Advances in Intelligent Systems and Computing 616

Florentino Fdez-Riverola
Mohd Saberi Mohamad
Miguel Rocha
Juan F. De Paz
Tiago Pinto *Editors*

11th International Conference on Practical Applications of Computational Biology & Bioinformatics

EXTRAS ONLINE

Springer

Advances in Intelligent Systems and Computing

Volume 616

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

More information about this series at http://www.springer.com/series/11156

Florentino Fdez-Riverola
Mohd Saberi Mohamad · Miguel Rocha
Juan F. De Paz · Tiago Pinto
Editors

11th International Conference on Practical Applications of Computational Biology & Bioinformatics

Springer

Editors
Florentino Fdez-Riverola
Escuela Superior de Ingeniería Informática
Universidad de Vigo
Ourense
Spain

Juan F. De Paz
Departamento de Informática y Automática
Universidad de Salamanca
Salamanca
Spain

Mohd Saberi Mohamad
Faculty of Computing
Universiti Teknologi Malaysia
Johor
Malaysia

Tiago Pinto
Departamento de Informática y Automática
Universidad de Salamanca
Salamanca
Spain

Miguel Rocha
Department de Informática
Universidade do Minho
Braga
Portugal

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-60815-0 ISBN 978-3-319-60816-7 (eBook)
DOI 10.1007/978-3-319-60816-7

Library of Congress Control Number: 2017943012

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Biological and biomedical researches are increasingly driven by experimental techniques that challenge our ability to analyze, process, and extract meaningful knowledge from the underlying data. The impressive capabilities of next-generation sequencing technologies, together with novel and ever-evolving distinct types of omics data technologies, have put an increasingly complex set of challenges for the growing fields of bioinformatics and computational biology. To address the multiple related tasks, for instance in biological modeling, there is the need to, more than ever, create multidisciplinary networks of collaborators, spanning computer scientists, mathematicians, biologists, doctors, and many others.

The International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) is an annual international meeting dedicated to emerging and challenging applied research in bioinformatics and computational biology. Building on the success of previous events, the 11th edition of PACBB Conference will be held on June 21–23, 2017, in the Polytechnic of Porto, Porto (Portugal). In this occasion, special issues will be published by the Interdisciplinary Sciences-Computational Life Sciences, Journal of Integrative Bioinformatics, Neurocomputing, Journal of Computer Methods and Programs in Biomedicine, Knowledge and Information Systems: An International Journal covering extended versions of selected articles.

This volume gathers the accepted contributions for the 11th edition of the PACBB Conference after being reviewed by different reviewers, from an international committee from 13 countries. PACBB'17 technical program includes 39 papers of 61 submissions spanning many different subfields in bioinformatics and computational biology.

Therefore, this event will strongly promote the interaction of researchers from diverse fields and distinct international research groups. The scientific content will be challenging and will promote the improvement of the valuable work that is being carried out by the participants. In addition, it will promote the education of young scientists, in a postgraduate level, in an interdisciplinary field.

We would like to thank all the contributing authors and sponsors, as well as the members of the Program Committee and the Organizing Committee for their hard

and highly valuable work and support. Their effort has helped to contribute to the success of the PACBB'17 event. PACBB'17 would not exist without your assistance.

Mohd Saberi Mohamad
Miguel P. Rocha
Juan F. De Paz
PACBB'17 Programme Co-chairs
Tiago Pinto
Florentino Fdez-Riverola
PACBB'17 Organizing Co-chairs

Organization

General Co-chairs

Mohd Saberi Mohamad	Universiti Teknologi Malaysia
Miguel Rocha	University of Minho, Portugal
Juan F. De Paz	University of Salamanca, Spain
Tiago Pinto	University of Salamanca, Spain
Florentino Fdez-Riverola	University of Vigo, Spain

Program Committee

Alejandro F. Villaverde	IIM-CSIC, Spain
Alexandre Perera Lluna	Universitat Politècnica de Catalunya, Spain
Alfonso Rodriguez-Paton	Universidad Politecnica de Madrid, Spain
Alfredo Vellido Alcacena	UPC, Spain
Alicia Troncoso	University Pablo de Olavide, Spain
Amin Shoukry	Egypt-Japan University of Science and Technology, Egypt
Amparo Alonso	University of A Coruña, Spain
Ana Cristina Braga	University of Minho, Portugal
Ana Margarida Sousa	University of MInho, Portugal
Anália Lourenço	University of Vigo, Spain
Armando Pinho	University of Aveiro, Portugal
Boris Brimkov	Rice University, USA
Carlos A.C. Bastos	University of Aveiro, Portugal
Carole Bernon	IRIT/UPS, France
Carolyn Talcott	Stanford University, USA
Daniel Glez-Peña	University of Vigo, Spain

David Hoksza	Charles University in Prague, Czech Republic
David Rodríguez Penas	IIM-CSIC, Spain
Eduardo Valente	IPCB, Spain
Eva Lorenzo Iglesias	University of Vigo, Spain
Fernanda Brito Correia	University of Aveiro, Portugal
Fernando De la Prieta	University of Salamanca, Spain
Fernando Diaz-Gómez	University of Valladolid, Spain
Filipe Liu	University of Minho, Portugal
Francisco Couto	University of Lisboa, Portugal
Gabriel Villarrubia	University of Salamanca, Spain
Gael Pérez Rodríguez	University of Vigo, Spain
Giovani Librelotto	Federal University of Santa Maria, Brasil
Gustavo Isaza	University of Caldas, Colombia
Gustavo Santos-García	University of Salamanca, Spain
Hugo López-Fernández	University of Vigo, Spain
Isabel C. Rocha	University of Minho, Portugal
Javier Bajo	Technical University of Madrid, Spain
Javier De Las Rivas	CSIC, Spain
João Ferreira	University of Lisboa, Portugal
Joel P. Arrais	DEI/CISUC University of Coimbra, Portugal
Jorge Vieira	IBMC, Porto, Portugal
José Antonio Castellanos Garzón	University of Salamanca, Spain
José Luis Oliveira	University of Aveiro, Portugal
Josep Gómez	Universitat Rovira i Virgili, Spain
Juan Ramos	University of Salamanca, Spain
Julio R. Banga	IIM-CSIC, Spain
Loris Nanni	University of Bologna, Italy
Lourdes Borrajo Diz	University of Vigo, Spain
Luis F. Castillo	University of Caldas, Colombia
Luis M. Rocha	Indiana University, USA
Mª Araceli Sanchís de Miguel	University of Carlos III, Spain
Manuel Álvarez Díaz	University of A Coruña, Spain
Marcelo Maraschin	Federal University of Santa Catarina, Florianopolis, Brazil
Marcos Martinez-Romero	Stanford University, UK
Maria Olivia Pereira	IBB - CEB Centre of Biological Engineering, Portugal
Martin Krallinger	CNIO, Spain
Martín Pérez-Pérez	University of Vigo, Spain
Masoud Daneshtalab	University of Turku, Finland
Miguel Reboiro	University of Vigo, Spain
Mohd Firdaus Raih	National University of Malaysia, Malaysia
Narmer Galeano	Cenicafé, Colombia

Nuno F. Azevedo	University of Porto, Portugal
Nuno Fonseca	CRACS/INESC, Porto, Portugal
Oscar Dias	CEB/IBB, Universidade do Minho, Portugal
Pablo Chamoso	University of Salamanca, Spain
Patricia González	University of A Coruña, Computer Architecture Group (GAC), Spain
Paula Jorge	IBB - CEB Centre of Biological Engineering, Portugal
Pedro G. Ferreira	Ipatimup - Institute of Molecular Pathology and Immunology of the University of Porto, Portugal
Pierpaolo Vittorini	University of L'Aquila, Italy
Ramón Doallo	University of A Coruña, Spain
René Alquezar Mancho	UPC, Spain
Rita Ascenso	Polytechnic Institute of Leiria, Portugal
Rita Margarida Teixeira Ascenso	ESTG – IPL, Portugal
Rosalía Laza	University of Vigo, Spain
Rui Camacho	University of Porto, Portugal
Sara C. Madeira	IST/INESC ID, Lisbon, Portugal
Sara Rodríguez	University of Salamanca, Spain
Sérgio Deusdado	Polytecnic Institute of Bragança, Portugal
Sergio Matos	DETI/IEETA, Portugal
Thierry Lecroq	University of Rouen, France
Valentin Brimkov	SUNY Buffalo State College, USA
Vera Afreixo	University of Aveiro, Portugal
Yingbo Cui	National University of Defense Technology, China

Organising Committee

Diogo Martinho	Polytechnic of Porto, Portugal
Filipe Sousa	Polytechnic of Porto, Portugal
João Soares	Polytechnic of Porto, Portugal
Luís Conceição	Polytechnic of Porto, Portugal
Nuno Borges	Polytechnic of Porto, Portugal
Sérgio Ramos	Polytechnic of Porto, Portugal

PACBB 2016 Sponsors

Contents

S2P: A Desktop Application for Fast and Easy Processing of 2D-Gel and MALDI-Based Mass Spectrometry Protein Data

Hugo López-Fernández[1,2,3(✉)], Jose E. Araújo[3], Daniel Glez-Peña[1,2], Miguel Reboiro-Jato[1,2], Florentino Fdez-Riverola[1,2], and José L. Capelo-Martínez[3]

[1] ESEI - Escuela Superior de Ingeniería Informática, Edificio Politécnico, Universidad de Vigo, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain
{hlfernandez,dgpena,mrjato,riverola}@uvigo.es
[2] CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain
[3] UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
{jeduardoaraujo,jlcapelom}@bioscopegroup.org

Abstract. 2D-gel electrophoresis is widely used in combination with MALDI-TOF mass spectrometry in order to analyse the proteome of biological samples. It can be used to discover proteins that are differentially expressed between two groups (e.g. two disease conditions) obtaining thus a set of potential biomarkers. Biomarker discovery requires a lot of data processing in order to prepare data for analysis or in order to merge data from different sources. This kind of work is usually done manually, being highly time consuming and distracting the operator or researcher from other important tasks. Moreover, doing this repetitive process in a non-automated, handling-based manner is error-prone, affecting reliability and reproducibility. To overcome these drawbacks, the *S2P*, an AIBench based desktop multiplatform application, has been specifically created to process 2D-gel and MALDI-mass spectrometry protein identification-based data in a computer-aided manner. *S2P* is open source and free to all users at http://www.sing-group.org/s2p.

Keywords: Protein identification · Data processing · Bioinformatics tools · Open source · 2D-gel · MALDI-TOF-MS · Protein data · Mascot identifications

1 Introduction

2D-gel electrophoresis and mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analysers, MALDI-TOF-MS, are widely used in conjunction in order to perform proteome analysis [1, 2]. In brief, while the comparison of 2D-gels allows obtaining a set of differentially expressed spots, MALDI-TOF-MS allows to identify the proteins separated in such spots.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_1

The scientific community is particularly interested in the challenging task of finding proteins that can be used to differentiate different conditions of health with the aim to aid in diagnosis, prognosis and new targeted therapies development [3–5]. In order to find such proteins, known as biomarkers, a typical experimental workflow combining 2D-gel and MALDI-TOF-MS can involve the following steps: (*i*) separation of the proteins present in a complex proteome; (*ii*) comparing the 2D-gels across samples to obtain the spots that were found expressed differentially; (*iii*) excising such spots and treating them for protein identification; (*iv*) linking the protein identifications to the 2D-gel spots; and (*v*) performing different types of data analysis to find out the potential biomarkers. Such workflow generates a large amount of data, which need to be processed before it can be properly analysed. A considerable part of the aforementioned data processing is usually carried out manually by laboratory researchers (e.g. using text editors and Excel). However, doing this repetitive process in a non-automated way presents important drawbacks: it is time consuming, it is error-prone, and it tends to lack reliability and reproducibility.

To overcome the aforementioned drawbacks we have developed the *S2P* software application (http://www.sing-group.org/s2p/), a free software that aims to help researchers overcoming these tedious but necessary data processing steps.

S2P has been created with two main goals in mind: to improve reproducibility and to save time. Nowadays, lack of reproducibility is a growing concern in science [6] and the *S2P* software aims to improve reproducibility by avoiding human errors due to manual data processing. For instance, this issue has been particularly important in recent genomics bioinformatics, where it has been demonstrated that gene name errors are widespread in the scientific literature due to the use of Excel [7, 8]. Through its user-friendly GUI interface, *S2P* dramatically reduces the time that researchers need to invest in order to get data ready for analysis. The usefulness of *S2P* is illustrated by a case study experiment that aims to establish a biomarker-based method to allow better diagnosis and monitoring of patients with bladder cancer.

The rest of the paper is structured as follows. Section 2 presents the case study and the most relevant implementation details. Section 3 reviews the results, showing how to use *S2P* to process the case study dataset. Finally, Sect. 4 concludes the paper and outlines future research work.

2 Materials and Methods

2.1 Case Study

As a case study, a dataset composed by 14 patients plus 1 healthy group of 6 individuals was used. Plasma samples from 7 anonymous patients diagnosed with bladder cancer, 7 anonymous patients diagnosed with lower urinary tract symptoms (LUTS) and 6 healthy individuals were collected following standard procedures. Both patients and healthy volunteers were informed about the project and their consent was obtained in written form. The local ethics committee approved the study. This experiment was developed as a proof of concept to differentiate bladder cancer from LUTS.

Once in the laboratory, the samples were centrifuged, and then the supernatant was withdrawn, aliquoted and stored at –80 °C until analysis. Most abundant proteins (MAPs) in plasma can mask or interfere with the detection of proteins belonging to the low-abundance protein fraction [9]. To avoid this problem, protein equalization from plasma samples was performed with dithiothreitol, DTT, according to the protocol described by Warder et al. [10] with minor modifications as described by Fernández et al. [11] and Araújo et al. [12–14]. This process was performed with five replicates for each patient. Then, the total protein content was determined using a Bradford protein assay [15].

Two dimensional gel electrophoresis separation was carried out by duplicate for each patient and for the healthy pool. Then, 2D-gels obtained for each patient and the pool of healthy volunteers were compared using the *Progenesis SameSpots* software v4.0 (NonLinear Dynamics) to find out the differentially expressed proteins. All spots of interest were excised and subjected to in-gel protein(s) digestion and then to protein fingerprint identification by mass spectrometry using MALDI-TOF-MS [16]. Finally, *S2P* was used to process the spots data (i.e. differentially expressed spots) obtained with the *SameSpots* software as well as to analyze them along with the protein identifications obtained from Mascot.

2.2 Implementation

S2P v1.0.0 is implemented in Java and it was constructed using the *AIBench* framework [17], which has been demonstrated to be suitable for rapid development of scientific applications [18, 19]. The Graphical User Interface (GUI) was constructed in Java Swing using freely available extensions such as *SwingX* or *GC4S*. *S2P* also makes use of several well-established open-source libraries such as *JFreeChart*, *charts4j*, *iText* and the *Apache Commons Mathematics library*.

The source code of the project is freely available at https://github.com/sing-group/S2P under a GNU GPL 3.0 License (http://www.gnu.org/copyleft/gpl.html). It is divided into three modules: (*i*) *core*, which contains the default implementation API, (*ii*) *gui*, which contains several reusable GUI components, and (*iii*) *aibench*, which contains a GUI application based on the *AIBench* framework.

3 Results and Discussion

With the goal of showing the main features of *S2P* as well as its usefulness to analyse real data, this section shows how it has been used to process and analyse the case study data presented.

Figure 1 illustrates the five main steps where *S2P* was used through the experiments: (*1*) to merge the *SameSpots* report into a single table where all samples can be compared; (*2*) to design the MALDI plate; (*3*) to load and filter the Mascot identifications; (*4*) to link the Mascot identifications with their corresponding spots using the MALDI plate; and (*5*) to examine and analyse spots data along with Mascot identifications. All data

needed to reproduce the steps explained below is available at http://www.sing-group.org/s2p/tutorial.html, along with a detailed quick-start tutorial that guides users using *S2P* for the first time.

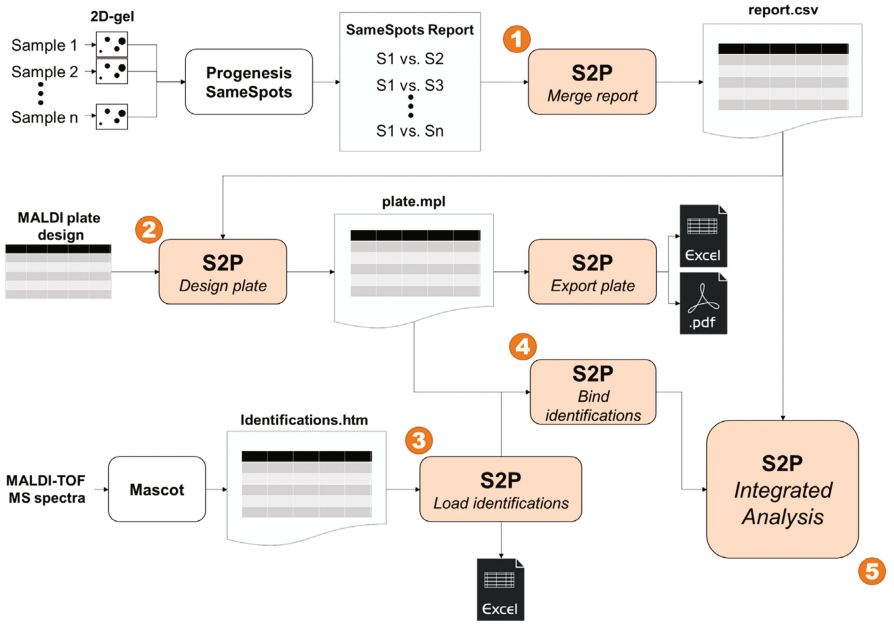


Fig. 1. Schematic S2P flow diagram.

The case study dataset was composed by 7 anonymous patients diagnosed with bladder cancer, 7 anonymous patients diagnosed with lower urinary tract symptoms (LUTS) and 6 healthy individuals that were pooled. The *Progenesis SameSpots* software was used to compare the 2D-gels corresponding to each individual against the health pool's 2D-gels to obtain the differentially expressed spots. These results were exported using the "Export report" option of *SameSpots*, which creates one HTML file per comparison (i.e. 14 files in this case). *S2P* was then used to parse and merge these reports into a single table with samples in columns and spots in rows (Step 1 of Fig. 1). This table was exported into a *comma-separated values* (CSV) file that can be easily reopen with *S2P* as well as external applications such as Excel, LibreOffice or R.

Then, these differentially expressed spots were first treated and then analysed through MALDI-TOF MS in order to identify their protein content. To do that, a dedicated sample treatment is done [16] and the pool of peptides obtained is spotted twice into a MALDI plate, which is then introduced into the MALDI apparatus for analysis. Usually, researchers fill a sheet with the position of the spots in the plate so that they can trace back where each spot was placed. This is important to know which spot is associated to each MALDI spectrum and, therefore, to know which Mascot identifications are associated to each spot. However, keeping a unique handwritten copy of this key information is risky as it can be lost or mislead and, most likely, it will be no way

to recover this information. For these two reasons, *S2P* incorporates a MALDI plate editor that allows the storage of digital copies of experiments' plates as well as print them into PDF files (Step 2 of Fig. 1). *S2P* also allows to automatically filling the plate using a set of previously loaded spots (Step 1 of Fig. 1), allowing the user to define parameters such as matrix dimensions (i.e. number of rows and columns) or the number of replicates of each spot. In our case study, *S2P* was used to create the MALDI plate and to obtain a printed copy of it that is used to guide the experimental work.

Once the MALDI-TOF MS analysis was done, the MALDI-based spectra of the digested protein(s) were submitted to Mascot in order to identify the proteins. Then, they were exported into a HTML file that was loaded into *S2P* in order to remove duplicated entries and exclude identifications with a Mascot score under 56 (Step 3 of Fig. 1). This processed list of Mascot identifications was exported into a CSV file so that it can be directly loaded into *S2P* later or used in other applications (e.g. Excel). Then, these Mascot identifications integrated with the spots data using the MALDI plate (Step 4 of Fig. 1) to know which identifications are associated with each spot.

Finally (Step 5 of Fig. 1), *S2P* allows an integrated analysis of the spots data and the Mascot identifications (Fig. 2). In the context of our case study, this option was firstly used to try to identify potential biomarkers of the two conditions of interest. When the healthy pool was compared with the bladder cancer patients, four differentially expressed spots present in at least 5 of 7 bladder cancer patients (Fig. 3A) were found. The corresponding proteins were: (*i*) Serum albumin (Spot Number [SN] = 137), (*ii*) Gelsolin (SN = 137), (*iii*) Fibrinogen gamma chain (SN = 337), (*iv*) Ig alpha-1 chain C region (SN = 360), (*v*) Ig alpha-2 chain C region (SN = 360) and (*vi*) Haptoglobin (SN = 266). When the healthy pool was compared with the LUTS patients, we found five differentially expressed spots that were present in at least 4 of 7 LUTS patients (Fig. 3B). The associated proteins were the following: (*i*) CD5 antigen-like (SN = 244), (*ii*) Heparin cofactor 2 (SN = 175 and SN = 190), (*iii*) Hemopexin (SN = 175), (*iv*) Serum albumin (SN = 192 and SN = 190) and (*v*) Inter-alpha-trypsin inhibitor heavy chain H4 (SN = 88).

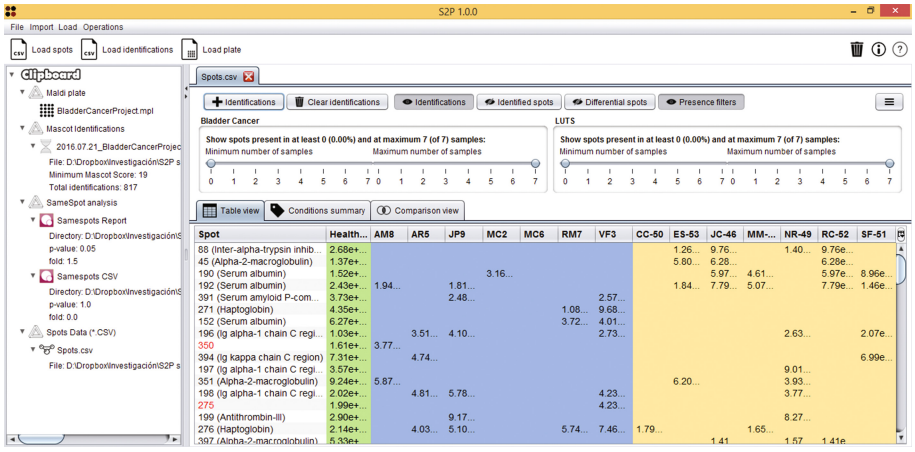


Fig. 2. Screenshot of the *S2P* integrated analysis window.

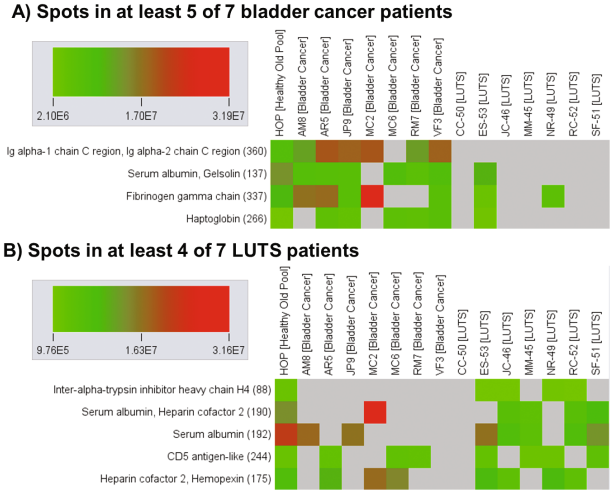


Fig. 3. Heat maps showing the differentially expressed spots.

As it can be seen in Fig. 3, a small set of candidate biomarkers was identified that can be associated to each disease. Due to this reason, a complementary approach was experimented: exporting all spots data from *Samespots* instead of exporting only those spots that were differentially expressed when each individual and the healthy pool were compared. This way, we used *S2P* to process these new dataset (analogously to step 1) and then to find spots whose average value were statistically different between bladder cancer and LUTS patients. Following this strategy, 40 differentially expressed spots (i.e. having t-test p-values corrected using Benjamini-Hochberg less than 0.05) between bladder cancer and LUTS were found, 27 of which have protein identifications associated (corresponding to 14 unique proteins). This also allowed us to compare the distribution of the expression values of each condition using box plots. For instance, Fig. 4 shows the box plots of the two spots identified in Fig. 3 that are differentially expressed between bladder cancer and LUTS patients. This information must be carefully analysed, but the usefulness of S2P to fast and accurate process and analyse data is thus proven.

Finally, it is important to remark that doing the steps described above manually took more than two weeks of handling. Now, with the help of *S2P* this data processing time

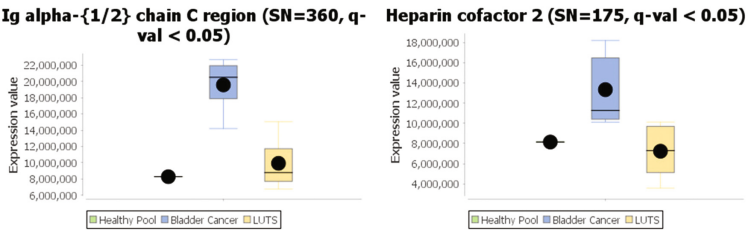


Fig. 4. Box plots of the differentially expressed spots.

has been dramatically reduced to a few minutes. Moreover, *S2P* offers the additional data analysis features shown that also allow researchers saving a lot of their valuable time.

4 Conclusions

S2P (http://www.sing-group.org/s2p/) is freely distributed under license GPLv3, providing a friendly graphical user interface designed to allow researchers saving time in data processing tasks related to 2D-gel electrophoresis and MALDI mass spectrometry protein identification-based data. The usefulness of S2P has been demonstrated by its application to a real experiment, where it notably speed up data processing as well as it improves experiment reproducibility and reliability. *S2P* is open to further extensions and we are currently developing support for more types of datasets.

Acknowledgements. This work has been partially funded by (*i*) the *"Platform of integration of intelligent techniques for analysis of biomedical information"* project (TIN2013-47153-C3-3-R) from Spanish Ministry of Economy and Competitiveness, (*ii*) the *"Discovery of biomarkers for bladder carcinoma diagnosis"* project from Nova Medical School, (*iii*) Unidade de Ciências Biomoleculares Aplicadas-UCIBIO, which is financed by national funds from FCT/MEC/Portugal (UID/Multi/04378/2013), and (*iv*) Consellería de Cultura, Educación e Ordenación Universitaria (Xunta de Galicia) and FEDER (European Union). H. López-Fernández is supported by a post-doctoral fellowship from Xunta de Galicia. J. L. Capelo acknowledges *Associação Cientifica ProteoMass* for financial support. J. E. Araújo acknowledges the financial support given by the Portuguese Foundation for Science and Technology under doctoral grant number SFRH/BD/109201/2015. SING group thanks CITI (*Centro de Investigación, Transferencia e Innovación*) from University of Vigo for hosting its IT infrastructure.

References

1. Susnea, I., Bernevic, B., Wicke, M., Ma, L., Liu, S., Schellander, K., Przybylski, M.: Application of MALDI-TOF-Mass spectrometry to proteome analysis using stain-free gel electrophoresis. In: Cai, Z., Liu, S. (eds.) Applications of MALDI-TOF Spectroscopy, pp. 37–54. Springer, Heidelberg (2012)
2. Martinez, J.L.C., Espiño, C.L., Santos, H.M.: Mass spectrometry-based proteomics: what is it expecting ahead? J. Proteomics **145**, 1–2 (2016)
3. Nagalla, S.R., Canick, J.A., Jacob, T., Schneider, K.A., Reddy, A.P., Thomas, A., Dasari, S., Lu, X., Lapidus, J.A., Lambert-Messerlian, G.M., Gravett, M.G., Roberts, C.T., Luthy, D., Malone, F.D., D'Alton, M.E.: Proteomic analysis of maternal serum in down syndrome: identification of novel protein biomarkers. J. Proteome Res. **6**, 1245–1257 (2007)
4. Thongboonkerd, V., Mcleish, K.R., Arthur, J.M., Klein, J.B.: Proteomic analysis of normal human urinary proteins isolated by acetone precipitation or ultracentrifugation. Kidney Int. **62**, 1461–1469 (2002)
5. Hsueh, C.-T., Liu, D., Wang, H.: Novel biomarkers for diagnosis, prognosis, targeted therapy and clinical trials. Biomark. Res. **1**, 1 (2013)
6. Baker, M.: Reproducibility: Seek out stronger science. Nature **537**, 703–704 (2016)

7. Zeeberg, B.R., Riss, J., Kane, D.W., Bussey, K.J., Uchio, E., Linehan, W.M., Barrett, J.C., Weinstein, J.N.: Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics **5**, 80 (2004)
8. Ziemann, M., Eren, Y., El-Osta, A.: Gene name errors are widespread in the scientific literature. Genome Biol. **17** (2016)
9. Anderson, N.L.: The Human Plasma Proteome: History, Character, and Diagnostic Prospects. Mol. Cell. Proteomics **1**, 845–867 (2002)
10. Warder, S.E., Tucker, L.A., Strelitzer, T.J., McKeegan, E.M., Meuth, J.L., Jung, P.M., Saraf, A., Singh, B., Lai-Zhang, J., Gagne, G., Rogers, J.C.: Reducing agent-mediated precipitation of high-abundance plasma proteins. Anal. Biochem. **387**, 184–193 (2009)
11. Fernández, C., Santos, H.M., Ruíz-Romero, C., Blanco, F.J., Capelo-Martínez, J.-L.: A comparison of depletion versus equalization for reducing high-abundance proteins in human serum. Electrophoresis **32**, 2966–2974 (2011)
12. Araújo, J.E., Santos, T., Jorge, S., Pereira, T.M., Reboiro-Jato, M., Pavón, R., Magriço, R., Teixeira-Costa, F., Ramos, A., Santos, H.M.: Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry-based profiling as a step forward in the characterization of peritoneal dialysis effluent. Anal. Methods **7**, 7467–7473 (2015)
13. Araújo, J.E., Jorge, S.: Teixeira e Costa, F., Ramos, A., Lodeiro, C., Santos, H.M., Capelo, J.L.: A cost-effective method to get insight into the peritoneal dialysate effluent proteome. J. Proteomics **145**, 207–213 (2016)
14. Araújo, J.E., Jorge, S., Magriço, R., Costa, T.E, Ramos, A., Reboiro-Jato, M., Fdez-Riverola, F., Lodeiro, C., Capelo, J.L., Santos, H.M.: Classifying patients in peritoneal dialysis by mass spectrometry-based profiling. Talanta **152**, 364–370 (2016)
15. Bradford, M.M.: A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. **72**, 248–254 (1976)
16. Oliveira, E., Araújo, J.E., Gómez-Meire, S., Lodeiro, C., Perez-Melon, C., Iglesias-Lamas, E., Otero-Glez, A., Capelo, J.L., Santos, H.M.: Proteomics analysis of the peritoneal dialysate effluent reveals the presence of calcium-regulation proteins and acute inflammatory response. Clin. Proteomics **11**, 17 (2014)
17. Glez-Peña, D., Reboiro-Jato, M., Maia, P., Rocha, M., Díaz, F., Fdez-Riverola, F.: AIBench: a rapid application development framework for translational research in biomedicine. Comput. Methods Programs Biomed. **98**, 191–203 (2010)
18. López-Fernández, H., Reboiro-Jato, M., Glez-Peña, D., Méndez Reboredo, J.R., Santos, H.M., Carreira, R.J., Capelo-Martínez, J.L., Fdez-Riverola, F.: Rapid development of Proteomic applications with the AIBench framework. J. Integr. Bioinforma. **8**, 171 (2011)
19. Reboiro-Jato, M., Glez-Peña, D., Méndez-Reboredo, J.R., Santos, H.M., Carreira, R.J., Capelo, J.L., Fdez-Riverola, F.: Building proteomics applications with the aibench application framework (2011)

Multi-Enzyme Pathway Optimisation Through Star-Shaped Reachable Sets

Stanislav Mazurenko[1(✉)], Jiri Damborsky[1,2], and Zbynek Prokop[1,2]

[1] Faculty of Science, Research Centre for Toxic Compounds in the Environment RECETOX, Loschmidt Laboratories, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
stan.mazurenko@gmail.com

[2] St. Anne's University Hospital, Pekarska 53, 656 91 Brno, Czech Republic

Abstract. This article studies the time evolution of multi-enzyme pathways. The non-linearity of the problem coupled with the infinite dimensionality of the time-dependent input usually results in a rather laborious optimization. Here we discuss how the optimization of the input enzyme concentrations might be efficiently reduced to a calculation of reachable sets. Under some general conditions, the original system has star-shaped reachable sets that can be derived by solving a partial differential equation. This method allows a thorough study and optimization of quite sophisticated enzymatic pathways with non-linear dynamics and possible inhibition. Moreover, optimal control synthesis based on reachable sets can be implemented and was tested on several simulated examples.

Keywords: Enzyme kinetics, Optimal control, Synthetic biology, Metabolic networks, Non-linear dynamics

1 Introduction

1.1 Multi-Enzyme Pathways

In this paper, we consider a set of chemical reactions catalysed by several enzymes. Such reactions take place inside cells and are also used in synthetic biology, e.g. in manufacturing of chemical compounds, biodegradation, medicine, etc. Currently, there are large databases of enzymes based on which pathways can be constructed to turn given substrates into desired products [1]. The enzyme kinetic optimisation of these processes is high on the agenda as it may lead to a substantial economy of time and consumables. Such optimisation may

This research was supported by the National Sustainability Programme of the Czech Ministry of Education, Youth and Sports (LO1214) and the RECETOX research infrastructure (LM2011028).

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-60816-7_2) contains supplementary material, which is available to authorized users.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_2

also provide insights into the evolution of cells since some studies suggest that optimal pathways are evolutionarily advantageous and can be predicted based on the genetic information of living cells [2].

We consider an n-step chemical reaction in which the state variables are the concentrations of metabolites produced and consumed in the course of the reaction:

$$S \xrightarrow{E_0} M_1 \xrightarrow{E_2} \dots \xrightarrow{E_{n-2}} M_{n-2} \xrightarrow{E_{n-1}} P.$$

The control here are the concentrations of enzymes E_i, the sum of which is limited from above. We will prove that under some general assumptions about the rate equations, one can expect the set of all the possible states of such systems to be star-shaped at any point in time. As a result, an optimisation of the pathway using star-shaped reachable sets [3] can be implemented to obtain the maximum concentration of the final product and the corresponding optimal profile of enzymes.

1.2 Mathematical Setup

For a pathway consisting of n consecutive steps, we will use the following notations: e_i is the concentration of the enzyme responsible for step i; x_i is the metabolite concentration; $f_i(x,t)$, $\boldsymbol{x} = (x_1, \dots, x_n)$, is the reaction rate per unit of the enzyme concentration e_i. We assume that f_i includes all the individual kinetic parameters such as k_{cat} and K_M and may depend on the concentrations of all the metabolites involved (e.g., systems with cross-inhibition are included). Moreover, the dependence of all the variables on t is implied in all the cases below, but we will omit this explicit notation for the sake of simplicity. In practice, all the rates f_i are non-linear, which significantly complicates any treatment of such systems.

According to enzyme kinetics, the time evolution of a multi-enzyme system over the time $t \in [0, T]$ can be described as follows:

$$\begin{cases} \dot{x}_1 = e_0 f_0(\boldsymbol{x}) - e_1 f_1(\boldsymbol{x}), \\ \dot{x}_2 = e_1 f_1(\boldsymbol{x}) - e_2 f_2(\boldsymbol{x}), \\ \dots \\ \dot{x}_{n-1} = e_{n-2} f_{n-2}(\boldsymbol{x}) - e_{n-1} f_{n-1}(\boldsymbol{x}), \\ \dot{x}_n = e_{n-1} f_{n-1}(\boldsymbol{x}). \end{cases} \tag{1}$$

In order to make sure that none of the concentrations becomes negative, we will require that for any metabolite i the rate f_{i-1} is non-negative and f_i is non-positive at $x_i = 0$. In other words, metabolite i is not consumed when its concentration is already zero.

