensures runtime security of the system. Compartments contain arbitrary software, e.g., a complete legacy operating system (Linux in our case), and may communicate only via well-defined interfaces. In particular, a malicious compartment cannot read arbitrary memory of other compartments. In our solution, EHR viewer, non-medical applications and wallet run in different compartments, and we assume that arbitrary software (including malware like Trojan horses and viruses) may be running in the non-medical compartment. Therefore, the security of our solution is based on trusted components (wallet and EHR viewer) that are executed in separated compartments, isolated from untrusted software that might be running simultaneously on the same platform.

In an earlier work [14], we have demonstrated the feasibility of the wallet architecture on a PC platform. In the PC-based implementation, the compartmentalization was realized by using the isolation property of virtual machines combined with the resource sharing control of an underlying microkernel. The wallet compartment is trusted, which

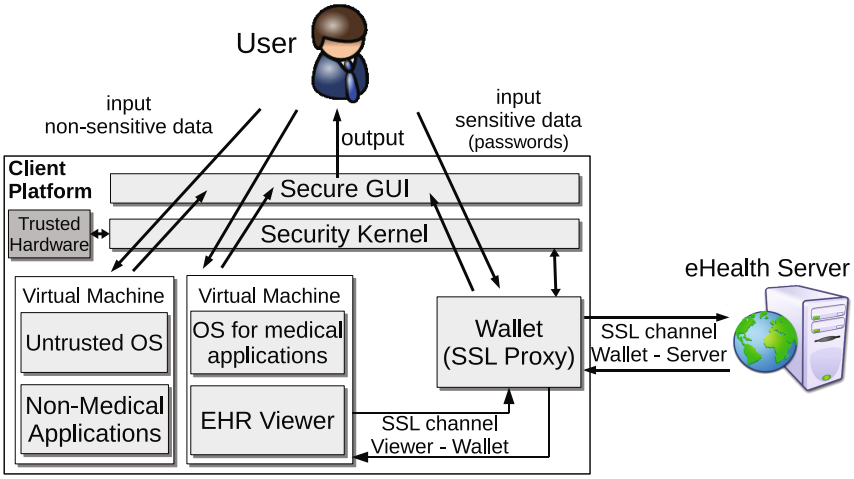


Fig. 3. Generic TruWallet Architecture

is motivated by the fact that the complexity of the wallet is much lower than that of an EHR viewer or a compartment containing several different applications. Moreover, users cannot install arbitrary software (which may be malicious or flawed) in the wallet compartment, but they may install arbitrary viewers or other tools into other compartments. To prevent unauthorized access by other users to the platform and, hence, the sensitive data, the security kernel requires an overall user authentication (e.g., a user password) to login into the whole system. In this way, the credentials stored by the wallet are bound to the corresponding user[2].

Trusted Computing Support. Trusted Computing (TC) hardware and TC-enabled software is used to provide *authenticated boot*, i.e., based on a "chain of trust", the integrity of the software stack including the Trusted Computing Base (TCB) can be verified later, e.g., before access to cryptographic keys is allowed. An alternative to authenticated boot which is usually used on for mobile platforms is *secure boot*: In this case, the system's integrity state is compared to reference values and can be started only if it has not been modified[3]. Moreover, TC hardware can be used for secure storage, i.e., encryption keys protected by the hardware can only be used if load-time integrity of the system is maintained. Thus, the credentials stored by the wallet are bound to the TCB to prevent an adversary from gaining access to the data by replacing software (e.g., booting a different OS). On the PC platform [14] we used a Trusted Platform Module (TPM) [38] as TC hardware for TruWallet. The TPM is a dedicated security chip that provides – among other features – cryptographic operations (e.g., key generation, encryption,

[2] In fact the security kernel has to provide comprehensive user access control as in typical operating systems, including system login and screen lock functionality, in order to prevent unauthorized access to the wallet. However, the details of those mechanisms are out of scope of this paper.

[3] Of course, it is important that these reference values are stored in a secure location, e.g., protected by security hardware, to avoid manipulations.

digital signatures), support for authenticated boot, and the possibility to bind cryptographic keys to the load-time integrity state of the system.

3.4 Mobile TruWallet

To implement our security architecture for mobile e-health scenarios, several building blocks for mobile environments are required:

- Trusted hardware for mobile platforms which supports features to protect cryptographic keys and verify the system integrity;
- a secure hypervisor layer for mobile platforms to provide isolated execution environments for applications;
- a security kernel with a secure GUI for mobile platforms to provide a trusted path between the user and applications, and with secure storage for applications;
- a trusted wallet (TruWallet) to handle authentication and protect the user's credentials.

In the following, we briefly introduce the first three building blocks, before we focus in more detail on the implementation of a trusted wallet on a mobile phone in the next section.

Trusted Hardware for Mobile Platforms. The architecture of TruWallet relies of trusted hardware for performing security critical operations. To instantiate TruWallet architecture on a mobile phone, we have to use mobile hardware security extensions instead of a TPM (which is not available on current phones). On mobile platforms, general-purpose secure hardware such as M-Shield [8] and TrustZone [4] is available. In this paper, we focus on M-Shield, because this hardware extension is available in some current mobile phones, including Nokia N900 (which we used for our prototype).

M-Shield provides a small amount of dedicated on-chip ROM and RAM as well as one-time programmable memory for device keys which are accessible only in a special execution mode of the main CPU – the Trusted Execution Environment (TrEE). A secure state machine (SSM) guarantees secure switching between both processor modes, thus the TrEE and normal execution environment are isolated from each other. M-Shield enables the TrEE on a device with the following features: (i) isolated secure code execution; (ii) secure boot; (iii) hardware-based secure storage.

Secure Hypervisor for Mobile Devices. Several microkernels for mobile and embedded devices have been implemented, for instance the commercially available L4 microkernels OKL4 [29] and PikeOS P4 [10]. These microkernels provide isolation between user space applications, just like their counterparts on other platforms (e.g., on PCs). Therefore, they can be used for a secure hypervisor layer for a security kernel on mobile phones. In particular, the seL4 microkernel has been formally verified for correctness [21], hence taking an important step towards building a formally verifiable security kernel on top of a microkernel.

Security Kernel with Secure GUI for Mobile Devices. Besides the isolation support, a security kernel usually has to include other security services as well. One prominent – and in our case needed – service is that of a secure graphical user interface system [31]. The main properties of a secure GUI system are (i) to protect user input/output from eavesdropping or manipulation by unauthorized applications, and (ii) to provide the user a trusted path that indicates with which application the user is currently interacting.

4 Wallet Prototype on Nokia N900

In order to demonstrate the feasibility of running a trusted wallet on a mobile phone, we have implemented Mobile TruWallet, a mobile version of trusted wallet, on a Nokia N900 device.

4.1 Mobile TruWallet Architecture

Instead of implementing a security kernel on a mobile device – for which we refer the reader to [10,21,32] – we used Maemo [27] as the basis for our mobile wallet. Maemo is a Linux-based operating system that provides standard process isolation and discretionary access control. Though Maemo does not provide the same security properties as a security kernel, we think it serves sufficiently to demonstrate the concepts of our approach. In a real product deployment, a security kernel implementation would be used instead.

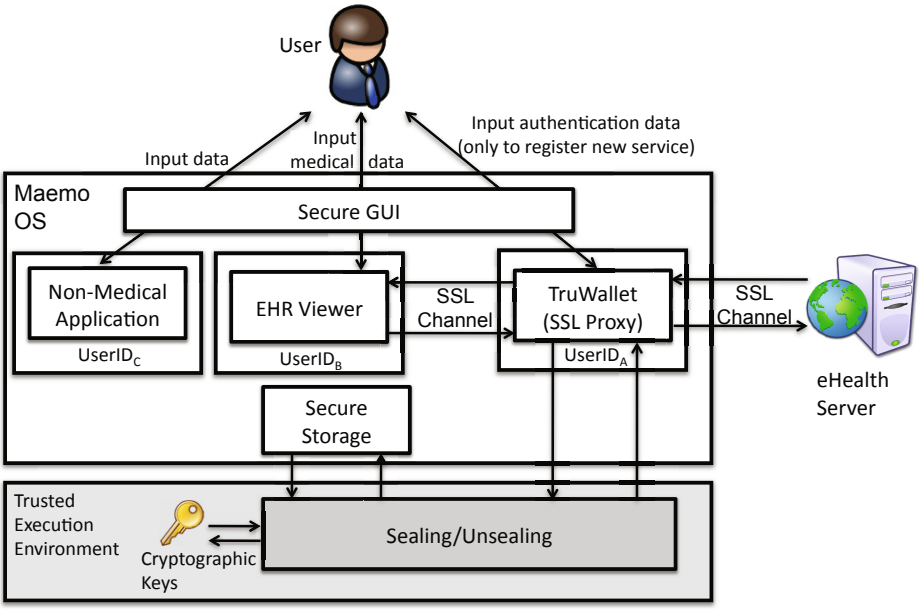


Fig. 4. Mobile TruWallet Architecture

The architecture of Mobile TruWallet we have implemented is depicted in Figure 4. As the figure shows, TruWallet resides on the operating system side, but also operates on secrets at the same time, e.g., maintains a TLS channel to the web-server and also performs authentication with the user passwords. However, our generic architecture assumes that TruWallet is isolated from the rest of the system. This assumption is reasonable to some extend in the context of the Maemo operating system because Maemo's security model is based on Discretionary Access Control (DAC) which enforces security by process ownership.