We will consider the following control set:

$$\boldsymbol{e} \in E = \left\{ (e_0, \dots, e_{n-1}) \,\middle|\, e_i \geq 0, i = 0..n-1, \sum_{i=0}^{n-1} e_i \leq E_{max} \right\},$$

which indicates that at any moment in time the total enzyme concentration must not exceed a certain predefined value E_{max}. This limitation, for example,

describes limited resources of a cell that force it to choose which enzyme to produce or maintain at any point in time.

As far as the starting points are concerned, we will consider the following two most wide-spread frameworks: (A) all $x_i(0) = 0$ and $f_0 \geq 0$ (there is a constant supply of the initial substrate); or (B) the initial concentration $x_1(0) = 1$, $x_i(0) = 0$ for $i = 2..n$, and $f_0 \equiv 0$ (the first metabolite is the initial substrate being consumed in the course of the reaction).

Finally, we will assume that the standard existence and uniqueness results hold for the solutions to (1) over the whole relevant time interval for any measurable input $\bar{\boldsymbol{e}} \in E$ [4,5], which is usually the case in enzyme kinetics since the state vector denotes real concentrations limited from above and below. We will provide some examples of such systems in the following sections.

1.3 Optimal Control

In this framework, several objectives for optimal control are possible. Usually, one is interested in maximizing the final product, which can be formulated either as the minimization of the transition time t_f to drive x_n to some predefined level [6] or by maximizing x_n at a fixed point in time [7]. Other definitions of the transition time are also possible [8–10]. Moreover, a multi-objective optimization problem can also occur [11]. For the sake of simplicity, we will be considering the maximization of the final product at a given point in time although more general target functions can also be used (see below).

There are two main groups of methods commonly used to find optimal solutions: the so-called direct and indirect methods. The former usually imply a transformation of the original problem into non-linear programming by time-discretization and approximation of the control variables either alone or together with the states (for a comprehensive review see [11]). The advantages include a great variety of solvers, a general applicability, and an intuitive implementation. Nonetheless, these methods require some preliminary proof of the existence and stability of the solution. Moreover, global optima finders are much more computationally expensive than local ones, and due to the innate infinite dimensionality, the costs of refining the grid are high. Finally, if the target function is changed, e.g. to account for other metabolites, the entire calculation has to be repeated.

The indirect methods suggest analytical treatment of the problem, e.g., by using Pontryagin's maximum principle [2,6,9,10]. The main advantages include a more comprehensive analysis of the system behaviour and simpler numerical methods. However, Pontryagin's maximum principle is only a necessary condition, and the exact analytical solutions are usually difficult to obtain even in the case of simple linear systems. The proof of a global maximum is again complicated, and any change of the model, e.g. addition of cross-inhibition, may completely invalidate the analysis.

In this article, we suggest an alternative indirect method based on exact reachable sets [12,13], i.e. the states of a multi-enzyme system reachable from the initial point for all the possible enzyme profiles. While this method is more

computationally intensive than the maximum principle, it provides the time-evolution of the system in full since all the possible states are analysed. This allows for some flexibility in choosing the target functional after the calculation of reachable sets. Optimal control synthesis may be implemented in various ways once the sets are calculated, and the global optimality is implied automatically. No change to the model will require any qualitative re-analysis. Moreover, geometric state constraints may be taken into account, which extends the applicability of the method to, e.g. the problems with metabolite constraints due to metabolite toxicity. Finally, given some relatively broad assumptions about the reaction rates, the reachable sets are star-shaped, which reduces the problem dimensionality by one and enhances its computational efficiency and applicability. The summary table comparing the approaches mentioned above is given in the Supplement (table S1).

2 Star-Shaped Reachable Sets

We will now briefly define reachable sets and their applications to optimization, provide the evolution theorems for star-shaped sets, and formulate the main theoretical result for the systems in question.

2.1 Reachable Sets and Optimization

Reachable sets provide an important tool for the analysis of the time evolution of systems as they demonstrate how systems might behave given every possible control input. In order to demonstrate a general idea, consider the following differential inclusion:

$$\dot{\boldsymbol{x}} \in F(t, \boldsymbol{x}), \;\; \boldsymbol{x}(t_0) \in X_0, \;\; t \in T = [t_0, t_1], \tag{2}$$

where X_0 is a compact subset of R^n and F is a continuous multivalued map from $T \times R^n$ to compact convex subsets of R^n. For instance, (1) can be formulated in the above terms if one takes the union of the right-hand side of the equations over $\boldsymbol{e} \in E$. This differential equation generates a bundle of trajectories; consequently, its behaviour may be translated into that of the bundle. Let the reachable set $X[t]$ be the set of all possible states of the system at time t. The intuitive strategy to find $X[t]$ by inserting different values from $F(t, \boldsymbol{x})$ may work only if an explicit analytical solution is available, which is hardly ever the case even for linear systems. However, under some general assumptions on F, the reachable set can be found as the solution to an evolutionary equation [14]. While this equation is usually difficult to solve, a great variety of methods has been developed to calculate such sets [12,13,15].

In this paper, we will use the fact that under some general assumptions (see the Supplement), inclusion 2 has reachable sets that are star-shaped [16,17], i.e. they are compact, and for any $\lambda \in [0, 1]$ the set $\lambda X[t] \subseteq X[t]$. Such sets are uniquely defined by their radial function:

$$r(\boldsymbol{l}, t) = r(\boldsymbol{l} | X[t]) = \max\{\lambda \geq 0 : \; \lambda \boldsymbol{l} \in X[t]\}$$

that is the viscosity solution to the following partial differential equation on an n-dimensional sphere S^n :

$$\frac{\partial r}{\partial t} = \rho\left(-\frac{\partial_s r}{\partial \boldsymbol{l}} + r\boldsymbol{l}\,\middle|\,\frac{1}{r}F(t, r\boldsymbol{l})\right), \tag{3}$$

where $\rho(\boldsymbol{l}|F) = \sup\{\sum_i l_i y_i | \boldsymbol{y} \in F\}$ is the support function. This result, together with viscosity methods [18,19], provides a powerful tool for an exact calculation of reachable sets, e.g. for multi-enzyme reactions as demonstrated below.

As soon as one calculates the reachable set $X[t]$, the optimal solution to maximizing x_n at time T is tantamount to finding the point in $X[T]$ with the maximal value of coordinate x_n. In general, any target function dependent only on the final metabolite concentrations can be used since given $X[T]$, the initial optimal control problem turns into a relatively simple optimization of the function over the set $X[T]$. And once the optimal point has been found, one may apply control synthesis strategies to find the control profile that will lead the system to this optimum [3].

2.2 Star-Shaped Sets Generated by Multi-Enzyme Pathway

We will now apply the results of the previous subsection to the multi-enzyme systems (1) for initial conditions (A), i.e. some constant supply of the substrate, and (B), in which the first substrate is being consumed without any supply. The direct adaptation of Assumption S to (1) leads to the following results:

Proposition 1. *Suppose for system (1) with initial condition (A) the rate functions $f_i(\boldsymbol{x})$ are Lipschitz-continuous with the constant independent of t. If for any $\lambda \in (0,1]$ and $\boldsymbol{x}$: $f_i(\lambda\boldsymbol{x}) \neq 0 \Rightarrow 0 \leq \lambda f_i(\boldsymbol{x})/f_i(\lambda\boldsymbol{x}) \leq 1$, the radial function of the reachability set $r(\boldsymbol{l}, t) = r(\boldsymbol{l}|X[t])$ is the pointwise limit of $r_\varepsilon(\boldsymbol{l}, t)$ for any $\boldsymbol{l} \in S^n$ and $t \in [0, T]$, where $r_\varepsilon(\boldsymbol{l}, t)$ is the viscosity solution to the following equation on $S^n \times [0, T]$:*

$$\frac{\partial r_\varepsilon}{\partial t} = E_{max} \max_i \left\{ f_i(r_\varepsilon \boldsymbol{l}) \left(\frac{1}{r_\varepsilon} \left(\frac{\partial_s r_\varepsilon}{\partial l_i} - \frac{\partial_s r_\varepsilon}{\partial l_{i+1}} \right) - l_i + l_{i+1} \right) \right\},$$

$$r_\varepsilon(l, 0) = \varepsilon \to +0.$$

(here for $i = 0$ symbols $\partial_s r/\partial l_i$ and l_i should be omitted).

As far as initial condition (B) is concerned, we will replace the coordinate x_1 with $x_1^* = x_1 - 1$. If in addition to the above we require that f_i is non-negative and non-decreasing in x_1, the following holds:

Corollary 1. *Suppose for (1) the initial concentration $x_1(0) = 1$, $x_i(0) = 0$ for $i = 2..n$, and $f_0 \equiv 0$. Moreover, suppose that in addition to the requirements of Proposition 1 on f_i, the f_i that depend on x_1 are non-negative and non-decreasing in x_1. Then for (1) with the new coordinate $x_1^* = x_1 - 1$ Proposition 1 holds.*

The proofs of the statements above are given in the Supplement.

2.3 Examples

Here we will list the examples of (1) relevant to the enzyme kinetics, for which Proposition 1 holds:

1. Linear mass-action kinetics $f_i(\boldsymbol{x}) = k_i x_i$;
2. Michaelis-Menten kinetics: $f_i(\boldsymbol{x}) = k_i x_i/(K_i + x_i)$, with substrate inhibition: $f_i(\boldsymbol{x}) = k_i x_i/(K_i + x_i + N_i x_i^2)$, or with cross-inhibition: $f_i(\boldsymbol{x}) = k_i x_i/(K_i + \sum_j N_{ij} x_j)$;
3. Power law $f_i(\boldsymbol{x}) = k_i x_i^c$ with $c \in (0, 1)$;

All the above functions may be present in any combination, thereby providing a significant flexibility for the model selection.

Moreover, the same enzyme can be used in different steps if the following additional requirement holds: for any enzyme e used in several reactions the value $\lambda f_i(\boldsymbol{x})/f_i(\lambda \boldsymbol{x})$ is independent of i for the respective i's. This will be the case, e.g. in Michaelis-Menten kinetics since the free enzyme, and consequently, the denominator of f_i, will be the same across the respective i's. Reversible reactions are also covered. In other cases when the star-shapedness cannot be guaranteed, one may still use general reachable set methods [13], albeit forgoing the advantage of the reduced dimensionality.

We will now proceed to several examples.

Example 1. The first example is a three-metabolite scheme with a constant supply of substrate zero, and it demonstrates the standard bang-bang optimal profile [2,9,10] (Fig. 1):

$$\begin{cases} \dot{x}_1 = \frac{0.1x_0}{1+x_0} e_0 - \frac{0.1x_1}{0.1+x_1} e_1, \\ \dot{x}_2 = \frac{0.1x_1}{0.1+x_1} e_1. \end{cases}, \; t \in [0, 1], \; x_0 \equiv 1, \; E_{max} = 10. \tag{4}$$

This switching between the two regimes stems from the intuitive fact that the rate of the reaction is increasing with the increase in x_1. As a result, the optimal strategy is to accumulate x_1 first and then to switch to production of x_2.

Example 2. The second example is a modification of the previous case with a substrate inhibition of enzyme e_1 (Fig. 2):

$$\begin{cases} \dot{x}_1 = \frac{0.1x_0}{1+x_0} e_0 - \frac{0.1x_1}{0.1+x_1+5x_1^2} e_1, \\ \dot{x}_2 = \frac{0.1x_1}{0.1+x_1+5x_1^2} e_1. \end{cases}, \; t \in [0, 1], \; x_0 \equiv 1, \; E_{max} = 10. \tag{5}$$

Now, the simple accumulation of x_1 will not yield an optimal solution; due to the inhibition, the reaction rate would decrease for large values of x_1. Hence, e_1 should be switched on earlier and not to its maximal value as can be seen from the optimal control synthesis in Fig. 2.

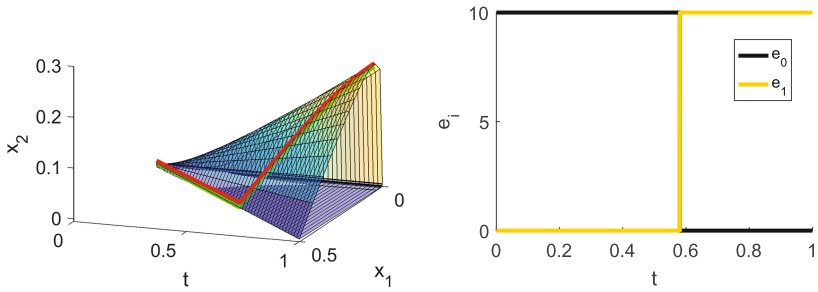


Fig. 1. The reachable tube of Example 1 (left) and the synthesized optimal control (right). The red line is the synthesized trajectory from the point with the maximal coordinate x_2 backward in time. The green line is the trajectory from the origin calculated with the filtered optimal control. The calculation time on a regular desktop was 7 s.

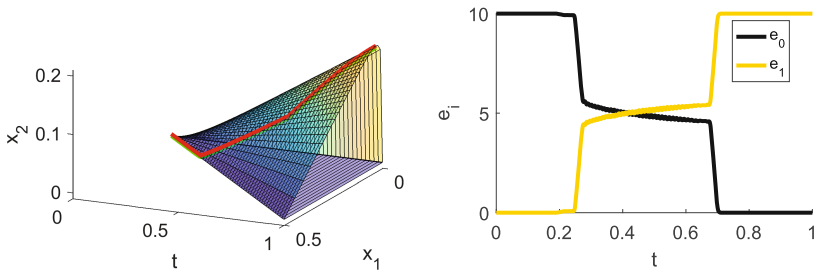


Fig. 2. The reachable tube of Example 2 (left) and the synthesized optimal control (right). The red line is the synthesized trajectory from the point with the maximal coordinate x_2 backward in time. The green line is the trajectory from the origin calculated with the filtered optimal control. The calculation time on a regular desktop was 7 s.

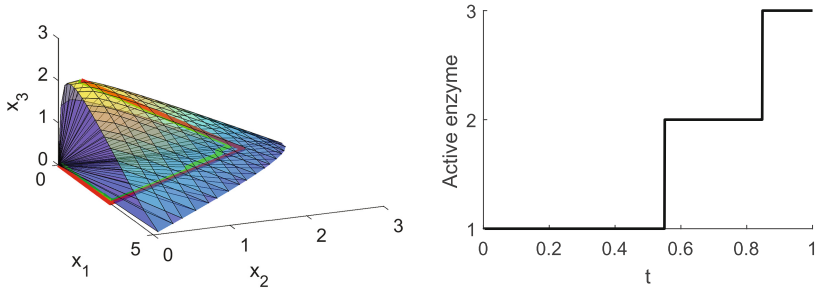


Fig. 3. The reachable set of Example 3 at t = 1 (left) and the synthesized optimal control (right). The red line is the synthesized trajectory from the point with the maximal coordinate x_3 backward in time. The green line is the trajectory from the origin calculated with the filtered optimal control. The calculation time on a regular desktop was 57 s.

Example 3. Finally, we will also consider a three-dimensional example to demonstrate the calculability of the method (Fig. 3):

$$\begin{cases} \dot{x}_1 = \frac{x_0}{1+x_0} e_0 - \frac{2x_1}{2+x_1} e_1, \\ \dot{x}_2 = \frac{2x_1}{2+x_1} e_1 - \frac{3x_2}{1+x_2} e_2. \\ \dot{x}_3 = \frac{3x_2}{1+x_2} e_2. \end{cases}, \; t \in [0,1], \; x_0 \equiv 1, E_{max} = 10. \tag{6}$$

In general, the curse of dimensionality leads to a significant increase in computational costs as the dimensionality of x increases, in contrast to direct methods that are sensitive to the dimensionality of the control vector. The star-shaped sets partially alleviate the problem by reducing the dimensionality by one, which is why a two-dimensional grid was used in this example. Thus, the calculations for systems with up to 5–6 state variables can be performed on a regular desktop in a reasonable time. Otherwise, approximation techniques, e.g., ellipsoidal calculus [12] or zonotopes [15], might be used.

3 Conclusions

In this work, we studied a multi-enzyme optimization problem. We demonstrated that under some general assumptions, the reachable sets of such a problem are star-shaped. Further, we constructed reachable sets using their radial function that is a viscosity solution to a certain partial differential equation. By doing so, we were able to visualize the time-evolution of the system given all possible enzyme profiles. Once calculated, the reachability tube provides means for optimal control synthesis. Finally, we considered several examples that verified results obtained by other authors using different techniques as well as provided some new insights into the behavior of more sophisticated multi-enzyme pathways, e.g. the ones with inhibition.

References

1. Carbonell, P., Parutto, P., Herisson, J., Pandit, S.B., Faulon, J.L.: XTMS: pathway design in an eXTended metabolic space. Nucleic Acids Res. **42**(W1), W389–W394 (2014)
2. Klipp, E., Heinrich, R., Holzhütter, H.G.: Prediction of temporal gene expression. Eur. J. Biochem. **269**(22), 5406–5413 (2002)
3. Mazurenko, S.: Partial differential equation for evolution of star-shaped reachability domains of differential inclusions. Set-Valued Variational Anal. **24**(2), 333–354 (2016)
4. Filippov, A.: On certain questions in the theory of optimal control. J. Soc. Ind. Appl. Math. Ser. A Control **1**(1), 76–84 (1962)
5. Aubin, J.P., Cellina, A.: Differential Inclusions: Set-Valued Maps and Viability Theory, vol. 264. Springer, Heidelberg (1984)
6. Bayon, L., Otero, J.A., Ruiz, M.M., Suárez, P.M., Tasis, C.: Sensitivity analysis of a linear and unbranched chemical process with n steps. J. Math. Chem. **53**(3), 925–940 (2015)

7. Dvorak, P., Kurumbang, N.P., Bendl, J., Brezovsky, J., Prokop, Z., Damborsky, J.: Maximizing the efficiency of multienzyme process by stoichiometry optimization. ChemBioChem **15**(13), 1891–1895 (2014)
8. Llorens, M., Nuño, J.C., Rodríguez, Y., Meléndez-Hevia, E., Montero, F.: Generalization of the theory of transition times in metabolic pathways: a geometrical approach. Biophys. J. **77**(1), 23–36 (1999)
9. Bartl, M., Li, P., Schuster, S.: Modelling the optimal timing in metabolic pathway activation – use of Pontryagin's Maximum Principle and role of the Golden section. Biosystems **101**(1), 67–77 (2010)
10. Oyarzun, D.A., Ingalls, B.P., Middleton, R.H., Kalamatianos, D.: Sequential activation of metabolic pathways: a dynamic optimization approach. Bull. Math. Biol. **71**(8), 1851–1872 (2009)
11. Hijas-Liste, G.M., Klipp, E., Balsa-Canto, E., Banga, J.R.: Global dynamic optimization approach to predict activation in metabolic pathways. BMC Syst. Biol. **8**(1), 1 (2014)
12. Kurzhanski, A.B., Varaiya, P.: Dynamics and Control of Trajectory Tubes. Theory and Computation. Birkhauser, Basel, (2014)
13. Mitchell, I.M., Bayen, A.M., Tomlin, C.J.: A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. IEEE Trans. Autom. Control **50**(7), 947–957 (2005)
14. Panasyuk, A.I., Panasyuk, V.I.: An equation generated by a differential inclusion. Math. Notes Acad. Sci. USSR **27**(3), 213–218 (1980)
15. Althoff, M., Stursberg, O., Buss, M.: Computing reachable sets of hybrid systems using a combination of zonotopes and polytopes. Nonlinear Anal. Hybrid Syst. **4**(2), 233–249 (2010)
16. Mazurenko, S.S.: Viscosity Solutions to Evolution of Star-Shaped Reachable Sets (2016). Submitted and is currently under review
17. Kurzhanski, A.B., Filippova, T.F.: On the theory of trajectory tubes - a mathematical formalism for uncertain dynamics, viability and control. In: Advances in Nonlinear Dynamics and Control, pp. 122–188. Birkhauser, Boston (1993)
18. Crandall, M.G., Lions, P.L.: Two approximations of solutions of Hamilton-Jacobi equations. Math. Comput. **43**(167), 1–19 (1984)
19. Souganidis, P.E.: Approximation schemes for viscosity solutions of Hamilton-Jacobi equations. J. Differ. Eqn. **59**(1), 1–43 (1985)

Automated Collection and Sharing of Adaptive Amino Acid Changes Data

Noé Vázquez[1,2], Cristina P. Vieira[3,4], Bárbara S.R. Amorim[3,5], André Torres[3,5], Hugo López-Fernández[1,2,6], Florentino Fdez-Riverola[1,2], José L.R. Sousa[3,4], Miguel Reboiro-Jato[1,2], and Jorge Vieira[3,4(✉)]

[1] ESEI – Escuela Superior de Ingeniería Informática, Edificio Politécnico Universidade de Vigo, Campus Universitario as Lagoas s/n, 32004 Ourense, Spain

[2] CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain

[3] Instituto de Investigação e Inovação em Saúde (I3S), Universidade do Porto, Rua Alfredo Allen, 208, 4200-135 Porto, Portugal
jbvieira@ibmc.up.pt

[4] Instituto de Biologia Molecular e Celular (IBMC), Rua Alfredo Allen, 208, 4200-135 Porto, Portugal

[5] Instituto Nacional de Engenharia Biomédica (INEB), Rua Alfredo Allen, 208, 4200-135 Porto, Portugal

[6] UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Lisbon, Portugal

Abstract. When changes at few amino acid sites are the target of selection, adaptive amino acid changes in protein sequences can be identified using maximum-likelihood methods based on models of codon substitution (such as codeml). Such methods have been used numerous times using a variety of different organisms but the time needed to collect the data and prepare the input files means that tens or a couple of hundred coding regions are usually analyzed. Nevertheless, the recent availability of flexible and ease to use computer applications to collect the relevant data (such as BDBM), and infer positively selected amino acid sites (such as ADOPS) means that the whole process is easier and quicker than before, but the lack of a batch option in ADOPS, here reported, still precluded the analysis of hundreds or thousands of sequence files. Given the interest and possibility of running such large scale projects, we also developed a database where ADOPS projects can be stored. Therefore, here we also present B+ that is both a data repository and a convenient interface to look at the information contained in ADOPS projects without the need to download and unzip the corresponding ADOPS project file. The ADOPS projects available at B+ can also be downloaded, unzipped, and opened using the ADOPS graphical interface. The availability of such a database ensures results repeatability, promotes data reuse with significant savings on the time needed for preparing datasets, and allows further exploration of the data contained in ADOPS projects effortlessly.

N. Vázquez, C.P. Vieira and B.S.R. Amorim—These authors contributed equally to this work.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_3

Keywords: ADOPS · Positive selection · B+ database · Open data

1 Introduction

Amino acid changes in protein sequences can be adaptive, and when changes at few amino acid sites are the target of selection they can be detected using maximum-likelihood methods based on models of codon substitution [1–3]. This approach has been applied numerous times to infer positively selected amino acid sites, such as at interleukin-3 (IL3), a protein associated with brain volume variation in general human populations [4], at formyl peptide receptors in mammals [5], at scorpion sodium channel toxins [6], at the *Mimulus* plant CENH3 protein [7], at the oyster *Crassostrea gigas* peptidoglycan recognition proteins [8], at host immune response genes [9, 10], the envelope glycoprotein of dengue viruses [11], the attachment glycoprotein of respiratory syncytial virus [12], measles virus haemagglutinin [13], influenza B virus hemagglutinin [14], HIV proteins [15], hemagglutinin-neuraminidase protein of Newcastle disease virus [16], *Trypanosoma brucei* proteins [17], at the vertebrate skeletal muscle sodium channel protein [18], at the p53 protein [19], the fruitless protein in *Anastrepha* fruit flies [20], CC chemokine receptor proteins [21], or at the proteins encoded by plant genes that are involved in gametophytic self-incompatibility specificity determination [22–25] to name just a few. Recently, it has been argued that pharma and biotech industries can successfully use the knowledge generated by such approach to tackle real-life problems [26].

Although maximum-likelihood methods based on models of codon substitution have been widely used to infer positively selected amino acid sites, the size of the average project is still relatively small mainly due to the time needed to collect the relevant coding sequences and prepare input files for the different software applications. The recent availability of computer applications such as BDBM (http://www.sing-group.org/BDBM/) greatly eases the preparation of large data sets. Moreover, the availability of the ADOPS [27] computer application allowed running in an automated way all the steps needed to infer positively selected amino acid sites, starting from a FASTA file with non-aligned coding sequences, but the lack of a batch option in this application still meant that it was not practical to run thousands of sequence files.

Here, we report the implementation of a batch option in the ADOPS software [27] that allows users to easily run large scale analyses involving thousands of genes, using moderate computer resources. Given this improvement, making ADOPS projects (especially large scale projects) available to the research community was the next logical step. Therefore, here we also present B+ (http://bpositive.i3s.up.pt/), a database that has been specifically designed to store and show the information contained in ADOPS project files. Although a database dedicated to positive selection inferences at the codon level has been published [28] it is dedicated to a specific group of organisms, and the possibility of reusing data is not as easy as with B+ and ADOPS. Both large and small ADOPS datasets can be submitted to B+ (as compressed tar.gz files) along with a description containing the details about how the project was performed. At present, the B+ database hosts the "Closely related Drosophila data set (2016)" that provides ADOPS projects

for 19652 Drosophila transcripts, 14.6% of which show signs of positive selection (1200 genes), although curated analyses must now be performed to validate these results.

2 ADOPS Batch Mode

Multiple instances of the ADOPS graphical user interface (GUI) can be launched simultaneously and thus multiple parallel processing of ADOPS projects is possible as long as enough memory is available, the required memory being dependent on the number of sequences used in the project and the total number of individual projects to be run. A single ADOPS batch project with 50 individual projects each with an average of 10 sequences per individual project runs in about 1–2 days. This means that even with limited computational power it is possible to run about 100 individual projects every two days.

In order to launch the new batch option implemented in ADOPS, the user launches the GUI and chooses the *'Create Batch Project'* option under the *'Project'* menu (Fig. 1). Then, the user gives the name and location of the folder that will contain the individual ADOPS project files. The base configuration can be changed at this time but if none is specified the default configuration stored in the *'system.conf'* file will be used. Finally, the user selects the FASTA files that will be used for the experiments and a new window is launched, showing the status of each individual ADOPS project (Fig. 1). The name of the experiment of each individual project will be named "batch".

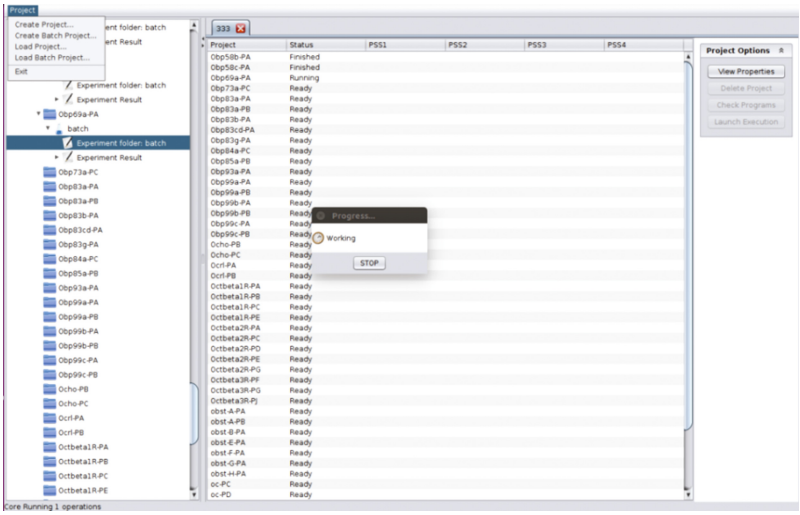


Fig. 1. The *'Create Batch Project'* option.

3 B+ Database Implementation

The B+ database is both a data repository and a convenient interface to browse the information contained in each ADOPS project interactively, without the need to download and unzip the corresponding ADOPS project. Thus, B+ allows the exploration of the data contained in ADOPS projects effortlessly.

B+ has been developed using the Laravel framework (https://laravel.com/) for web development. For a richer user interface, the Bootstrap framework (http://getbootstrap.com/) and the jQuery library (https://jquery.com/) have also been used. Figure 2 details the architecture of the B+ database and Fig. 3 shows the user interface.

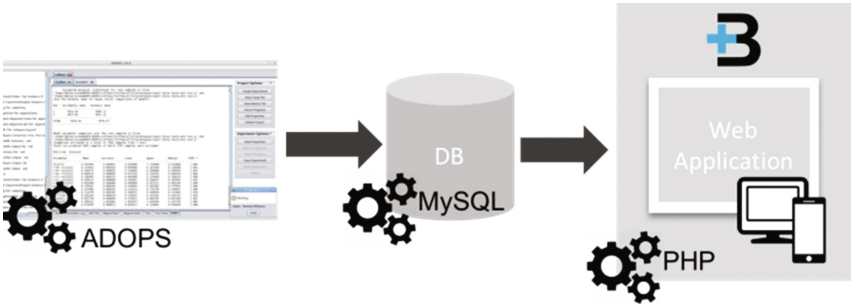


Fig. 2. Data architecture of the integration of ADOPS and other technological solutions in the development of the B+ repository. The MySQL technology allows the integration of the ADOPS results in order to allow through PHP programming its web availability.

B+ repository is divided into three visualization levels, flowing from the general to the detail. The first level is a broad data view introducing each dataset available under the platform, the second is a tabular view showing the content of each dataset and finally an individual download link with access to a detailed view of the selected record. The default view has a table arrangement of ten rows per page that can be refreshed using the "Number of entries" field. The search is provided at the right top corner of the interface and it is executed in the Database on the server side to provide maximum performance. The pagination is also handled in the server side to minimize the transfer of unnecessary data to the client. The search matches full and partial words using name and description fields of the database. The detailed view of each record is structured in a tabular view. The first tab is a viewer for positively selected amino acid sites that can be configured dynamically to match user preferences. It also allows downloading a PDF or PNG file with the result. Another tab that includes a viewer is the so called "Tree View". Using PhyD3 JavaScript library (https://phyd3.bits.vib.be/), shows a phylogram for each tree available in the record. It can be also configured and the result can be downloaded in PNG or SVG formats. Figure 3 details the full information on a specific record including several different datasets views. B+ repository is available at http://bpositive.i3s.up.pt/ and its source code is publicly available at https://github.com/sing-group/bpositive, under a GNU GPL 3.0 Open Source License (http://www.gnu.org/copyleft/gpl.html).

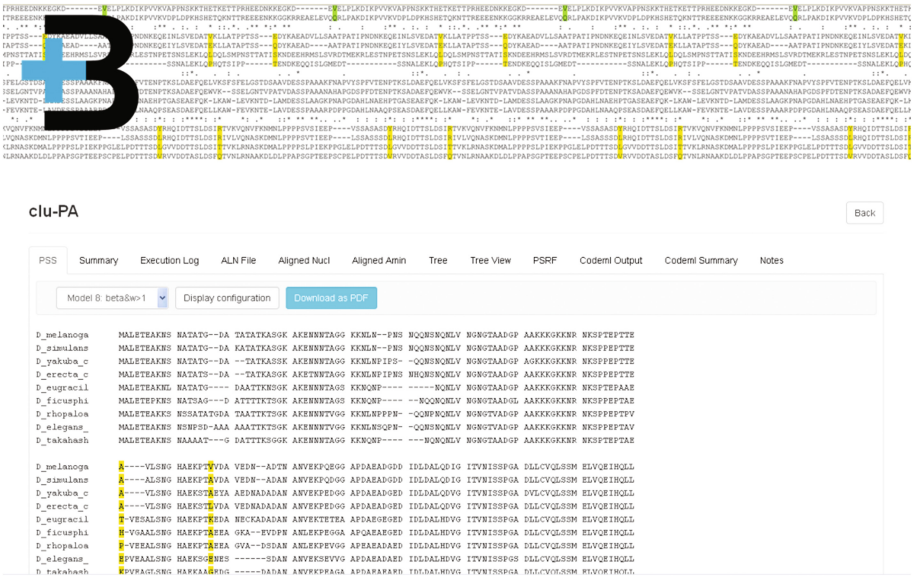


Fig. 3. Screenshot of the B+ database. All ADOPS projects tabs can be viewed in the B+ repository after selecting the gene/transcript of interest.

The first large scale data set available at B+ is the "Closely related Drosophila data set (2016)". In brief, ADOPS projects for 19652 Drosophila transcripts were generated (the details on how the sequence data was obtained and the analyses performed is provided at the B+ database under the project description), 14.6% of which show signs of positive selection (1200 genes), although human curated analyses must now be performed to validate these automatic inferences.

While ADOPS is intended to be a flexible and easy to use pipeline aimed at making robust inferences on positively selected amino acid sites, the information contained in the B+ database may serve many other purposes. For instance, since a Bayesian phylogenetic tree is always generated and the corresponding NEWICK tree file saved, a robust tree for the relationship of the species analyzed using applications such as CLANN [29] that allows the construction of supertrees from partially overlapping species datasets can be easily performed. Moreover, ADOPS projects always provide the nucleotide and protein sequences in FASTA format (aligned and non-aligned) that can be used for many other types of analyses. It should be noted that the "notes.txt" (the information is shown in the notes tab) file under the folder with the name of the ADOPS experiment is a convenient way to store plain text results obtained with additional software applications and that may help the user with the interpretation of the data.

The ADOPS projects available at B+ can be downloaded, unzipped, and opened using the ADOPS GUI. Therefore, the availability of such a database ensures the results repeatability, promotes data reuse with significant savings on the time needed for preparing datasets, and allows further exploration of the data contained in ADOPS projects effortlessly. In the new ADOPS version, there is also an option for adding new

sequences to a given project, a tool that is certainly useful when not all the sequences that a given researcher needs are contained in the original ADOPS project.

4 Conclusion

The ADOPS batch option allows running hundreds or even thousands of projects in a short period of time without human intervention. B+ is both a data repository and a convenient interface to look at the information contained in ADOPS projects. The ADOPS projects can be downloaded, unzipped, and opened using the ADOPS GUI (https://www.sing-group.org/ADOPS/). Therefore, researchers can repeat the analyses, reuse the sequence and phylogenetic trees data, and make novel analyses without losing time on input file preparation. B+ currently holds a large dataset but more will be soon available. Furthermore, the research community is welcome to contribute with other projects as well, even with small datasets. B+ will increase the repeatability of published analyses on the inference of positively selected amino acid sites, as well as making article reading more interactive.

Acknowledgements. This article is a result of the project Norte-01-0145-FEDER-000008 - Porto Neurosciences and Neurologic Disease Research Initiative at I3S, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). This work has been also funded by the "Platform of integration of intelligent techniques for analysis of biomedical information" project (TIN2013-47153-C3-3-R) from Spanish Ministry of Economy and Competitiveness. SING group thanks CITI (Centro de Investigación, Transferencia e Innovación) from University of Vigo for hosting its IT infrastructure. H. López-Fernández is supported by a post-doctoral fellowship from Xunta de Galicia.

References

1. Yang, Z.H., Nielsen, R.: Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**(4), 409–418 (1998)
2. Yang, Z.H.: PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**(5), 555–556 (1997)
3. Yang, Z.H.: PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**(8), 1586–1591 (2007)
4. Li, M., Huang, L., Li, K.Q., Huo, Y.X., Chen, C.H., Wang, J.K., Liu, J.W., Luo, Z.W., Chen, C.S., Dong, Q., et al.: Adaptive evolution of interleukin-3 (IL3), a gene associated with brain volume variation in general human populations. Hum. Genet. **135**(4), 377–392 (2016)
5. Muto, Y., Guindon, S., Umemura, T., Kohidai, L., Ueda, H.: Adaptive evolution of formyl peptide receptors in mammals. J. Mol. Evol. **80**(2), 130–141 (2015)
6. Zhang, S., Gao, B., Zhu, S.: Target-driven evolution of scorpion toxins. Sci. Rep. **5** (2015). Article No: 14973, doi:10.1038/srep14973
7. Finseth, F.R., Dong, Y.Z., Saunders, A., Fishman, L.: Duplication and adaptive evolution of a key centromeric protein in Mimulus, a genus with female meiotic drive. Mol. Biol. Evol. **32**(10), 2694–2706 (2015)

8. Zhang, Y., Yu, Z.N.: The first evidence of positive selection in peptidoglycan recognition protein (PGRP) genes of Crassostrea gigas. Fish Shellfish Immun. **34**(5), 1352–1355 (2013)
9. Jiggins, F.M., Kim, K.W.: A screen for immunity genes evolving under positive selection in Drosophila. J. Evol. Biol. **20**(3), 965–970 (2007)
10. Morales-Hojas, R., Vieira, C.P., Reis, M., Vieira, J.: Comparative analysis of five immunity-related genes reveals different levels of adaptive evolution in the virilis and melanogaster groups of Drosophila. Heredity **102**(6), 573–578 (2009)
11. Twiddy, S.S., Woelk, C.H., Holmes, E.C.: Phylogenetic evidence for adaptive evolution of dengue viruses in nature. J. Gen. Virol. **83**, 1679–1689 (2002)
12. Woelk, C.H., Holmes, E.C.: Variable immune-driven natural selection in the attachment (G) glycoprotein of respiratory syncytial virus (RSV). J. Mol. Evol. **52**(2), 182–192 (2001)
13. Woelk, C.H., Jin, L., Holmes, E.C., Brown, D.W.G.: Immune and artificial selection in the haemagglutinin (H) glycoprotein of measles virus. J. Gen. Virol. **82**, 2463–2474 (2001)
14. Shen, J., Kirk, B.D., Ma, J.P., Wang, Q.H.: Diversifying selective pressure on influenza B virus Hemagglutinin. J. Med. Virol. **81**(1), 114–124 (2009)
15. Yang, W., Bielawski, J.P., Yang, Z.H.: Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. **57**(2), 212–221 (2003)
16. Gu, M., Liu, W.J., Xu, L.J., Cao, Y.Z., Yao, C.F., Hu, S.L., Liu, X.F.: Positive selection in the hemagglutinin-neuraminidase gene of Newcastle disease virus and its effect on vaccine efficacy. Virol. J. **8**, 150 (2011)
17. Emes, R.D., Yang, Z.H.: Duplicated paralogous genes subject to positive selection in the genome of Trypanosoma brucei. Plos One **3**(5), e2295 (2008)
18. Lu, J., Zheng, J.Z., Xu, Q.G., Chen, K.P., Zhang, C.Y.: Adaptive evolution of the vertebrate skeletal muscle sodium channel. Genet. Mol. Biol. **34**(2), 323–328 (2011)
19. Khan, M.M.G., Ryden, A.M., Chowdhury, M.S., Hasan, M.A., Kazi, J.U.: Maximum likelihood analysis of mammalian p53 indicates the presence of positively selected sites and higher tumorigenic mutations in purifying sites. Gene **483**(1–2), 29–35 (2011)
20. Sobrinho, I.S., de Brito, R.A.: Evidence for positive selection in the gene fruitless in Anastrepha fruit flies. BMC Evol. Biol. **10**, 293 (2010)
21. Metzger, K.J., Thomas, M.A.: Evidence of positive selection at codon sites localized in extracellular domains of mammalian CC motif chemokine receptor proteins. BMC Evol. Biol. **10**, 139 (2010)
22. Vieira, C.P., Charlesworth, D., Vieira, J.: Evidence for rare recombination at the gametophytic self-incompatibility locus. Heredity **91**(3), 262–267 (2003)
23. Nunes, M.D.S., Santos, R.A.M., Ferreira, S.M., Vieira, J., Vieira, C.P.: Variability patterns and positively selected sites at the gametophytic self-incompatibility pollen SFB gene in a wild self-incompatible Prunus spinosa (Rosaceae) population. New Phytol. **172**(3), 577–587 (2006)
24. Vieira, J., Morales-Hojas, R., Santos, R.A.M., Vieira, C.P.: Different positively selected sites at the gametophytic self-incompatibility pistil S-RNase gene in the Solanaceae and Rosaceae (Prunus, Pyrus, and Malus). J. Mol. Evol. **65**(2), 175–185 (2007)
25. Vieira, J., Santos, R.A.M., Ferreira, S.M., Vieira, C.P.: Inferences on the number and frequency of S-pollen gene (SFB) specificities in the polyploid Prunus spinosa. Heredity **101**(4), 351–358 (2008)
26. Anisimovam, M.: Darwin and Fisher meet at biotech: on the potential of computational molecular evolution in industry. BMC Evol. Biol. **15**, 76 (2015)

27. Reboiro-Jato, D., Reboiro-Jato, M., Fdez-Riverola, F., Fonseca, N.A., Vieira, J.: On the development of a pipeline for the automatic detection of positively selected sites. Adv. Intel. Soft Comput. **154**, 225−+ (2012)
28. Nickel, G.C., Tefft, D., Adams, M.D.: Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucleic Acids Res. **36**, D800–D808 (2008)
29. Creevey, C.J., McInerney, J.O.: Clann: investigating phylogenetic information through supertree analyses. Bioinformatics **21**(3), 390–392 (2005)

ROC632: An Overview

Catarina Santos[1(✉)] and Ana Cristina Braga[2]

[1] University of Minho, Braga, Portugal
cfssantos_13@hotmail.com
[2] Algoritmi Centre, University of Minho, Braga, Portugal
acb@dps.uminho.pt

Abstract. The present paper aims to analyze and explore the ROC632 package, specifying its main characteristics and functions. More specifically, the goal of this study is the evaluation of the effectiveness of the package and its strengths and weaknesses. This package was created in order to overcome the lack of information concerning incomplete time-to-event data, adapting the 0.632+ bootstrap estimator for the evaluation of time dependent ROC curves. By applying this package to a specific dataset (DLBCLpatients), it becomes possible to assess tangible data, determining if it is able to analyze complete and incomplete data efficiently and without bias.

Keywords: ROC632 package · 0.632+ bootstrap · ROC curves

1 Introduction

The ROC632 package is a R package currently available in version 0.6. It was created by Yohann Foucher and it was first published in December 27th of 2013 [5] on CRAN repository.

This package was created in order to overcome the lack of information concerning incomplete time-to-event data (in this case, patient death [5]), adapting the 0.632+ bootstrap estimator for the evaluation of time dependent ROC curves, where the results do not depend on the incidence of the event [6].

This package allows estimation of prognostic capacity of microarray data and it relies on four main functions: *ROC, AUC, boot.ROC* and *boot.ROCt* [5,6].

The information listed above was adapted from [5,6] and further details about the features in which this package is based can be found in Sects. 1.1, 1.2 and 1.3.

1.1 ROC Curves

A ROC curve is a plot used to evaluate the relationship between sensitivity and one minus specificity (*false positive rate* (FPR)) [2,4,9].

Thus, *sensitivity*, or *true positive rate*, is the proportion of true positives (TP), i.e., the correctly classified positives divided by the true positives plus those who

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_4

should be classified as such (FN) [2–4,9,10]. In medical terms, sensitivity can also be perceived as the ability to identify diseased patients from a given sample [2].

Similarly, *specificity*, or *true negative rate*, is the proportion of correctly classified negatives (true negatives - TN) from all the expected negatives [2–4,9,10], which can also be interpreted as the capacity to disregard all healthy individuals from a given sample [2]. Specificity can be formulated as $TN/(TN + FP)$, from which we can obtain the false positive rate: $FPR = 1 - specificity$.

Accuracy, which is a key parameter for any test, [2–4,13] is the number of correctly classified objects out of all the given objects [13], i.e., the proportion of true positives and true negatives (correctly classified elements) in a sample [2–4], which implies that test accuracy is measured by sensitivity and specificity [2].

1.2 Area Under the ROC Curve

The most common and most important index able to attain the essential features of a ROC curve is the Area Under the ROC Curve (AUC) [4,9,10], which reduces it to a single scalar value [3,10]. The AUC can be computed by the trapezoidal method [3,9,10].

The value of the AUC ranges from 0 to 1, because it belongs to the *unit square* [3,4]. The ideal value for the AUC is 1 [2], meaning that every positive scored higher that every negative [4]; inversely, an AUC of 0 means that every negative scored higher than every positive and that the test has no accuracy [2,4] and thus should be discarded. Despite starting at 0, one should only consider AUC values ranging from 0.5 to 1, because the diagonal line (see Sect. 1.1) has an area of 0.5 [3].

Accuracy can also be estimated through the AUC: a test with an AUC value below 0.5 has *no accuracy*, between 0.5 and 0.7 has *low accuracy*, ranging from 0.7 to 0.9 has *moderate accuracy* and above 0.9 has *high accuracy*, emphasizing that "the greater the AUC, the better the test" [2,3].

1.3 The Bootstrap Method

Finding a method for validating predictive models and obtaining an unbiased performance has been a target of discussion by multiple authors [11], [16]. Although there are several approaches to estimate the error rate of a prediction rule, such as the jackknife (leave-one-out) method and cross-validation, the bootstrap method has been considered the most efficient throughout the years, as it is capable of directly assessing the variability, returns higher accuracy and it is able to calculate the variance of a point estimate of prediction error [1,14,15].

The *bootstrap method* separates the available data into two sets: the *training set*, which is used to obtain a predictive model and the *test set*, used to evaluate its performance [4,5,8], [16]. It draws random instances with replacement (resampling) from the original dataset, which means that some sets can be used multiple times and some might not be used at all, although, typically, eventually all the data will be selected at least once [10,14,15].

The *0.632 bootstrap estimator*, or *0.632 bootstrap resampling variable*, evaluates independent data, estimated on a per-subject basis [1,14].

However, since this method can still be biased, another estimator, the *0.632+ estimator*, was created to improve it [1,14,15].

2 Materials and Methods

This paper focused mainly on the exploration of the ROC632 package, using the *DLBCL* dataset, created by Rosenwald et al. [12]. The *DLBCLpatients* dataset and the DLBCLgenes matrix concern, respectively, 240 patients affected by a diffuse large b-cell lymphoma (DLBCL) treated with anthracycline based therapy and their clinical information, based on the scientific discoveries made by Rosenwald [12].

The assessment of the *boot.ROC* function, which builds a model relying on logistic regression with lasso penalty and deals only with complete data, consisted on varying parameters, that is, assigning different values to its arguments, namely the lasso penalty (*lambda1* – tuning parameter) and/or the number of bootstrap iterations (*N.boot*) and, in the cases where lambda was *NULL*, the fold for cross-validation (*fold.cv*). Then, the significance of the difference between the *apparent*, *cross-validation*, *0.632* and *0.632+* bootstrap curves for each condition was estimated. The study of the *boot.ROCt* function, which is capable of dealing with censored data and draws a model by applying the Cox model with lasso penalty, included the process listed for the first function and the modification of the prognostic limit for which the variable is evaluated (*pro.time* argument).

Both functions return a vector (*Coef*) with the regression coefficients obtained in the logistic or Cox model with lasso penalty (*boot.ROC* and *boot.ROCt*, respectively). These coefficients were used to determine the significant features for each result (those whose coefficient is nonzero), which could be related to the emergence of this type of lymphoma.

Since the ROC632 package does not calculate the standard error of the AUC, three functions were added to enable comparison between them. The first function calculates the *standard error* of a given curve, the second function determines the *z score* of two compared curves and the third function estimates the *p value* of the difference between the curves.

```
"st_error" = function (A) {
Q1 = A / (2 - A)
Q2 = (2 * A ^2) / (1 + A)
sqrt((A*(1-A)+(ndead-1)*(Q1-A^2)+(nalive-1)*(Q2-A^2))/(ndead*nalive))
}
"z" = function(A1, A2, se1, se2) (A1-A2)/sqrt(se1^2 + se2^2)
"pval" = function(z) {
if (z < 0) p = 2*pnorm(z,lower.tail = T )
else p = 2*pnorm(z,lower.tail = F)
return (p)
}
```

R Script 1.1. Standard Error; Z Score and P value Functions

3 Results and Discussion

3.1 Evaluation of the *boot.ROC* Function

As mentioned in Sect. 2, the *boot.ROC* function constructs a lasso penalized model for complete data according to a scoring system using logistic regression and estimates the resulting traditional four curves: apparent, cross-validation, 0.632 and 0.632+. The produced results depend on patient survival.

The boot.ROC function returns a list with 10 elements. *Coef* is a replica of *Model@penalized*, that is, the regression coefficients for the penalized co-variables (features and status). *Signature* is the score for each patient, obtained by the sum of the regression (*Coef*) multiplied by the values of the features. *Lambda* is the value of the lasso penalty and *AUC* is the mean of the area under the curve for the four estimators. This function also generates 4 important data frames for the false positive and false negative rates (obtained from the ROC function's argument *cut.values*): *ROC.Apparent* for the apparent estimator, *ROC.CV* for the cross-validation estimator, *ROC.632* for the 0.632 bootstrap estimator and *ROC.632p* for the 0.632+ bootstrap estimator (these data frames generate each point of the traditional curve). The last element is *Model*, which is a penfit object with 15 elements (see reference [8] for details).

The evaluation of this function consisted of varying the values of its arguments, specifically assessing its performance by assigning different values to the lasso penalty (*lambda1*), ranging from three fixed values – 10, 15 and 20 – to variable lambda values (*lambda1* = NULL) while increasing the number of bootstrap iterations (*N.boot*) from 2, 50 and 100 to 1000. The *precision* argument (generates the points for each curve) remained constant throughout the process, set as a vector of 4 values (0.05, 0.35, 0.65 and 0.95).

Fixed Lambda Values. It was verified that there were no significant alterations in the curves by altering the number of bootstrap iterations within the same tuning parameter (lambda). As such, only the two extreme values (2 and 1000 iterations) were included in Fig. 1. This similarity, however, was not verified for the different lasso penalties. For 2 bootstrap iterations, there were significant fluctuations between the apparent curves for a lasso penalty of 10 and 15 ($p < 0.002$) and 10 and 20 ($p < 0.0003$). For 1000 bootstrap iterations, there were also significant changes for the apparent curves between the 10 and 20 lasso penalties ($p < 0.01$).

For the three fixed lambdas, it was found that increasing the lasso penalty decreased the range of values present in the signature, that regardless of the tuning parameter more patients were assigned a negative than a positive score and that the signature was invariable with the increasing bootstrap iterations within the same lambda.

The significant features were identified by matching the unique ID in Rosenwald's NEJM_13Web_13Fig1data dataset (available at http://llmpp.nih.gov/DLBCL) to the nonzero coefficients obtained in each result. For lasso penalties of 10, 15 and 20, 42, 24 and 10 significant features were found, respectively. For the latter case, according to the NCBI database and publications [7], the 10

features are well-known overexpression transcriptional factors, genes or cell types related to the diffuse large-b-cell or other similar types of lymphoma, while the results for the other lambdas included hypothetical proteins and other non-cancer related genes. These results indicate that increasing the lasso penalty renders higher accuracy in highlighting risk factors (high influence on patient survival).

Lastly, it was ascertained that the number of bootstrap iterations had no influence in the level of fit (log likelihood) of the produced model. However, this level was directly proportional to the lasso penalty (tuning parameter of 20 produced the best results), which implies that a higher lasso penalty could be linked to a higher accuracy in determining risk factors and patient survival.

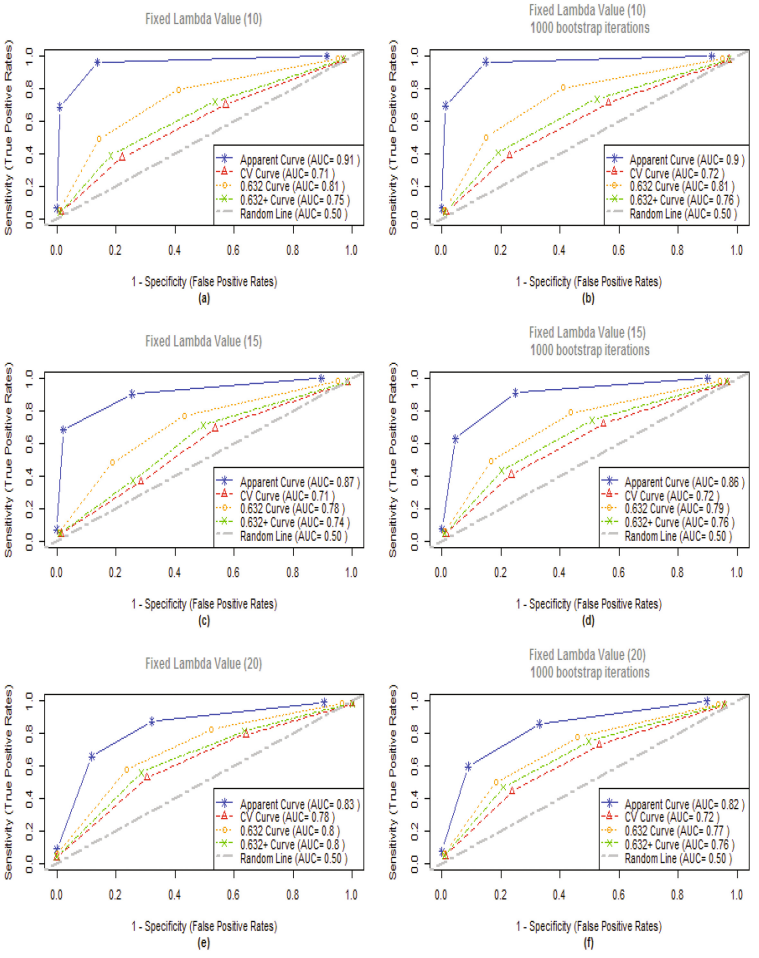


Fig. 1. ROC curves for a fixed tuning parameter of 10 (*boot.ROC* function).

Variable Lambda Values. In this subsection, the *boot.ROC* function's argument *lambda1* was set to NULL, which means that the value for the lasso penalty is generated by cross-validation by re-estimating the tuning parameter and selecting features at each bootstrap iteration. The fold for cross-validation was set to 5 (default), 10 and 20 and the number of bootstrap iterations was increased from 2 to 10.

Figure 2(a) demonstrates the results with higher test accuracy (higher AUC values) attained for the variable lasso penalty, found for the 5-fold at 2 bootstrap iterations (with a lesser overestimation of the apparent curve). Figure 2(b) shows the calculated signature for this fold, with values ranging from −1.008426 to 0.5911469 (narrower range than the observed for the fixed lambdas).

The level of fit of the model (*Log likelihood*) had increasingly negative values for the 20, 10 and 5 cross-validation folds, which implies that there is better adjustment for smaller folds.

The significant feature search for the variable lambdas yielded highly specific matches, revealing features that are all characteristic of the diffuse large b-cell lymphoma signature [7,11,12] and a better performance than the observed for the fixed lambda values. For 2 and 10 bootstrap iterations, 4 significant features were acquired, from which three are coincident with the ones found for the fixed 20 lasso penalty.

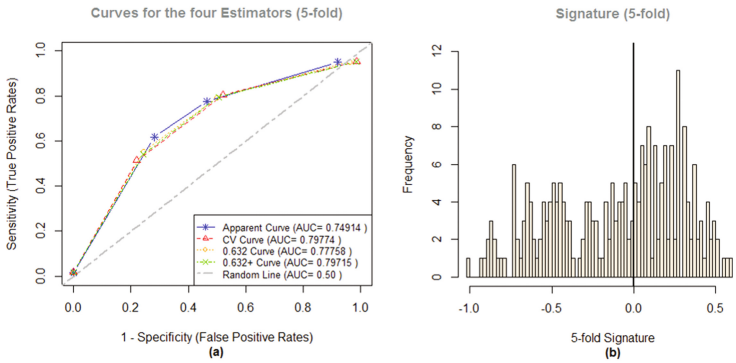


Fig. 2. Produced curves and signature for a variable tuning parameter (5-fold).

3.2 Assessment of the *boot.ROCt* Function

The *boot.ROCt* function constructs a lasso penalized model using the Cox's proportional hazards model, given a calculated signature (see Sect. 2 for details) and estimates the corresponding time-dependent curve. Since this function is capable of dealing with censored data, the argument *status* from the *boot.ROC* function is replaced by the binary argument *failure*, where 0 means right-censoring and 1 implies the event (in this case, death) took place. Right-censoring, for the newDLBCLpatients dataset, since the maximum follow up time is 21.8 years, is

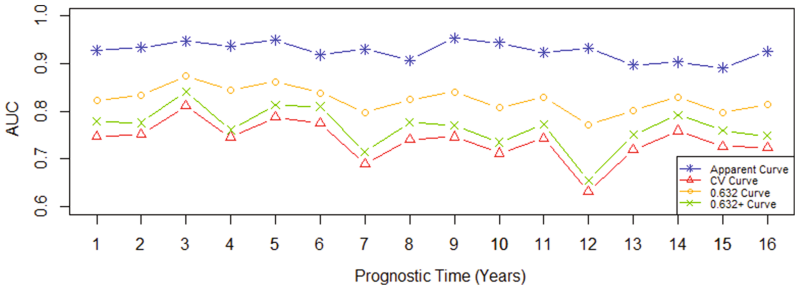


Fig. 3. Time dependent ROC curves for variable **(a)** and fixed **(b)** tuning parameters.

perceived as the patient leaving the study before that time period or surviving after it [5,6,10].

This function returns a list similar to the one described in Subsect. 3.1. The assessment of the *boot.ROCt* function included the process listed for the *boot.ROC* function, while varying the maximum prognostic time (*pro.time*) for which each variable is evaluated (from 1 to 16 years) and changing the lasso penalty (*lambda1*) from variable to fixed (15). The proportion of nearest neighbors (*prop*) was kept constant at a value of 0.02 during the course of the study and only 2 bootstrap iterations (*N.boot*) were considered.

Figure 3 illustrates the performance of the four estimators for right-censored data (*Apparent*, *Cross-Validation*, *0.632* and *0.632+*) by varying the argument *pro.time* from 1 to 16 years, according to a variable lasso penalty.

Using the functions described in Sect. 2, there were significant differences ($p < 0.05$) between the curves for fixed and variable penalties: in the apparent curve, for prognostic times of 9 and 14 years, in the cross-validation curve and in the 0.632+ bootstrap curve, for 3 and 12 years. These differences are due to the small number of observations for those specific years, leading the model to under- or overestimate those values depending on the used approach.

The most negative log likelihood (−812.4193) and lowest number of significant features (which ranged from 4 to 32) for the variable penalty were found for a prognostic time of eleven years, which means that those particular models were the most accurate in predicting patient outcome. For the fixed lambda value of 15, independently of the prognostic time, a log likelihood of −739.6779 was achieved and 68 relevant genes were found, indicating low reliability.

4 Conclusions

From the results described throughout Sect. 3, the strengths and weaknesses of the ROC632 package could be highlighted and, independently of the condition or the function in use, some patterns in the outcomes were identified.

The most significant disadvantages of this package are that it does not calculate the standard error for the estimated curves, forcing the user to calculate

it using an additional method and that the number of patients assigned to the training and test sets is not explicitly shown in the results.

The apparent curve seemed to be overly optimistic and the cross-validation curve considerably pessimistic. These results were expected, since these curves only represent the training and test sets, respectively. The 0.632 and 0.632+ bootstrap curves had an overall similar performance, with marginally lower values for latter, since the 0.632+ estimator's performance depends on the amount of overfitting, whereas the former has a constant weight [1,14,15]. Hence, the 0.632+ estimator provided the best results with the least variance (similar results within the same condition – fixed or variable lambda) and bias (no over- or underestimation) and it should thus be used in future analysis.

Finally, the signature created was able to create an overall efficient prognosis for up to 10 years, being capable of attributing a higher score to patients who survived the longest (and were still alive). However, since most patients had died after that time point (only 26 patients had survived), the predictions for longer timeframes were considerably erroneous, with patients who had survived having the same score as patients who did not. Hence, although this scoring system could be highly accurate for well documented data, with as many observations possible, it isn't advised for small datasets where the data is overly repetitive in some cases and missing in others and the number of examples is limited.

References

1. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Nat. Acad. Sci. **99**(10), 6562–6566 (2002)
2. Collinson, P.: Of bombers, radiologists and cardiologists: time to ROC. Heart **80**, 215–217 (1998)
3. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
4. Flack, P.: ROC analysis. In: Encyclopedia of Machine Learning, 1 edn., pp. 869–874. Springer (2011)
5. Foucher, Y.: ROC632: construction of diagnostic or prognostic scoring system and internal validation of its discriminative capacities based on ROC curve and 0.633+ bootstrap resampling, R package version 0.6 (2013). https://cran.r-project.org/web/packages/ROC632/index.html
6. Foucher, Y., Danger, R.: Time dependent ROC curves for the estimation of true prognostic capacity of microarray data. Stat. Appl. Genet. Mol. Biol. **11**(6), 1 (2012)
7. Geer, L.Y., Marchler-Bauer, A., Geer, R.C., et al.: The NCBI BioSystems database. Nucleic Acids Res. **38**(Database), D492–D496 (2009)
8. Goeman, J., Meijer, R., Chaturvedi, N.: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalized Estimation, R package version 0.9-47 (2013). https://cran.r-project.org/web/packages/penalized/index.html
9. Gonen, M.: Receiver Operating Characteristic (ROC) Curves (Paper 210-31). SUGI 31 Proceedings, pp. 1–18. SAS Institute Inc. (2006)
10. Krzanowski, W.J., Hand, D.J.: ROC Curves for Continuous Data. CRC Press, Boca Raton (2009)

11. Liu, H., Li, J., Wong, L.: Use of extreme patient samples for outcome prediction from gene expression data. Bioinformatics **21**(16), 3377–3384 (2005)
12. Rosenwald, A., Wright, G., Chan, W.C., et al.: The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. N. Engl. J. Med. **346**(25), 1937–1947 (2002)
13. Sammut, C., Webb, G.: Encyclopedia of Machine Learning. Springer (2011)
14. Steyerberg, E.W., Harrell, F.E., Borsboom, G.J., et al.: Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. **54**(8), 774–781 (2001)
15. Vu, T., Sima, C., Braga-Neto, U.M., Dougherty, E.R.: Unbiased bootstrap error estimation for linear discriminant analysis. J. Bioinf. Syst. Biol. **2014**(1), 1–15 (2014)

Processing 2D Gel Electrophoresis Images for Efficient Gaussian Mixture Modeling

Michal Marczyk(✉)

Data Mining Group, Institute of Automatic Control, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
michal.marczyk@polsl.pl

Abstract. In modern molecular biology the most commonly used method to distinct proteins present in complex sample is two-dimensional gel electrophoresis. Unfortunately, the quality of the gel image is reduced by the presence of non-linear background signal, spikes, streaks and other artefacts. The main components of gel image are protein spots. To properly distinguish spots, mostly in overlapping regions, mixture modeling can be performed. Due to many signal impurities the estimation of model parameters is inadequate. In this study, by using two fragments of real gel image and a set of synthetic data, three background correction methods with four image filtering methods were collated and the quality of spot detection based on mixture modeling was checked. The presented results prove that efficient modeling of 2D gel electrophoresis images must be preceded by proper background correction and noise filtering. A two-step Otsu algorithm was the best method for removing background signal. There was no single favorite from filtering methods, but using 2D matched filtering leads to good results despite the background correction method used.

Keywords: 2D gel electrophoresis · Image filtering · Mixture modeling

1 Introduction

Proteomics is a branch of science that attempts to characterize proteins, compare variations in their expression levels between phenotypes, study their interactions with other proteins and identify their functional roles. 2D gel electrophoresis (2DGE) is a measurement technique commonly used in proteomics for separation of proteins in a complex sample, finding post-translational modifications or discovering protein biomarkers by analyzing series of samples [1]. The basic principle of the technique is to separate proteins based on their weight and pH gradient. A result of measuring single sample is a grayscale image with light background and dark spots. The intensity of each spot represents the amount of the protein in the analyzed sample. In a single gel there may be even thousands of protein spots visible, so the automatic analysis of gel image is necessary. Despite that 2DGE is a very popular technique, there are evident drawbacks. The overall quality of the gel image suffers due to artefacts, inhomogeneous background

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_5

and high level of noise. Also, the proteins with similar molecular properties can form streaks or clusters in which spots are hardly identifiable.

Methods for detecting spots in gel image can be grouped into three categories: local maximum methods, segmentation methods and model-based approaches. Among the last group of methods the interesting technique is mixture modeling that enables detecting spots that are hidden in the clusters of overlapping spots [2, 3]. Fitting the mixture model allows for achieving higher sensitivity in detecting spots and better overall performance of the spot detection than using local maximum or segmentation methods [2]. Signal modeling is robust to small pixel intensity variations, thus providing consistent and reliable spot quantity estimates [3]. A single spot shape is the most commonly approximated by Gaussian normal distribution [4]. A serious problem in mixture modeling is that real data are often contaminated, so the estimation of the model parameters for proper modeling may be unsuccessful. To correctly identify true spots and prevent the identification of artefacts as spots using Gaussian mixture modeling, the signal processing must be performed at first.

The background signal in gel image is not uniform, but consists of local regions of elevated pixel-intensities. The simplest method is to find global or local minima in the image and set a constant threshold value e.g. based on image intensity histogram [5]. Other approach is to approximate the background using a polynomial function. Iterative polynomial fitting preserves too elevated estimate of the background surface [6]. The commonly used method, called the 'rolling ball' algorithm, uses a circular disc as a structural element in the morphological opening operation [7]. It is a very efficient solution, but it fails to remove streaks from the image. Rapid, random changes in the intensity of neighboring image pixels and random artefacts can be reduced by the process called filtering. In conventional filters used in 2DGE a window of predefined size is placed at each pixel in the image and the value of this pixel is then determined by some relationship or function with respect to the surrounding pixels defined by the window [8, 9]. Adaptive filters are constructed to smooth images similar to linear filters, but at the same time preserve significant discontinuities. A more sophisticated denoising methods are based on the wavelet transformation [10].

The aim of this paper is to prove that proper background correction and noise filtering can improve the quality of detection of protein spots in 2DGE when Gaussian mixture modeling is used. Two fragments of real dataset and a set of synthetic images were examined. Three background correction methods were collated with four image filtering methods and the sensitivity plus the false discovery rate (FDR) of spot detection were calculated. Also, goodness of fit of the mixture model was checked.

2 Materials and Methods

2.1 Data

There are no properly annotated 2D gel images of full human proteome freely available. Some efforts were done by Raman et al. [11], but the annotation data are stored as a low-quality PNG image, so it is hard to distinguish spots in the overlapping regions. Thus two fragments (Fig. 1) of real gel image given by Raman et al. [11] were chosen and the

existing annotation of true spots was improved by manual inspection of the image. First fragment (size – 93 × 91 pixels) is called R1 in the following text and it contains 51 annotated spots. In R1 the background signal has low intensity compared to spots intensity and there are no streaks visible. Second fragment (size – 97 × 118 pixels) is called R2 in the following text and it contains 44 true spots. In R2 the background signal has higher intensity than in R1 and the streaks are observed.

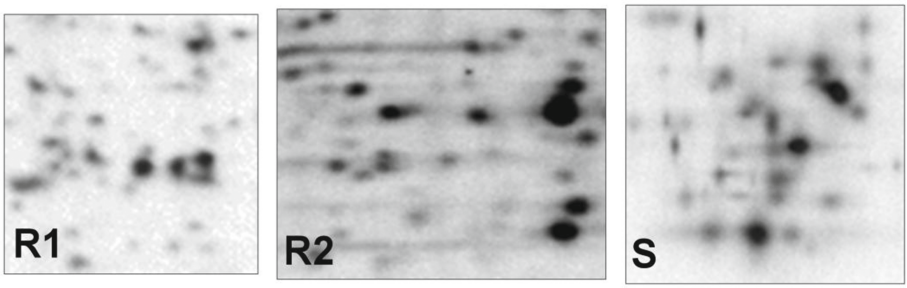


Fig. 1. Two fragments of real 2D gel image (R1 and R2) and an example of synthetic data (S).

The synthetic data were created based on the models of 2DGE image background, additive noise, streaks and true spots, proposed in [2]. Background and noise model parameters were estimated using the real gel image from Raman et al. [11]. A spot model is based on diffusion principles observed in 2D gel electrophoresis [4]. The spread of each spot is calculated based on its intensity. Synthetic dataset is called S in the following text. It contains 100 gel images (size – 100 × 100) with 50 spots each. An example of the synthetic gel image is presented in Fig. 1.

2.2 Processing Methods

Three methods for removing background signal were implemented and compared. Iterative polynomial fitting (IPF) [6] is performed for each line separately in horizontal and vertical direction. First, a polynomial curve is fitted to the signal values and the signal above the curve is removed. Next, a new approximation is performed and the method is repeated until convergence. The background is a mean value from all fits. A polynomial degree of 4 is used in this study. In rolling ball algorithm (RB) [7] a circular disc with size larger than the largest spot in the image is used as a structural element in the operation of morphological opening. A disc radius equal to 15 was used for processing all images within the study. Two-step Otsu thresholding (OTSU) [5] searches for the optimal separation of the histogram into two groups based on statistical measures of between and within variances. The threshold is identified first by searching for the overall threshold based on the full histogram and then the thresholding procedure is applied only to the intensity values lower than the first threshold.

Four methods for noise filtering were implemented and compared. In median filtering (MEDF) [8] a window of predefined size is placed at each pixel in the image and the signal value is averaged by calculating median intensity of the neighborhood pixels. The

size of the moving window is varied from 3 × 3 to 9 × 9. In median modified Wiener filter (MMWF) [9] a window of predefined size is placed at each pixel in the image and a local kernel median around each pixel is calculated, that includes local image variance, to estimate new value. The size of the moving window is varied from 3 × 3 to 9 × 9. In 2D matched filtering (2DMF) histogram of image intensity is first divided into overlapping fragments using Otsu thresholding. Then, in each fragment 2D matched filtering using Gaussian function is performed. The size and the standard deviation of Gaussian function is found adaptively by using Peak signal-to-noise ratio measure. The number of fragments was varied from 2 to 20 and the overlap from 0 to 20%. 2D Wavelet denoising (WD) [10] is performed in the wavelet domain and ensures efficient elimination of noise, without deteriorating the significant high frequency features. Decomposition was made using the two-dimensional wavelet basis function and the soft-thresholding policy was applied. A MATLAB implementation of all processing methods and all datasets analyzed in the study are available to freely download along with the 2DGMMgel software from the following website: http://zaed.aei.polsl.pl/index.php/pl/oprogramowanie-zaed.

2.3 Gaussian Mixture Modeling

In 2DGE the most commonly used spot model is a Gaussian function. This choice is inspired by the 3D shape of the spot and by general considerations on diffusion processes following gel image creation. To efficiently estimate parameters of several dozen components of the model a modified version of expectation-maximization (EM) algorithm was proposed [2]. Initial conditions for EM were set to true peak positions and spread. To provide regularization of the model Bayesian information criterion [12] was used for estimating the final number of model components in backward elimination scheme. An example of background correction, filtering and Gaussian mixture modeling of real gel image R2 is presented in Fig. 2.

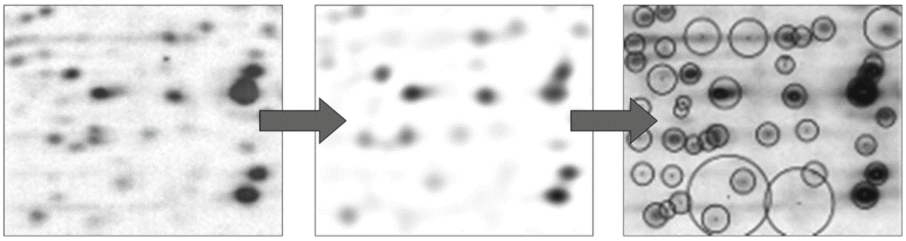


Fig. 2. Result of background correction (left), noise filtering (middle) and Gaussian mixture modeling (right) using IPF and 2DMF methods on R2 dataset.

3 Results and Discussion

Three methods for background correction and four methods for image filtering were collated to check if they can improve the results of Gaussian mixture modeling of gel

image. Also, the performance indices for the case where no method was used, either to correct background or remove noise, were calculated. In total, twenty different pre-processing scenarios were competed within the study.