We achieved process isolation on Maemo by creating a Mobile TruWallet process under a unique UserID and defining restrictive access rights to that UserID. Note that for this prototype, we rely on the standard Unix/Linux discretionary access control security framework, and there is always the threat that an administrator (root account) with the super-user access rights is compromised. However, as mentioned before, we implemented the wallet as if it was running on a security kernel. This approach allows us to concentrate on the wallet-specific aspects for the prototype (i.e., performance, user interface, compatibility to the mobile web browser and web sites, etc.). In a later stage, the wallet can be easily adapted to a system like the L4-based security kernel on the N900 [32].

The Nokia N900 device is based on M-Shield secure hardware. We utilize M-Shield functionality for the secure boot, and we also implement a secure storage functionality on top of M-Shield. Note that even if an attacker could compromise the operating system, all cryptographic keys are protected by the security hardware and cannot be disclosed or copied by the attacker.

Only authenticated programs, so-called protected applications (PAs), can be executed within the TrEE of M-Shield. However, protected applications have to be authorized, i.e., certified, by the device's M-Shield stakeholder, most likely the device manufacturer. For our prototype, where it is unrealistic to obtain a certificate of the PA from the device manufacturer, we use a different approach: We reuse the general purpose APIs available for the M-Shield TrEE. This approach allows third parties to leverage the TrEE. For instance, the On-board Credentials platform (ObC) [23] provides the means to develop programs for the TrEE without the involvement of the manufacturer. In our implementation, we build the secure storage functionality of Mobile TruWallet on top of ObC.

4.2 ObC Architecture

Figure 5 illustrates the ObC architecture. The core component of the ObC platform – which resides in the dedicated RAM and can be executed in secure environment – is an *interpreter*. The interpreter provides a virtualized environment where "credential programs", i.e., scripts developed by third parties, can be executed. The credential programs are written using (a subset of) Lua scripting language [26] or in assembler. When a credential program is executed, the interpreter isolates it from secrets that are stored within the TrEE and from the execution of other credential programs.

The interpreter makes use of a Crypto Library which provides an interface for commonly used cryptographic primitives. It provides a sealing/unsealing function for ObC programs, which can be used to protect secret data stored persistently outside the secure

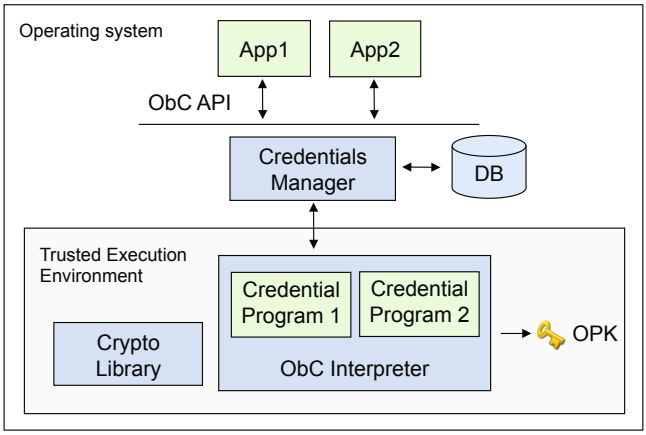


Fig. 5. On-board Credential architecture

environment. Sealed data is encrypted with a key which is protected by the TrEE, and the ObC platform controls the usage of this key. A device-specific symmetric key called ObC platform key (OPK) is used for the sealing/unsealing functionality.

Credential Manager is a mediator between OS side applications and components that reside within the TrEE. It provides an API for third-party developed applications. Using the API the applications can execute credential programs, and create and use new asymmetric keys. The credential manager maintains a database, in which credentials and credential programs are stored in a cryptographically sealed way.

A more detailed description of the ObC architecture can be found in [23,22].

4.3 Implementation

In our prototype, the wallet is implemented in the C programming language, contains about 2600 lines of code, and runs as separate process with a unique UserID. For the SSL/TLS proxy, we use Paros [30], which is an open-source implementation in Java, and it executes as a process with the same UserID as the wallet. We define restrictive access rights on this UserID so that other processes cannot access the data or code of the wallet.

Accessing Health Records. The wallet uses the `libxml` library to parse web sites and web forms in order to search for password fields. Whenever it finds such fields, these forms are put into a cache and are disabled before they are shown in the web browser. This prevents the user from accidentally typing passwords into a potentially malicious or faked web site. Instead users just provide their user name and simply click the submit or login button in the mobile web browser.

Hence, when physicians want to access a health record from the e-health server, they simply open the EHR viewer browser on the phone and click the login button. The wallet replaces then automatically the disabled password field with the actual password of the physician's account on the e-health server. This process runs transparently, so the

physician just sees the EHR viewer application, and when the login is completed he can access the health records on the server.

Registration. Before physicians can use the wallet to login to websites like the e-health server, they have to register their account in the wallet on the phone. Therefore, the wallet also looks for registration forms. When the user tries to access a website with the browser for the first time, the wallet asks the user for an existing password or it can create a new one.

Once the password has been provided (or newly created), the wallet stores the credentials in a specific file. During runtime, the access to this file is only granted to the UserID of the wallet. Hence, other programs cannot read or modify the stored credentials. When the device is going to be shut off, this file is sealed using the ObC platform as mentioned before.

Users can view a list of the stored accounts in wallet. We realized the user interface based on the Hildon GUI framework [18] on Maemo.

Interoperability. We have tested our wallet implementation with several public websites, like web e-mail services, eBay, Amazon, etc. Registration and login work transparently and without noticeable performance overhead for the user. Hence, it should be easy to integrate web-based e-health services on our platform. Special-purpose EHR viewers or other medical applications can be supported as well as long as they use SSL/TLS and web-based login procedures. Other authentication protocols could also be integrated, but may require some effort to adapt the wallet.

5 Related Work

Previous works on secure operating systems, e.g., [13,20,33], have shown how to achieve strong isolation for secure execution and to have less complexity for the trusted computing base, i.e., the code that all security relies upon. The concept of a *security kernel* [5] incorporates all relevant functionality needed to enforce the security into a kernel that is isolated and protected from tampering by other software and small enough to be verifiable for its correctness and security. While earlier systems suffered mostly from poor performance in those days, recent CPU hardware technology, especially their virtualization support, and the development of efficient microkernel software architectures [24] allow for the realization of security kernels with low performance overhead while maintaining compatibility to existing applications. For example, Turaya [12] and the OpenTC security architecture [36] are research efforts that take advantage of these technologies to develop a security kernel on modern CPU hardware.

The Turaya Trusted Mobile Desktop [32] implements a security kernel with a secure user interface for mobile devices. Its TCB consists of a hypervisor layer and a trusted software layer. The hypervisor layer is implemented on top of an L4 microkernel, which has been ported to the Nokia N900 mobile phone. The Trusted Software Layer contains a number of security services, such as a secure graphical user interface (called TrustedGUI), a virtual private network (VPN) client, and a file encryption service. An implementation such as the Turaya Trusted Mobile Desktop can be used as an underlying security kernel for our architecture.

The protection of health records on smartphones has been addressed in other works [3,15]. However, the focus of these works is the encryption of the health records and storing the encrypted records directly on the mobile device. This approach is orthogonal to ours, as we do not consider to store the health records on the phone, but rather to protect the viewing and the access to them. In both cases, health records have to be shown in plaintext on the device at some point of time. Our architecture ensures their runtime protection by executing the EHR viewer in a separate application environment. In addition, our approach protects the credentials by leveraging trusted hardware functionality, whereas the approaches of [3,15] employ a software-only solution.

6 Conclusions and Future Work

Mobile access to electronic medical data is an emerging application scenario with strong security and privacy requirements that is rapidly gaining practical importance. Existing systems suffer from a lack of appropriate protection for security- and privacy-critical data and applications. Moreover, standard operating systems do not use existing hardware security features of mobile platforms to their full extent.

To enable secure mobile access to electronic health records containing privacy-sensitive data, we propose an e-health security architecture which protects the user's authentication credentials as well as the sensitive medical data. Our architecture is based on commonly available trusted hardware components, a security kernel, and a trusted wallet. In this paper, we introduce our comprehensive security architecture, discuss security building blocks on mobile phones, and present our implementation of Mobile TruWallet on a commodity smartphone.

Since our Mobile TruWallet prototype demonstrates the feasibility of the architecture on mobile phones, future work includes the integration of all security building blocks (i.e., the use of hardware security features, a security kernel consisting of a secure hypervisor and a trusted software layer, and the Mobile TruWallet authentication solution) into one system.

Acknowledgements. This work was partially funded by the German federal state North Rhine-Westphalia and supported by the European Regional Development Fund under the project RUBTrust/MediTrust. Further, the author Alexandra Dmitrienko was supported by the Erasmus Mundus External Co-operation Window Programme of the European Union.

References

1. Aggarwal, M., Vennon, T.: Study of BlackBerry proof-of-concept malicious applications. Technical Report White paper, SMobile Global Threat Center (January 2010)
2. Agreiter, B., Alam, M., Hafner, M., Seifert, J.P., Zhang, X.: Model driven configuration of secure operating systems for mobile applications in healthcare. In: Proceedings of the 1st International Workshop on Mode-Based Trustworthy Health Information Systems (2007)
3. Akinyele, J.A., Lehmann, C.U., Green, M.D., Pagano, M.W., Peterson, Z.N.J., Rubin, A.D.: Self-protecting electronic medical records using attribute-based encryption. Cryptology ePrint Archive, Report 2010/565 (2010), `http://eprint.iacr.org/2010/565`

4. Alves, T., Felton, D.: TrustZone: Integrated hardware and software security. Technical report, ARM (July 2004)
5. Anderson, J.P.: Computer security technology planning study. Technical Report ESD-TR-73-51, AFSC, Hanscom AFB, Bedford, MA, AD-758 206, ESD/AFSC (October 1972)
6. Android Open Source Project. Project website (2010), http://www.android.com
7. Apple Inc. iOS website (2010), http://www.apple.com/iphone/ios4
8. Azema, J., Fayad, G.: M-Shield^{TM}mobile security technology: making wireless secure. Texas Instruments White Paper (February 2008), http://focus.ti.com/pdfs/wtbu/ti_mshield_whitepaper.pdf
9. Benelli, G., Pozzebon, A.: Near Field Communication and Health: Turning a Mobile Phone into an Interactive Multipurpose Assistant in Healthcare Scenarios. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2009. CCIS, vol. 52, pp. 356–368. Springer, Heidelberg (2010)
10. Brygier, J., Fuchsen, R., Blasum, H.: PikeOS: Safe and secure virtualization in a separation microkernel. Technical report, Sysgo (September 2009)
11. Dmitrienko, A., Hadzic, Z., Löhr, H., Sadeghi, A.-R., Winandy, M.: A security architecture for accessing health records on mobile phones. In: Proceedings of the 4th International Conference on Health Informatics (HEALTHINF 2011), pp. 87–96. SciTePress (2011)
12. EMSCB Project Consortium. The European Multilaterally Secure Computing Base (EMSCB) project (2005-2008), http://www.emscb.org
13. Fraim, L.: SCOMP: A solution to the multilevel security problem. IEEE Computer, 26–34 (July 1983)
14. Gajek, S., Löhr, H., Sadeghi, A.-R., Winandy, M.: TruWallet: Trustworthy and migratable wallet-based web authentication. In: The 2009 ACM Workshop on Scalable Trusted Computing (STC 2009), pp. 19–28. ACM (2009)
15. Gardner, R.W., Garera, S., Pagano, M.W., Green, M., Rubin, A.D.: Securing medical records on smart phones. In: Proceedings of the 1st ACM Workshop on Security and Privacy in Medical and Home-Care Systems, SPIMACS 2009, pp. 31–40. ACM (2009)
16. Google Android. Security and permissions (2010), http://developer.android.com/intl/de/guide/topics/security/security.html
17. Han, D., Park, S., Lee, M.: THE-MUSS: Mobile U-Health Service System. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2008. CCIS, vol. 25, pp. 377–389. Springer, Heidelberg (2008)
18. Hildon Application Framework. Project website (2010), http://live.gnome.org/Hildon
19. Iozzo, V., Weinmann, R.-P.: Ralf-Philipp Weinmann & Vincenzo Iozzo own the iPhone at PWN2OWN (March 2010), http://blog.zynamics.com/2010/03/24/ralf-philipp-weinmann-vincenzo-iozzo-own-the-iphone-at-pwn2own/
20. Karger, P.A., Zurko, M.E., Bonin, D.W., Mason, A.H., Kahn, C.E.: A VMM security kernel for the VAX architecture. In: Proceedings of the IEEE Symposium on Research in Security and Privacy, Oakland, CA, pp. 2–19. IEEE Computer Society, Technical Committee on Security and Privacy (May 1990)
21. Klein, G., Elphinstone, K., Heiser, G., Andronick, J., Cock, D., Derrin, P., Elkaduwe, D., Engelhardt, K., Kolanski, R., Norrish, M., Sewell, T., Tuch, H., Winwood, S.: seL4: Formal verification of an OS kernel. In: Proceedings of the 22nd ACM Symposium on Operating Systems Principles, Big Sky, MT, USA. ACM Press (October 2009)
22. Kostiainen, K., Dmitrienko, A., Ekberg, J.-E., Sadeghi, A.-R., Asokan, N.: Key Attestation from Trusted Execution Environments. In: Acquisti, A., Smith, S.W., Sadeghi, A.-R. (eds.) TRUST 2010. LNCS, vol. 6101, pp. 30–46. Springer, Heidelberg (2010)

23. Kostiainen, K., Ekberg, J.-E., Asokan, N., Rantala, A.: On-board credentials with open provisioning. In: ASIACCS 2009: Proceedings of the 4th International Symposium on Information, Computer, and Communications Security, pp. 104–115. ACM (2009)
24. Liedtke, J.: On microkernel construction. In: Proceedings of the 15th ACM Symposium on Operating Systems Principles (SOSP 1995), Copper Mountain Resort, Colorado (December 1995); Appeared as ACM Operating Systems Review 29.5
25. Loscocco, P., Smalley, S.: Integrating flexible support for security policies into the Linux operating system. In: Proceedings of the FREENIX Track: 2001 USENIX Annual Technical Conference, pp. 29–42. USENIX Association (2001)
26. Lua. Project website (2010), http://www.lua.org
27. Maemo. Project website (2010), http://maemo.org
28. Microsoft. Windows mobile website (2010), http://www.microsoft.com/windowsmobile
29. Open Kernel Labs. OKL4 project website (2010), http://okl4.org
30. Paros. Project website (2010), http://www.parosproxy.org
31. Picciotto, J., Epstein, J.: Trusting X: Issues in building Trusted X window systems –or– what's not trusted about X? In: 14th National Computer Security Conference (1991)
32. Selhorst, M., Stüble, C., Feldmann, F., Gnaida, U.: Towards a Trusted Mobile Desktop. In: Acquisti, A., Smith, S.W., Sadeghi, A.-R. (eds.) TRUST 2010. LNCS, vol. 6101, pp. 78–94. Springer, Heidelberg (2010)
33. Shapiro, J.S., Smith, J.M., Farber, D.J.: EROS: a fast capability system. In: Proceedings of the 17th ACM Symposium on Operating Systems Principles (SOSP 1999), Kiawah Island Resort, near Charleston, Sout Carolina, pp. 170–185 (December 1999); Appeared as ACM Operating Systems Review 33.5
34. Sunyaev, A., Leimeister, J.M., Krcmar, H.: Open security issues in german healthcare telematics. In: Proceedings of the 3rd International Conference on Health Informatics, HEALTHINF 2010, pp. 187–194. INSTICC (2010)
35. Symbian Foundation Community. Project website (2010), http://www.symbian.org
36. The OpenTC Project Consortium. The Open Trusted Computing (OpenTC) project (2005-2009), http://www.opentc.net
37. Felton, D., Alves, T.: TrustZone: Integrated Hardware and Software Security (July 2004), http://www.arm.com/pdfs/TZ%20Whitepaper.pdf
38. Trusted Computing Group. TPM Main Specification, Version 1.2 rev. 103 (July 2007), http://www.trustedcomputinggroup.org
39. Vennon, T.: Android malware. A study of known and potential malware threats. Technical Report White paper, SMobile Global Threat Center (February 2010)
40. Vouyioukas, D., Kambourakis, G., Maglogiannis, I., Rouskas, A., Kolias, C., Gritzalis, S.: Enabling the provision of secure web based m-health services utilizing xml based security models. Security and Communication Networks 1(5), 375–388 (2008)

Similarity Grouping of Human Sleep Recordings Using EEG and ECG

Amro Khasawneh[1], Sergio A. Alvarez[2], Carolina Ruiz[1],
Shivin Misra[1], and Majaz Moonis[3]

[1] Department of Computer Science, Worcester Polytechnic Institute
100 Institute Road, Worcester, MA 01609 U.S.A.
[2] Department of Computer Science, Boston College
140 Commonwealth Avenue, Chestnut Hill, MA 02467 U.S.A.
[3] Department of Neurology, University of Massachusetts Medical School
55 Lake Avenue North, Worcester, MA 01655 U.S.A.
alvarez@cs.bc.edu, ruiz@cs.wpi.edu

Abstract. Characterizing variations in sleep stage composition is important in the scientific study of sleep. We use clustering, a form of unsupervised machine learning, to seek naturally occurring types within a collection of records that describe the sleep stage composition of 244 all-night human sleep studies. The results uncover a hierarchy of sleep composition types differentiated primarily by sleep efficiency or total sleep time and by the relative proportion of slow-wave sleep. The potential significance of these sleep type clusters for sleep medicine is suggested by associations between sleep type and health-related variables such as body-mass index, smoking frequency, and heart disease. EEG and ECG features, including spectral power distribution and measures of heart-rate variability, differ significantly among sleep types. The EEG signal provides sufficient information for an approximate reconstruction of the sleep type clusters, while ECG alone is found to be insufficient.

Keywords: Sleep architecture, Hypnogram, Clustering, Machine learning.

1 Introduction

The work of Loomis et al in the 1930's [19] showed that human sleep may be segmented into identifiable stages based on electrical potentials measured on the surface of the scalp. Contemporary all-night human sleep studies follow a similar approach, but employ a larger set of physiological signals. This approach, polysomnography, uses not only electroencephalography (EEG), which records electrical brain potentials, but also electrocardiography (ECG), which records heart potentials, electrooculography (EOG), which records eye movements, and electromyography (EMG), which records chin muscle movements [16]. Following standard rules of sleep scoring, a sleep technician scores these polysomnographic recordings into sleep stages one epoch (typically 30 sec) at a time, resulting in a sequence of sleep stage labels known as a hypnogram. Fig. 1 shows a hypnogram, from one of the polysomnographic recordings used for the present paper, that is labeled according to the classical Rechstchaffen and Kales (R&K) staging

A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2011, CCIS 273, pp. 380–394, 2013.
© Springer-Verlag Berlin Heidelberg 2013

standard [29]. A distinction is made between Rapid Eye Movement (REM) sleep [1], which is associated with dreaming [6], and non-REM (NREM) sleep. The latter is further subdivided into stages corresponding to light sleep (NREM 1 and 2) and deep, or slow-wave sleep (NREM 3, 4). As illustrated in Fig. 1, sleep typically follows an overall temporal pattern of stages, with several cycles involving alternation between REM and NREM sleep during the night, with a greater fraction of slow-wave sleep during the first half of the night, and a greater fraction of lighter stage NREM 2 sleep during the second half of the night. Occasional stage fragmentations, including periods of wakefulness after the beginning of sleep, are also observed.