3.1 Spot Detection Performance

Comparison of different methods using real and synthetic datasets was provided by calculating three measures of spot detection quality. Sensitivity is the number of true spots detected divided by the number of all true spots in the image. False discovery rate (FDR) is the number of detected spots that do not correspond to true spots divided by the number of all detected spots. F1 is the harmonic mean of 1-FDR and sensitivity. Higher values of F1 score shows better performance of the method. F1 is also used to optimize parameters of all methods and the best results are presented.

Table 1. Results of spot detection in two fragments of real image (R1 and R2). First column – background correction method used, second column – filtering method used.

Back. corr.	Filtering	F1		Sensitivity		FDR	
		R1	R2	R1	R2	R1	R2
None	None	0,760	0,674	0,745	0,659	0,224	0,310
IPF		0,720	0,795	0,706	0,795	0,265	0,205
RB		0,792	0,750	0,784	0,750	0,200	0,250
OTSU		**0,990**	0,884	**0,980**	0,864	**0,000**	0,095
None	MEDF	0,720	0,674	0,706	0,659	0,265	0,310
IPF		0,720	0,828	0,706	0,818	0,265	0,163
RB		0,760	0,805	0,745	0,795	0,224	0,186
OTSU		0,980	0,847	0,961	0,818	**0,000**	0,122
None	MMWF	0,720	0,674	0,706	0,659	0,265	0,310
IPF		0,700	0,805	0,686	0,795	0,286	0,186
RB		0,760	0,800	0,745	0,773	0,224	0,171
OTSU		0,960	0,847	0,941	0,818	0,020	0,122
None	2DMF	0,887	0,837	0,843	0,818	0,065	0,143
IPF		0,920	0,955	0,902	**0,955**	0,061	0,045
RB		0,878	0,864	0,843	0,864	0,085	0,136
OTSU		0,926	**0,966**	0,863	**0,955**	**0,000**	**0,023**
None	WD	0,840	0,674	0,824	0,659	0,143	0,310
IPF		0,780	0,828	0,765	0,818	0,204	0,163
RB		0,812	0,805	0,804	0,795	0,180	0,186
Otsu		**0,990**	0,874	**0,980**	0,864	**0,000**	0,116

In R1 the number of estimated spot locations ranged from 44 to 50 (the number of true spots equals to 51), while in R2 from 41 to 44 (the number of true spots equals to 44). Results of spot detection in real images are presented in Table 1. The maximum value of F1 in R1 dataset was achieved after using OTSU method connected with no filtering or filtering by the wavelet decomposition. The maximum sensitivity was obtained for the same pairs of methods. Four combinations of methods gave no false positives in R1, namely OTSU + None, OTSU + MEDF, OTSU + 2DMF and OTSU + WD. In R2 the best overall performance was obtained after using OTSU

method connected with 2DMF filtering. The maximum sensitivity was given for IPF + 2DMF and OTSU + 2DMF. Only one spot was falsely detected after using OTSU + 2DMF combination. For each method an average F1 score can be calculated to show its robustness. In such comparison, from background correction methods, OTSU gave the highest average F1 score and from filtering methods, 2DMF gave the best result. These outcomes are the same for R1 and R2 datasets.

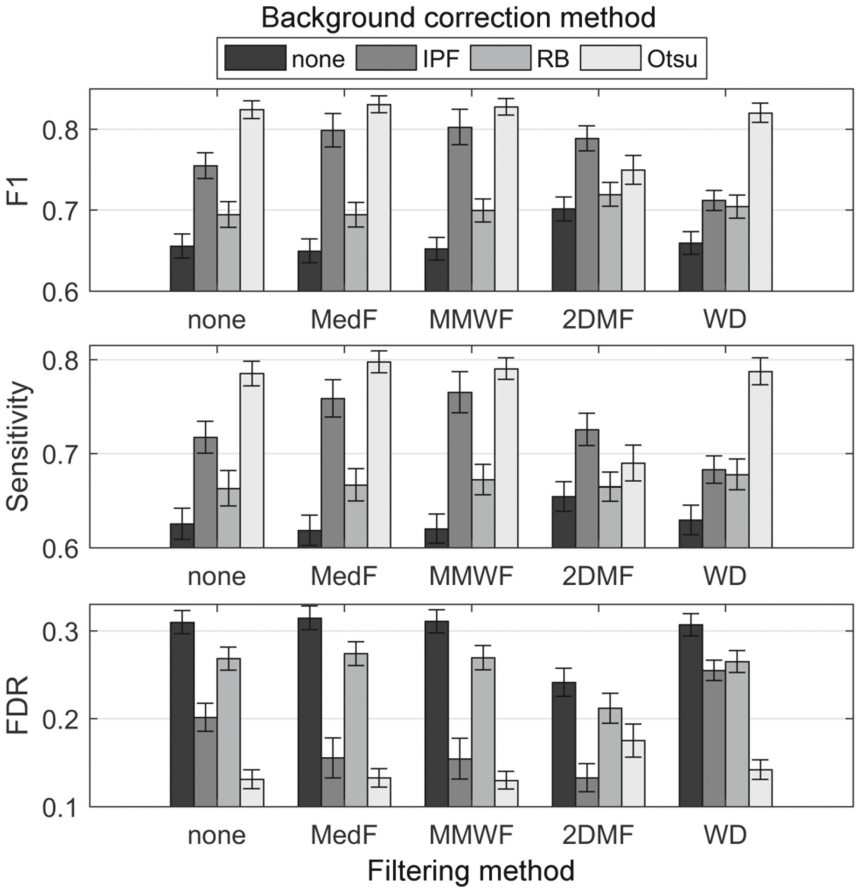


Fig. 3. Results of spot detection in synthetic data. The error bars indicate 95% confidence intervals.

By generating 100 synthetic images it was possible to calculate the indices of spot detection quality with 95% confidence intervals (Fig. 3). The maximum average F1 score was obtained for OTSU background correction and MEDF filtering, but the result is statistically not better than the one obtained by the following pairs of methods: OTSU + None, OTSU + MMWF, OTSU + WD. The highest sensitivity was obtained for the same methods. The lowest FDR was acquired after using the following methods: OTSU + None, OTSU + MMWF, IPF + 2DMF and OTSU + WD. From background

correction methods using OTSU leads to the highest average F1 score. From filtering methods MMWF gave the highest F1 score. 2DMF leads to lower FDR despite the background correction method used, but also lower sensitivity, in comparison to other filtering methods.

3.2 Goodness of Model Fit

Bayesian information criterion [12] was used to measure how well a mixture model is fitted to data after image processing (Fig. 4). The index takes into account both the statistical goodness of fit and the number of parameters that have to be estimated by imposing a penalty for increasing the number of parameters. The model with the lowest BIC is preferred. The best fit was obtained after using OTSU background correction and 2DMF filtering. Among background removal methods, using OTSU gave the best results. Comparing filtering methods, using 2DMF gave the best fit. The result is not surprising, since 2DMF uses the same Gaussian shape to perform image filtering as the probability distribution function used in the mixture model.

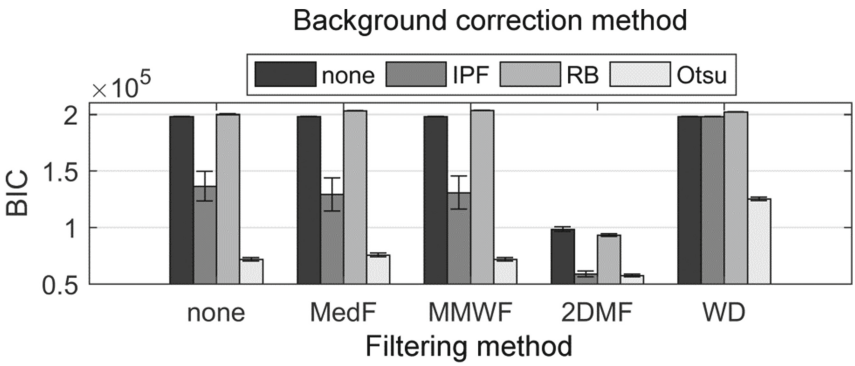


Fig. 4. Goodness of model fit to synthetic images after different processing methods.

4 Conclusions

Using two fragments of real gel image and artificially created dataset it was shown that background correction and noise filtering are necessary to improve the quality of protein spots detection in 2DGE when using Gaussian mixture modeling. A background correction method based on global OTSU thresholding gave the best results in all analyses, but since the comparison was performed on fragments of gel images, results for whole gel image may be different. The most universal method for image filtering, that also guarantees the best model fit and low FDR level, was 2DMF. According to an overall performance of spot detection there was no single favorite from compared filtering methods.

Acknowledgments. This work was financially supported by internal grant of Silesian University of Technology 02/010/BKM16/0047/33 and partially by BKM17 grant. All calculations were carried out using GeCONiI infrastructure (POIG.02.03.01-24-099/13).

References

1. Magdeldin, S., Enany, S., Yoshida, Y., Xu, B., Zhang, Y., Zureena, Z., Lokamani, I., Yaoita, E., Yamamoto, T.: Basics and recent advances of two dimensional- polyacrylamide gel electrophoresis. Clin. Proteomics **11**, 16 (2014)
2. Marczyk, M.: Mixture modeling of 2D Gel Electrophoresis spots enhances the performance of spot detection. IEEE Trans. Nanobiosci. **16**, 91–99 (2017)
3. Tsakanikas, P., Manolakos, E.S.: Protein spot detection and quantification in 2-DE gel images using machine-learning methods. Proteomics **11**, 2038–2050 (2011)
4. Rogers, M., Graham, J., Tonge, R.P.: Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images. Proteomics **3**, 887–896 (2003)
5. Peer, P., Corzo, L.G.: Local pixel value collection algorithm for spot segmentation in two-dimensional gel electrophoresis research. Comp. Func. Genom. (2007). Article ID 89596
6. Faergestad, E.M., Rye, M., Walczak, B., Gidskehaug, L., Wold, J.P., Grove, H., Jia, X., Hollung, K., Indahl, U.G., Westad, F., van den Berg, F., Martens, H.: Pixel-based analysis of multiple images for the identification of changes: a novel approach applied to unravel proteome patterns of 2-D electrophoresis gel images. Proteomics **7**, 3450–3461 (2007)
7. Rye, M.B., Faergestad, E.M., Martens, H., Wold, J.P., Alsberg, B.K.: An improved pixel-based approach for analyzing images in two-dimensional gel electrophoresis. Electrophoresis **29**, 1382–1393 (2008)
8. Kaczmarek, K., Walczak, B., de Jong, S., Vandeginste, B.G.: Preprocessing of two-dimensional gel electrophoresis images. Proteomics **4**, 2377–2389 (2004)
9. Cannistraci, C.V., Montevecchi, F.M., Alessio, M.: Median-modified Wiener filter provides efficient denoising, preserving spot edge and morphology in 2-DE image processing. Proteomics **9**, 4908–4919 (2009)
10. Daszykowski, M., Stanimirova, I., Bodzon-Kulakowska, A., Silberring, J., Lubec, G., Walczak, B.: Start-to-end processing of two-dimensional gel electrophoretic images. J. Chromatogr. A **1158**, 306–317 (2007)
11. Raman, B., Cheung, A., Marten, M.R.: Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. Electrophoresis **23**, 2194–2202 (2002)
12. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)

Improving Document Prioritization for Protein-Protein Interaction Extraction Using Shallow Linguistics and Word Embeddings

Sérgio Matos(✉)

DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
aleixomatos@ua.pt

Abstract. Understanding of biological processes, associated to disease or pharmacological action for example, requires the analysis of large amounts of interconnected information. Protein interaction networks form part of this puzzle, and extracting this information from the scientific literature is an important but challenging task.

In this work, we present a supervised classification approach for identifying and ranking literature documents that contain information regarding protein interactions. We studied the use of word embedding together with simple chunking features, and show that the combination of these features with baseline bag-of-words can lead to similar or even improved results when compared to the use of features based on deep linguistic parsing. When applied to the BioCreative III Article Classification Task dataset, our approach achieves an area under the precision-recall curve of 0.70 and a Matthew's correlation coefficient of 0.56.

Keywords: Protein-protein interactions · Literature retrieval · Machine learning · Word embeddings

1 Introduction

The identification of protein-protein interactions (PPIs) is of utmost importance for biomedicine, since the understanding of disease, pharmacological and other processes requires the analysis of networks formed by these relations. Several databases maintain manually curated protein-protein interaction data but, since the primary source for identifying PPIs is the scientific literature, keeping these databases up-to-date is a demanding and expensive task. Therefore, the use of named-entity recognition (NER) and relation extraction methods in assisted curation workflows has been evaluated, and shown to expedite this work [6,14]. Even when information extraction methods is not applied, document prioritization or triage is a required step, in order to obtain articles that have more likelihood of containing information.

Several works have addressed the problem of document prioritization for protein-protein interactions. Sumoela and Andrade [13] proposed a classification and ranking model to evaluate the entire MEDLINE database, the largest

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_6

repository of scientific literature in the life sciences, with respect to any topic of interest. Their method is based on selecting words that commonly convey meaning, namely nouns, verbs, and adjectives, and relies on the different frequencies of these discriminating words between a set of relevant articles and a reference set. This approach is also behind the MedlineRanker web-service [4], which allows to retrieve a list of articles ranked by similarity to a training set defined by the user. One possibility, as referred by the authors, is to use a list of document identifiers obtained from a PPI database, therefore getting as result other articles related to that same topic. Marcotte et al. [9] proposed a log likelihood scoring function to identify articles discussing PPIs, using a feature set composed of 83 discriminating words selected from a training set of 260 MEDLINE abstracts involving yeast proteins. They reported an accuracy of 77%, with a recall around 55%, when articles with a log likelihood score of 5 or higher were selected.

Retrieval and extraction of PPI related information has been a major focus of recent shared evaluations in the biomedical domain, namely in the BioCreative challenges. Lan et al. [8] compared the use of Bag-of-Words (BoW), interaction trigger words and protein Named Entities (NEs) features in a Support Vector Machine (SVM) classifier, applied to the BioCreative-II PPI task data. Their best result, when using a single classifier, was obtained with a feature set containing BoW features and protein NEs co-occurring with interaction trigger words (F-score of 77%). Abi-Haidar et al. [1] tested three classifiers in the same data set: SVM, Variable Trigonometric Threshold classifier (VTT), and a nearest neighbor classifier with singular value decomposition (SVD) applied for feature selection. They reported a top F-score of 78% using the VTT classifier with a feature set of 650 discriminating words. The latest PPI Article Classification Task (ACT), part of the BioCreative III Challenge (BC-III), counted with 52 submissions from ten participating teams [7]. Most teams applied some sort of machine learning technique, the best results being obtained using Support Vector Machines, Maximum Entropy or Large Margin classifiers. The top performing teams used various levels of lexical analysis, including Part-of-Speech (PoS) tagging and Named Entity Recognition (NER), and the best team overall also used dependency parsing to extract the textual features used for classification. Additionally, various teams used the manually assigned MeSH terms, which are indexing terms that provide information regarding the article's subject. The best AUC iP/R (area under the interpolated precision-recall curve) was 0.680 and the highest MCC (Matthew's correlation coefficient) was 0.553, with an accuracy of 89.2% and an F-score of 61.4% [5]. This lower result, as compared to results obtained on the BC II dataset, reflects the highly unbalanced nature of the test set used (15% positive documents), which represents more closely the real scenario.

In this work, we evaluated the use of word embedding features, together with simple chunking features, for prioritization of MEDLINE articles containing protein-protein interaction information. We follow a classification based ranking approach, in which the class probabilities obtained from the classifier are used for ranking the results, and show that classifiers trained with such distributional

and shallow parsing features can achieve state-of-the-art results, without the use of dependency parsing or indexing terms. The paper is organized as follows: the next section describes the methods and data used, followed by the presentation of results, and finally the conclusions.

2 Methods

This section describes the data and methods used. Text processing and classification tasks were implemented in Python, using the Scikit-learn machine-learning library scikit-learn and the Natural Language Toolkit [2]. Neji [3], a framework for biomedical concept recognition, was used for identifying protein mentions and for writing the annotated documents in CoNLL format. Word embedding models were obtained with the gensim framework [11].

2.1 Data

We used the dataset from the BioCreative III protein-protein interaction, article classification task (ACT) [7]. This corpus is composed of manually annotated MEDLINE abstracts, containing 2280 documents in the training set, 4000 in the development set, and 6000 in the test set. The training set has the same number of positive and negative examples, while the development and test sets are highly unbalanced, with around 15–17% positive examples, which reflects the expected real scenario.

2.2 Feature Extraction

We compared the use of various features, starting with the common token n-grams, with n varying between 1 and 5. These bag-of-words (BoW) features encode common words or phrases used to express protein interactions, for example 'interacts with', 'affects', or 'binds'. The different frequencies of these words in PPI articles versus non-PPI articles is sufficient to learn a classifier with reasonable results. We applied simple tokenization and stopwords filtering, using the MEDLINE stop list, before extracting the n-grams.

Protein-protein interactions, and relations in general, are commonly referenced by particular linguistic constructions. To model this linguistic information, we chose to use chunking instead of deeper linguistic parsing, to evaluate if a simpler and less computationally demanding approach could lead to good results. To extract the patterns, we first identified protein mentions using Neji[1], and encoded the results in CoNLL format, including the chunking information. Protein mentions were identified through a conditional-random fields (CRF) model trained on the BioCreative II gene mention recognition corpus [12]. We then iterated through the chunks to obtain the sequence of tags and words representing each sentence. In this process, we applied the following transformations: (a) if a

[1] Available from https://github.com/BMDSoftware/neji.

protein entity was matched for any token in a noun-phrase (NP), the tag would be replaced by a placeholder PTN, otherwise the NP tag would be used; (b) for verb-phrases, we replaced each word in the phrase by its lemma, using NLTK's WordNet lemmatizer, and produced the sequence of lemmas, rather than the tag VP. Using lemmas allows combining different lexical variations in the same grammatical pattern, leading to better generalization, while keeping linguistic information regarding the verbs used in the sentences. We then extracted n-gram features from these sequences. For example, the sentence *"Here, we show that Schizosaccharomyces pombe CHD remodellers, the Hrp1 and Hrp3 paralogs physically interact with the histone chaperone Nap1."* (PMID 17510629), can be represented as

```
Here, [we]/NP [show]/VP that [Schizosaccharomyces pombe CHD
remodellers]/NP, [the Hrp1/PTN and Hrp3/PTN paralogs]/NP
physically [interact]/VP with [the histone/PTN chaperone/PTN
Nap1/PTN]/NP.
```

where NP represents a noun-phrase, VP a verb-phrase, and PTN identifies tokens that were recognized as protein entities. The corresponding pattern extracted for this sentence is `NP show NP PTN interact PTN`.

We also evaluated the use of word embedding vectors as global features for the classification. We used gensim's implementation of word2vec [10] to create a model for 12 million abstracts from MEDLINE for the years 2000 to 2015, containing around 600 thousand distinct words. We used a window of 100 and a vector size of 500 for these experiments.

When using embedding vectors for classification, it is usual to sum or average the vectors of the words that appear in the text. This however does not take into consideration prior information available in a supervised setting, as in this case. We therefore used a weighted average of the word vectors, using as weights the coefficients of a linear regression between the term-document matrix and the label of each document in the training set. Preliminary cross-validation results on the training set showed that this weighting produced better results than the unweighted average, or the sum of vectors. Similarly, preliminary validation also showed that the best results were obtained by representing the words in the term-document matrix by their term-frequency inverse-document-frequency (tf-ifd) weight, rather than raw counts, log frequency, or idf alone.

We compared different weighted combinations of these features, which were normalized to unit norm before being encoded on a single feature vector.

2.3 Document Classification

Different classifiers were considered, namely logistic regression (LR), passive-aggressive (PA), Ridge classifier (RC), and linear support-vector machine with stochastic gradient descent (SGD) learning. We applied grid search through cross-validation on the training set, to select the n-grams to use for the BoW and chunking features and to select the best combination of feature weights.

3 Results

Table 1 shows the 5-fold cross validation results on the training set, comparing the best parameters identified for each classifier. The best results for each classifier were obtained by combining the three types of features, with a weight of 1.0 for the BoW features, and varying between 0.5 and 0.75 for the chunking and WE features. In terms of n-grams, the results were mixed. The LR classifier

Table 1. 5-fold cross validation accuracy on the training set. BoW: bag-of-word features; NLP: NER and chunking features; WE: word embedding features.

Features	Classifier			
	PA	LR	RC	SGD
BoW	0.860	0.856	0.861	0.857
BoW+NLP	0.870	0.864	0.867	0.870
BoW+WE	0.867	0.865	0.869	0.868
NLP+WE	0.864	0.867	0.867	0.870
BoW+NLP+WE	**0.873**	**0.871**	**0.874**	**0.875**

Table 2. Evaluation results on the test set. BoW: bag-of-word features; NLP: NER and chunking features; WE: word embedding features. AUC: Area under the curve; Acc: Accuracy; MCC: Matthew's correlation coefficient; P@Full R: Precision at full recall.

Model	Features	Metrics			
		AUC	Acc.	MCC	P @ Full R
PA	BoW	0.664	0.872	0.533	0.157
	BoW+NLP	0.687	0.879	0.549	0.157
	BoW+NLP+WE	0.689	0.884	0.558	**0.183**
		+0.025	+0.012	+0.025	+0.026
LR	BoW	0.642	0.881	0.512	0.161
	BoW+NLP	0.677	0.884	0.532	0.159
	BoW+NLP+WE	**0.700**	**0.885**	**0.562**	0.165
		+0.058	+0.004	+0.050	+0.004
RC	BoW	0.660	0.871	0.518	0.156
	BoW+NLP	0.684	0.882	0.556	0.154
	BoW+NLP+WE	0.691	0.881	0.560	0.162
		+0.031	+0.010	+0.042	+0.006
SGD	BoW	0.643	0.886	0.493	0.153
	BoW+NLP	0.674	0.891	0.533	0.152
	BoW+NLP+WE	0.686	0.883	0.540	0.163
		+0.043	−0.003	+0.047	+0.010

produced better results with only 1-gram BoW features and 2-gram chunking features, and achieved similar results with 1-gram for both types of features, while for the remaining classifiers better results were obtained with 3- or 4-gram chunking features and 2- or 3-gram BoW features.

Table 2 shows the results obtained on the test set of the BioCreative III article classification task, illustrating the improvements provided by the chunking and word embedding features. Comparing to the current state-of-the-art results, reported in the official BioCreative task, our results show an improvement in terms of area under the precision-recall curve (0.700 vs. 0.680) and of Matthew's correlation coefficient (0.562 vs. 0.551), without the use of dependency parsing or document indexing features.

4 Conclusions

We present results for the prioritization of scientific articles containing information regarding protein-protein interactions, following a classification based ranking approach. We evaluated the use of simple patterns extracted from shallow linguistic parsing (chunking), together with results from named entity recognition, and the use of word embedding features. Our results show that both feature types improve the classification and ranking performance over the BoW baseline features, and over current state-of-the-art results, which rely on linguistic parsing features.

A current limitation of this work is that we have not considered full texts of the articles, where most PPI information is available, and which could improve the classification performance. Also, MEDLINE abstracts are manually indexed with terms from the MeSH vocabulary and previous works have used this information, showing that these terms are informative regarding PPI document prioritization. Although we expect that our ranking results could also be improved with this information, not requiring it allows the classifier to be effectively applied to recently published articles that have not yet been indexed.

Acknowledgments. This work was supported by Portuguese National Funds through FCT - Foundation for Science and Technology, in the context of the project IF/01694/2013. Sérgio Matos is funded under the FCT Investigator programme.

References

1. Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Rechtsteiner, A., Verspoor, K., Wang, Z., Rocha, L.M.: Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. Genome Biol. **9**(2), 1 (2008)
2. Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions, pp. 69–72. Association for Computational Linguistics (2006)
3. Campos, D., Matos, S., Oliveira, J.L.: A modular framework for biomedical concept recognition. BMC Bioinf. **14**(1), 281 (2013)

4. Fontaine, J.F., Barbosa-Silva, A., Schaefer, M., Huska, M.R., Muro, E.M., Andrade-Navarro, M.A.: Medlineranker: flexible ranking of biomedical literature. Nucleic Acids Res. **37**(suppl 2), W141–W146 (2009)
5. Kim, S., Wilbur, W.J.: Classifying protein-protein interaction articles using word and syntactic features. BMC Bioinf. **12**(8), 1 (2011)
6. Krallinger, M., Valencia, A., Hirschman, L.: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. **9**(2), 1 (2008)
7. Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatr-aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L., Iannuccelli, M., et al.: The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinf. **12**(8), 1 (2011)
8. Lan, M., Tan, C.L., Su, J.: Feature generation and representations for protein-protein interaction classification. J. Biomed. Inform. **42**(5), 866–872 (2009)
9. Marcotte, E.M., Xenarios, I., Eisenberg, D.: Mining literature for protein-protein interactions. Bioinformatics **17**(4), 359–363 (2001)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
11. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta, May 2010. http://is.muni.cz/publication/884893/en
12. Smith, L., Tanabe, L.K., nee Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., et al.: Overview of biocreative ii gene mention recognition. Genome Biol. **9**(2), 1 (2008)
13. Suomela, B.P., Andrade, M.A.: Ranking the whole medline database according to a large training set using text indexing. BMC Bioinf. **6**(1), 1 (2005)
14. Wang, Q., Abdul, S.S., Almeida, L., Ananiadou, S., Balderas-Martínez, Y.I., Batista-Navarro, R., Campos, D., Chilton, L., Chou, H.J., Contreras, G., Cooper, L., Dai, H.J., Ferrell, B., Fluck, J., Gama-Castro, S., George, N., Gkoutos, G., Irin, A.K., Jensen, L.J., Jimenez, S., Jue, T.R., Keseler, I., Madan, S., Matos, S., McQuilton, P., Milacic, M., Mort, M., Natarajan, J., Pafilis, E., Pereira, E., Rao, S., Rinaldi, F., Rothfels, K., Salgado, D., Silva, R.M., Singh, O., Stefancsik, R., Su, C.H., Subramani, S., Tadepally, H.D., Tsaprouni, L., Vasilevsky, N., Wang, X., Chatr-Aryamontri, A., Laulederkind, S.J.F., Matis-Mitchell, S., McEntyre, J., Orchard, S., Pundir, S., Rodriguez-Esteban, R., Van Auken, K., Lu, Z., Schaeffer, M., Wu, C.H., Hirschman, L., Arighi, C.N.: Overview of the interactive task in biocreative v. Database 2016 (2016). http://database.oxfordjournals.org/content/2016/baw119.abstract

K-Means Clustering with Infinite Feature Selection for Classification Tasks in Gene Expression Data

Muhammad Akmal Remli[1], Kauthar Mohd Daud[1], Hui Wen Nies[1], Mohd Saberi Mohamad[1(✉)], Safaai Deris[2], Sigeru Omatu[3], Shahreen Kasim[4], and Ghazali Sulong[5]

[1] Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
akmalmuhd@gmail.com, kautharmohdaud@yahoo.com, hwnies2@live.utm.my, saberi@utm.my

[2] Faculty of Creative Technology & Heritage, Universiti Malaysia Kelantan, Locked Bag 01, Bachok, 16300 Kota Bharu, Kelantan, Malaysia
safaai@umk.edu.my

[3] Department of Electronics, Information and Communication Engineering, Osaka Institute of Technology, Osaka, 535-8585, Japan
omatu@rsh.oit.ac.jp

[4] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Batu Paha, Malaysia
shahreen@uthm.edu.my

[5] School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
ghazali@spaceutm.edu.my

Abstract. In the bioinformatics and clinical research areas, microarray technology has been widely used to distinguish a cancer dataset between normal and tumour samples. However, the high dimensionality of gene expression data affects the classification accuracy of an experiment. Thus, feature selection is needed to select informative genes and remove non-informative genes. Some of the feature selection methods, yet, ignore the interaction between genes. Therefore, the similar genes are clustered together and dissimilar genes are clustered in other groups. Hence, to provide a higher classification accuracy, this research proposed k-means clustering and infinite feature selection for identifying informative genes in the selected subset. This research has been applied to colorectal cancer and small round blue cell tumors datasets. Eventually, this research successfully obtained higher classification accuracy than the previous work.

Keywords: Gene expression data · K-means clustering · Infinite feature selection · Cancer classification · Small round blue cell tumors · Informative genes · Artificial intelligence

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_7

1 Introduction

The advent of microarray technology has benefited researchers in conducting large-scale experiments on thousands of genes by analyzing the variation of interactions among genes. In line with that, the biological datasets, such as cancer datasets, have been increasing rapidly despite the various biological experiments being conducted. The biological dataset can be considered as a high dimensional data since it consists of thousands of genes. Particularly, in cancer detection, despite thousands of genes, merely a small subset of genes, known as informative genes, is correlated with the respective diseases [15]. Besides, the physicians and other related researchers faced problems in accurately determining the disease.

Therefore, machine learning (ML) methods have been applied in analyzing and classifying gene expression data into different subclasses [17]. The ability of machine learning methods to efficiently discover and identify patterns and relationships among genes has made it a popular tool among researchers [13]. Before applying the ML methods, genes expression data, which consists of several genes and different types of samples, will be pre-processed by removing noise, missing or duplicated genes [7, 8]. The preprocessing of gene expression data including normalization will improve the accuracy and quality of the results. Clustering is applied in order to find the interaction among genes [3]. The selection of a clustering technique can control the behavior of a grouped data [7, 8]. Prior to the classification techniques, feature selection or gene selection has been considered as a de facto in reducing the data dimensionality as well [4, 7, 8].

As mentioned earlier, the genes that associate with a specific disease are consist of a small subset of genes which is known as informative genes. Therefore, feature selection is used to identify the important and related informative genes for the classification, while removing non-informative and redundant genes [16].

Based on method conducted in [9], we proposed k-means clustering with infinite feature selection in colorectal cancer (CRC) and small round blue cell tumors (SRBCT) datasets. K-means clustering and silhouette width are used to split and validate the clusters. Then, the selected sub-clusters are analyzed using infinite feature selection to select the informative genes. Lastly, the informative genes are used to train a classifier using linear Support Vector Machine (SVM) to obtain the accuracy. The differences between the proposed method and previous method [9] is shown in Table 1.

Table 1. Comparative table showing differences between previous and proposed method.

Methods	Garzón and González [9]	K-means clustering with infinite feature selection	Statnikov [20]
Clustering	Agnes	K-means clustering	–
Clustering validation	Silhouette width		–
Feature selection	Signal-to-noise (S2 N)	Infinite feature selection	–
Classification	Stratified ten-fold cross validation		
Classifier	Linear SVM		*K*-nearest neighbors

The rest of the paper is organized as follows. Datasets and methods used are described in Sect. 2. Section 3 provides the results of our proposed method and comparison with previous researches. The conclusion of the research is presented in Sect. 4.

2 Materials and Methods

2.1 Datasets and Tools

The datasets used in this research were colorectal cancer (CRC) and small round blue cell tumors (SRBCT). CRC dataset is obtained from http://genomics-pubsprinceton.edu/oncology/affydata/index.html [1, 9, 16]. The gene expression values of data are normalized based on z-scores over all samples to a mean zero and variance one [9]. This gene expression data consists of 2000 genes and 62 samples. It comprises of 40 tumour samples and 22 normal samples. The SRBCT dataset can be downloaded at http://research.nhgri.nih.gov/microarray/Supplement/ [12, 16]. This dataset contains 2303 genes and 40 samples. There are 29 Ewing sarcoma (EWS) and 11 Burkitt lymphoma (BL) samples [12].

Clustering and feature selection are performed using MATLAB 2014b, while Feature Selection Library (FSLib) is applied for selecting informative genes [18]. Meanwhile, the classification task is run in R version 3.3.2.

2.2 Centroid Clustering Analysis (CCA-I)

This stage performs centroid clustering analysis of the target dataset. The input of this stage is the list of genes from gene expression data and the output is the number of k clusters. Centroid clustering based on k-means [14] was chosen to cluster the list of genes. This technique used k-means for partitioning the genes into k number of clusters. Each sample in gene expression data is assigned to a cluster by minimizing the distance between each gene and the mean location of its assigned cluster. Unlike hierarchical clustering, centroid based clustering using k-means give good result in shorter time on a large dataset [5]. However, in k-means, the number of k cluster need to be determined beforehand. In order to find the best k cluster from target dataset and avoiding local minima, k-means is run several iterations [21]. In this work, k-means was run 50 iterations, which each iteration corresponds to the different number of k. The result for 50 iterations is passed to the next stage for cluster validation.

2.3 Clustering Validation (CV-II)

The result from k-means clustering was validated by using silhouette analysis [19]. In this stage, there are two steps. The first step is to validate the best number of k by obtaining mean silhouette value from each of k. The mean silhouette values represent the uniform separation of genes. The highest mean silhouette is chosen as the best k to be used in the second step. The second step runs k-means again with the best k number of clusters. Then, a silhouette graph is plotted to observe which cluster has a larger

silhouette value. Finally, all genes belong to the best cluster are selected for feature selection in the next stage.

2.4 Feature Selection (FS-III)

Selecting informative genes in unsupervised learning scenario is a challenging problem, due to the absence of classes to guide the classification. The main purpose of this stage is to select informative genes in the subset. The output of this stage is the list of informative genes. Filter based feature selection using Infinite Feature Selection (Inf-FS) algorithm is applied [18]. To apply Inf-FS, each gene is represented as a node in the graph while the relationship between them are shown as weighted edges. Each node is assigned an l length number of genes. The path with highest centrality scores was selected, as it consists of more informative genes against other paths. Next, each of the informative genes is assigned a score (weight), where the genes with the highest score are selected for classification.

2.5 Classification (C-IV)

Classification is the last process of this research, which is to evaluate the classification performance based on the selected genes obtained from the previous stages. Linear Support Vector Machine (SVM) is used as a classifier to train and test the datasets [6]. Stratified ten-fold cross validation is applied to assess the accuracy of classification on the selected gene subset. To do the comparison with Garzón and González [9], this research also needs the same classification evaluation. To show the relationship between genes and cancers, the lists of genes are further validated using a Google Scholar literature search (https://scholar.google.com/) [10].

3 Results and Discussion

This section discusses the classification results in terms of accuracy and number of selected genes in the subset.

3.1 Accuracy and Number of the Selected Genes in the Subset

The result achieved from stage I (CCA-I) and stage 2 (CV-II) are shown in Figs. 1 and 2, respectively. Figure 1 depicts the mean silhouette values obtained for each number of k. It should be noted that the k-means ran with a different number of k, starting from 2 until 50. Based on the figure, mean silhouette values have a peak at k equals to 2, which can be indicated that the distribution of this dataset may truly be partition into two groups.

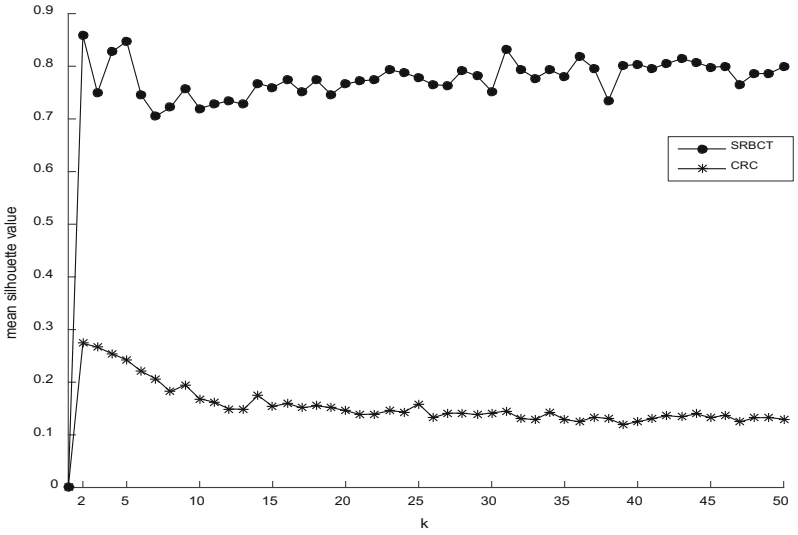


Fig. 1. Mean silhouette value obtained from each of k from k-mean clustering. The highest mean silhouette value is at $k = 2$ for CRC and SRBCT datasets.