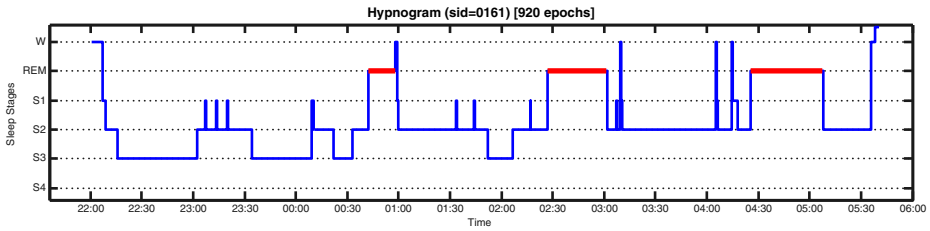


Fig. 1. Hypnogram from the present study, staged by Rechstchaffen and Kales standard

1.1 Scope of the Paper

The present paper considers the grouping of staged sleep studies by similarity, and the characterization of such groups in terms of features of the EEG and ECG time signals. Work by the authors is described that shows that staged human sleep studies may be grouped naturally into a small number of distinct clusters, or sleep types, each of which is characterized by a different sleep stage composition. Sleep efficiency and time in slow wave sleep are particularly important in differentiating among sleep types. Associations are found between sleep type and health-related factors such as age, Body-Mass Index (BMI), and smoking, which supports the objectivity and medical relevance of these sleep types. Previous work by the authors [13] only considers all-night summaries of staged sleep studies, while full polysomnographic recordings comprise more detailed electroencephalogram (EEG) and electrocardiogram (ECG) time series data. The present paper also addresses the underlying EEG and ECG data, focusing on variations in summary statistics of the EEG and ECG signals among sleep composition types. It is found that heart rate, Hjörth activity and mobility, and normalized low frequency content of ECG all differ significantly among sleep types. These results provide further evidence that sleep types constitute objective and medically relevant concepts. Furthermore, it is found that analysis of EEG signals alone, in the absence of ECG data, is sufficient for approximate reconstruction of sleep type based on unstaged sleep data. On the other hand, ECG data alone does not provide accurate sleep type classification. This paper is a revised version of the authors' work [14].

1.2 Related Work

There has been substantial interest in studying the effects on sleep architecture of various health-related and lifestyle factors. Previous works consider variations in sleep

composition associated with factors such as medication [32], smoking [36], the practice of yoga [34], body composition [28], handedness [27], autism spectrum conditions [18], and psychiatric disorders [30]. Work that relates sleep stage structure to subjective assessments of sleep quality include [3] and [12]. In contrast with the present paper, these works use groupings in sleep structure that are explicitly guided by existing sleep measures, such as the Pittsburgh sleep quality index [4], or measures extracted from the Karolinska Sleep Diary [12]. The work described in the present paper is the only one of which we are aware that addresses a general description of intrinsic sleep types based on sleep architecture itself. Employing unsupervised clustering, the present work uncovers a natural hierarchy of sleep types based on differences in measures of overall sleep stage composition, particularly sleep efficiency and time in slow wave sleep. These sleep types also exhibit characteristic differences in several features of the EEG and ECG time signals. Statistically significant associations between sleep type and health-related indicators, including age and BMI, support the medical objectivity of the sleep types described.

2 Methodology

2.1 Data

Study data were extracted from 244 recorded polysomnographic overnight sleep studies performed in the Sleep Disorder Center at Day Kimball Hospital in Putnam, CT. Data were anonymized prior to analysis. Summary statistics for patients associated with the dataset appear in Table 1. A total of 122 male and 122 female adult subjects are represented. It should be noted that the mean Body Mass Index (BMI) of the subjects in the present study is in the obese range according to international standards [26]. This point is important in interpreting some of the results obtained (e.g., section 3.3).

Table 1. Summary statistics of the sleep dataset used in this paper

	Age (years)	BMI ($kg\,m^{-2}$)	ESS (score)	BDI (score)	Mean SaO_2 (%)	Heart rate (bpm)
Male (n=122) $\mu \pm \sigma$	47.4±15.1	33.7±8.1	7.6±5.4	11.5±8.8	93.5±2.9	68.5±11.3
Female (n=122) $\mu \pm \sigma$	48.4±14.5	33.7±8.3	7.1±4.8	13.0±7.8	94.6±1.9	70.8±9.6
Overall (n=244) $\mu \pm \sigma$	47.9±14.8	33.7±8.2	7.4±5.1	12.2±8.3	94.1±2.5	69.7±10.5
min–max	20–85	19.2–64.6	0–23	0–48	70.2–97.9	46–99

Staging. The available staging information for the sleep studies used in the present study follows the Rechstchaffen and Kales (R&K) recommendations [29], which were the standard at the time that these studies were conducted, prior to 2007. In 2007, the American Academy of Sleep Medicine (AASM) revised their staging recommendations [9]. The rationale for the new staging recommendations is discussed in [31]. We briefly discuss the relationship between the two standards below.

A recognizable feature of the AASM system as compared with the R&K system is the use of only three non–REM stages N1, N2, N3 instead of four. A new AASM N3 stage essentially replaces R&K NREM stages 3 and 4 and corresponds to slow wave sleep (SWS). The AASM system also revises the rule for scoring stage N2, by eliminating the "3 minute rule" that allowed continuation of NREM stage 2 labeling for up

to three minutes in the absence of the K–complex and sleep spindle EEG features that characterize stage 2. The AASM standard also includes clarified specifications for electrode placement. The variation in stage content between the AASM and R&K systems is studied in [24]. Because of the new guidelines for scoring stage 2, AASM N2 content in a given sleep recording, as a fraction of total sleep time, is found to be 4.9% less than R&K NREM2 content, with the increase balanced by greater AASM N1 and N3 content: N1 content is 2.8% greater than NREM 1, and N3 content is 2.4% greater than NREM 3, 4 combined [24]. No statistically significant differences are found in [24] between R&K and AASM stagings in total sleep time, sleep efficiency, or REM duration.

The relatively small differences between the AASM and R&K systems in terms of relative time spent in each stage (with combined time in R&K stages NREM 3, 4 corresponding to time in AASM stage N3) suggest that the results of this work will be similarly relevant to sleep studies staged by the AASM system.

Descriptive Variables for Clustering. In order to group sleep studies by similarity, we first develop a concrete representation of sleep studies as points in a seven-dimensional space, the dimensions of which correspond to different summary measurements of sleep stage composition. These measurements are listed in Table 2. Total sleep time in minutes, fraction of time in bed spent sleeping, fraction of time in bed awake, and fraction of sleep period time in each of the sleep stages NREM $1, 2$, in SWS (NREM$3 + 4$), and REM, are used. See Fig. 2 for an illustration. Only the variables described in Table 2 are used for clustering. A dataset is constructed in which each polysomnographic sleep study is summarized as a feature vector consisting of the values of the variables in Table 2 for that study. Clustering is performed over this dataset as described in section 2.2. Additional descriptive variables are used to study associations of other factors with sleep stage composition. These variables describe health history information such as age, Body Mass Index (BMI), and habitual smoking.

Physiological Signal Variables. We use the C3-A2 EEG time signal, sampled at a rate of 200 Hz. The signal is first bandpass-filtered ($0.5 - 40$ Hz). EEG spectral variables are obtained by applying a short-time Fourier transform on overlapping 30-second windows. EEG spectra are then binned into δ ($0.5 - 4$ Hz), θ ($4 - 7$ Hz), α ($8 - 12$ Hz) and β ($12 - 30$ Hz) ranges. Summary descriptors of the EEG spectral variables are generated by taking the all-night mean, median, and standard deviation of the collection of epoch-specific spectra for each spectral band, as well as by computing other measures such as spectral entropy and Hjörth activity, mobility, and complexity.

For the ECG-related variables, RR intervals, the time durations between consecutive R peaks in the QRS complexes of the ECG time signal, are extracted and likewise described in terms of all-night summary variables including the mean RRm, standard deviation SDRR, entropy RRentr, autocorrelation RRxcorr, and mean absolute linear predictability error LPCerror of the sequence of RR interval durations, Hjörth activity, mobility, and complexity measures, the standard deviation SDSD of the sequence of differences between successive RR intervals, the root mean squared difference RMSSD of successive RR intervals, the fraction pNNx of consecutive RR intervals that differ by more than x milliseconds [22] for x = $10, 20, 30, 40, 50$, and the lengths SD1, SD2

of the principal axes of the Poincaré plot, which provides graphical information on parasympathetic nervous system activity and sympathovagal balance [10]. The Lomb-Scargle method [23] is used to compute spectral content of the RR interval sequence, and normalized RR spectral components within low frequency ($0.04 - 0.15$ Hz) and high frequency ($0.15 - 0.4$ Hz) bands are extracted. See [20] for further information on the above and other commonly used measures of Heart Rate Variability (HRV).

Table 2. Summary descriptors of sleep composition

Variable name	range of values
Total sleep time (TST)	$0 - 500$ min
Sleep efficiency (TST / Time-in-bed)	$0 - 1$
Percentage of sleep period time (SPT) in each of the stages NREM1, NREM2, SWS (NREM3 + 4), REM, wake	$0 - 100$ %

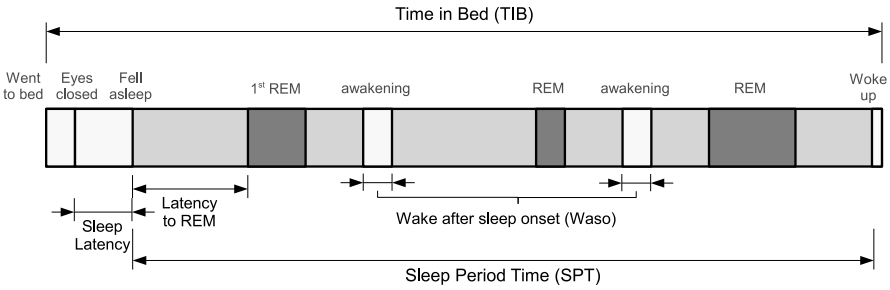


Fig. 2. Summary-level descriptors of sleep composition. NREM sleep lightly shaded.

2.2 Clustering

We consider whether sleep stage composition may be described in terms of a small number of groups, each defined by a mixture of Gaussian distributions. Unsupervised clustering is employed for this purpose. Expectation-maximization (EM) [7,25] is used to compute maximum likelihood estimates of the mixture parameters. Experiments are carried out with the EM clustering implementation in the Weka machine learning toolkit [8], version 3.7.1. This implementation initializes the Gaussian mixture parameters by k-means clustering. Unsupervised EM clustering is applied to the set of sleep composition feature vectors consisting of the values of the summary variables of Table 2 for the available staged PSG studies, yielding a predetermined number, k, of clusters. The values $k = 3,4,5$ are considered in the present paper. Consistency of clustering results is evaluated by using two metrics, the cluster purity and the normalized mutual information (NMI) [21,33]. Each of these measures the agreement between two clusterings on a scale of 0 to 1, with higher values indicating better agreement.

2.3 Statistical Significance

Statistical significance is assessed by using a χ^2 test for independence in the case of nominal variables; for continuous variables, a t-test or Wilcoxon signed rank test

are used for pairwise comparisons, and ANOVA or a Kruskal-Wallis test for multifactor comparisons. Control of the increased risk of false positives that is associated with simultaneous multiple tests of significance is accomplished by using the Benjamini-Hochberg method [2]. Given n individual findings with corresponding p-values $p_1 < p_2 < \cdots < p_n$, and given a desired overall level of significance (p-value) p, the Benjamini-Hochberg procedure declares as significant the first k findings, where k is the largest index i, $1 \leq i \leq n$, for which $p_i i/n < p$. The Benjamini-Hochberg method provides a rigorous bound on the false discovery rate (FDR), the fraction of predicted positives that are actually negatives. Indeed, the procedure as described here guarantees an overall FDR below the desired level p [2]. The FDR is distinct from the traditional type I error rate, or familywise error rate (FWR), which is the probability of one or more false positives, regardless of the number of positive predictions. The FDR is generally considered to be a better choice of significance criterion than the FWR for exploratory data analysis tasks in which there are a large number of findings to evaluate.

3 Results

We discuss the results of EM clustering of the sleep composition instance data as described in section 2, beginning with a summary of the authors' prior work [13], and show that the clusters found are stable, that they can be described by sleep efficiency and fraction of sleep time in slow wave sleep, and that the family of clusterings for $k = 3,4,5$ has a hierarchical structure. Health-related association results suggest that the clusters represent medically meaningful groups of distinct sleep behaviors. We then proceed to describe new results involving differences in EEG and ECG variables among the sleep type clusters. We find significant differences among clusters in EEG spectral band content, mean heart rate, Hjörth activity and mobility of both the EEG and ECG signals, and ECG normalized low frequency content, among others.

3.1 Salient Properties of Sleep Type Clusters

Stability. As with any hill-climbing algorithm, the EM clustering algorithm may converge to a non-global local maximum of the likelihood of the input data. Therefore, it is important to determine the variation of the clustering results with the choice of initial values for the mixture parameters. Randomization of these parameter values was performed, repeating the clustering procedure 10 times. The results show high means and low standard deviations of standard measures of clustering agreement (0.95 ± 0.11 (purity), 0.93 ± 0.16 (NMI)), indicating very little variation with initial conditions [13].

Cluster Summary Statistics. Kruskal-Wallis multiway comparison analysis of the clusters found by the EM algorithm for $k = 3,4,5$ clusters indicate significant differences among clusters in the means of all sleep composition variables used for clustering (Benjamini-Hochberg FDR $p < 0.01$, cf. section 2.3). Because of these differences in sleep composition, the clusters are referred to as *sleep types*. Wilcoxon pairwise test results confirm that the mean values of total sleep time, sleep efficiency, and fraction of sleep period time in NREM stages 2 and SWS, and in stages REM and wake, differ

significantly (FDR $p < 0.01$) between all pairs of the three clusters in the case $k = 3$. However, the only variables found to be significantly different among all pairs of clusters for $k = 4, 5$ are sleep efficiency and fraction of time in SWS.

Visualization of Sleep Types in Sleep Composition Space. Exploratory data analysis was carried out, comparing different combinations of the descriptive summary variables to visualize the clustering. Sleep efficiency and fraction of sleep period time in SWS were found to provide good results. Fig. 3 displays sleep types in terms of these two variables, for three prespecified numbers of EM clusters: $k = 3$ (left), $k = 4$ (center), and $k = 5$ (right). Classification models were constructed to predict the cluster label based

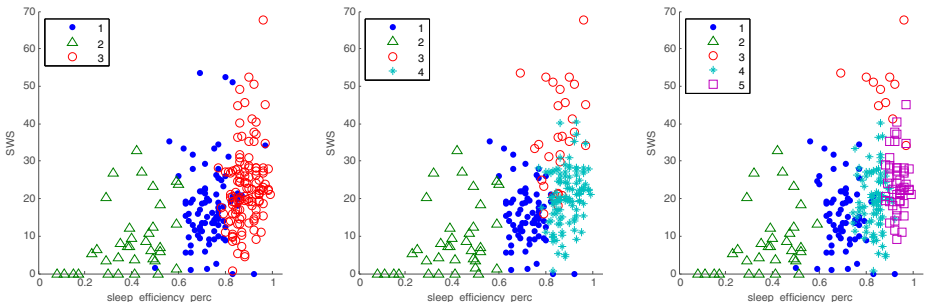


Fig. 3. Sleep types in terms of sleep efficiency and fraction of SWS, $k = 3, 4, 5$

on sleep efficiency and SWS. Pruned decision trees achieved a classification accuracy of at least 0.91 for $k = 3, 4, 5$, confirming that these variables provide good separation among clusters in all cases. The classification results have implications discussed below.

Hierarchical Structure of Sleep Types. Fig. 3 suggests that the family of clusterings for $k = 3, 4, 5$ has an approximately hierarchical structure: clusters 1 and 2 are relatively stable across the family, while the cluster labeled 3 in the leftmost image ($k = 3$) splits into the two clusters labeled 3 and 4 in the middle image ($k = 4$); in turn, the cluster labeled 4 in the middle image ($k = 4$) splits into the two clusters labeled 4 and 5 in the rightmost image ($k = 5$). Further evidence of the existence of this hierarchical structure is provided by a visualization of the decision boundaries between cluster regions that are found using Linear Discriminant Analysis (LDA). See Fig. 4. Classification accuracies are 90%, 86%, and 86% for $k = 3, 4, 5$, respectively. Use of alternative classifiers, such as naive Bayes or decision trees, yields a very similar structure. Despite some variation in the sizes and boundaries of the clusters for $k = 3, 4, 5$, support for the hierarchical structure of sleep types, including the stability of clusters 1 and 2 as conceptual entities across different values of k, derives from the summary statistics of the clusters [13]. Sleep efficiency and stage composition of clusters 1 and 2 are seen to remain nearly constant across the values $k = 3, 4, 5$. Cluster 2 has the lowest sleep efficiency across values of k, followed by cluster 1. The remaining clusters, which we loosely associate with subgroups of cluster 3 for $k = 3$, consistently have higher sleep efficiency than clusters 2 and 1. The transition from $k = 3$ to $k = 4$ produces an approximate subdivision

of cluster 3 into a new cluster 3 that is SWS–heavy, and a cluster 4 that is stage NREM2–heavy. Stage composition of these two clusters is similar in other regards. Likewise, the transition from $k = 4$ to $k = 5$ generates a new cluster, 5, characterized by higher sleep efficiency and SWS content than cluster 4, but similar NREM2 content.

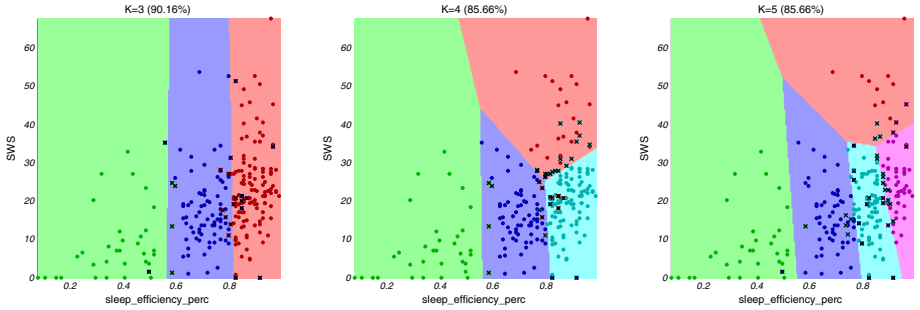


Fig. 4. LDA decision boundaries in stage composition space. Classification accuracies along top.