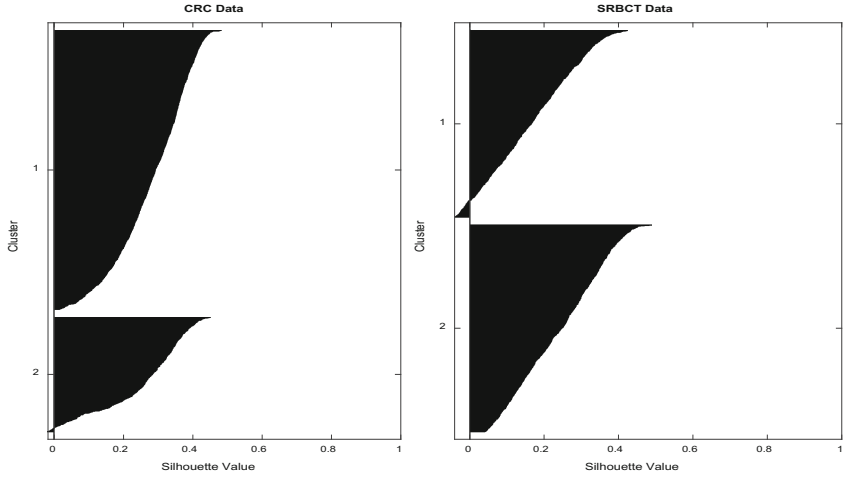


Fig. 2. Silhouette plot based on 2 clusters from k-means clustering.

Silhouette plot in Fig. 2 depicts the silhouette values for each of the clusters. Based on the figure, the biggest number of genes in the first cluster having a larger silhouette value. This indicates that the cluster is separated from the neighboring cluster. Meanwhile, the second cluster contains fewer genes with low silhouette values and a few genes with negative values. For CRC dataset, 1419 genes in cluster one and 581 genes in cluster two. For SRBCT dataset, 1093 genes in cluster one and 1210 genes in cluster

two. Genes in cluster one of CRC and cluster two of SRBCT datasets are selected as the input for feature selection (FS-III) stage.

In stage 3, the Inf-FS algorithm returns the ranked number of genes based on weight scores for 1419 genes for CRC and 1210 genes for SRBCT datasets. From the ranking, merely 50 and 45 genes are selected as the informative genes for CRC and SRBCT datasets, respectively.

Table 2 shows the comparative table between previous method [9] and proposed method based on CRC and SRBCT datasets. For CRC dataset, the proposed method has a higher classification accuracy compared to the previous method. Nonetheless, the proposed method obtained a lesser number of informative genes than the previous method. K-means clustering was a good method for recognizing the hidden patterns from datasets, although it was not often used for classification problems [11, 22]. With the effort of the Inf-FS algorithm, it has performed effectively in ranking the informative genes [18]. Thus, a small subset of selected genes can be used to build classifiers with a very high classification rate [2]. However, work in [20] provides a higher classification accuracy compared with the proposed method, as they did not select any genes and directly conducted the classification task.

Table 2. Comparative table showing accuracy and number of informative genes in the subset.

Datasets	Methods	Accuracy (%)	Number of informative genes in the subset
CRC	Garzón and González [9]	88.710	76
	Proposed Method (K-means clustering with Inf-FS)	**Mean: 89.434** **Best: 90.178**	50
SRBCT	Proposed Method (K-means clustering with Inf-FS)	Mean: 72.750 Best: 74.375	45
	Statnikov [20]	**86.90**	2308

Note: **Bold**: The best results.

3.2 List of the Selected Genes

From the previous stage (FS-III and C-IV), 50 and 45 informative genes were obtained from CRC and SRBCT datasets, respectively. These selected genes are further validated for identifying the gene markers. Hence, 15 and 11 informative genes are validated as CRC and SRBCT gene markers. The full list of informative genes can be found at the following link: https://drive.google.com/open?id=0B_G_-pnPRD1CeXVKeEJfbEJ-XYms

4 Conclusion

Classification on gene expression of patient samples has much focused in cancer diagnosis and treatment. This paper has presented a proposed method using K-means

clustering and infinite feature selection. To deal with the high dimensionality of data and low classification accuracy, this proposed method can identify informative genes with biological insights. Based on the experimental results of this research, the proposed method has higher classification accuracy compared with the previous work [9]. Hence, this proposed method has successfully identified 26 gene markers for CRC and SRBCT.

Acknowledgements. We would like to thank Universiti Teknologi Malaysia for funding this research through GUP Research Grants (grant numbers: Q.J130000.2528.12H12 and Q.J130000.2528.11H05). This research is also funded by Malaysian Ministry of Higher Education under a fundamental research grant (grant number: 1559).

References

1. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. **96**(12), 6745–6750 (1999)
2. Au, W.H., Chan, K.C., Wong, A.K., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **2**(2), 83–101 (2005)
3. Bajo, J., De Paz, J.F., Rodríguez, S., González, A.: A new clustering algorithm applying a hierarchical method neural network. Logic J. IGPL (2010). doi:10.1093/jigpal/jzq030
4. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014). doi:10.1016/j.ins.2014.05.042
5. Cebeci, Z., Yildiz, F.: Comparison of K-means and Fuzzy C-means algorithms on different cluster structures. J. Agric. Inform. **6**(3), 13–23 (2015). http://doi.org/10.17700/jai.2015.6.3.196
6. Chan, W.H., Mohamad, M.S., Deris, S., Corchado, J.M., Omatu, S., Ibrahim, Z., Kasim, S.: An improved gSVM-SCADL2 with firefly algorithm for identification of informative genes and pathways. Int. J. Bioinform. Res. Appl. **12**(1), 72–93 (2016)
7. Corchado, J.M., De Paz, J.F., Rodríguez, S., Bajo, J.: Model of experts for decision support in the diagnosis of leukemia patients. Artif. Intell. Med. **46**(3), 179–200 (2009)
8. De Paz, J.F., Bajo, J., Vera, V., Corchado, J.M.: MicroCBR: a case-based reasoning architecture for the classification of microarray data. Appl. Soft Comput. **11**(8), 4496–4507 (2011)
9. Garzón, J.A.C., González, J.R.: A gene selection approach based on clustering for classification tasks in colon cancer. ADCAIJ: Adv. Distrib. Comput. Artif. Intell. J. **4**(3), 1–10 (2015)
10. Haynes, W.A., Higdon, R., Stanberry, L., Collins, D., Kolker, E.: Differential expression analysis for pathways. PLoS Comput. Biol. **9**(3), e1002967 (2013)
11. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
12. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7**(6), 673–679 (2001)
13. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine Learning Applications in Cancer Prognosis and Prediction. Computational and Structural Biotechnology Journal **13**, 8–17 (2015). doi:10.1016/j.csbj.2014.11.005. Elsevier B.V.

14. Macqueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, no. 233, pp. 281–297 (1967). http://doi.org/citeulike-article-id:6083430
15. Mohamad, M., Omatu, S., Deris, S., Misman, M., Yoshioka, M.: Selecting informative genes from microarray data by using hybrid methods for cancer classification. Artif. Life Robot. **13**(2), 414–417 (2009). doi:10.1007/s10015-008-0534-4
16. Moorthy, K., Mohamad, M.S.: Random Forest for Gene Selection and Microarray Data Classification. Bioinformation **7**(3), 142–146 (2011). doi:10.6026/97320630007142
17. Önskog, Jenny, Freyhult, Eva, Landfors, Mattias, Rydén, Patrik, Hvidsten, Torgeir R.: Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. BMC Bioinform. **12**(1), 390 (2011). doi:10.1186/1471-2105-12-390
18. Roffo, G., Melzi, S., Cristani, M.: Infinite feature selection. In: Proceedings of the IEEE International Conference on Computer Vision, 11–18 December, pp. 4202–4210 (2016). http://doi.org/10.1109/ICCV.2015.478
19. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987). doi:10.1016/0377-0427(87)90125-7
20. Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics **21**(5), 631–643 (2005)
21. Vattani, A.: k-means requires exponentially many iterations even in the plane. Discrete Comput. Geom. **45**(4), 596–616 (2011). doi:10.1007/s00454-011-9340-1
22. Zheng, B., Yoon, S.W., Lam, S.S.: Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. **41**(4), 1476–1482 (2014)

Classification of Colorectal Cancer Using Clustering and Feature Selection Approaches

Hui Wen Nies[1], Kauthar Mohd Daud[1], Muhammad Akmal Remli[1], Mohd Saberi Mohamad[1(✉)], Safaai Deris[2], Sigeru Omatu[3], Shahreen Kasim[4], and Ghazali Sulong[5]

[1] Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
hwnies2@live.utm.my, kautharmohdaud@yahoo.com, akmalmuhd@gmail.com, saberi@utm.my

[2] Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Locked Bag 01, Bachok, 16300 Kota Bharu, Kelantan, Malaysia
safaai@umk.edu.my

[3] Department of Electronics, Information and Communication Engineering, Osaka Institute of Technology, Osaka, 535-8585, Japan
omatu@rsh.oit.ac.jp

[4] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Malaysia
shahreen@uthm.edu.my

[5] School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
ghazali@spaceutm.edu.my

Abstract. Accurate cancer classification and responses to treatment are important in clinical cancer research since cancer acts as a family of gene-based diseases. Microarray technology has widely developed to measure gene expression level changes under normal and experimental conditions. Normally, gene expression data are high dimensional and characterized by small sample sizes. Thus, feature selection is needed to find the smallest number of informative genes and improve the classification accuracy and the biological interpretability results. Due to some feature selection methods neglect the interactions among genes, thus, clustering is used to group the similar genes together. Besides, the quality of the selected data can determine the effectiveness of the classifiers. This research proposed clustering and feature selection approaches to classify the gene expression data of colorectal cancer. Subsequently, a feature selection approach based on centroid clustering provide higher classification accuracy compared with other approaches.

Keywords: Cancer classification · Gene expression data · Feature selection · Clustering · Artificial intelligence · Bioinformatics

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_8

1 Introduction

The dimensionality of biological datasets, such as cancer datasets, has been increasing rapidly. The high dimensional of biological datasets has hindered the process of transcribing structural information into functional genomics [19]. Hence, considering the complexity and combinatorial problem of codifying the biological data, it has resulted in the emergence of microarray technology [24]. The existence of microarray has enabled the researchers in biology and chemistry to acquire information and understanding about the genes expression profile in a parallel-large-scale way [3]. However, in the cancer diagnosis study, the chances of accurately predict a patient having a cancer is low. Owing to the fact that biological data consists of thousands of genes, merely a small subset of genes, known as informative genes, are correlated with the respective diseases [23, 29]. Therefore, machine learning methods which can characterize and identify normal and tumor data, have been applied to the gene expression data by analyzing and classifying into different subclasses [6, 25]. Besides, feature selection does not consider the structure among the features, unlike clustering which can cluster the similar features into the same cluster [12].

Presently, several of the techniques in clustering, feature selection, and classification have been applied to the gene expression data [10]. Work in [13] proposed Recursive Feature Elimination (RFE) in selecting the informative genes of colorectal cancer and successfully obtained 98% accuracy with only 4 informative genes. A hybrid method, GASVM-II is proposed in [23] for selecting the informative genes and improving the classification accuracy. Meanwhile, combination clustering techniques; hierarchical clustering and self-organizing maps have been proposed in [4]. The quality of the selected data can determine the effectiveness of a classifier. An appropriate classifier is important to effectively derive reliable information from data and improve the classification accuracies [21].

Since Garzón and González [11] has performed well in classification using clustering and feature selection, we proposed to test different methods for clustering, feature selection, and classifiers in colorectal cancer (CRC) dataset. Together, we have proposed four methods. The differences between previous and proposed methods are represented in Table 1.

Table 1. Differences between previous and proposed methods.

Methods	Garzón and González [11]	Proposed Method (1)	Proposed Method (2)	Proposed Method (3)	Proposed Method (4)
Clustering	Agnes	K-Means clustering	Agglomerative hierarchical clustering	Agnes	
Clustering validation	Silhouette width				
Feature selection	Signal-to-noise (S2N)	Infinite feature selection (Inf-FS)	Correlation-based feature selection (CFS) with forward search		
Classification	Stratified ten-fold cross validation				
Classifier	Linear SVM				Naïve Bayes

The distribution of the paper is organized as follows. Section 2 describes the methods and dataset. In Sect. 3, the results and comparison of our proposed method with previous studies. Section 4 provides the conclusion of this research.

2 Material and Methods

This section describes the details of dataset and methods used in this research.

2.1 Dataset and Tools

The dataset used in this research is colorectal cancer (CRC). This dataset is gene expression data. The gene expression values of data are normalized based on z-scores over all samples to a mean zero and variance one [11]. The CRC dataset is available to download at http://genomics-pubs.princeton.edu/oncology/affydata/index.html [2, 11, 24]. The CRC dataset was analyzed with an Affymetrix oligonucleotide array [2]. It consists of 2000 genes and 62 samples of classes whereby there are 40 tumor samples and 22 normal samples [2, 11].

For clustering methods, both k-means and agglomerative hierarchical clustering are performed in MATLAB 2014b. R package cluster is used for Agnes clustering. Infinite feature selection (Inf-FS) library in MATLAB is used for feature selection. Another feature selection, namely Correlation-based Feature Selection (CFS) with forward search, can be found at R package Biocomb. All the classification tasks are used in R version 3.3.2.

2.2 Clustering

Clustering has been proven effective in the medical area which has inherently overlapping information [20]. Agglomerative hierarchical clustering, Agnes (Agglomerative Nesting), and k-means clustering are used for clustering the similar genes, in this research. Clustering can be categorized into agglomerative and divisive. Agglomerative hierarchical clustering and Agnes belong to agglomerative, while k-means clustering belongs to divisive. The details of these methods are further described as follows.

Agglomerative hierarchical clustering is a bottom-up clustering method where each cluster has sub-clusters [1, 9]. It starts with each point (or genes) as a cluster. The nearest pair of clusters will be agglomerates (merged) together into one cluster. This step is repeated until merely left one big cluster. In agglomerative hierarchical clustering, the similarity between clusters is denoted as the measured distance between clusters [9].

Agnes clustering method places genes into a cluster and constructs the distance matrix using Euclidean distance [1, 11, 17]. Then, the clusters are merged together with an unweighted pair-group average method and repeated until the favour number of cluster is achieved [1].

K-means method is based on centroid clustering [18]. It computes the distance between gene expression data and cluster centre. Then, it assigns a gene to its closest cluster centre. Finally, it moves each centre to the mean of its assigned gene [33]. To

find the best k, k-means was run 50 times and mean silhouette values are obtained for each k. Therefore, the clusters of genes from Sect. 2.2 are further validated in Sect. 2.3.

2.3 Clustering Validation

The gene clusters obtained were further validated using Silhouette width. Silhouette width values measure the degree of confidence for each gene clusters [15, 27]. The well-clustered genes have values nearly to +1, whereas the poorly clustered genes have values nearly to −1 [15, 27]. One way of choosing the appropriate number of clusters (k) is to select the k value with larger Silhouette width value [15].

For k-means clustering, each cluster is validated using the silhouette value analysis. After the best k is found, the silhouette plot is used to determine which cluster has better silhouette value [30]. In this plot, the higher number of genes in the respective cluster is selected for the gene selection stage.

Thus, the best clusters for every clustering method with the largest values of Silhouette width are further used for feature selection (in Sect. 2.4).

2.4 Feature Selection

Feature selection is useful to select informative genes and remove non-informative genes, in order to reduce the high dimensionality of data and provide higher classification accuracy [5].

Correlation-based feature selection (CFS) with forward search acts as a filter method of feature selection. CFS evaluates a subset of genes with the individual predictive ability of each gene and the degree of redundancy between genes [14, 31].

Regarding infinite feature selection (Inf-FS) [26], all individual genes from the cluster are ranked based on a path among gene distributions. Then, the number of genes based on ranking is chosen to test the classification accuracy.

Hence, the classification performance of the selected gene subsets obtained from Sect. 2.4 will be further evaluated in Sect. 2.5.

2.5 Classification

Classification is used to train and test the data that obtained from feature selection (in Sect. 2.4). In this research, Naïve Bayes and linear support vector machine (SVM) are used as a classifier to demonstrate the advantages and disadvantages of feature selection [31]. Stratified ten-fold cross validation is used in this research to evaluate the performance of classifiers, in order to assess the accuracy of the experiments [7, 28].

3 Results and Discussion

This section discusses the classification results in terms of accuracy and number of the selected genes in the subset. Table 2 shows the comparative table between previous method [11] and proposed methods based on the CRC dataset.

Table 2. Comparative table showing accuracy and number of informative genes in the subset.

Method	Accuracy (%)	Number of informative genes
Garzón and González [11]	88.710	76
Proposed Method (1)	**Mean: 89.434** **Best: 90.178**	50
Proposed Method (2)	Mean: 79.282 Best: 80.353	10
Proposed Method (3)	Mean: 82.019 Best: 82.138	26
Proposed Method (4)	Mean: 83.347 Best: 85.788	26

Note: **Bold:** The best results. Proposed Method (1): k-means clustering with Inf-FS, Proposed Method (2): Agglomerative hierarchical clustering with CFS (forward search), Proposed Method (3): Agnes with CFS (forward search), Proposed Method (4): Agnes with CFS (forward search) (classifier: Naïve Bayes).

K-means clustering and Inf-FS (Proposed Method 1) has the highest classification accuracy among the methods. K-means clustering is an unsupervised learning algorithm, but it was a good method for recognizing the hidden patterns from the dataset [16, 33]. Inf-FS can effectively and correctly rank the most informative genes to gain more biological insights [26]. Hence, Inf-FS can perform the ranking step in an unsupervised manner [26] that related to the k-means clustering. Agglomerative hierarchical clustering and CFS with forward search (Proposed Method 2) provides the lowest accuracy. Agglomerative hierarchical clustering does not guarantee that, within dendrogram, the similarity is maximized [8]. However, k-means clustering is also one of the improved methods of agglomerative hierarchical clustering, which intended to improve the clustering quality. Hence, k-means clustering tends to be faster and produce inferior results compared with agglomerative hierarchical clustering [8]. The classification accuracy of CFS with forward search and Naïve Bayes (Proposed Method 4) is better than CFS with forward search and linear SVM (Proposed Method 3). This is because CFS with forward search is normally used with Naïve Bayes to detect dependencies immediately [31].

Finally, the list of informative genes in the best subsets for CRC dataset has been included as supplementary and available at the following link: https://drive.google.com/drive/folders/0B_G_-pnPRD1CanhBMWVQRXRwc0k.

4 Conclusion

Cancer classification has been improved to provide useful information based on the ability of microarray technology [10]. There is also a need for a more accurate classification and diagnosis of cancer, which can help the cancer patients having earlier treatment and therapies. This research aims to identify the most informative gene subset and to provide a more accurate cancer classification by comparing different approaches of clustering and feature selection. This research has proposed four different approaches of clustering and feature selection to use for classification task in colorectal cancer. Among the proposed methods, the first proposed method based on k-means clustering

and infinite feature selection (Inf-FS) has provided the highest classification accuracy. Besides, k-means clustering has been succeeded in providing a good clustering result [18] and Inf-FS has also performed effectively in a high ranking on the informative genes [26]. Hence, linear SVM has applied to the method. SVM is efficient in separating the classes linearly [22] and mapping the given training set in a possibly high-dimensional feature space [32].

Acknowledgements. We would like to thank Universiti Teknologi Malaysia for funding this research through GUP Research Grants (grant numbers: Q.J130000.2528.12H12 and Q.J130000.2528.11H05). This research is also funded by Malaysian Ministry of Higher Education under a fundamental research grant (grant number: 1559).

References

1. Aliahmadipour, L., Eslami, E.: GHFHC: generalized hesitant fuzzy hierarchical clustering algorithm. Int. J. Intell. Syst. **31**, 855–871 (2016)
2. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci. **96**(12), 6745–6750 (1999)
3. Arakawa, Y., Shimada, M., Utsunomiya, T., Imura, S., Morine, Y., Ikemoto, T., Mori, H., Kanamoto, M., Iwahashi, S., Saito, Y., Takasu, C.: Gene profile in the spleen under massive partial hepatectomy using complementary DNA microarray and pathway analysis. J. Gastroenterol. Hepatol. **29**, 1645–1653 (2014). doi:10.1111/jgh.12573
4. Bajo, J., De Paz, J.F., Rodríguez, S., González, A.: A new clustering algorithm applying a hierarchical method neural network. Logic JIGPL **19**, 304–314 (2010)
5. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014). doi:10.1016/j.ins.2014.05.042
6. Campo, L., Aliaga, I.J., De Paz, J.F., García, A.E., Bajo, J., Villarubia, G., Corchado, J.M.: Retreatment predictions in odontology by means of CBR systems. Comput. Intell. Neurosci. **2016**, 39 (2016)
7. Chan, W.H., Mohamad, M.S., Deris, S., Corchado, J.M., Omatu, S., Ibrahim, Z., Kasim, S.: An improved gSVM-SCADL2 with firefly algorithm for identification of informative genes and pathways. Int. J. Bioinf. Res. Appl. **12**(1), 72–93 (2016)
8. Chen, T.S., Tsai, T.H., Chen, Y.T., Lin, C.C., Chen, R.C., Li, S.Y., Chen, H.Y.: A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray. In: Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2005, pp. 405–408. IEEE, December 2005
9. Davidson, I., Ravi, S.S.: Agglomerative hierarchical clustering with constraints: theoretical and empirical results. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 59–70. Springer, Heidelberg, October 2005
10. De Paz, J.F., Bajo, J., López, V.F., Corchado, J.M.: Biomedic organizations: an intelligent dynamic architecture for KDD. Inf. Sci. **224**, 49–61 (2013)

11. Garzón, J.A.C., González, J.R.: A gene selection approach based on clustering for classification tasks in colon cancer. ADCAIJ Adv. Distrib. Comput. Artif. Intell. J. **4**(3), 1–10 (2015)
12. Ghalwash, M.F., Cao, X.H., Stojkovic, I., Obradovic, Z.: Structured feature selection using coordinate descent optimization. BMC Bioinf. **17**(1), 158 (2016)
13. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**, 389–422 (2002). doi:10.1023/A:1012487302797
14. Hall, M.A.: Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato) (1999)
15. Hancer, E., Karaboga, D.: A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number. Swarm Evol. Comput. **32**, 49–67 (2016)
16. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
17. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344. John Wiley & Sons, Hoboken, NJ, USA (1990)
18. Kavya, D.S., Desai, C.D.: Comparative Analysis of K means clustering sequentially and parallely. Int. Res. J. Eng. Technol. **3**(4), 2311–2315 (2016)
19. Kelly, D.L., Rizzino, A.: DNA microarray analyses of genes regulated during the differentiation of embryonic stem cells. Mol. Reprod. Dev. **56**, 113–123 (2000)
20. Khanmohammadi, S., Adibeig, N., Shanehbandy, S.: An improved overlapping k-means clustering method for medical applications. Expert Syst. Appl. **67**, 12–18 (2017)
21. Kothandan, R., Biswas, S.: Identifying microRNAs involved in cancer pathway using support vector machines. Comput. Biol. Chem. **55**, 31–36 (2015)
22. Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., de Mendonça, A.: Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res. Notes **4**(1), 299 (2011)
23. Mohamad, M., Omatu, S., Deris, S., Misman, M., Yoshioka, M.: Selecting informative genes from microarray data by using hybrid methods for cancer classification. Artif. Life Robot. **13**, 414–417 (2009). doi:10.1007/s10015-008-0534-4
24. Moorthy, K., Mohamad, M.S.: Random forest for gene selection and microarray data classification. Bioinformation **7**, 142–146 (2011). doi:10.6026/97320630007142
25. Önskog, J., Freyhult, E., Landfors, M., Rydén, P., Hvidsten, T.R.: Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. BMC Bioinf. **12**, 390 (2011). doi:10.1186/1471-2105-12-390
26. Roffo, G., Melzi, S., Cristani, M.: Infinite feature selection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4202–4210 (2015)
27. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
28. Seetha, H., Murty, M.N., Saravanan, R.: Classification by majority voting in feature partitions. Int. J. Inf. Decis. Sci. **8**(2), 109–124 (2016)
29. Tarek, S., Elwahab, R.A., Shoman, M.: Cancer classification ensemble system based on gene expression profiles. In: 2016 5th International Conference on Electronic Devices, Systems and Applications (2016)
30. Vattani, A.: k-means requires exponentially many iterations even in the plane. Discrete Comput. Geom. **45**(4), 596–616 (2011)

31. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F., Mewes, H.W.: Gene selection from microarray data for cancer classification—a machine learning approach. Comput. Biol. Chem. **29**(1), 37–46 (2005)
32. Zaki, N.M., Deris, S., Illias, R.: Application of string kernels in protein sequence classification. Appl. Bioinf. **4**(1), 45–52 (2005)
33. Zheng, B., Yoon, S.W., Lam, S.S.: Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. **41**(4), 1476–1482 (2014)

Development of Text Mining Tools for Information Retrieval from Patents

Tiago Alves[1,2(✉)], Rúben Rodrigues[1], Hugo Costa[2], and Miguel Rocha[1]

[1] Centre Biological Engineering, University of Minho, 4710-057 Braga, Portugal
tiago_alves26@hotmail.com
[2] Silicolife Lda, 4715-387 Braga, Portugal

Abstract. Biomedical literature is composed of an ever increasing number of publications in natural language. Patents are a relevant fraction of those, being important sources of information due to all the curated data from the granting process. However, their unstructured data turns the search of information a challenging task. To surpass that, Biomedical text mining (BioTM) creates methodologies to search and structure that data. Several BioTM techniques can be applied to patents. From those, Information Retrieval is the process where relevant data is obtained from collections of documents. In this work, a patent pipeline was developed and integrated into @Note2, an open-source computational framework for BioTM. This integration allows to run further BioTM tools over the patent documents, including Information Extraction processes as Named Entity Recognition or Relation Extraction.

Keywords: Biomedical text mining · Patents · Information retrieval task · PDF to text conversion · @Note2

1 Introduction

Huge amounts of information are generated every day. In the life sciences, the number of publications, reports and patents available on databases is increasing considerably [1,2]. Patents are validated documents representing the intellectual property rights of an invention, being important sources of information due to their novelty nature, with exclusive data that is not published in other scientific literature [3,4]. So, exploring them is critical to understand several biological fields [3,5]. However, the access to these documents is limited. There are some systems able to extract some patent sections. This is the case with SureChEMBL, a tool that searches for chemicals and their structure on patents [6].

Patent documents are available in numerous databases. Those which have grant protection only for specific countries can be used for localized searches. For general-purpose searches, worldwide databases with patents with international protection are a more viable option. The j-PlatPat from Japan Patent Office (JPO) or PatFT from the United States Patent and Trademark Office (USPTO) are databases included in the former group, while the PATENTSCOPE from

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_9

World Intellectual Property Organization (WIPO) or esp@cenet from European Patent Office (EPO) are included in the latter [4].

For instance, the WIPO database has 2.7 million patents registered only in 2014 [7–9]. Since these large amounts of data are available in an unstructured nature without annotations about the text structure and available entities, the search and extraction of relevant information is a difficult and time-consuming task, impossible to be done manually [7]. To exploit these data, automating that process, the Biomedical Text Mining (BioTM) field emerged [10]. It is based on different knowledge areas such as statistics, artificial intelligence or management science, combined with text analytic components as Information Retrieval (IR), Information Extraction (IE) or Natural Language Processing (NLP) [11]. From these, IR allows to obtain relevant information resources (e.g. papers or patents) from an extensive collection of documents, and IE allows the extraction of pertinent information from these documents [12].

To apply BioTM techniques, text files are usually the input. However, patent documents are typically accessed in Portable Document Format (PDF) files, coming from encrypted image files, usually BMP, TIFF, PNG or GIF. So, the conversion of these files into machine-coded, readable, editable and searchable data is mandatory. For that, methods as Optical Character Recognition (OCR) are used [13]. The process can be summarized in two main processes: character extraction, where learned patterns are applied to delimit words or individual letters; and character recognition, where words are identified [14].

Several BioTM platforms has been developed by the scientific community. @Note2[1], developed by the University of Minho and the SilicoLife company is among these efforts. As a Java multi-platform BioTM Workbench, @Note2 uses a relational database and is based on a plug-in architecture, allowing the development of new tools/methodologies in the BioTM field [15].

Structurally, @Note2 is organized into core libraries and user interface tools. The core libraries are organized in three main functional modules: the Publication Manager Module (PMM), which can search documents on online repositories (IR Search process) and download their respective full-text documents (IR Crawling process); the Corpora Module (CM), responsible for corpora management, creating and applying IE processes to them with a manual curation system; and the Resources Module (RM), which allows the management of lexical resources to be used in IE processes. The user interface tools allow a simples interaction with the user to configure and use @Note2's functionalities [15].

Here, the objective was to develop a pipeline, a new plug-in to @Note2, able to make patent data amenable to be searched and used as an information source for the IE processes already available in @Note2 and BioTM in general.

2 Patent Pipeline Development

The patent pipeline can be organized into four different tasks. It can search for patent IDs, retrieve patent metadata, download the published patent PDF file,

[1] http://anote-project.org/.

and, finally, apply PDF to text conversion methodologies to those files. Each task was structured into a module with specific inputs and outputs. Thus, sources to search and retrieve patent IDs, to search for metadata about each patent and to return the patent file(s) in PDF format were configured as components of the *search sources module*, *metainformation sources module* and *retrieval sources module*, respectively. The used PDF to text conversion methodologies were organized in the *PDF conversion module* (Fig. 1).

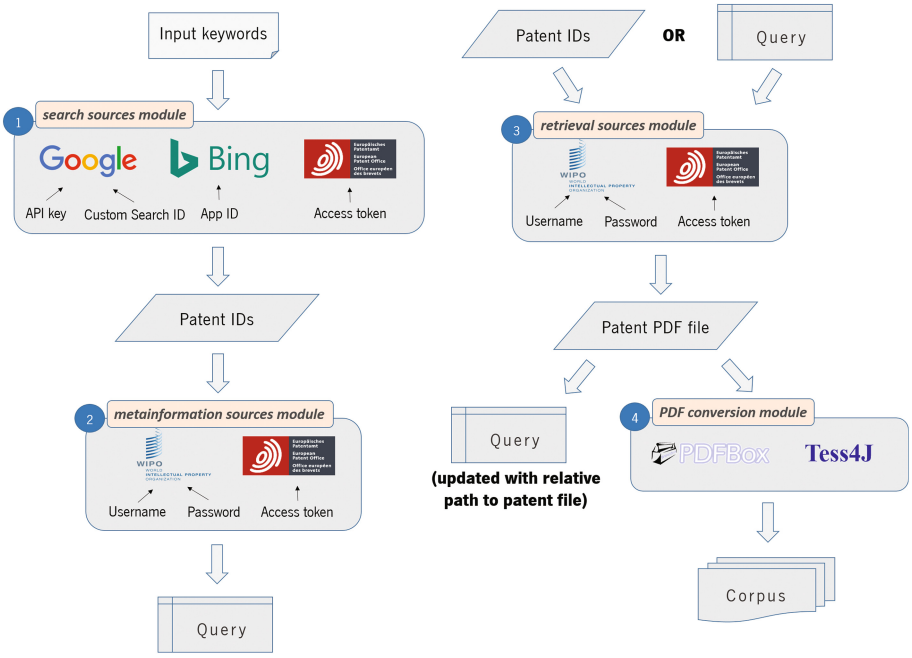


Fig. 1. Summary of the designed patent pipeline (numbers represent the process flow).

To get any result using the first three modules, specific access keys resulting from the services registration are required to get access to servers and retrieve the requested data. To start the search process, input keyword(s) are required, which may be biomedical entities as chemicals, genes, diseases, among others. These keywords are then processed by the *search sources module*. Into this module, two popular search engines (the *Custom Search* API from Google and the *Bing Search* API from Microsoft) and the *Open Patent Services (OPS) web services* API from EPO were used. The two first were configured to retrieve patent IDs from Google Patents, with around 87 millions of patents from 17 countries [16]. The result is the union of the patent IDs returned by all components.

The *metainformation sources module* returns the invention title, authors, publication date, a link to a patent database entry (if available) and the abstract to each patent. When available, the description and claims are also extracted.

To avoid repetitions, the patent family is extracted and only one ID is used to retrieve metadata, being the others saved as external references. That data is then stored into *query*, a data structure from @Note2 to save the document information (Fig. 2). Two different services were configured: the *PATENTSCOPE web service* API from WIPO and the *OPS web service* API from EPO.

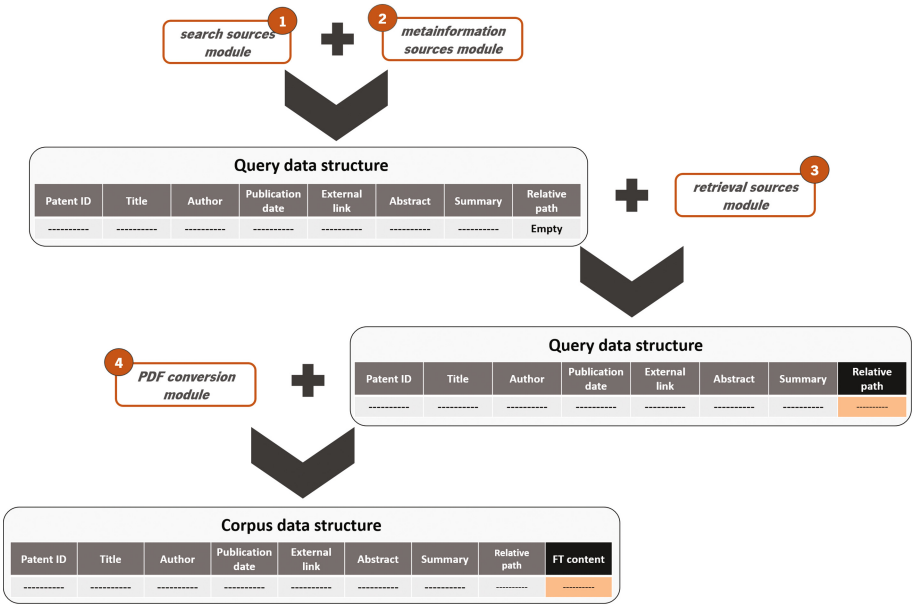


Fig. 2. Creation and update process for *query* and *corpus* data structures. The numbers represent the modules of the pipeline and their flow. The orange *query* data field represents the update process of the original *query*, while the orange *corpus* data field represents the field that turns the *corpus* into a different data structure.

The *retrieval sources module* returns the patent PDF files, saving their path into the *query* (Fig. 2). This module uses the same APIs from the previous with different configurations. Both metainformation and PDF retrieval modules use a sequential architecture. The first takes all the patents, while the next components receive only the ones that did not get any result. That process is repeated until all patents are processed or all components were used.

The *PDF conversion module* takes all the files from the previous module, extracting their text. As shown in Fig. 2, this allows the creation of a *corpus*, allowing to run IE methods, for instance, NER or RE. In this module, alongside with *Apache PDFBox* library (already implemented on @Note2) it was configured the *Tess4J*, version 3.2.1 (developed by Quan Nguyen) implementing *Tesseract*, an OCR algorithm from Google, and also a hybrid method combining these two methodologies. The *Apache PDFBox* allows to extract the Unicode

text available on PDF documents. The hybrid method allows a previous PDF treatment, improving their quality to be processed by *Tess4J* system.

On @Note2, patent handling features were inserted in different core libraries. The patent ID search and metadata retrieval were added as new IR Search processes called "Patent Search", while the patent PDF file download was added as a new IR Crawling process, and the new PDF to text conversion methods were put into the Corpora Module as a pre-processing method (Fig. 3).

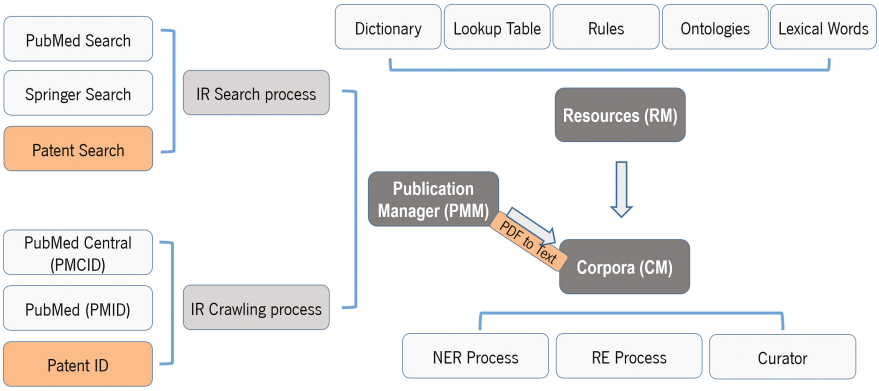


Fig. 3. @Note2 structure with patent pipeline implementations. The orange boxes represent the new components added.