Relationships between Sleep Type and Health History Factors. For $k = 3$, Kruskal-Wallis multiway analysis reveals significant differences among clusters in mean patient age, collar size, BMI, smoking habit, and heart disease, at the Benjamini-Hochberg FDR level $p < 0.05$. These results show that sleep types are meaningfully connected with overall health.

Pairwise Significant Health History Differences. A Wilcoxon test shows that the cluster with highest sleep efficiency (cluster 3 on the left in Fig. 3) has significantly lower mean patient age (42.2 ± 12.7 vs 52.0 ± 14.3 and 57.0 ± 15.2), collar size (15.5 ± 2.1 vs 16.3 ± 2.1 and 16.5 ± 1.6), and habitual smoking than clusters 1 and 2, respectively. Body Mass Index (BMI) differs significantly between the two clusters at opposite ends of the sleep efficiency scale (2 and 3). Heart disease is significantly more frequent in cluster 2 than in the two clusters with higher sleep efficiency.

Pairwise comparison tests for $k = 4$ refine the above results. The mean ages in cluster 3 (47.0 ± 14.7) and 4 (42.0 ± 12.0), the clusters into which cluster 3 for $k = 3$ approximately splits, are significantly lower than in cluster 2 (57.5 ± 15.2), the cluster with lowest sleep efficiency. Age is also significantly lower in cluster 4 than in the cluster with second lowest sleep efficiency, cluster 1 (age 52.1 ± 14.7). Collar size is significantly lower in cluster 4 (15.6 ± 2.1) than in cluster 2 (16.5 ± 1.6). Smoking is significantly less frequent in cluster 4 than in cluster 2. Heart disease is significantly more frequent in cluster 2 than in any other cluster.

3.2 EEG Characteristics of Sleep Type Clusters

EEG Summary Statistics. For $k = 3$, ANOVA / Kruskal-Wallis multiway comparison finds significant differences among clusters in the variables listed in Table 3. Given the stage composition differences among clusters described in section 3.1, the corresponding differences in spectral content seen in table 3 are not entirely unexpected. For example, SWS content increases from cluster 2 to cluster 3 to cluster 1, which is the same

cluster ordering obtained according to δ band (slow wave) spectral content. Likewise, α band spectral content is highest in cluster 2, which corresponds to the fact that cluster 2 has the lowest sleep efficiency, and hence the highest occurrence of wakefulness after sleep onset. Hjörth activity indicates overall signal amplitude, and, accordingly, Hjörth activity and total power similarly vary according to the proportion of wakefulness and light sleep during the night.

EEG Summary Statistics for k=4,5. The EEG results for $k = 4, 5$ are qualitatively similar to those for $k = 3$, but with a smaller number of variables that differ significantly among clusters. Relative low-frequency (δ band) power is lowest in cluster 2, while high-frequency (β band) and total power are both highest in cluster 2. However, the level of statistical significance of the differences among clusters is reduced as k increases, and in one case the threshold $p = 0.05$ is crossed: for $k = 4$ and $k = 5$, the difference in mean β power between cluster 2 and cluster 1 is no longer significant at the level $p < 0.05$.

Table 3. EEG variables that differ significantly among sleep types, $k = 3$

	Cluster 1	Cluster 2	Cluster 3	Kruskal-Wallis p-value
No. instances	87	36	121	to 3 digits:
mu_deltaRelPower	0.550±0.079	0.438±0.112	0.634±0.069	0.000*
mu_thetaRelPower	0.100±0.022	0.088±0.026	0.106±0.028	0.000*
mu_alphaRelPower	0.110±0.045	0.141±0.073	0.084±0.026	0.000*
mu_betaRelPower	0.115±0.041	0.135±0.042	0.092±0.036	0.000*
mu_spectralEntropy	0.568±0.049	0.611±0.081	0.516±0.048	0.000*
sd_deltaRelPower	0.211±0.032	0.226±0.037	0.184±0.027	0.000*
sd_alphaRelPower	0.081±0.037	0.091±0.042	0.064±0.025	0.000*
sd_totalPower	2.076E5±12.530E5	2.489E5±9.980E5	0.603E5±2.429E5	0.000*
sd_medianFrequency	6.777±0.422	7.215±0.414	6.286±0.445	0.000*
sd_spectralEntropy	0.136±0.025	0.146±0.026	0.125±0.021	0.000*
md_deltaRelPower	0.575±0.094	0.414±0.141	0.654±0.082	0.000*
md_alphaRelPower	0.089±0.037	0.125±0.077	0.071±0.023	0.000*
md_betaRelPower	0.105±0.044	0.134±0.045	0.080±0.040	0.000*
md_spectralEntropy	0.567±0.053	0.629±0.091	0.520±0.056	0.000*
mu_hjorthMobility	0.230±0.035	0.265±0.049	0.197±0.032	0.000*
mu_hjorthComplexity	2.524±0.384	2.407±0.762	2.714±0.353	0.000*
sd_hjorthActivity	5.315E5±34.771E5	5.816E5±25.234E5	1.434E5±6.339E5	0.000*
sd_hjorthMobility	0.084±0.019	0.090±0.023	0.073±0.014	0.000*
md_hjorthMobility	0.218±0.039	0.269±0.055	0.189±0.036	0.000*
md_hjorthComplexity	2.360±0.390	2.134±0.776	2.544±0.352	0.000*
md_thetaRelPower	0.097±0.023	0.084±0.028	0.102±0.030	0.004*
md_hjorthActivity	0.197E5±0.173E5	0.219E5±0.346E5	0.258E5±0.429E5	0.005*
sd_betaRelPower	0.071±0.021	0.073±0.021	0.064±0.017	0.010*
mu_totalPower	0.388E5±2.102E5	0.573E5±2.333E5	0.150E5±0.267E5	0.042*

Sleep Type Classification based on EEG Alone. MultiDimensional Scaling (MDS) reveals considerable separation among clusters in the space described by the EEG variables, though less so than for the stage composition attributes used by EM to produce the clusters originally (see section 3.1). In order to determine if EEG provides enough information to characterize the sleep type clusters, we apply Linear Decision Analysis (LDA) to the δ band power and spectral entropy attributes for $k = 3, 4, 5$ clusters. The result exhibits the characteristic hierarchical structure discussed in section 3.1, as shown

in Fig. 5. Classification accuracies for $k = 3, 4, 5$ are 62%, 50%, and 42%, respectively. Slightly higher LDA classification accuracies (75%, 59%, and 54%, respectively) are obtained for the attribute pair consisting of mean δ relative power and median Hjörth complexity (not shown). These results confirm that the EEG attributes allow relatively accurate sleep type labeling if full polysomnography is not available.

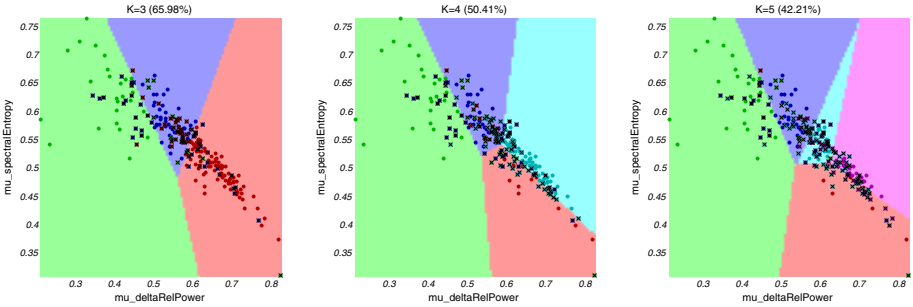


Fig. 5. LDA decision boundaries in EEG space. Classification accuracies along top.

3.3 ECG and Heart Rate Variability (HRV)

ECG Summary Statistics. In the case $k = 3$, ANOVA / Kruskal-Wallis multiway comparison finds significant differences among clusters in the variables listed in Table 4. The mean duration of RR intervals, that is, the time lapse between successive R peaks in the ECG signal (corresponding to overall heart rate), occupies the top position on this list. However, many of the remaining variables with the greatest significance, that is, with lowest p-values in Table 4, represent measures of the difference between consecutive RR intervals. Examples include SD2, SDRR, and S. In contrast, the values of variables such as SD1 that measure longer term variation in RR interval duration, do not differ significantly among sleep types. Consistent with this, the variables pNNx that measure the fraction of consecutive RR intervals that differ by more than x milliseconds are more significant the smaller the value of x, and the variable pNN50 with the largest value of x does not even meet the significance threshold $p < 0.05$.