3 Results

The pipeline is materialized by a plug-in allowing patent search in Google Patents and esp@cenet repositories. A graphical interface was made to set @Note2 Preferences where credentials can be saved. The main wizard includes two steps (Fig. 4): the keywords and the *query* name input pane; and the configurations pane, where the previous defined configurations can be edited (Fig. 5).

To test the system, data from the 1000 patents with the longest abstracts from the BioCreative V CHEMDNER task were used (IDs, titles and abstracts). The abstract was tokenized and compared with the tokens from our PDF to text conversion. In this comparison, we used the Smith-Waterman algorithm, a Dynamic Programming algorithm to evaluate the matches. This allows calculating performance metrics as precision, recall and F1 values (based on the number of tokens that match exactly on the texts). Alongside the accuracy calculation, it is possible infer the amount of conversion errors, as well as verify the number of documents correctly downloaded.

Complete metadata were extracted for 917 patents (91,7%). From the remaining 83, 76 were filled partially. Then, also 993 patent PDF files were correctly obtained (99,3%). For both processes, the success rate was limited due to repositories coverage and to restrictions imposed by the use of free credentials.

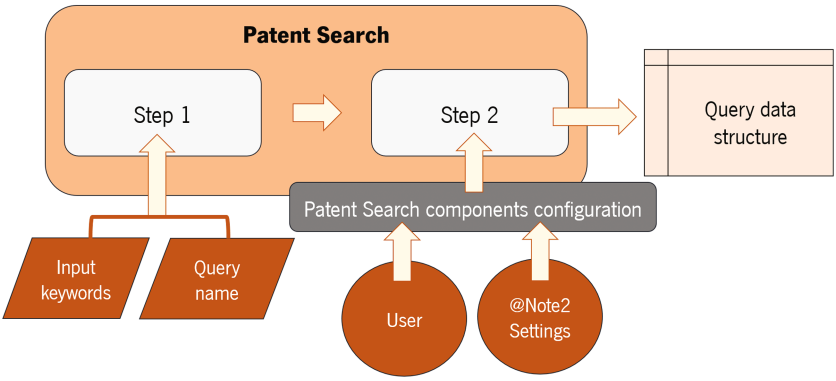


Fig. 4. @Note2 Patent Search plug-in. The pipeline uses input keywords, the *query* name and configurations provided by the user or by @Note2 settings to search for patent IDs and to download patent metadata.

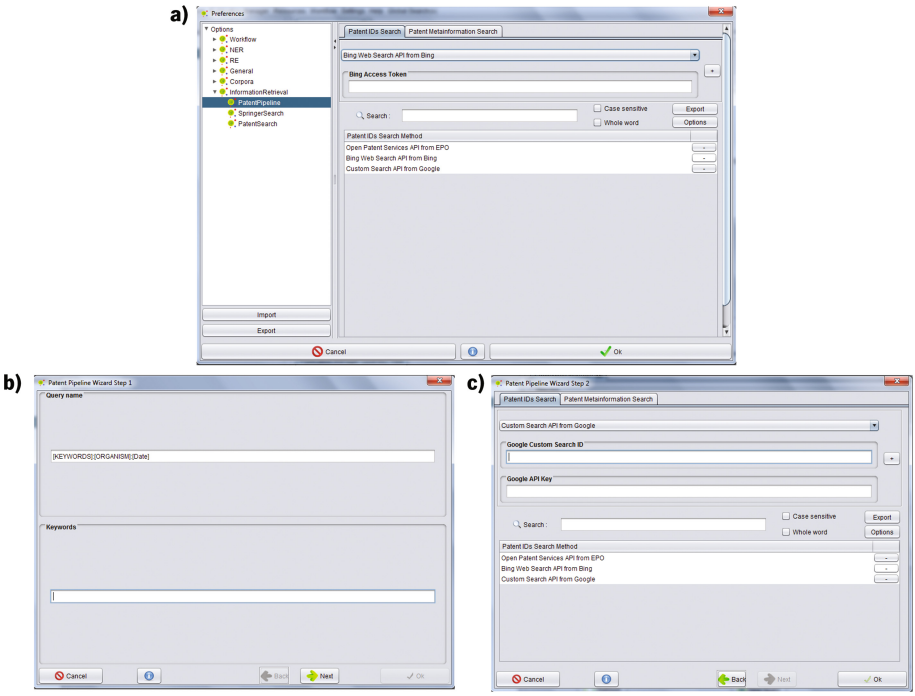


Fig. 5. @Note2 Patent Search GUI. (a) panel for @Note2 preferences; (b) and (c) Steps 1 and 2 from the Patent Search Wizard, respectively.

From the PDF to text evaluation (Fig. 6), the precision values showed a small variance being high in all documents (mean around 95%), while recall values were higher than 80% for 75% of all documents. However, 94 documents returned a recall value under 10% representing old patents (some patents before the 1970s) with only some drawings and a brief description, being the full text data absent. As expected, this led to a high standard deviation (around 30%) which can be also explained by the presence of a high number of chemical structures or formulas that are omitted in the BioCreative task abstract text or simply are converted to noise. The F1 measure summarizes the system capacity to transform most of the PDF files into readable text. Since some patent files have more than 200 pages, to process 1000 patents, the whole pipeline took around 3 days using a PC with an i7 960 @ 3.2 GHz processor and 16 GB of RAM.

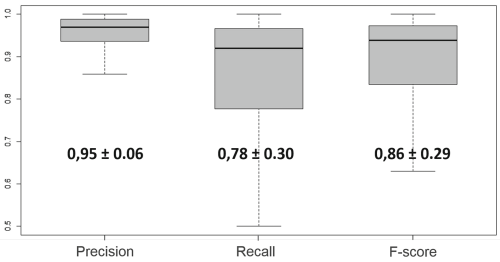


Fig. 6. Boxplots for the evaluation metrics of the PDF to text conversion process. The mean and standard deviation are given in bold.

4 Conclusions

Recently, patents have been a target for BioTM techniques since they are a great source of information for many fields. Based on @Note2, IR Search and Crawling processes were designed and implemented, allowing the search and retrieval of patent information and respective documents. Also, new improvements were made to the @Note2 PDF to text conversion system. Testing these processes with a set of 1000 patents from a BioCreative V task shows that nearly all PDFs were correctly downloaded with respective metadata. Using the new PDF to text system on that documents, we got around 85% of F-score.

The main innovation of this work was the creation of new IR processes applied to patents surpassing common problems related to searching and retrieving those documents, allowing also the posterior implementation of several IE techniques to those texts. Since @Note2 is an open-source software, this framework opens doors to the community to take advantage of all sections from the published patents with biological relevance more easily and without the need to expend large amounts of time browsing several databases. To @Note2, the integration of these tools allows developing an extensive set of text mining pipelines over patents, which were only possible for scientific articles so far.

Some improvements can still be made, namely reducing the processing time and adding new components in each module using the designed architecture.

Acknowledgments. This work is co-funded by the North Portugal Regional Operational Programme, under "Portugal 2020", through the European Regional Development Fund (ERDF), within project SISBI- Ref^aNORTE-01-0247-FEDER-003381. This study was also supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte (NORTE-01-0145-FEDER-000004) funded by European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte.

References

1. Faro, A., Giordano, D., Spampinato, C.: Combining literature text mining with microarray data: advances for system biology modeling. Brief Bioinform. **13**(1), 61–82 (2012)
2. Klinger, R., Kolarik, C., Fluck, J., Hofmann-Apitius, M., Friedrich, C.M.: Detection of IUPAC and IUPAC-like chemical names. Bioinformatics **24**(13), i268–i276 (2008)
3. WIPO, Guidelines for Preparing Patent Landscape Reports (2015)
4. Latimer, M.T.: Patenting inventions arising from biological research. Genome Biol. **6**(1), 203 (2005)
5. WIPO, WIPO Guide to Using Patent Information (2015)
6. Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S.A., Pettersson, J., Goncharoff, N., Hersey, A., Overington, J.P.: Surechembl: a large-scale, chemically annotated patent document database. Nucleic Acids Res. **44**(D1), D1220–D1228 (2016)
7. Wu, C., Schwartz, J.M., Brabant, G., Peng, S.L., Nenadic, G.: Constructing a molecular interaction network for thyroid cancer via large-scale text mining of gene and pathway events. BMC Syst. Biol. **9**(Suppl. 6), S5 (2015)
8. Lu, Z.: Pubmed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford), vol. 2011, p. baq036 (2011)
9. WIPO, World Intellectual Property Indicators, 2015th edn. World Intellectual Property Organization - Economics and Statistics Division (2015)
10. Cohen, K.B., Hunter, L.: Getting started in text mining. PLoS Comput. Biol. **4**(1), e20 (2008)
11. Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., Fast, A.: Practical text mining and statistical analysis for non-structured text data applications. Academic Press (2012)
12. Krallinger, M., Valencia, A.: Text-mining and information-retrieval services for molecular biology. Genome Biol. **6**(7), 224 (2005)
13. Asif, A.M.A.M., Hannan, S.A., Perwej, Y., Vithalrao, M.A.: An overview and applications of optical character recognition. Int. J. Adv. Res. Sci. Eng. **3**(7) (2014)
14. Holley, R.: How good can it get? analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine **15** (2009)
15. Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., Ferreira, E.C., Rocha, I., Rocha, M.: @note: a workbench for biomedical text mining. J. Biomed. Inform. **42**(4), 710–720 (2009)
16. Google, About google patents (2017)

How Can Photo Sharing Inspire Sharing Genomes?

Vinicius V. Cogo[1(✉)], Alysson Bessani[1], Francisco M. Couto[1], Margarida Gama-Carvalho[2], Maria Fernandes[3], and Paulo Esteves-Verissimo[3]

[1] LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
vvcogo@fc.ul.pt
[2] Faculty of Sciences, BioISI – Biosystems & Integrative Sciences Institute, University of Lisbon, Campo Grande, Lisbon, Portugal
[3] SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg City, Luxembourg

Abstract. People usually are aware of the privacy risks of publishing photos online, but these risks are less evident when sharing human genomes. Modern photos and sequenced genomes are both digital representations of real lives. They contain private information that may compromise people's privacy, and still, their highest value is most of times achieved only when sharing them with others. In this work, we present an analogy between the privacy aspects of sharing photos and sharing genomes, which clarifies the privacy risks in the latter to the general public. Additionally, we illustrate an alternative informed model to share genomic data according to the privacy-sensitivity level of each portion. This article is a call to arms for a collaborative work between geneticists and security experts to build more effective methods to systematically protect privacy, whilst promoting the accessibility and sharing of genomes.

Keywords: Privacy · Data sharing · Biology and genetics

1 Introduction

We live in a world plenty of connected devices and services that stimulate and simplify data sharing, which promote the acceptance of exposure risks. Nowadays, the general public recognizes several privacy risks in sharing photos on the Internet. This was promoted by the public widespread dissemination of some information leakages that caused severe privacy harms, which made users start to demand more privacy guarantees to continue sharing their data on online platforms [14].

Solutions for photo sharing already faced several privacy-related conflicts and policies changes, and life sciences can learn from them. An analogy between sharing photos and genomes may increase people's awareness on privacy risks and contributes to avoid future leakages that could damage people's willingness

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_10

to share genomic data. We emphasize that this comparison is reasonable since sequenced genomes and modern photos are digitized records of real lives. Both contain private information that may compromise people's privacy, and most of the times, their highest value is only achieved when shared with others.

Human genome is privacy sensitive since it contains personal information, and researchers need the access to large collections of genomes to accelerate medical breakthroughs. The ethical appeal for disclosure stimulates altruistic individuals to donate biological samples for medical and genomic research. However, this point of view must coexist with the ethical discussion on the risks to donors' privacy and encourage the development of secure models to share genomic data [1].

Privacy and data sharing are not mutually exclusive. Properly discussing and defending privacy encourages the responsible data sharing and extends donors' engagement and trust in researches. Recent publications corroborate with the ideas that clearly informing donors about the privacy risks of their choices does not affect negatively their willingness in donating samples [12], and that there is a need for balancing data access and privacy in genomics [19].

In this article, we propose an analogy between privacy aspects of sharing photos and sharing genomes, which contributes to clarify the privacy risks in the latter. Additionally, we illustrate possible advances in sharing genomes with an alternative informed model to share genomic data according to the privacy-sensitivity of their portions. These two contributions promote the accessibility and sharing of human genomes, whilst advocates their responsible management considering the privacy of sample donors.

2 An Analogy Between Sharing Photos and Sharing Genomes

We defined an analogy by comparing the similarities and features of the processes of sharing photos and sharing genomes, which is based on the following aspects:

- Some portions of data are more privacy-sensitive than others.
- One's data may affect the privacy of others.
- Systematically detecting the privacy-sensitive portions of data is feasible.
- After classifying the portions, decide how to share them.
- The impact of data sharing is unpredictable.

On each topic, we first describe it from the perspective of sharing photos and then we present the analogy on how does it apply in sharing human genomes. Note the present analogy is non-exhaustive since further discussions from the community may identify other similarities in the future.

2.1 Some Portions of Data Are More Privacy-Sensitive Than Others

Some elements in photos (e.g., faces and places) may disclose sensitive information about the people that own or are depicted in them, such as identity, ancestry,

health, behavior, preference, possession, and location. Similarly, genomes contain portions of sequences that contain more critical information (e.g., predisposition to a disease, parental correlation) about their donors and their relatives [15]. Authors of recent publications managed to compromise donors' privacy by targeting specific portions of human genomes, such as short-tandem repeats [7], disease-related genes [16], and genomic variations [8]. These elements may disclose information, for example, related to identity, ancestry, and health.

2.2 One's Data May Affect the Privacy of Others

Photos portraying other individuals may compromise their privacy, as well as photos containing elements related to controversial topics may affect the privacy and safety of owners' relatives (e.g., [10]). In human genomes, some information is hereditary (e.g., Y chromosome from father to son), and thus compromising the privacy of one subject genome can also affect his relatives [15].

2.3 Systematically Detecting the Privacy-Sensitive Portions of Data Is Feasible

Detecting the privacy-sensitive elements in photos includes recognising faces [9], activities [17], texts [11], signs and other location-specific elements [6]. Recently, we proposed a method that detects the privacy-sensitive portions of human genomes by comparing small DNA portions against a knowledge database of privacy-sensitive genomic sequences [5]. In both cases (photos and genomes), the detection compares small elements against large databases of known patterns. Although those detection methods contribute to privacy protection by differentiating sensitive information, the challenges remain mostly in building comprehensive knowledge databases and querying them efficiently.

2.4 After Classifying the Portions, Decide How to Share Them

Regarding photo sharing there are two distinct options: (1) enable the share if the person concludes it does not compromise his/her privacy nor the privacy of others, and (2) share a portion of the photo, which the person believes it does not compromise anyone's privacy, while keeping private or obfuscating the remaining sensitive portions for the general public. Excluding these two options there is always the possibility to not share the photo. Recent publications proposed alternative informed models to share photos considering the privacy-sensitivity of their portions [9,18]. Similar to photos, every human genome contains some privacy-sensitive portions. We advocate that sharing certain portions of data is more attractive than sharing nothing, and those privacy-sensitive portions may still be shared in a controlled way (e.g., using the cryptographic methods discussed in [15]). In the next section, we propose an alternative informed model to share genomic data considering the privacy-sensitivity of their portions.

2.5 The Impact of Data Sharing Is Unpredictable

Sharing photos may have an immediate impact in the lives of a small number of people related or depicted on them. However, the global impact of a shared photo is unpredictable. For instance, a photo can be considered meaningful to history independently from depicting everyday-life or epic moments. Additionally, several quotidian applications we use rely on common user-contributed content, as well as some news we read depend on participatory journalism. The contribution of sharing each data is little, but all these incremental collaborations have a huge impact. The same happens with human genomic data, where the highest value of photos and genomes is most of times achieved only when sharing them with others. The individual altruism in contributing to medical and genomic studies has an extreme importance on the breakthroughs in health-related areas.

3 An Informed Model to Share Genomic Data

With all the previously mentioned aspects in mind, we call attention to the opportunities a hybrid solution can bring to balance data access and privacy of genomic data [5,19]. Our proposal is to use the referred detection method [5], as mentioned in Sect. 2.3, to identify and differentiate the privacy-sensitive sequences of human genomes from the remaining portions. This enables one to keep the small privacy-sensitive portions (i.e., less than 12%, conservatively [5]) of human genomes under a strict access control list, and make the remaining portions directly accessible to researchers and projects, according to the rights defined at their registry in the data repository. The completeness of this method evidences that there is already a large body of knowledge on the privacy sensitiveness of human genomes and that the discovery of novel privacy-sensitive sequences is unlikely using current methods (e.g., [7,8,16]). In this section, we introduce this model and its main internal components, as well as place it in the ecosystem and describe our vision on how should players interact with it.

3.1 Players and Interactions with the Model

There are four main players in the ecosystem of genomic data sharing. *Sample donors* donate biological material to a sample manager and inform their preferences on data sharing (if any). *Sample managers* receive, manipulate, sequence, store, and provide these biological specimens and their resulting data. Research projects are study proposals, encompasses one or more researchers, and have well-defined goals that require access to data associated with specific samples. *Researchers* are entities within projects that consume data from the storage system according to donors preferences and other permission rules. *Auditors* are stakeholders (e.g., governments, investors, donors, and data managers) that want to verify when and which researchers accessed specific data sets.

Donors fill consent forms at their registry to comply with regulations and to inform their preferences on data sharing. Donors should be free to customize

their informed preferences to state they want to automatically participate in projects related to specific topics (i.e., a blanket consent). They should also inform they want to contribute with their samples to additional specific projects they sympathize with (i.e., opt-in). Additionally, donors could delegate the decision of participating in which projects to data controllers acting on behalf of groups of individuals. Exceptionally, donors could separately forbid the use of the non-sensitive portions of their genomes by specific projects they disagree, or may require to re-categorise some non-sensitive portions as privacy-sensitive (i.e., opt-out). The per-project opt-out dissuades an eventual retraction of all genomes from the platform if an isolated misuse happens [4]. When a donor dies, the sharing preferences may become open or be kept the same, while his/her relatives gain the ability to explicitly customize them.

In the envisioned model, researchers should register themselves in the system and propose projects that are approved in the same way and with the same responsibilities it is currently done in biobanks and other repositories. Projects (i.e., groups of researchers) may start working with all non-sensitive sequences immediately, must wait for a short period to start using the automatically authorised privacy-sensitive portions, and have the option to request access to the privacy-sensitive portions of other genomes of interest. The utility of sequenced data is kept intact to authorised researchers in this model, which complements other approaches from the literature (e.g., [2]).

3.2 Internal Components

This data sharing model can be adapted to different legal, geographic, and organizational regulations. Additionally, this model, as depicted in Fig. 1, is completely independent of the protocols and technologies necessary to implement it. In the following, we describe four components that are of extreme importance to this model, but others can be integrated to them if needed in the future.

Evolution Module. The knowledge database from the DNA privacy detector can be automatically updated to address future attacks as new privacy-sensitive sequences are identified [5]. Thus, the detection method is generic and evolvable—i.e., it does not become outdated since public databases can be automatically tracked for updates as they evolve. An *evolution module* in this system architecture should allow the stored data sets to be re-analyzed at any moment and attested again for their privacy-sensitivity. As soon as a new privacy-sensitive sequence is identified, the data sets updated, access rules are adapted accordingly, and the access history is logged for future inquiries.

Storage. *Storage* components should retain and provide the large amount of genomic data coming from life-sciences institutions. Storage infrastructures encompass several options from private data centers to public clouds. Data from human genomes in the envisioned model is stored according to the privacy-sensitivity of its portions. The privacy-sensitive portions of human genomes must be stored in infrastructures with appropriate levels of security and dependability, while the non-sensitive ones can stored in more affordable infrastructures.

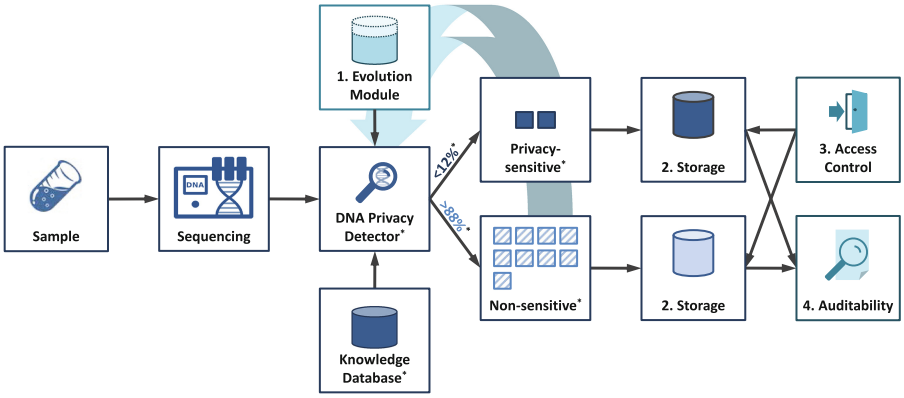


Fig. 1. Overview of the hybrid data sharing model. This model considers that genomes have their privacy-sensitive portions differentiated* from the remaining ones.

Noticeable, this hybrid model improves the cost efficiency of any storage system since it reduces the percentage of data requiring strong security and dependability premises.

The level of security and dependability depend on the use of encryption, information dispersal, data replication, etc. Choosing the best fit is orthogonal to this model and depends also on the legal constraints defined by regulators from the region of the sample manager. Restrictive regulations may impede sending data to infrastructures in other countries, while less restrictive ones may allow the use of standard encryption and public clouds. For instance, the storage solution from the BiobankCloud project already considers this range of options and provides data storage in private repositories, in single public clouds, and in multiple clouds (i.e., a cloud-of-clouds) [3].

Access Control. Access control establishes a differential access to users, accordingly to their roles and analysis. An *access control* solution should verify and permit researchers to access the different portions of the genomes they are allowed to. Additionally, the access control complements the evolution module by automatically updating the lists and rules according to the data sets' version.

There are three main factors to authenticate an access request: something the user knows (e.g., a password), has (e.g., a token), or is (e.g., biometrics). Combinations of them can be used to increase the difficulty for an illegitimate user having access to a resource. For instance, the BiobankCloud platform [3] requires each user to authenticate with his/her password and an one-time password generated using a mobile phone or a Yubikey. Additionally, cryptographic solutions complement access control mechanisms since an attacker that circumvents the access control does not obtain the data in clear.

Auditability. Auditability is the relative ease of auditing a system or an environment, acts as a deterrent measure, and complements preventive ones, such

as security, dependability, and privacy-protection. An *auditability* component should enable stakeholders to assess at any moment exactly who accessed what data in a chronological order. Auditors should access only some metadata about the files, the access logs, and access control rules—i.e., they do not need to access the whole data sets of genomic data. The auditability component complements the evolution module by allowing the detection of who has read previous versions of a data set that was re-analyzed because it could contain previously unknown privacy-sensitive sequences. Accountability supplements auditability by ensuring all actors and actions performed on the data have been persistently recorded as evidence [13]. The system must keep an indelible tamper-proof track of data accessed by researchers, in order to detect, analyze, and sanction misuses.

4 Final Remarks

In this work, we presented an analogy between privacy aspects of sharing photos and sharing genomes, and proposed an informed model to share genomic data according to the privacy-sensitivity of their portions. The analogy contributes to advancing the privacy-perception in sharing genomes by comparing it to some well-known examples and threats from sharing photos. The informed model motivates the discussion of novel solutions for sharing genomic data considering their privacy-sensitivity.

Notwithstanding, there are many open questions (related to this model and the problems identified in the analogy) that deserve further investigation and discussion within the community, namely:

- How to provide this data sharing model without incurring in unreasonable increased management effort?
- How can public clouds be securely used in this model to reduce the costs of creating and maintaining private storage infrastructures (e.g., in biobanks)?
- Which additional type of genomic data, beyond those discussed in [5], can be considered privacy-sensitive and should thus be detected?
- There are methods that associate genomic information and photographic records (e.g., selecting individuals in a database using the association between specific SNPs and the probability of an individual having brown or blue eyes [20]). Understanding the impact of those associations on subjects' privacy may contribute for more complete protection methods.

Currently, there is a great investment to advance from conventional to precision medicine, which can succeed only if we embrace genomic data sharing in a secure and controlled environment. This article is a call to arms for geneticists and security experts, to work together and build better and more effective methods to systematically protect privacy, whilst improving the accessibility and sharing of genomic data. Our model can even be accommodated in a linked data or beacon service perspective, sharing sensitive data only means that we need to be aware of what and how we share to make it safe and useful for everyone.

Acknowledgements. This work was partially supported by the Fundação para a Ciência e para a Tecnologia, through the LaSIGE (UID/CEC/00408/2013) and BioISI research units (UID/MULTI/04046/2013), by the Fonds National de la Recherche Luxembourg (FNR) through the PEARL grant FNR/P14/8149128, and by the European Commission, through the BiobankCloud (FP7-ICT-317871) and SUPER-CLOUD (H2020-ICT-643964) projects.

References

1. Allen, A.L.: What must we hide: the ethics of privacy and the ethos of disclosure. Thomas L. Rev. **25**, 1 (2012)
2. Ayday, E., Raisaro, J.L., Hengartner, U., et al.: Privacy-preserving processing of raw genomic data. In: Data Privacy Management and Autonomous Spontaneous Security, pp. 133–147. Springer (2014)
3. Bessani, A., et al.: Biobankcloud: a platform for the secure storage, sharing, and processing of large biomedical data sets. In: Proceedings of the DMAH 2015 (2015)
4. Brenner, S.E.: Be prepared for the big genome leak. Nature **498**(7453), 139–139 (2013)
5. Cogo, V.V., Bessani, A., Couto, F.M., et al.: A high-throughput method to detect privacy-sensitive human genomic data. In: Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society, pp. 101–110. ACM (2015)
6. Doersch, C., Singh, S., Gupta, A., et al.: What makes paris look like paris? ACM Trans. Graph. **31**(4), 101:1–101:9 (2012)
7. Gymrek, M., McGuire, A.L., Golan, D., et al.: Identifying personal genomes by surname inference. Science **339**(6117), 321–324 (2013)
8. Homer, N., Szelinger, S., Redman, M., et al.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genet. **4**(8), e1000167 (2008)
9. Ilia, P., Polakis, I., Athanasopoulos, E., et al.: Face/off: preventing privacy leakage from photos in social networks. In: Proceedings of the CCS 2015, pp. 781–792. ACM (2015)
10. Jones, S., Norton-Taylor, R.: 'Congrats to Uncle C'–how his wife's Facebook page exposed new MI6 head. The Guardian, July 2009
11. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern Recogn. **37**(5), 977–997 (2004)
12. Kaufman, D.J., Murphy-Bollinger, J., Scott, J., et al.: Public opinion about the importance of privacy in biobank research. Am. J. Hum. Genet. **85**(5), 643–654 (2009)
13. Ko, R.K.: Data accountability in cloud systems. In: Security, Privacy and Trust in Cloud Systems, pp. 211–238. Springer (2014)
14. Kokolakis, S.: Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. Comput. Secur. **64**, 122–134 (2015)
15. Naveed, M., Ayday, E., Clayton, E.W., et al.: Privacy in the genomic era. ACM Comput. Surv. (CSUR) **48**(1), 6 (2015)
16. Nyholt, D.R., Yu, C.E., Visscher, P.M.: On Jim Watson's APoE status: genetic information is hard to hide. Eur. J. Hum. Genet. **17**, 147–149 (2009)
17. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
18. Ra, M.R., Govindan, R., Ortega, A.: P3: toward privacy-preserving photo sharing. In: Proceedings of the USENIX Symposium on NSDI 2013, pp. 515–528 (2013)

19. Vayena, E., Gasser, U.: Between openness and privacy in genomics. PLoS Med. **13**(1), e1001937 (2016)
20. Walsh, S., Liu, F., Ballantyne, K.N., et al.: Irisplex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci. Int. Genet. **5**(3), 170–180 (2011)

An App Supporting the Self-management of Tinnitus

Chamoso Pablo[1], De La Prieta Fernando[1], Eibenstein Alberto[2], Tizio Angelo[2], and Vittorini Pierpaolo[2(✉)]

[1] IBSAL/BISITE Research Group, University of Salamanca, Calle Espejo 12, Edificio I+D+i, 37007 Salamanca, Spain
{chamoso,fer}@usal.es
[2] University of L'Aquila, Delta 6, Via G. Petrini, 67100 Coppito, L'Aquila, Italy
pierpaolo.vittorini@univaq.it

Abstract. Tinnitus is an annoying ringing in the ears, in varying shades and intensities. Tinnitus can affect a patient's overall health and social well-being (e.g., sleep problems, trouble concentrating, anxiety, depression and inability to work). Usually, the diagnostic procedure of tinnitus passes through three steps, i.e., audiological examination, psychoacoustic measurement, and disability evaluation. All steps are performed by physicians, by using dedicated hardware/software and administering questionnaires. The paper reports on the results of a one-year running project whose aim is to directly support patients in such a diagnostic procedure, and in particular on an Android app that controls an ad-hoc developed device and automate both the execution of the audiometric examinations and the administration of the questionnaires that measure the disability induced by the tinnitus.

Keywords: Tinnitus · App · Audiometry · Acufenometry

1 Introduction

Tinnitus is known to be a complex of annoying ringing/buzzing/hissing in the ears, in varying shades and intensities [1]. Recent statistics about epidemiology of tinnitus reports on a minimum prevalence of 6% and a maximum of 28–30% [2–5]. It may cause sleep problems, trouble concentrating, ongoing depression and inability to work [6,7] and therefore affecting a patient's overall health and social well-being [8,9]. Usually, the diagnostic procedure is performed by physicians – by using dedicated hardware/software and administering questionnaires – and takes place in terms of accurate audiological examinations, psychoacoustic measurements of tinnitus, and evaluations of disability. The paper reports on (part of) the results of a one-year running project whose aim is to directly support patients with tinnitus in such a diagnostic procedure. The project resulted in the development of an hardware device that executes (a part of) the audiological and psychoacoustic examinations needed to diagnose tinnitus, and an

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_11

ad-hoc developed app that controls the device and automates both the execution of the examinations and the administration of the questionnaires that measure the disability induced by tinnitus. The paper mainly focuses on the app development and evaluation, whereas a short description of the device is given for completeness.

The novelty of our project with respect to the available literature is twofold: (i) it directly addresses patients instead of health professionals and (ii) provides an integrated tool able to perform the audiological and psychoacoustic measurements as well as the evaluation of disability. It is worth remarking that devices controlled by apps are not common and only available to physicians (see, e.g., [10,11] for devices concerning the audiological examination), while apps are instead available though not including many features needed for a thorough diagnostic procedure (i.e., they include only a simplified audiological examination without the psychoacoustic measures or the evaluation of disability). Figure 1 summarizes the main concept of the project. The paper is organized as follows. Section 2 introduces the necessary background on tinnitus, i.e., the audiological and psychoacoustic examinations, the questionnaires for measuring the impact of tinnitus in different aspect of quality of life. Section 3 briefly describes the ad-hoc developed hardware device. Section 4 – the main contribution of the paper – discusses the app and how the automated examinations are implemented. Section 5 reports on an experiment concerning the quality of the automated reporting procedure. Finally, Sect. 6 ends the paper with a discussion on the future work.

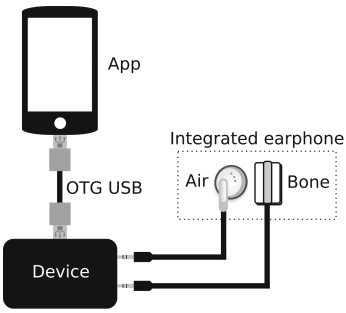


Fig. 1. Project concept

2 Background

2.1 Audiometry

A Pure Tone Audiometry (PTA) is the procedure that uses pure tones – sounds having a single specific frequency – to assess an individual's hearing [12]. The general procedure for a pure tone audiometry goes as follows. The patient is instructed to listen carefully for a beeping sound (pure tone): when heard, even if very softly, he/she is asked to raise his or her hand. Pure tones are then presented to the patient, initially at an intensity level that it is assumed can be heard quite well. After the patient demonstrates a good understanding of the task, the intensity (loudness) of the tone is decreased in 10 to 15 dB steps, until the patient no longer responds. The intensity is then raised in 5 dB steps, decreased again and increased again in 5 dB steps, until the patient responds. The lowest audible intensity is then defined as the patient's threshold for the particular frequency. This method is described as the "modified Hughson-Westlake ascending-descending paradigm". This routine is repeated for all test frequencies in one ear, then again in the other

ear. Such a procedure then establishes a threshold curve for each ear called audiogram. Depending on the transducer through which the stimuli is presented, the audiometry can be either air-conducted or bone-conducted. An air conducted signal is defined as a sound wave travelling through air. This mode of signal presentation assesses the entire auditory system: the outer ear, ear canal, tympanic membrane, middle ear system, cochlea, auditory nerve, auditory brainstem, and auditory cortex. A deficit in one or more of these areas may result in a measurable hearing loss when testing via air conduction. Thus, when a hearing loss is measured during air conduction, further tests become necessary to determine which part(s) of the auditory system are dysfunctional. If there is any degree of hearing loss measured at any frequency in either ear, bone conduction pure tone testing must be performed [13]. Bone conduction pure tone testing stimulates the cochlea directly, bypassing the outer and middle ear. This type of testing is used to determine whether a hearing loss measured via air conduction is reflective of a cochlear/neural deficit or an outer or middle ear dysfunction. If bone conduction pure tone thresholds agree with air conduction thresholds, the loss is determined to be related to the cochlea or higher neural processes and is termed "sensorineural". If, on the other hand, bone conduction thresholds are better than air conduction thresholds, a "conductive" hearing loss is present. If bone conduction thresholds indicate a hearing loss but one which is less severe than by air conduction thresholds, the loss is termed a "mixed" hearing loss, i.e., there is both a sensorineural and a conductive component.

2.2 Acufenometry

Acufenometry aims at determining the frequency and intensity of the tinnitus, by asking the patient to compare the frequency of a test-sound (i.e., a pure tone) with that of the tinnitus. The procedure for acufenometry goes as follows. Two tones are presented alternately to both ears so that each tone is heard 4–5 times; the frequency is increased or decreased until the patient finds out the one closest to the tinnitus. A pure tone at the previously identified frequency is firstly sent to the other side ear. Then, the intensity is increased by 5 dB until the patient hears it. In this way the "threshold of perception" of a signal is established and taken as the reference level of 0 dB. By increasing now the intensity by 5 dB steps, the patient is asked to report when the sound level completely masks the tinnitus. The frequency and intensity reported in such a way by the patient represent the result of the acufenometry.

2.3 Questionnaires

– *Pittsburgh Sleep Quality Index (PSQI).* The Pittsburgh Sleep Quality Index (PSQI) is a self-administered questionnaire which assesses sleep quality and disturbances over a 1-month time interval [14]. It is made up nineteen individual items that measure the subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleep medication, and daytime dysfunction over the last month. A complex procedure

results in a score: if it is ≤5, a good sleep quality is detected; if it is >5, a poor sleep quality is instead revealed;
- *Khalfa Hyperacusis Questionnaire.* The Khalfa Hyperacusis Questionnaire is a tested and validated tool for hyperacusis. It is made up of 14 questions, each with four possible answers (i.e., "no", "rarely", "often" and "always"). The scoring procedure yields to a total score: if greater than 28, it is a severe hyperacusis; if greater than 16, it is a mild hyperacusis; in the remaining case, the result is normal [15];
- *Tinnitus Handicap Inventory (THI).* The Tinnitus Handicap Inventory (THI) [16] self-administered questionnaire evaluates the impact of tinnitus on the quality of life. It is made up of 25 questions, each with 3 possible answers (i.e., "no", "sometimes" and "yes"). According to the score, it identifies five different grades of disorder, i.e., very slight tinnitus, mild tinnitus, moderate tinnitus, severe tinnitus and catastrophic tinnitus.