The mean duration of RR intervals is found by a t test to be significantly higher in cluster 3 than in both cluster 1 and cluster 2. Recall that the latter two clusters are the ones with lower sleep efficiency. Note also that mean RR interval duration increases with sleep efficiency across the three clusters. Hjörth activity likewise is significantly higher in cluster 3 than in both cluster 1 and cluster 2. Overall heart rate variability as measured by the standard deviation SDRR of the sequence of RR intervals, is lowest in cluster 2, intermediate in cluster 1, and highest in cluster 3, that is, it increases with sleep efficiency (although the difference in SDRR between clusters 1 and 2 is not significant at the level $p < 0.05$). Also, RRxcorr, the autocorrelation of the RR interval sequence, is significantly lower in cluster 3 than in cluster 2, and linear prediction error LPCerror is correspondingly higher. Significance statements are at an overall Benjamini-Hochberg FDR level $p < 0.05$. Decreased heart rate variability is known to be associated with an increased risk of coronary heart disease and mortality from multiple causes [5]. Thus,

Table 4. ECG variables that differ significantly among sleep types, $k = 3$

	Cluster 1	Cluster 2	Cluster 3	Kruskal-Wallis p-value
No. instances	87	36	121	to 3 digits:
heartrate_mean	72.046 ± 9.749	74.722 ± 14.561	66.438 ± 8.507	0.000*
RRm	0.852 ± 0.123	0.840 ± 0.179	0.924 ± 0.125	0.000*
pNN10	70.565 ± 18.404	68.674 ± 17.588	77.529 ± 12.783	0.003*
SD2	0.101 ± 0.041	0.087 ± 0.039	0.111 ± 0.040	0.004*
SDRR	0.075 ± 0.031	0.065 ± 0.028	0.083 ± 0.030	0.004*
hjorthActivity	0.007 ± 0.006	0.005 ± 0.004	0.008 ± 0.006	0.004*
SDNNindex	0.053 ± 0.027	0.046 ± 0.022	0.058 ± 0.025	0.008*
S	0.009 ± 0.008	0.008 ± 0.007	0.012 ± 0.010	0.008*
SDANNindex	0.048 ± 0.022	0.043 ± 0.023	0.054 ± 0.021	0.009*
pNN20	47.104 ± 22.316	45.909 ± 22.383	55.495 ± 19.315	0.010*
nLF	33.031 ± 7.146	30.938 ± 7.788	34.781 ± 7.402	0.017*
pNN30	32.744 ± 21.494	32.761 ± 20.596	40.317 ± 20.122	0.018*
hjorthMobility	0.482 ± 0.190	0.590 ± 0.241	0.508 ± 0.161	0.022*
RRentr	5.080 ± 0.590	4.828 ± 0.688	5.169 ± 0.379	0.030*
LPCerror	0.025 ± 0.017	0.024 ± 0.014	0.028 ± 0.016	0.032*
FreqRatio	1.944 ± 1.308	1.514 ± 1.167	1.920 ± 1.141	0.034*
pNN40	23.671 ± 19.357	24.008 ± 17.906	29.595 ± 18.917	0.034*
RRxcorr	0.996 ± 0.005	0.997 ± 0.003	0.997 ± 0.003	0.046*

our HRV findings are consistent with the distribution of heart disease among the clusters described in section 3.1. These results support the view that, for $k = 3$, the sleep type clustering is linearly ordered by overall health and sleep quality.

Relationship between Frequency Ratio Behavior and Study Population. It is interesting that FreqRatio, the ratio between low and high frequency power, is significantly lower in cluster 2, which has the lowest sleep efficiency among the three clusters, than in the other two clusters. This is surprising at first sight, since cluster 2 has the greatest proportion of wakefulness after sleep onset, and hence one would expect increased sympathetic nervous system activity in this cluster than in the others, and hence a higher ratio of low to high frequency power. However, as found in [35], this expected behavior is reversed in patients that have suffered a myocardial infarction. Since cluster 2 contains the greatest incidence of heart disease among the clusters as discussed in section 3.1, this reversal is likely to contribute to the lowered FreqRatio in cluster 2. It has also been found that body fat mass is negatively correlated with low frequency ECG content in obese populations [15]. This may be an additional contributing factor to the lowered FreqRatio values in cluster 2, as many of the patients in our study are obese (Table 1).

Significant Differences among Clusters for $k = 4,5$. For $k = 4$, mean RR interval duration, standard deviation of the RR interval sequence, ratio of low to high frequency power, and ECG Hjörth activity are again significantly higher in the cluster with the highest sleep efficiency, cluster 4, than in the two clusters with the lowest sleep efficiency, 1 and 2. The findings for $k = 3,4$ persist for $k = 5$, as mean RR interval duration, standard deviation of the sequence of RR intervals, ratio of low to high frequency power, and ECG Hjörth activity are significantly higher in cluster 5, which has the highest sleep efficiency, than in the three clusters with the lowest sleep efficiency, clusters $2,1,3$. Autocorrelation of the RR sequence is significantly lower in cluster 5 than in clusters $2,1,3$.

Sleep Type Classification based on ECG. An attempt to discriminate among the sleep type clusters in terms of ECG-related variables yields poor results, with considerable overlaps between clusters. An assignment of cluster labels by LDA classification, using mean RR interval duration and Hjörth mobility as predictive variables, results in a classification accuracy of 47% for $k = 3$ clusters, decreasing to 32% for $k = 5$. Equally significantly, while the hierarchical structure of the clustering family is reflected in the relationship between the cases $k = 3, 4$, the structure breaks down for $k = 5$. See Fig. 6. As expected based on the discussion in section 3.1 and section 3.2, the LDA-predicted cluster that is most closely associated with intermediate sleep efficiency in the case $k = 3$ persists relatively unchanged for $k = 4$, while the cluster with the highest sleep efficiency splits into two clusters from $k = 3$ to $k = 4$. However, in the transition from $k = 4$ to $k = 5$, the LDA-predicted cluster with intermediate sleep efficiency changes substantially. Comparable results are obtained using other pairs of ECG variables as predictors. In view of these results, it is apparent that the existing ECG data alone is insufficient to fully characterize the family of sleep composition types.

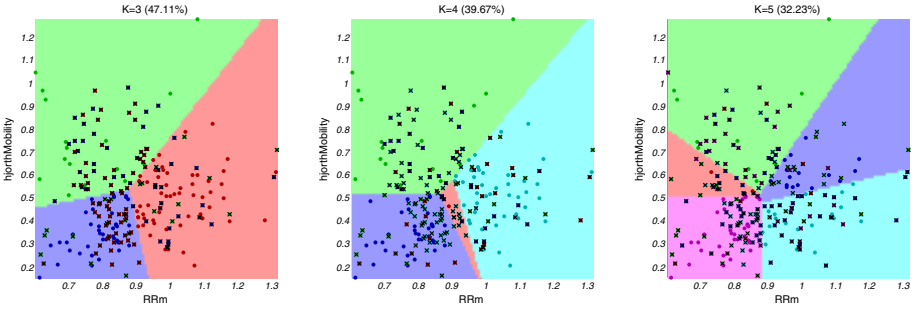


Fig. 6. LDA decision boundaries in ECG space. Cluster discrimination is poor.

Since sleep type classification relies on staged sleep recordings, a natural route toward ECG-based sleep type classification is to first construct a sleep stager that uses ECG signals only, with neither EEG nor EMG. Classification of ECG time signals into the two classes sleep and wake has been addressed without additional data [17] and with the aid of respiratory signals [11]. However, we are not aware of techniques that successfully perform full sleep staging based only on ECG. The classification results based on ECG variables described in the preceding paragraph are therefore consistent with the current state of the art. It is an open problem whether alternative descriptions of ECG time series will allow accurate sleep type predictions from ECG data alone.

4 Conclusions and Future Work

This paper has described the results obtained through unsupervised clustering of a collection of all-night human polysomnographic studies. A hierarchy of sleep composition types has been revealed, determined primarily by sleep efficiency, and subordinately by time in SWS. Associations of the sleep type clusters with health-related variables

have been described, as have some of the differences in EEG and ECG signals among clusters. Statistically significant differences have been found among clusters in BMI, age, and heart disease incidence, supporting the medical objectivity of sleep types. A comparative analysis of sleep types in control populations would be of interest for future work. Significant differences have also been found in δ, θ, α, and β EEG spectral band content, and it has been shown that relatively accurate prediction of sleep type is possible based on EEG alone. Analysis of the ECG signals has revealed significant differences among sleep types in RR interval duration, Hjörth activity and mobility, and in overall heart rate variability as measured for example by the standard deviation of the sequence of RR intervals. These findings are consistent with the health-related associations described above, in particular with heart disease and obesity. Despite the ECG findings, we have found that only limited information about sleep type may be extracted from ECG recordings alone based on the variables considered in the present paper. Future work should further explore the possibility of determining sleep type based on alternative descriptions of the ECG signals. Work in progress by the authors of the present paper involves modeling the dynamics of sleep stage transitions, and in particular studying the differences in dynamics among sleep type clusters.

References

1. Aserinsky, E., Kleitman, N.: Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. Science 118(3062), 273–274 (1953)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 57(1), 289–300 (1995)
3. Bonnet, M.H., Johnson, L.C.: Relationship of arousal threshold to sleep stage distribution and subjective estimates of depth and quality of sleep. Sleep 1(2), 161–168 (1978)
4. Buysse, D., Reynolds III, C., Monk, T., Berman, S., Kupfer, D.: The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. Psychiatry Research 28(2), 193–213 (1989)
5. Dekker, J.M., Crow, R.S., Folsom, A.R., Hannan, P.J., Liao, D., Swenne, C.A., Schouten, E.G.: Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: The ARIC study. Circulation 102(11), 1239–1244 (2000)
6. Dement, W., Kleitman, N.: The relation of eye movements during sleep to dream activity: An objective method for the study of dreaming. Journal of Experimental Psychology 53, 339–346 (1957)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1–38 (1977)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
9. Iber, C., Ancoli-Israel, S., Chesson, A.L., Quan, S.F.: The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications. American Academy of Sleep Medicine, Westchester, Illinois, USA (2007)
10. Kamen, P.W., Krum, H., Tonkin, A.M.: Poincaré plot of heart rate variability allows quantitative display of parasympathetic nervous activity in humans. Clin. Sci. (Lond.) 91(2), 201–208 (1996)

11. Karlen, W., Mattiussi, C., Floreano, D.: Sleep and wake classification with ECG and respiratory effort signals. IEEE Transactions on Biomedical Circuits and Systems 3(2), 71–78 (2009)
12. Keklund, G., Akerstedt, T.: Objective components of individual differences in subjective sleep quality. J. Sleep Res. 6(4), 217–220 (1997)
13. Khasawneh, A., Alvarez, S.A., Ruiz, C., Misra, S., Moonis, M.: Discovery of sleep composition types using expectation-maximization. In: Proc. 23rd IEEE International Symposium on Computer-Based Medical Systems (CBMS 2010), Perth, Australia (October 2010)
14. Khasawneh, A., Alvarez, S.A., Ruiz, C., Misra, S., Moonis, M.: EEG and ECG characteristics of human sleep composition types. In: Traver, V., Fred, A., Filipe, J., Gamboa, H. (eds.) Proc. Fourth International Conference on Health Informatics (HEALTHINF 2011), in Conjunction with the Fourth International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2011), pp. 97–106. SciTePress (January 2011)
15. Kim, J.A., Park, Y.-G., Cho, K.-H., Hong, M.-H., Han, H.-C., Choi, Y.-S., Yoon, D.: Heart rate variability and obesity indices: Emphasis on the response to noise and standing. J. Am. Board Fam. Med. 18(2), 97–103 (2005)
16. Kryger, M.H., Roth, T., Dement, W.C.: Principles and Practice of Sleep Medicine, 4th edn. Elsevier Saunders, Philadelphia (2005)
17. Lewicke, A.T., Sazonov, E.S., Corwin, M.J., Schuckers, S.A.C.: Reliable determination of sleep versus wake from heart rate variability using neural networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks, IJCNN 2005, vol. 4 (2005)
18. Limoges, E., Mottron, L., Bolduc, C., Berthiaume, C., Godbout, R.: Atypical sleep architecture and the autism phenotype. Brain 128(5), 1049–1061 (2005)
19. Loomis, A.L., Harvey, E.N., Hobart, G.A.: Cerebral states during sleep, as studied by human brain potentials. Journal of Experimental Psychology 21(2), 127–144 (1937)
20. Malik, M., et al.: Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation 93(5), 1043–1065 (1996)
21. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008); Web publication at informationretrieval.org
22. Mietus, J.E., Peng, C.-K., Henry, I., Goldsmith, R.L., Goldberger, A.L.: The pNNx files: re-examining a widely used heart rate variability measure. Heart 88(4), 378–380 (2002)
23. Moody, G.B.: Spectral analysis of heart rate without resampling. Computers in Cardiology 20, 715–718 (1993)
24. Moser, D., Anderer, P., Gruber, G., Parapatics, S., Loretz, E., Boeck, M., Kloesch, G., Heller, E., Schmidt, A., Danker-Hopfe, H., Saletu, B., Zeitlhofer, J., Dorffner, G.: Sleep classification according to AASM and Rechtschaffen & Kales: Effects on sleep scoring parameters. Sleep 32(2), 139–149 (2009)
25. Neal, R., Hinton, G.E.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Learning in Graphical Models, pp. 355–368. Kluwer Academic Publishers (1998)
26. World Health Organization. Obesity: preventing and managing the global epidemic. Report of a WHO Consultation. WHO Technical Report Series 894. World Health Organization (2000)
27. Propper, R.E., Christman, S.D., Olejarz, S.: Home-recorded sleep architecture as a function of handedness II: Consistent right- versus consistent left-handers. J. Nerv. Ment. Dis. 195(8), 689–692 (2007)
28. Rao, M.N., Blackwell, T., Redline, S., Stefanick, M.L., Ancoli-Israel, S., Stone, K.L.: Association between sleep architecture and measures of body composition. Sleep 32(4), 483–490 (2009)

29. Rechtschaffen, A., Kales, A. (eds.): A Manual of Standardized Terminology, Techniques, and Scoring System for Sleep Stages of Human Subjects. US Department of Health, Education, and Welfare Public Health Service – NIH/NIND (1968)
30. Benca, R.M., Obermeyer, W.H., Thisted, R.A., Gillin, J.C.: Sleep and psychiatric disorders. a meta-analysis. Arch. Gen. Psychiatry 49(8), 651–668 (1992)
31. Silber, M.H., Anconi-Israel, S., Bonnet, M.H., Chokroverty, S., Grigg-Damberger, M.M., Hirshkowitz, M., Kapen, S., Keenan, S.A., Kryger, M.H., Penzel, T., Pressman, M.R., Iber, C.: The visual scoring of sleep in adults. Journal of Clinical Sleep Medicine 3(2), 121–131 (2007)
32. Smith, S.S., Dingwall, K., Jorgensen, G., Douglas, J.: Associations between the use of common medications and sleep architecture in patients with untreated obstructive sleep apnea. Journal of Clinical Sleep Medicine 2(2), 156–162 (2006)
33. Strehl, A.: Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining. PhD thesis, The University of Texas at Austin (May 2002)
34. Sulekha, S., Thennarasu, K., Vedamurthachar, A., Raju, T.R., Kutty, B.M.: Evaluation of sleep architecture in practitioners of Sudarshan Kriya yoga and Vipassana meditation. Sleep and Biological Rhythms 4(3), 207–214 (2006)
35. Vanoli, E., Adamson, P.B., Ba-Lin, Pinna, G.D., Lazzara, R., Orr, W.C.: Heart rate variability during specific sleep stages: A comparison of healthy subjects with patients after myocardial infarction. Circulation 91(7), 1918–1922 (1995)
36. Zhang, L., Samet, J., Caffo, B., Punjabi, N.M.: Cigarette smoking and nocturnal sleep architecture. Am. J. Epidemiol. 164(6), 529–537 (2006)

Author Index

Abate, Francesco 173
Abidi, Syed Sibte Raza 335
Acquaviva, Andrea 173
Almeida, V. 246
Aloise, Fabio 273
Alvarez, Sergio A. 380
Andry, François 301
Ansermino, J. Mark 31
Apalkov, Andrey 15
Araújo, Tiago 233
Arvas, Mikko 126
Astolfi, Laura 273

Babiloni, Fabio 273
Bader, Rainer 74
Bersani, Alberto Maria 155
Borges, Carla M. 80
Borges, E. 246
Brown, B.H. 191

Cabeleira, M. 246
Cantoni, Virginio 97
Cardoso, J. 246
Cincotti, Febo 273
Cobo, Antonio 312
Correia, C. 246
Correia, Miguel V. 80

Dell'Acqua, Guido 155
Denaï, M. 191
Díez, Francisco Javier 312
Dmitrienko, Alexandra 365
Doncescu, Andrei 141
Dumont, Guy 31

Elhabiby, Mohamed 57
El-Sheimy, Naser 57
Ewald, Hartmut 74

Fallani, Fabrizio De Vico 273
Fargetta, Marco 353
Fernandez, Izaskun 312
Fernández, Roemi 15
Ficarra, Elisa 173
Fritsche, Andreas 74

Galjan, Wjatscheslaw 43
Gamboa, Hugo 233
Garcia-Remesal, Miguel 312
Gatti, Riccardo 97
Giordano, Daniela 353
Gow, Jennifer 31
Grando, M.A. 285

Hadzic, Zecir 365
Hafkemeyer, Kristian M. 43
Hendler, Talma 219

Inoue, Katsumi 141
Intrator, Nathan 219
Ishihata, Masakazu 141

Jimenez-Castellanos, Ana 312

Kameya, Yoshitaka 141
Karlen, Walter 31
Khasawneh, Amro 380
Kluess, Daniel 74
Kortazar, Xabier García de 312
Kovář, Bohumil 110
Krautschneider, Wolfgang H. 43
Krens, Robin 324

Liu, Yisi 258
Löhr, Hans 365
Lombardi, Luca 97

Macii, Enrico 173
Maganti, Hari Krishna 205
Maglione, Anton Giulio 273
Mahfouf, Mahdi 191
Mao, Chunhong 3
Maojo, Victor 312
Matassoni, Marco 205
Matos, P. 246
Mattia, Donatella 273
Meir-Hasson, Yehudit 219
Mills, G.H. 191
Misra, Shivin 380
Mittelmeier, Wolfram 74
Moonis, Majaz 380

Nabeshima, Hidetomo 141
Nguyen, Minh Khoa 258
Nicholson, Daren 301
Nunes, Neuza 233

Otero, Abraham 15

Palacios, Francisco 15
Panoutsos, George 191
Pereira, H.C. 246
Pereira, T. 246
Perez-Rey, David 312
Petersen, Chris 31
Pino, Carmelo 353
Pitkänen, Esa 126
Pozo, Francisco del 312

Rousu, Juho 126
Ruiz, Carolina 380
Ruther, Catherine 74

Sadeghi, Ahmad-Reza 365
Salazar, Antonio J. 80
Samuri, Suzani Mohamad 191
Sato, Taisuke 141
Schier, Jan 110
Schonenberg, M.H. 285
Schroeder, Dietmar 43
Shukla, Maulik 3
Silva, Ana S. 80
Sobral, Bruno 3
Sourina, Olga 258
Spampinato, Concetto 353
Spruit, Marco 324
Stefano, Angela Di 353
Stewart, Samuel Alan 335
Sullivan, Dan 3
Synnaeve, Gabriel 141

Teskey, Wesley 57
Timm, Ulrich 74
Tomasik, Jakob M. 43
Toppi, Jlenia 273

Urbanus, Nathalie 324

van der Aalst, W. 285
Vecchiato, Giovanni 273
Viejo, Elisa 312

Wan, Lin 301
Wang, Qiang 258
Winandy, Marcel 365

Zhang, Chengdong 3
Zhdanov, Andrey 219