3 The Device

During the project we designed and developed a device responsible for generating a pure tone associated to the audiometry and acufenometry processes. The hardware device is connected as a peripheral to the patient's smartphone by using the USB On-The-Go (OTG) connection, where the smartphone assumes the role of power supplier (OTG A-device) and the developed hardware assumes the role of power consumer (OTG B-device). This means that the device does not need its own battery, which is itself a cost-reducing benefit, but still maintains its characteristic of mobile device. The device receives from the app the frequency and the intensity of the sound to be emitted, and whether the sound should be sent via air or bone conduction. As a consequence, the device generates the pure-tone at the required frequency and intensity, and send it to the required transceiver allocated in an integrated earphone. For more information about the device, you can refer to [17].

4 The App

The app is available for smartphones and tablets running Android 4.4 and above. It is written in Java using Android Studio, and is available for download at http://vittorini.univaq.it/tinnitus/. The app includes the functionalities needed for the clinical evaluation described in Sect. 2, takes advantage of the device described in Sect. 3 to implement the automated audiometry (Subsect. 4.1) and acufenometry (Subsect. 4.2), and finally proposes and automatically scores the questionnaires for sleep quality, hyperacusis and the impact of tinnitus in life (Subsect. 4.3).

4.1 Automated Audiometry with Reporting

The app implements the Hughson-Westlake process described in Subsect. 2.1, for both air and bone conduction audiometry, with the following two exceptions: (i) the patient touches a button placed in the centre of the smartphone when

he/she hears the sound, instead of raising an hand; (ii) the decrement/increment of intensity is not performed[1]. The process returns a matrix of intensities, i.e., when the patient heard the sound, for both ears, for both ways (i.e., air and bone), for all investigated frequencies, that is given in input to an automated audiometric reporting procedure. Such an automated audiometric reporting procedure works as follows. Initially, for both ears and all investigated frequencies, recalls the intensities reported for both air and bone conduction. Then, it deduces:

- *Type of problem.* The decision concerning the type of problem is almost straightforward. According to [18] and as briefly reported in Subsect. 2.1: if both intensities are below 25 dB, there is no hearing loss; else, if bone conduction is worse than air conduction, there is an error in the audiometry; if bone conduction agrees with air conduction (i.e., the difference is <20 dB) then the problem is considered sensorineural; else the problem is considered conductive.
- *Severity of problem.* The severity of the hearing loss is deduced in terms of the thresholds defined in [18], i.e., $[25, 40) =$ mild, $[40, 70) =$ medium, $[70, 90) =$ serious, $\geq 90 =$ severe.

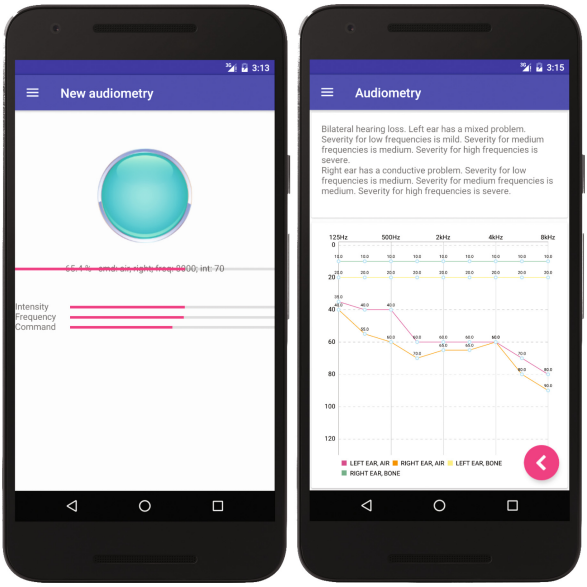


Fig. 2. Snapshots of an audiometry in progress (left) and of an audiogram with the automated reporting (right)

Given the above, the automated procedure sums up all such information in terms of frequencies ranges, i.e., low/medium/high frequencies. Therefore, a summative interpretation of the phenomenon (for each ear) is given. Firstly, if all severities are normal, the algorithm concludes that the audiometry (for that ear) is normal. Otherwise, if all types are sensorineural, then the ear has a sensorineural problem; if all types are conductive, then the ear has a conductive problem; otherwise, the problem is a mixed one. Figures 2(a, b) show the interfaces for executing the audiometry (on the centre

[1] Such a step is present in the Hughson-Westlake process so to ensure that the intensity level reported by a patient was actual and that the patient was not "cheating" to the physician. In our case, since the app is autonomously used by a patient, this step was considered unnecessary.

the button that the patient has to press to signal that he/she heard a sound; on the bottom the progresses of the audiometry) and for showing the results (on the top the automated reporting; on the bottom the audiogram).

4.2 Automated Acufenometry

The app implements the acufenometry described in Subsect. 2.2, without detecting the "threshold of perception" level. Figure 3 shows on the left the interface used to perform the acufenometry: the switches placed on the top can be used to select which ear has the tinnitus, the horizontal/vertical arrows change the frequency/intensity of the tinnitus, while the central button can be tapped to confirm that the emitted sound actually resembles the tinnitus.

4.3 Questionnaires

The PSQI, Khalfa and THI questionnaires are implemented and automatically scored in the app. A simple ad-hoc XML document defines and enable the app to show the questionnaires. The document allows to define the possible variable types used in the questionnaire, as a specialization of numeric, time and categoric types, and the list of questions placed in the questionnaire, their types and if required or optional. The automated scoring is instead performed by an ad-hoc class. For example, Fig. 3 shows on the right the PSQI questionnaire (in Italian language) as proposed by the app.

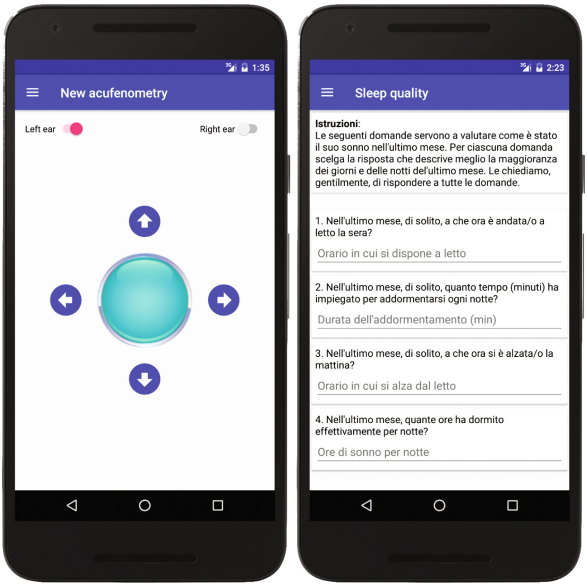


Fig. 3. Snapshots of (a) an audiometry in progress, (b) an audiogram with the automated reporting, (c) an acufenometry in progress, (d) the PSQI questionnaire

5 Results

The quality of the automated audiometric reporting procedure was evaluated by comparing it with human reporting. In particular, we selected the audiometries performed on patients with tinnitus in the "Otorinolaringoiatria" ward of

the Hospital of L'Aquila (Italy), during the period October, 2013 – July, 2016. The archive consisted of a total of 89 audiometries: 3 conductive, 11 mixed, 55 sensorineural, the remaining 20 did not highlighted any hearing loss. Given that conductive and mixed audiometries were not very frequent, the sample we used for the comparison was made up of all conductive and mixed audiometries, plus a 20% sample of the sensorineural ones (i.e., other 11 audiometries). Accordingly, a total of 25 audiometries were used for the comparison. The results show that all the automated audiometric reporting are correct, few of them were even more detailed than the human ones. Besides that, a possible improvement in the automated reporting came out: an audiometry showing a problem only for one frequency is usually reported as an "acoustic hole". Since such a wording is not provided by the app, it might be added in the next release.

A usability testing was also carried out. The choice of performing one type of usability evaluation over another has to be established in relation to the stage of the project, what to evaluate, the available experts, as well as the time constraints and the available resources of the project [19]. Since our project is in its first release, coherently with the state of the art, we decided to perform (i) an heuristic evaluation to generate the initial number of potential usability problems [19]. For it, we decided to ask to one usability expert to generate the initial number of potential usability problems. The expert used a check-list specifically designed to evaluate mobile interfaces, that reuses 69% of literature heuristics, the rest deriving from best-practices and recommendations for mobile interfaces [20]. For space constraints, we report only the main results, i.e., the app already has a good usability, but we should (i) add a clear back/undo button, (ii) more clearly show the goals of each functionality (i.e., audiometry, acufenometry and questionnaires), (iii) implement a font scaling feature[2], (iv) add a search facility, even if it may be not necessary given the shallow navigational structure.

6 Discussion and Future Work

The paper presented a new device and app regarding the self-management of tinnitus, with a specific focus on the app and the automated audiometry and reporting. As for the automated reporting, the results showed that the app is correctly able to summarize the results of an audiometry without mistakes. However, the contribution discussed in the paper represent only part of the work produced during the project, and more work is planned for the next months.

In particular, it is worth reporting on the next usability testing steps, since giving the device and app directly to patients requires a careful investigation about whether the system can be proficiently used by the patients or not. After revising the app with respect to the suggestions coming from the expert, we will inquire primary stakeholders, i.e., patients. We aim at gathering quantitative data through the following metrics: (i) Single Ease Question (SEQ) [21], (ii) Expectation Measure (EM) [22], and (iii) SUS [23]. As known, the first two metrics are used to understand the user experience in performing specific tasks,

[2] The font size may be too narrow for some users (especially elderly) if not scalable.

while the latter regards a general view on the usability. In particular, the SEQ is a 5-point rating scale metric that is used to assess how a task was found easy/difficult to accomplish by users. The higher the value of the response, the easier the task is. We will ask patients to rate the three tasks of: (T1) performing an audiometry, (T2) performing an acufenometry, and (T3) completing a questionnaire. The EM is instead a measure of comparing the results of the SEQ with how easy or difficult the user thought a task was going to be. So, before the users actually did any of the tasks, we will ask patients to rate how easy/difficult they expect each of the tasks to be, based simply on their understanding of the tasks. Finally, the SUS is a reliable tool for measuring usability, consisting of a 10 item questionnaire with five response options per item. It can be used to evaluate a wide variety of products and services, including mobile apps. The tool rates a system with a score ranging from 0 to 100. The higher the score, the more usable the system is.

References

1. Del Bo, M., Giaccai, F., Grisanti, G.: Manuale di audiologia, 3 edizione edn. Elsevier (1995)
2. Ahmad, N., Seidman, M.: Tinnitus in the older adult: epidemiology, pathophysiology and treatment options. Drugs Aging **21**(5), 297–305 (2004)
3. Axelsson, A., Ringdahl, A.: Tinnitus-a study of its prevalence and characteristics. Br. J. Audiol. **23**(1), 53–62 (1989)
4. Pilgramm, M., Rychlick, R., Lebisch, H., Siedentop, H., Goebel, G., Kirchhoff, D.: Tinnitus in the Federal Republic of Germany: a representative epidemiological study. In: Proceedings of the Sixth International Tinnitus Seminar, pp. 64–67. The Tinnitus and Hyperacusis Centre, London (1999). bibtex: pilgramm_tinnitus_1999
5. Martines, F., Bentivegna, D., Di Piazza, F., Martines, E., Sciacca, V., Martinciglio, G.: Investigation of Tinnitus patients in Italy: clinical and audiological characteristics. Int. J. Otolaryngol. **2010**, e265861 (2010)
6. Moring, J., Bowen, A., Thomas, J., Bira, L.: The emotional and functional impact of the type of Tinnitus sensation. J. Clin. Psychol. Med. Settings **23**, 310–318 (2015)
7. American Tinnitus Association: Impact of Tinnitus (2016)
8. Zenner, H.P.: A systematic classification of Tinnitus generator mechanisms. Int. Tinnitus J. **4**(2), 109–113 (1998)
9. Aazh, H., McFerran, D., Salvi, R., Prasher, D., Jastreboff, M., Jastreboff, P.: Insights from the first international conference on Hyperacusis: causes, evaluation, diagnosis and treatment. Noise Health **16**(69), 123–126 (2014)
10. INVENTIS: Piccolo Portable Audiometer (2016)
11. SHOEBOX: SHOEBOX Portable Audiometer and Diagnostic Screening (2016)
12. Huizing, H.C.: Pure tone audiometry. Acta oto-laryngologica **40**(1–2), 51–61 (1951). bibtex: huizing_pure_1951
13. American Society for Legal History Association, others: Guidelines for manual pure-tone threshold audiometry (2005). bibtex: association_guidelines_2005
14. Smyth, C.A.: Evaluating sleep quality in older adults: the Pittsburgh Sleep Quality Index can be used to detect sleep disturbances or deficits. Am. J. Nurs. **108**(5), 42–50 (2008). quiz 50–51

15. Khalfa, S., Dubal, S., Veuillet, E., Perez-Diaz, F., Jouvent, R., Collet, L.: Psychometric normalization of a hyperacusis questionnaire. ORL J. Otorhinolaryngol. Relat. Spec. **64**(6), 436–442 (2002)
16. Newman, C.W., Jacobson, G.P., Spitzer, J.B.: Development of the Tinnitus handicap inventory. Arch. Otolaryngol. Head Neck Surg. **122**(2), 143–148 (1996)
17. Chamoso, P., Prieta, F., Eibenstein, A., Santos-Santos, D., Tizio, A., Vittorini, P.: A device supporting the self management of Tinnitus. In: Rojas, I., Ortuño, F. (eds.) IWBBIO 2017. LNCS, vol. 10209, pp. 399–410. Springer, Cham (2017). doi:10.1007/978-3-319-56154-7_36
18. Rossi, G.: Trattato di otorinolaringoiatria. Minerva Medica (1997)
19. Nielsen, J., Mack, R.L.: Usability Inspection Methods. Wiley, New York (1994)
20. Gómez, R.Y.: Cascado Caballero, D., Sevillano, J.L.: Heuristic evaluation on mobile interfaces: a new checklist. Sci. World J. **2014** (2014). Article id e434326
21. Sauro, J., Dumas, J.S.: Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1599–1608. ACM (2009)
22. Albert, W., Tullis, T.: Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes (2013). bibtex: albert2013measuring
23. Brooke, J.: SUS-a quick and dirty usability scale. Usability Eval. Ind. **189**, 194 (1996)

Anthropometric Data Analytics: A Portuguese Case Study

António Barata[1(✉)], Lucília Carvalho[2], and Francisco M. Couto[1]

[1] LASIGE, Faculdade de Ciências Da Universidade de Lisboa, Lisbon, Portugal
apbarata@gmail.com, fcouto@di.fc.ul.pt
[2] Hospital de Egas Moniz, Centro Hospitalar de Lisboa Ocidental, Lisbon, Portugal
lcmcarvalho@chlo.min-saude.pt

Abstract. Large amounts of information are systematically generated throughout the course of scientific research and progress. In our case, observations representing the Portuguese population within the central-southern region of Portugal were collected throughout various foetal autopsy procedures. Gestational age (GA) and measured distances and weights of numerous anthropometric features and organs, respectively, were recorded per singleton (24 variables in total). This work seeks to elaborate on the accuracy of different foetal parameters in terms of GA estimation, making use of principal component analysis (PCA) and regression techniques. We created a dataset of 450 foetuses, ranging from 13 to 42 weeks of age, to compute both PCA and regression models. Initial exploratory analysis shed light onto which variables are most explanatory in terms of foetal development, and are thus most likely suitable for predictive rolls. We produced clusters of models, based on coefficient of determination (R^2) values, by comparing the squared sum of residuals between models (significance level $\alpha = 0.05$). Models comprised of linear combinations of different variables exhibited significantly higher values of R^2 (p-value ≤ 0.05) when compared to single variable models. Across all regressions, crown-heel length (CHL), crown-rump length (CRL), and foot length (FL) are constantly present within the cluster of best predictors of gestational age. Depending on the type of regression analysis applied, body weight (Body), hand length (HL) also fall onto the same category.

Keywords: Foetopathology · Foetus · Prediction · Estimate · Gestational age · Crown-rump length · Crown-heel length · Foot length

1 Introduction

Performing rigorous estimations of gestational age is invaluable for correct diagnosis and optimum treatment of disease during the neonatal period. GA prediction is an essential tool for parental counselling and to plan for appropriate perinatal care. It is also a prime requisite for foetal autopsy, particularly in situations of criminal abortion, alleged infanticide, and medically-terminated pregnancies. Previous peer-reviewed studies have elaborated on the accuracy of different foetal parameters in gestational age prediction [1], particularly head circumference (HC), HL, FL, CRL, and CHL [2, 5].

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_12

Model analysis and hypothesis tests may help determine not only how different measurements and weights are linked to foetal developmental age, but also which variables might be classified and ordered in terms of their predictive capabilities.

In regards to anthropometric data analytics, other published papers often approach the validity of different measured variables for conceptual age estimation [6, 10], and the quantitative standards of those measurements for foetal and neo-natal autopsy [11]. Regression analysis and model fitting are widely accepted and used in this field of work, hence being viewed as reliable tools for knowledge production [12]. Other relevant publications may also be found, discussing the relationship between different methods of analysis and discriminating regression properties, enabling model validation for subsequent selection [13, 14]. Currently, the application of analytical and statistical methods for the evaluation of information is accomplished with the use of data manipulative software [15, 16]. For these computer programs to be beneficial, however, all data must be made digitally available. Without a proper data frame, analysis of data becomes tedious and/or unfeasible.

Based on foetal autopsy records, we created a dataset of 450 individuals, each comprised of 24 foetal parameters. PCA produced results indicating CHL, CRL, and FL variables as the most explanatory in terms of total data variance. By comparing regressions models, Body and HL parameters were also found to be significantly viable measurements for GA estimation. Background information regarding related work is discussed in Sect. 2. The following section describes the methodological approaches used, while Sect. 4 presents the results of said methods. Discussion of obtained results and final remarks pertain to the 5th and final Section of this paper.

2 Case Study

For several years, the foetopathology department of Hospital de Egas Moniz, has been conducting the analysis and evaluation of foetal mortality cases pertaining to the central-southern region of Portugal. Each foetal autopsy produces a physical report file containing, amongst other relevant medical information, measurements and weights of the foetus. Whenever a foetopathology instance is concluded, the file is then archived within a dossier. This type of information processing and storage does not permit direct access to harboured values in more than a few cases at a time. Reports are regarded independently of each other, making any data study laborious and time-consuming.

To address this challenge, we developed a database representing foetal autopsy records. Each report had to be manually inserted, due to discrepancies of cursive between files, excluding the use of optical character recognition (OCR) software. A total of 450 individuals between the ages of 13 and 42 inclusive were inserted into the database.

3 Methods

Given the format of each autopsy report file in this work, a database was constructed and algorithms to store, retrieve, and manipulate information were devised. Python was applied as the programming language for these tasks mainly due to its extensive libraries and packages, notably the SQLite3, NumPy, and SciPy modules [17, 19]. IBM's SPSS software [20] was also utilized due to its inbuilt statistical applications, concretely PCA and variable selection algorithms for multiple linear regression.

3.1 Data Structure

24 quantitative variables were selected to represent each foetal autopsy case. Retrieved according to autopsy protocol, the extensive list of recorded foetal parameters follows: GA, CHL, CRL, HC, chest circumference (CC), abdominal circumference (AC), FL, HL, middle finger length (MFL), intercommissural distance (ID), philtrum length (PL), inner canthal distance (ICD), outer canthal distance (OCD), left palpebral fissure width (LPFW), right palpebral fissure width (RPFW), left ear length (LEL), right ear length (REL), body, kidneys, thymus, spleen, liver, lungs, and adrenals. Paired organs are represented by their combined weight. Units comprise of week (GA), centimetre (distances and lengths), and gram (organ and body weights). Additionally, GA values consist of observed occurrences, reported throughout every case file, and not mere value estimations.

3.2 Initial Exploration and Modelling

SPSS was used to conduct the initial PCA, which would provide foresight onto possible outcomes of successive regression models. Computed extraction communalities, loadings, explained variance per component, and adequacy parameters were consequently inspected. Computation of multiple linear regression models was performed through the same IBM software. GA was selected as the dependent variable, while the remaining 23 features were used as predictors. All available regression algorithms for variable selection (Enter, Stepwise, Remove, Backward, and Forward) were utilized and their outputs taken into consideration. Models were selected based on statistically significant coefficient values ($\alpha = 0.05$), as well as Durbin-Watson and R^2 values. Standardized and un-standardized β-weights were also a point of interest for later model comparison. In total, 5 different k^{th} degree polynomial regression functions were fit onto each of the 23 variables, for $k \in \{1, 2, 3, 4, 5\}$. Each variable dataset consisted of pairs of variable-age points, where each pair represents the gestational age and recorded variable value of a singleton foetus. The NumPy module polyfit() function was used to output each single variable model. R^2 and estimated parameter values were recorded for all regressions presenting a significant p-value for the null hypothesis that the estimated coefficients are equal to zero.

3.3 Model Comparison

Regression models were compared based on each model's proportion of variance in the dependent variable predictable by the independent variable. The F-statistic was selected and computed using the squared sum of residuals (SSR) and degrees of freedom of the models being compared [21]. A significance level of $\alpha = 0.05$ was established. The SciPy module stats.f.cdf() function was used to compute p-values. Each multiple linear regression model was compared to all other multiple and polynomial models. Polynomial models were compared to other polynomial models if and only if both models pertained to the same polynomial degree. The resulting p-values were stored for later interpretation.

4 Results

4.1 Principal Component Analysis

For our dataset, the Kaiser-Meyer-Olkin (KMO) index for sampling adequacy had a value of 0.973 while the p-value corresponding to the χ^2-statistic associated with Bartlett's test of homoscedasticity was below 5×10^{-4}. PCA produced only one significant component (eigenvalue ≥ 1) explaining 93.486% of total data variance. Communality and loading values for all variables are shown below (Table 1).

Table 1. Communality and loading values per variable. Darker shades representing lower values. Table spliced due to size constraints.

	Communality	Loading		Communality	Loading
CRL	0.963	0.981	**Kidneys**	0.804	0.897
CHL	0.956	0.978	**Lungs**	0.800	0.894
FL	0.946	0.972	**RPFW**	0.800	0.894
GA	0.937	0.968	**LPFW**	0.781	0.884
HC	0.931	0.965	**ICD**	0.743	0.862
Body	0.925	0.962	**Spleen**	0.695	0.834
REL	0.924	0.961	**Adrenals**	0.694	0.833
LEL	0.918	0.958	**Thymus**	0.679	0.824
AC	0.908	0.953	**PL**	0.651	0.807
OCD	0.897	0.947	**CC**	0.572	0.756
MFL	0.872	0.934	**HL**	0.460	0.678
Liver	0.847	0.921	**ID**	0.406	0.637

4.2 Multiple Linear Regression Models

Across all variable selection methods for regression, outputs presenting models with non-significant variable coefficients were excluded (Enter and Remove). The Backward

selection algorithm was discarded for presenting the same output as the Forward approach, while yielding a Durbin-Watson statistic further away from 2. Stepwise and Forward algorithms produced models with Durbin-Watson values of 1.961 and 1.958, respectively, and similar coefficients of determination ($R^2 \approx 0.953$). Both regressions share 5 retained variables, one exclusive variable each. Only statistically significant variable coefficients are present in either model (p-value ≤ 0.05) (Table 2).

Table 2. Standardized β-weights and variables selected by each regression algorithm method.

	Body	FL	CHL	CRL	REL	Lungs	Adrenals
Stepwise	0.402	0.310	0.266	-	0.157	-0.070	-0.087
Forward	0.384	0.384	-	0.199	0.163	-0.069	-0.083

4.3 Polynomial Regression Models

A collection of 115 single variable-based models for GA estimation were generated, 5 different degree polynomial regressions for each of the 23 independent variables. All models were retained after checking the statistical significance of each model's estimated parameters (p-value ≤ 0.05). R^2 values were stored for model comparison (Table 3).

Table 3. R^2 values computed for all polynomial regressions. Polynomial degrees are represented by numbers 1 through 5, for each variable-derived model. Darker shades representing lower values. Table spliced due to size constraints.

	1	2	3	4	5
CHL	0.931	0.942	0.943	0.943	0.944
FL	0.927	0.940	0.942	0.945	0.945
Body	0.868	0.937	0.942	0.942	0.942
CRL	0.931	0.936	0.938	0.940	0.940
HL	0.410	0.917	0.930	0.934	0.936
HC	0.896	0.911	0.914	0.916	0.917
REL	0.893	0.902	0.904	0.907	0.907
LEL	0.885	0.891	0.895	0.896	0.896
Kidneys	0.734	0.876	0.877	0.881	0.881
CC	0.503	0.871	0.883	0.898	0.899
MFL	0.849	0.864	0.917	0.917	0.920
AC	0.840	0.840	0.852	0.853	0.857

	1	2	3	4	5
Liver	0.759	0.840	0.842	0.843	0.843
OCD	0.834	0.835	0.854	0.857	0.860
Lungs	0.720	0.808	0.813	0.814	0.816
Spleen	0.623	0.791	0.833	0.847	0.849
RPFW	0.730	0.759	0.800	0.803	0.809
Thymus	0.608	0.756	0.816	0.820	0.820
LPFW	0.711	0.738	0.777	0.779	0.784
ICD	0.710	0.726	0.742	0.750	0.751
ID	0.363	0.715	0.722	0.777	0.787
Adrenals	0.589	0.681	0.689	0.691	0.692
PL	0.595	0.598	0.606	0.606	0.608

4.4 Comparison and Clustering

In terms of multiple linear regression, both previously selected models exhibited no statistically significant difference between them. In contrast, when either model was compared to any of the 115 polynomial regression models, a recurring p-value ≤ 0.05 was systematically observed.

By clustering models presenting no significant difference between other variable models, and creating different variable clusters based on statistical evidence for divergence, a goodness of fit hierarchy was established. CHL, CRL, and FL were the only single parameter-based regressions to be present in the top tier throughout all polynomial degrees. The hierarchical dissimilarities were most evident between 1st degree polynomial regressions and the remaining polynomial degree models.

Notably, body weight was placed alongside CHL, CRL, and FL as best GA estimators for any polynomial degree ≥ 2; HL was also classified in such terms for any polynomial degree ≥ 3. Generally, linear measurements outperformed weights in estimating GA. In addition, PCA and 1st degree polynomial clustering output the same variable hierarchy in terms of communality/loading values and R^2.

The following tables represent the outcome of polynomial regression clustering. Due to hierarchical ambiguity and/or redundancy, 3rd and 4th degree polynomial regression models were not included. Lower R^2 model clusters were also excluded due to size limitations (Tables 4, 5 and 6).

Table 4. 1st degree polynomial regression goodness of fit clusters and ordered R^2. Darker shades representing lower values. Only top predictive variable clusters are present. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s). For example, while AC and OCD models (first cluster centres) are statistically indistinguishable from MFL and one another, both have a significantly worse fit when compared to any other given model; MFL (second cluster centre) is statistically identical to Body, and both AC and OCD models, and significantly different from every other model.

0.931	CRL				
0.931	CHL				
0.927	FL				
0.896				HC	**HC**
0.893				REL	**REL**
0.885			LEL	**LEL**	LEL
0.868		Body	**Body**	Body	
0.849	MFL	**MFL**	MFL		
0.840	**AC**	AC			
0.834	**OCD**	OCD			

Table 5. 2^{nd} degree polynomial regression goodness of fit clusters and ordered R^2. Darker shades representing lower values. Only top predictive variable clusters are present. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s). Comparatively to the previous table, Body is now indistinguishable from any of the top 4 predictors.

0.942	CHL						
0.940	FL						
0.937	Body						
0.936	CRL						
0.917						HL	**HL**
0.911					HC	**HC**	HC
0.902				REL	**REL**	REL	
0.891			LEL	**LEL**	LEL		
0.876	Kidneys	Kidneys	**Kidneys**	Kidneys			
0.871	CC	**CC**	CC				
0.864	**MFL**	MFL	MFL				

Table 6. 5^{th} degree polynomial regression goodness of fit clusters and ordered R^2. Darker shades representing lower values. Only top predictive variable clusters are present. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s). Comparatively to the previous table, HL is now indistinguishable from any of the top 5 predictors.

0.945	FL				
0.944	CHL				
0.942	Body				
0.940	CRL				
0.936	HL				
0.920				MFL	**MFL**
0.917				HC	**HC**
0.907		REL	REL	**REL**	REL
0.899		CC	**CC**	CC	
0.896	LEL	**LEL**	LEL	LEL	
0.881	**Kidneys**	Kidneys			

5 Discussion and Final Remarks

In our case of 450 foetal autopsy cases, findings suggest that across all variables, CHL, CRL, and FL are the most appropriate candidate foetal parameters for GA estimation. For any degree of polynomial regression, these variables were always displayed within

the significantly highest R^2 cluster. The same variables were also selected by multiple linear regression, exhibiting positive standardized β-weights ≥ 0.199 (ascendingly ordered CRL, CHL, and FL), and presented the highest PCA communality and loading values. Body weight, HC, HL, and ear length are also noteworthy candidate variables for either presenting high PCA communality and loading values, or having significantly meaningful β and/or R^2 values.

Accurately estimating foetal gestational age is essential for pregnancy management. As a further matter, GA estimation during autopsy procedures is key in assessing legal and criminal abortion cases. During these events, the estimation of gestational age depends on the foetal parameters used. Measurements of various foetal anthropometric features are frequently used for this purpose. Consistent with previously published work, CHL, CRL, and FL are found to be the most reliable sources of information for estimating developmental age. In cases where such measurements are impossible to obtain, other foetal features can be utilized (albeit less reliable) such as HL, HC, body weight, and ear length.

As our database evolves, and different foetal features are recorded, different studies can emerge. By analysing features such as cause of death and family background, in association with measurements and weights, machine learning algorithms can be executed to create a pathology prediction tool. This approach would be useful for early diagnosis of disease, aiding professionals and family in taking the appropriate action.

Acknowledgements. This work was supported by FCT through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013.

References

1. Hern, W.M.: Correlation of fetal age and measurements between 10 and 26 weeks of gestation. Obstet. Gynecol. **63**(1), 26–32 (1984)
2. Gandhi, D., Masand, R., Purohit, A.: A simple method for assessment of gestational age in neonates using head circumference. Pediatrics **3**(5), 211–213 (2014)
3. Kumar, G.P., Kumar, U.K.: Estimation of gestational age from hand and foot length. Med. Sci. Law **34**, 48–50 (1994)
4. Mercer, B.M., Sklar, S., Shariatmadar, A., Gillieson, M.S., D'Alton, M.E.: Fetal foot length as a predictor of gestational age. Am. J. Obstet. Gynecol. **156**(2), 350–355 (1987)
5. Patil, S.S., Wasnik, R.N., Deokar, R.B.: Estimation of gestational age using crown heel length and crown rump length in India. Int. J. Healthcare Biomed. Res. **2**(1), 12–20 (2013)
6. Selbing, A., Fjällbrant, B.: Accuracy of conceptual age estimation from fetal crown-rump length. J. Clin. Ultrasound **12**(6), 343–346 (1984)
7. Scheuer, J.L., MacLaughlin-Black, S.: Age estimation from the pars basilaris of the fetal juvenile occipital bone. Int. J. Osteoarchaeol. **4**(4), 377–380 (1994)
8. Scheuer, J.L., Musgrave, J.H., Evans, S.P.: The estimation of late fetal and perinatal age from limb bone length by linear and logarithmic regression **7**(3), 257–265 (1980)
9. Chikkannaiah, P., Gosavi, M.: Accuracy of fetal measurements in estimation of gestational age. J. Pathol. Oncol. **3**(1), 11–13 (2016)

10. Gupta, D.P., Saxena, D.K., Gupta, H.P., Zaidi, Z., Gupta, R.P.: Fetal femur length in assessment of gestational age in thirds trimester in women of northern India (Lucknow, UP) and a comparative study with Western and other Asian countries. J. Clin. Prac. **24**(4), 372–375 (2013)
11. Archie, J.G., Collins, J.S., Lebel, R.R.: Quantitative standards for fetal and neonatal autopsy. Am. J. Clin. Pathol. **126**(2), 256–265 (2006)
12. Sherwood, R.J., Meindl, R.S., Robinson, H.B., May, R.L.: Fetal age: methods of estimation and effects of pathology. Am. J. Phys. Anthropo. **113**(3), 305–315 (2000)
13. Andrews, D.T., Chen, L., Wentzell, P.D., Hamilton, D.C.: Comments on the relationship between principal components analysis and weighted linear regression for bivariate data sets. Chemometr. Intell. Lab. Syst. **34**(2), 231–244 (1996)
14. Nadaraya, E.A.: On estimating regression. Theory Probab. Appl. **9**(1), 141–142 (1964)
15. R Core Team: R: a language and environment for statistical computing, version 3.3.2. R Foundation for Statistical Computing, Vienna (2016)
16. Eaton, J.W., et al.: GNU Octave Version 301 Manual: A High-Level Interactive Language for Numerical Computations. CreateSpace Independent Publishing Platform, Seattle (2009)
17. Oliphant, T.E.: Python for scientific computing. Comput. Sci. Eng. **9**(3), 10–20 (2007)
18. Millman, K.J., Aivazis, M.: Python for scientists and engineers. Comput. Sci. Eng. **13**(2), 9–12 (2011)
19. Walt, S., Colbert, S.C., Varoquaux, G.: The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**(2), 22–30 (2011)
20. IBM Corp. IBM SPSS Statistics for Windows, version 24.0. IBM Corp., Armonk (2016)
21. Judd, C.M., McClelland, G.H., Ryan, C.S.: Data Analysis: A Model Comparison Approach. Routledge, London (2011)

Reverse Inference in Symbolic Systems Biology

Beatriz Santos-Buitrago[1], Adrián Riesco[2], Merrill Knapp[3], Gustavo Santos-García[4(✉)], and Carolyn Talcott[5]

[1] Bio and Health Informatics Lab, Seoul National University, Seoul, South Korea
bsantosb@snu.ac.kr
[2] Universidad Complutense de Madrid, Madrid, Spain
ariesco@ucm.es
[3] Biosciences Division, SRI International, Menlo Park, USA
merrill.knapp@sri.com
[4] University of Salamanca, Salamanca, Spain
santos@usal.es
[5] Computer Science Laboratory, SRI International, Menlo Park, USA
clt@csl.sri.com

Abstract. Cell dynamics is intrinsically concurrent, since many different biochemical reactions might take place simultaneously in a cell. Productive symbolic mathematical models of cell biology can be developed by modeling such biochemical reactions with rewrite rules. Analyses and predictions of biological facts can be obtained from such models. The authors have previously published several approaches for searching along cellular signaling networks. In this paper, we introduce a novel reverse inference system by applying narrowing techniques. Moreover, we propose a new general architecture which allows an extendible set of tools for direct and reverse inference by using rewriting logic.

Keywords: Symbolic systems biology · Signal transduction · Pathway logic · Rewriting logic · Maude · Narrowing

1 Symbolic Systems Biology

Symbolic systems biology pursues to explore biological processes as whole systems instead of small and independent elements. The objective is to define formal models closer to the biologists mindsets [22]. It is equally important to be able to compute with, analyze and reason about these networks of biomolecular interactions at multiple levels of detail. Such models may suggest new insights and understanding of complex biological mechanisms.

Pathway Logic development has been funded in part by NIH BISTI R21/R33 grant (GM068146-01), NIH/NCI P50 grant (CA112970-01), and NSF grant IIS-0513857. This work was partially supported by NSF grant CNS-1318848. Research was supported by Spanish projects Strongsoft TIN2012-39391-C04-04, TRACES TIN2015-67522-C3-3-R, and Comunidad de Madrid project N-Greens Software-CM (S2013/ICE-2731).

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_13

A computational analysis of cellular signaling networks has been presented by several models to simulate, as close as possible, responses to specific stimuli [1,23]. Symbolic models provide a language which allows us to represent system states and change mechanisms such as reactions. These languages provide a well-defined semantics and different analysis tools based on this underlying semantics.

Executable models are a natural way for modeling processes [16,17]. An executable model defines system states and rules specifying the manners in which the state may progress and change along time. This can be seen as the simulation of the system behavior. Moreover, characteristics of processes can be established in connected logical languages and verified using formal analysis tools. From the definition of a model, we can define specific system configurations and carry out many kinds of analysis: forward simulation, forward and backward search, model checking, and meta analysis.

Biological interactions can be handled with rule-based modeling in a natural way. In addition, the underlying combinatorial complexity and rule-based systems can cover all the important subjects for these biological interactions [10,11]. Some relevant rule-based modeling approaches are Pathway Logic [8,9], Kappa [7], and BioNet-Gen [2]. A more detailed description of Pathway Logic will be given below in Sect. 3.

The rest of the paper is organized as follows: we show the main characteristics of rewriting logic and Pathway Logic in Sects. 2 and 3. Some abbreviated notes of the modeling of signal transduction networks with rewriting logic is presented in Sect. 4. The use of meta-level and reflexion in the rewriting logic is described in Sect. 5, which also includes our contribution for reverse inference in signaling pathways. Finally, Sect. 6 presents some conclusions and the future work.

2 Rewriting Logic

Rewriting logic constitutes a logic of change or becoming [13]. It facilitates users to specify the dynamic features of systems in a general meaning. It can deal in a natural manner with states and with highly nondeterministic concurrent computations. Rewriting logic has good properties as a flexible and general semantic framework for giving semantics to a broad spectrum of languages and models of concurrency [14].

A theory in rewriting logic consists of an equational theory, that allows the user to specify sorts, constructors and function symbols, and equality between terms. Rewriting logic extends this equational theory by incorporating the notion of *rewrite rules*, which depict transitions between states.

The deduction rules of rewriting logic facilitates us a sound reasoning about which general concurrent transitions are possible in a system satisfying such a description. From a computational point of view, each rewriting step can be seen as a parallel local transition in a concurrent system. From a logical point of view, each rewriting step is a logical entailment in a formal system.

Rewriting logic is efficiently implemented in Maude [4,5]. Maude is a high-level declarative language and high-performance system which supports equational and

rewriting logic computation for a wide range of applications. In the case of Maude, the underlying equational theory is *membership equational logic* [3]. In this equational logic, the user can provide equations and define membership axioms stating the members of a sort.

Maude provides several analysis tools for rewrite theories: rewrite computation, breadth-first search, linear temporal logic model checking, inductive theorem prover, and many others. Using these features, it is possible to study how our system behaves, to check whether it is possible to reach a certain state from an initial one, and analyze if our system verifies some temporal properties.

The design of Maude has good characteristics in aspects of simplicity, performance, and expressiveness. A very wide range of applications should be naturally expressible with Maude. Besides this generality in expressing both deterministic and nondeterministic computations, additional expressiveness is gained by the following features: equational pattern matching, user-definable syntax and data, types, subtypes, and partiality, generic types and modules, and support for objects.

3 Pathway Logic

The idea of Pathway Logic is to develop executable formal models of biomolecular processes [8,9,19]. Using the Maude system, formal executable models of processes can be represented and analyzed.

A Pathway Logic model consists of a specification of an initial state and a collection of rules together with the underlying data type specifications. An initial state contains cell components and locations. A collection of rules forms a knowledge base. Such executable models reflect the possible ways a system can evolve. Logical inference and analysis techniques can: (1) simulate possible ways a system could evolve, (2) build pathways in response to queries, and (3) think logically about dynamic assembly of complexes, cascading transmission of signals, feedback-loops, cross talk between subsystems, and larger pathways.

Pathway Logic system has been used to curate models of signal transduction, protease signaling in bacteria, metabolic processes in mycobacterium tuberculosis, glycosylation pathways, and response of the immune system to a generic pathogen [15,21].

The Pathway Logic Assistant is an application that facilitates an interactive visual representation of Pathway Logic models [20]. Graphs show nodes for model rules and components, and edges which connect reactant components to rules and rules to product components. Using Petri nets, the Pathway Logic Assistant provides interactively algorithms for answering reachability queries. In this way, the results are displayed naturally for biologists.

Pathway Logic Assistant can define the path graph of a given initial state from an executable model. Its nodes are the reachable states. Its rules connect these nodes. Paths correspond to possible ways a system can evolve. An execution strategy selects a specific path of the possible solutions.

4 Modeling Cellular Signaling Networks with Rewriting Logic

We briefly describe in this section the Maude specification of Pathway Logic and how to use it. In the next section we will show how to manipulate this specification to perform different analyses. First, data types like proteins, amino acids, and genes are defined as Maude `sorts`, while functions on these sorts are defined by means of `ops`. For example, we define constants for amino acids as:

```
sorts Protein AminoAcid Gene.
ops A C T Y S K P N L M V I F D E R H Q W G : -> AminoAcid [ctor].
```

where the attribute `ctor` indicates that these constants are *constructors*. Relations between sorts are stated by means of subsorts. For example, we can indicate that amino acids are a particular case of proteins as:

```
subsort AminoAcid < Protein.
```

Given a multi-set (a `Soup`) of elements like the proteins, amino acids, and genes above, we define `Location`s to specify the elements in different places of the cell, like the nucleus or the cytoplasm:

```
op {_|_} : LocName Soup -> Location [ctor format (n d d d d d)].
```

Finally, dishes are defined as wrappers of `Soup`, which in this case are not isolated elements but different locations:

```
op PD : Soup -> Dish [ctor].
```

Chemical reactions are defined on sets of locations by means of rewrite rules (with syntax `rl`), that stand for transitions in a concurrent system. For example, we can say that if Bleomycin is found on the outside of the cell (location `XOut`) and we can observe the chromatin (location `CHR`), then DNA strand break (`DSB`) will appear in the chromatin:

```
rl [1770.DSB.irt.Bleomycin] : {XOut | xout Bleomycin} {CHR | chr}
=>                            {XOut | xout Bleomycin} {CHR | chr  DSB}.
```

where the variables `xout` and `chr` stand for any other element that might appear in the corresponding location. Now, we can use the `rew` command to ask Maude to apply rules and check the reachable states from particular dishes. The following example shows the result after applying 5 rewrite steps to an initial dish with Bleomycin outside the cell, the protein H2ax in the chromatin, an empty set of elements in the plasma membrane (`CLm`), no elements stuck to the inside of the plasma membrane (`CLi`), no elements in the cytoplasm (`CLc`), and some extra elements in the nucleus (`NUc`):

```
rew [5] PD({XOut | Bleomycin}{CHR | H2ax}{CLm | empty}{CLi | empty}
          {CLc | empty}{NUc | Tip60 Atm Chek2 Prkdc}).
result Dish: PD({CLm | empty}{CLi | empty}{XOut | Bleomycin}{CLc | empty}
   {NUc | Atm Chek2 [Prkdc - phos(T 2609) phos(S 2056)] [Tip60 - act]}
   {CHR | DSB [H2ax - phos(S 140)]})
```

It is easy to see that, among the elements in the result we have `DSB` and phosphorus bound to `H2ax` in the chromatin.

However, since several different rules can be applied to the same dish to obtain different results, the `rew` command does not provide much information. To solve this problem Maude provides the `search` command, which performs a breadth-first search looking for the pattern given in the command. For example, we can check whether a dish without `DSB` and Bleomycin is reachable from the one above in 10 steps as:

```
search [1,10] PD({XOut | Bleomycin}{CHR | H2ax}{CLm | empty}{CLi | empty}
    {CLc | empty}{NUc | Tip60 Atm Chek2 Prkdc})
    =>* PD({XOut | empty}{CHR | H2ax} S:Soup).
No solution
```

where the variable `S:Soup` abstracts the rest of elements, which are not relevant in this case, and the search option `=>*` stands for zero or more steps. Maude indicates that there are not reachable states which fulfill this condition.

5 Beyond Maude Commands: Reverse Inference

Rewriting logic is reflective [6], which means that it can be faithfully interpreted in itself. Maude efficiently implements reflection in the `META-LEVEL` module [5] [Chap. 14], that provides functions for moving up and down modules and terms, for manipulating modules, and for executing terms at the metalevel. Hence, using the metalevel it is not only possible to perform the rewrite and search shown in the previous section (by means of `metaRewrite` and `metaSearch`, respectively), but also reason on the results. In this way, we implemented a general function that searches for all the common terms in multiple searches [18], hence describing how the subsequent searches modify the results.

Full Maude [5] [Part II] is an extension of Maude created in Maude itself. Full Maude offers an even more powerful module algebra than the one provided in Core Maude, It includes special features for parsing Maude modules. An explicit own module database, combined with the meta-level features, allows us to introduce, remove, modify, and analyze the modules defined by the user. The syntax of existing features can be changed. New kinds of modules and commands can be incorporated.

Full Maude is built on top of the Loop Mode [5] [Chap. 17], which provides a mechanism to read the modules and commands introduced by the user, and to show him the results generated by these commands. These properties facilitate the creation of further extensions, either for extra syntactic constructs, like the

Maude strategy language [12], or new commands, like the `narrowing` search currently available for symbolic execution [4] [Chap. 16].

Thus far, the analysis performed in Pathway Logic has been restricted to those commands available in Core Maude.[1] In [18], we presented a mechanism for going beyond these commands by implementing a function that used Maude metalevel for (i) applying the `metaSearch` command for finding all the possible solutions and (ii) traversing these solutions, keeping those terms that are equal and abstracting (i.e. using a new constant that stands for no similarities) those terms that are different. The main drawback of this function is that it is defined ad-hoc for our analyses, and hence the user must meta-represent the module, the initial term, and the pattern and the condition in the command and, once the result is shown, take it down to see the similarities.

Building up on this idea, in [15] we implemented a new function for using *narrowing*. Narrowing is a generalization of term rewriting that allows to execute terms with variables by replacing pattern matching by unification, for some *unconditional rewriting logic theories without memberships* and that is available in Full Maude. That is, while the standard `search` command in Core Maude allows the user to perform *forward search* as shown in Sect. 4, narrowing allows the user to perform searches starting from terms with variables as:

```
rew [1] PD({XOut | S:Soup} {CHR | P:Protein})
    ~>* PD({XOut | S:Soup} {CHR | DSB H2ax}).
Solution 1
S:Soup --> Bleomycin
P:Protein --> H2ax
```

The possible symptoms leading to the final state are thus studied, while the standard search just gives us the possible outcomes from particular initial states. However, as indicated above, narrowing can only be used under certain requirements: all the operators must have a particular set of attributes (e.g. associative but not commutative operators cannot be used), rules must be unconditional, and all of them must be declared at the top (in our case the top operator is the one for `Dish`, `PD`). Although the current implementation of Pathway Logic fulfills the first two requirements, it fails to fulfill the third one, as shown for the rule in Sect. 4. For this reason, and given that the structure for dishes is fixed, it is possible to implement a function at the meta-level that modifies the rules in the modules and uses the modified module into the `metaNarrowSearch` command. However, the architecture underlying this function was again ad-hoc and was difficult to allow further modifications.

This is why we have developed a general architecture that allows new commands in an easy and scalable way. We have defined a `PLE-SYNTAX` module (from Pathway Logic Extended) where new commands can easily be defined.

[1] The Pathway Logic Assistant allows the user to perform more complex analyses by combining different formalisms, but within the rewriting logic it is also restricted to Core Maude. Hence, our extension would also improve the range of analyses supported by the PLA.

In particular, we can add the `top` command for transforming the rules in a given module to the top or `common` for computing similarities. Then, the `PLE-COMMAND-PROCESSING` module is expected to import all those modules defining the behavior of the commands above. Then the user must define a function for parsing each command and returning the expected result. Next, the `PLE-DATABASE-HANDLING` module is in charge of defining rewrite rules for controlling the interaction between the user and the Loop. Each command in the syntax should have one or more rules that will use the parsing commands in the `PLE-COMMAND-PROCESSING` and update the current state of the loop. Finally, the `PLE` module deals with the I/O attributes in the loop, consuming the input from the user and showing the results generated by the rules in the `PLE-DATABASE-HANDLING` module. This structure is available at https://github.com/ariesco/pathway.

We give a glimpse of our code below. The function `rule2Top` is in charge of defining the terms of each rule at the top. The first equation indicates that, if the terms are defined at the level of locations (checked by extracting the type from the meta-reduced term, to make sure we get the least one) instead of dishes, then we use the auxiliary function `encapsulateLocs` to transform it. Otherwise (indicated by the `owise` attribute) the rule is defined at the top and we return the same function:

```
op rule2Top : Module Rule -> Rule.
ceq rule2Top(M, rl T => T' [AtS] .) = rl NT => NT' [AtS].
 if getType(metaReduce(M, T)) == 'Locations /\
    NT  := encapsulateLocs(M, T) /\
    NT' := encapsulateLocs(M, T') .
eq rule2Top(M, R) = R [owise] .
```

The auxiliary function `encapsulateLocs` just returns the term obtained by normalizing the dish where a variable for matching all possible locations has been added:

```
op encapsulateLocs : Module Term -> Term.
eq encapsulateLocs(M, T) = getTerm(metaNormalize(M,
                                   'PD['__[T, 'L@#:Locations]])).
```

6 Conclusions

The growth of genomic sequence information combined with technological advances in the analysis of global gene expression has revolutionized research in biology and biomedicine [17]. Various models for the computational analysis of cellular signaling networks have been proposed to simulate responses to specific stimuli [23]. Symbolic models are based on formalisms that provide a language to represent the states of a system; mechanisms to model their changes and tools for analysis based on computational or logical inference.

Pathway Logic [19] formalizes models that molecular biologists can use to think about signaling pathways and their behavior, allowing them to computationally formulate questions about their dynamics and outcomes. Pathway Logic is based on rewriting logic and Maude. Rewriting logic procedures are powerful symbolic methods that can be applied in order to understand naturally the dynamics of complex biological systems. As a consequence of the reflexion of rewriting logic [6], an important feature of Maude is its metalevel, that allows us to manipulate Maude modules and terms as standard data [5].

In this work we reveal the application of a rewriting logic procedure based in logic language Maude to the dynamic modeling of biological signaling pathways. On the one hand, our system allows us to perform reverse inference. By using narrowing, we can obtain the initial states that reach to desired final states. On the other hand, we propose a general structure which can be used as basis for further commands, like extensions for dealing with strategies or adding stochastic information for the rules. In conclusion, we travel through complex and dynamic cellular signaling processes by using a logical system.

References

1. Asthagiri, A.R., Lauffenburger, D.A.: A computational study of feedback effects on signal dynamics in a mitogen-activated protein kinase (MAPK) pathway model. Biotechnol. Prog. **17**(2), 227–239 (2001)
2. Blinov, M.L., Faeder, J.R., Goldstein, B., Hlavacek, W.S.: BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. Bioinformatics **20**(17), 3289–3291 (2004)
3. Bouhoula, A., Jouannaud, J.P., Meseguer, J.: Specification and proof in membership equational logic. Theor. Comput. Sci. **236**, 35–132 (2000)
4. Clavel, M., Durán, F., Eker, S., Escobar, S., Lincoln, P., Martí-Oliet, N., Meseguer, J., Talcott, C.: Maude Manual (Version 2.7), March 2015. http://maude.cs.illinois.edu/w/images/1/1a/Maude-manual.pdf
5. Clavel, M., Durán, F., Eker, S., Lincoln, P., Martí-Oliet, N., Meseguer, J., Talcott, C.L.: All about maude - a high-performance logical framework: how to specify, program and verify systems in rewriting Logic. Lecture Notes in Computer Science, vol. 4350. Springer (2007). http://dx.doi.org/10.1007/978-3-540-71999-1
6. Clavel, M., Meseguer, J., Palomino, M.: Reflection in membership equational logic, many-sorted equational logic, Horn logic with equality, and rewriting logic. Theor. Comput. Sci. **373**(1–2), 70–91 (2007)
7. Danos, V., Laneve, C.: Formal molecular biology. Theor. Comput. Sci. **325**(1), 69–110 (2004)
8. Eker, S., Knapp, M., Laderoute, K., Lincoln, P., Meseguer, J., Sönmez, M.K.: Pathway logic: symbolic analysis of biological signaling. In: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E. (eds.) Proceedings of the 7th Pacific Symposium on Biocomputing, PSB 2002, Lihue, Hawaii, USA, 3–7 January 2002, pp. 400–412, January 2002. http://helix-web.stanford.edu/psb02/eker.pdf
9. Eker, S., Knapp, M., Laderoute, K., Lincoln, P., Talcott, C.: Pathway logic: executable models of biological networks. In: Gadducci, F., Montanari, U. (eds.) Proceedings of the Fourth International Workshop on Rewriting Logic and its Applications, WRLA 2002, Pisa, Italy, 19–21 September 2002. Electronic Notes in Theoretical Computer Science, vol. 71, pp. 144–161. Elsevier (2004)

10. Faeder, J.R., Blinov, M.L., Hlavacek, W.S.: Rule-based modeling of biochemical systems with BioNetGen. Methods Mol. Biol. **500**, 113–167 (2009)
11. Hwang, W., Hwang, Y., Lee, S., Lee, D.: Rule-based multi-scale simulation for drug effect pathway analysis. BMC Med. Inform. Decis. Mak. **13**(Suppl 1), S4 (2013)
12. Martí-Oliet, N., Meseguer, J., Verdejo, A.: Towards a strategy language for Maude. In: Martí-Oliet, N. (ed.) Proceedings of the Fifth International Workshop on Rewriting Logic and its Applications, WRLA 2004, Barcelona, Spain, 27 March–4 April 2004. Electronic Notes in Theoretical Computer Science, vol. 117, pp. 417–441. Elsevier (2004)
13. Meseguer, J.: Conditional rewriting logic as a unified model of concurrency. Theor. Comput. Sci. **96**(1), 73–155 (1992)
14. Meseguer, J.: Twenty years of rewriting logic. J. Log. Algebr. Program. **81**(7–8), 721–781 (2012)
15. Riesco, A., Santos-Buitrago, B., De Las Rivas, J., Knapp, M., Santos-García, G., Talcott, C.: Epidermal growth factor signaling towards proliferation: modeling and logic inference using forward and backward search. BioMed Res. Int. **2017**, 11 (2017)
16. Santos-García, G., De Las Rivas, J., Talcott, C.L.: A logic computational framework to query dynamics on complex biological pathways. In: Saez-Rodriguez, J., Rocha, M.P., Fdez-Riverola, F., De Paz Santana, J.F. (eds.) 8th International Conference on Practical Applications of Computational Biology & Bioinformatics, PACBB 2014, 4–6 June 2014, Salamanca, Spain. Advances in Intelligent Systems and Computing, vol. 294, pp. 207–214. Springer (2014)
17. Santos García, G., Talcott, C.L., De Las Rivas, J.: Analysis of cellular proliferation and survival signaling by using two ligand/receptor systems modeled by pathway logic. In: Abate, A., Safránek, D. (eds.) Hybrid Systems Biology - Fourth International Workshop, HSB 2015, Madrid, Spain, 4–5 September 2015. Revised Selected Papers. Lecture Notes in Computer Science, vol. 9271, pp. 226–245. Springer (2015)
18. SantosGarcía, G., Talcott, C.L., Riesco, A., SantosBuitrago, B., De Las Rivas, J.: Role of nerve growth factor signaling in cancer cell proliferation and survival using a reachability analysis approach. In: Mohamad, M.S., Rocha, M.P., Fdez Riverola, F., Mayo, F.J.D., De Paz, J.F. (eds.) 10th International Conference on Practical Applications of Computational Biology & Bioinformatics, PACBB 2016, 1–3 June 2016. Advances in Intelligent Systems and Computing, vol. 477, pp. 173–181. Springer (2016)
19. Talcott, C.L.: Pathway Logic. In: Bernardo, M., Degano, P., Zavattaro, G. (eds.) Formal Methods for Computational Systems Biology, 8th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2008, Bertinoro, Italy, 2–7 June 2008, Advanced Lectures. Lecture Notes in Computer Science, vol. 5016, pp. 21–53. Springer (2008)
20. Talcott, C.L., Dill, D.L.: The pathway logic assistant. In: Plotkin, G. (ed.) Proceedings of the Third International Workshop on Computational Methods in Systems Biology, pp. 228–239 (2005)
21. Talcott, C.L., Eker, S., Knapp, M., Lincoln, P., Laderoute, K.: Pathway Logic modeling of protein functional domains in signal transduction. In: Altman, R.B., Dunker, A.K., Hunter, L., Jung, T.A., Klein, T.E. (eds.) Proceedings of the 9th Pacific Symposium on Biocomputing, PSB 2004, Fairmont Orchid, Hawaii, USA, 6–10 January 2004, pp. 568–580. World Scientific, January 2004
22. Tenazinha, N., Vinga, S.: A survey on methods for modeling and analyzing integrated biological networks. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(4), 943–958 (2011)
23. Weng, G., Bhalla, U.S., Iyengar, R.: Complexity in biological signaling systems. Science **284**(5411), 92–96 (1999)

Skin Temperature Monitoring to Avoid Foot Lesions in Diabetic Patients

A. Queiruga-Dios[1(✉)], J. Bullón Pérez[2], A. Hernández Encinas[1], J. Martín-Vaquero[1], A. Martínez Nova[3], and J. Torreblanca González[4]

[1] Department of Applied Mathematics, University of Salamanca, Salamanca, Spain
{queirugadios,ascen,jesmarva}@usal.es
[2] Department of Chemical and Textile Engineering, University of Salamanca, Salamanca, Spain
perbu@usal.es
[3] Department of Nursing, University of Extremadura, Plasencia, Spain
podoalf@unex.es
[4] Department of Applied Physics, University of Salamanca, Salamanca, Spain
torre@usal.es

Abstract. Foot temperature monitoring is of great importance in diabetic patients, as they are prone to complications such as peripheral neuropathy and vascular insufficiency. In recent years, the study of different non-invasive procedures to monitor healthy indicators is growing, due to the advances in mobile devices, micro-sensors, and also wireless sensors. The health monitoring systems are used by medical staff and also by patients when they are out of the hospital, in their personal environment. This paper presents a preliminary work to identify the specific points on the feet where the temperature sensors should be positioned. We have developed an statistical analysis of the data obtained by a thermal camera from healthy people.

Keywords: Wearable · Foot temperature · Statistical analysis

1 Introduction

The study of wearable Sensor-Based Systems for health monitoring, known as Wearable Health-Monitoring Systems (WHMS) is getting more importance in the industry, the medical specialists, and the scientific and research community in recent years [14]. Motivated by the increase in health expenditures and because of the recent technological advances in miniature biosensor devices, health monitoring systems can play a significant role in reducing hospitalization, excessive work load on the medical staff, the time of consultation, waiting lists, etc. There is a need to monitor patients health status while they are out of the hospital.

The monitoring of physiological parameters like blood pressure, heart rate, body and skin temperature, oxygen saturation, respiration rate, electrocardiogram, etc. is done by bio-micro-sensors or wearable medical systems that may

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_14

include a huge amount of components: sensors, wearable materials, smart textiles, actuators, power supplies, wireless communication modules and links, control and processing units, interface for the user, software, and advanced algorithms for data extracting and decision making [14].

Diabetic disease is one of the most important health problems, both because of its extraordinary frequency and because of its enormous socio-economic repercussions. One of the most feared problems, as it affects the quality of life of diabetic patients, is the appearance of ulcers in their feet, as a sequel to two of the most common chronic complications of this disease: peripheral neuropathy and vascular insufficiency. The combination of these factors, neuropathy and angiopathy, together with the high risk of infection and the intrinsic and extrinsic pressures due to bone malformations in the feet, are the final triggers of the diabetic foot. To be more precise, peripherical neuropathy is the most important risk factor that could develop foot ulcers in diabetic patients [5].

The prevalence of ulcers varies according to sex, age and population from 2.4% to 5.6%. It has been estimated that at least 15% of diabetics will suffer from foot ulcerations during their lifetime. It is also estimated that about 85% of diabetics suffering from amputations have previously had an ulcer [9].

World Health Organization estimates that the prevalence of Diabetes Mellitus (DM) at the start of the twenty-first century was 2.1% of the world's population. That is, about 125 million people. The foot temperature monitoring can significantly limit the rates of re-ulceration in diabetes, as the use of simple temperature measurement devices, like thermometers, serves as prevention tools, helping patients identify potentially damaging limbs inflammation ([2,10,11]). The combination of routine measurements and the use of thermal techniques may improve the quality of research in diabetes and facilitate the detection, monitoring and control of diabetic foot problems [4]. To avoid limbs damages, we consider the possibility of a prototype as a smart sock that will contain several sensors to take a more real temperature data. These sensors will be placed into the sock plants. The proposed device is of a wearable sensor-based system for foot temperature monitoring capable of continuously or intermittently measuring the foot temperature of the patient at one or more locations of the foot. That device will be possible using smart textiles (also known as electronic or e-textiles), i.e. "textiles that can detect and react to stimuli and conditions of the environment, as well as mechanical, thermal, chemical, electrical or magnetic stimuli" [6]. It is therefore the physical integration of an intelligent system with a textile substrate is a system to monitor physiological signals.

2 Wearable Systems for Foot Temperature Monitoring

Systems that monitor health can be classified into 3 categories [3]: Remote health monitoring systems (RHMS) are those with remote access or those that can send data to or from a remote location; mobile health monitoring systems (MHMS) refer to mobile phones, PDAs, handheld devices, etc.; wearable health-monitoring systems (WHMS), as the name implies, these are devices that can be

worn and used by patients and which consist of WHMS, RHMS and/or MHMS. When sensors become integrated into a garment or complement are called wearables, a term that is also used to refer to garments and accessories that have integrated sensors. Thus, these are the ones that give rise to intelligent fabrics, smart-textiles or e-textiles.

These devices must meet certain strict medical criteria and operate under ergonomic constraints and significant hardware limitations. Research in this area focuses mainly on producing clothing with features that contribute to improve or facilitate the lives of its users.

There are certain professions in which the worker must be very active and danger-prone, such as soldiers, firefighters, police personnel, miners, divers or astronauts in space. Considering the mobility and vulnerability, it is important to monitor the health status and the geo-location of the workers to ensure the completion of the assigned work. For this reason, there has been a significant change in the development of clothing as the introduction of sensors into the clothing worn by staff. A wearable physiological monitoring system consists of a garment with embedded sensors, data acquisition and processing hardware with the required embedded software, and a remote monitoring station to study the health of the wearer.

Conventional sensors and medical instruments can not be used as wearable applications of physiological monitoring, since they are difficult to carry for long periods, and they cause discomfort to the wearer. Thus, the gels used on the electrodes cause irritation when they dry out or when are used for a long time. In addition, the contact resistance between the electrode and the skin changes over time, thereby degrading the quality of the signal obtained. In conventional monitoring systems there are many cables for the acquisition of physiological signals and the system is too bulky to be used in wearable applications [13].

Many wearable physiological monitoring systems have been developed: a wristband that is a wearable monitoring and medical alert system for high cardiac-respiratory patients [1]. To perform an electrocardiogram (ECG) and measure blood pressure requires the attention of the subject. Vital parameters are not transmitted continuously. A new and discreet wearable, multi-parameter system was developed as an ambulatory physiological monitoring system for space and earth applications called BodyGuard, which had the ability to continuously record the 2-lead ECG readings, the respiratory rate through impedance plethysmography, cardiac rhythm, hemoglobin oxygen saturation, body or ambient temperature, etc. [12]. The Georgia Tech intelligent T-shirt known as Georgia Tech Wearable Motherboard (GTWM) was characterized by a wearable motherboard that incorporated numerous vital parameters into clothing and could be easily and comfortably worn by soldiers ([8,15]), but it used conventional electrodes to perform the ECG, which became corrupted with noise during movement of the subject. In addition, it did not perform stress measures, which is considered a vital measure.

In addition to other similar wearables, fabric-based wearable systems have also been used, such as the MagIC (*Maglietta Interattiva Computerizzata*), which

measured cardiorespiratory and movement signals in patients with heart problems. The MagicIC system was tested on subjects who moved freely, whether at home, at work, while driving or riding a bicycle, and also in microgravity conditions during a parabolic flight. Preliminary results showed good signal quality over most periods when measurements were taken and a correct identification of arrhythmia events and correct estimation of mean heart rate [7].

Therapeutic footwear, to monitor foot skin temperatures, were suggested in [5] to prevent foot ulcers in at-risk patients with diabetes. Several authors suggest that the identification of temperature asymmetries could be consider as important factor to identify early signs of disease ([17,18]). Recently a team at the University of Manchester has developed a system that measures temperature in diabetic feet to study the etiology of diabetic foot ulcerations [16]. A insole with the temperature sensors is connected to the myRIO, which records the temperature measurements and stores them in a USB.

3 Statistical Data Analysis

The collected data studied in this paper come from the thermal indices, taken to 70 healthy people, in the sole and the dorsal of the right and left foot as indicated in Fig. 1.

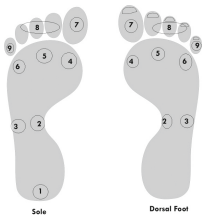


Fig. 1. Position indices.

The plant and dorsal areas correspond to the same position, except for the number one that is only in the sole: (1) heel, (2) medial midfoot, (3) lateral midfoot, (4) first metatarsal head, (5) central metatarsal heads, (6) fifth metatarsal head, (7) first finger, (8) central fingers, (9) fifth finger. The data were taken with a thermal FLIR E60bx camera that take images with the following characteristics: Resolution: 320 × 240 pixels; total pixels: 76,800; thermal sensitivity: $< 0.045\,^{\circ}$C; accuracy: $\pm 2\,^{\circ}$C or $\pm 2\%$ of reading; temperature range: $-4\,^{\circ}$F to $+248\,^{\circ}$F ($-20\,^{\circ}$C to $+120\,^{\circ}$C). The thermal images obtained with a thermographic should be interpreted from the different materials and circumstances that influence the temperature readings. Some of those factors are the thermal conductivity, the reflection and the emissivity. Emissivity is the efficiency with which an object emits infrared radiation. This is highly dependent on material properties. It is essential to set the right emissivity in the camera. In other case

the temperature measurements will not be correct. The FLIR thermal imaging cameras have predefined emissivity settings for lots of materials, and the rest can be found in an emissivity table. The right emissivity setting for human skin is around 0.97. There are six key requirements that should be evaluated to work with a infrared camera: The camera resolution or image quality (a resolution of 320×240 or 640×480 pixels deliver superior image quality), the thermal sensitivity (the difference in temperatures that can detect), the accuracy (the margin of error within which the camera will operate, the current industry standard for accuracy is $\pm 2\%$ / $\pm 2\,^{\circ}\mathrm{C}$), the camera functions (emissivity and reflected temperature values), the software, and the training demands.

We initially had establish different points for taking temperature measures: 9 indices for the foot's sole and 8 for the dorsal part, of course, on both feet. We call them with IJK, being I = P (Sole) or I = D (Dorsal); J = D (Right) or J = I (Left). K takes values from 1 to 9 according to the index concerned (e.g. PD1 is the index of the right foot plant corresponding to the heel).

There is a correlation higher than 0.8 between the plantar and dorsal indexes at the same position, so it should therefore be sufficient to consider one or the other. In our case we have analyzed only the plantar.

We started with a basic statistical study of the collected data, which it has been calculated the mean value (mean), standard deviation (sd), typical error (se mean), interquartile rate (IQR), coefficient of variation (cv), skewness, kurtosis, and the quartiles. Figure 2 represents the different Box Plot from all the data from the right and left soles, and right and left dorsal feet.

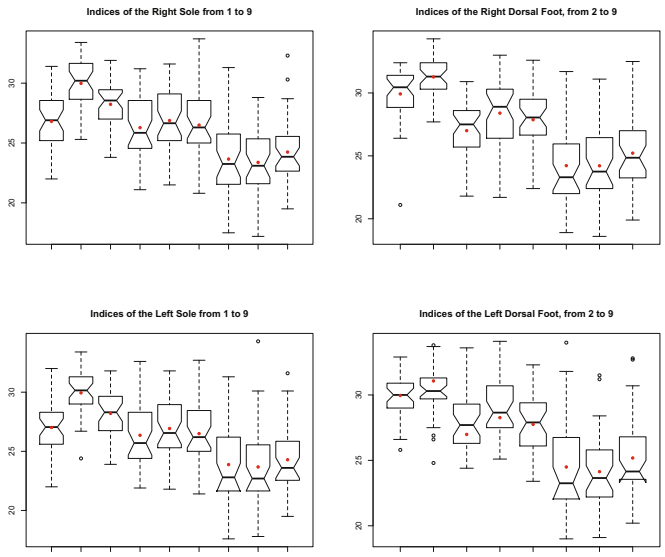


Fig. 2. Different Box Plot from collected data.

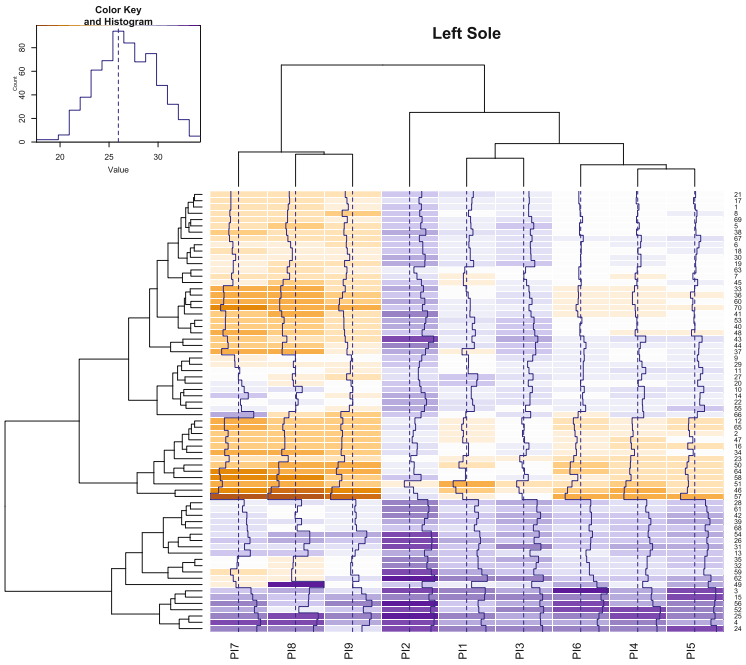


Fig. 3. Indices heat map combined with a dendogram.

Some basic statistical analysis shows that the coefficient of variation is small in all cases. The higher values (10 %) correspond to the data on the fingers. The standard errors for the mean (se mean), IQR, and sd, practically in all cases the greatest value corresponds to the index 7 (big toe), followed by the 8 and 9 (the other fingers). So maybe the finger data should not be used to extrapolate the results. The lower errors correspond to indices 3 and 2 (central part of the foot), as well as the lower IQR and sd. It is necessary to take into account that the diabetic people have special sensitivity in the toes, reason why it is necessary to try to improve the taking of measures in them.

In our study we are going to place a series of sensors in a sock in order to take measurements of temperatures in different places of the feet. With the arrangement of the clusters given by a dendogram we can get an idea of where to put them.

We use a heat map combined with a dendogram (Fig. 3), which is a way of grouping items based on distance or similarity between them. As a result of the cluster calculation, the rows of the heat map are rearranged to correspond with it. This order give us a idea of where to put the sensors in the sock. In all cases the first ones that are joined are the indices 5 and 6 and these with the 4 (except in the left plant that are 4-5 and then 6). The following are 8 and 9 and these two with 7 (except for the left dorsal that are 7-8 and then 9). In the plant the

indices 1-3 are joined, and the same occurs in the dorsal for 2-3. As can be seen, index 2 (medial midfoot) is the most important of the plant.

4 Conclusions and Further Work

We have analyzed the data collected from healthy individuals. This temperature data was gathered with the help of a thermal camera. We have recorded temperature data from different points of feet, as feet could be affected by peripheral neuropathy and autonomic neuropathy in diabetic patients.

We are developing a prototype inside a sock that will contains several sensors to take a more real temperature. As these sensors will be placed into the sock plants, we have focused our analysis in the data from the sole. From the result of the dendogram, and thinking about 4 sensors, these would be place in the medial midfoot (index 2); the fingers (indices 7, 8, and 9); between the heel and the medial lateral midfoot (indices 1 and 3); and in the metatarsal heads (indices 4, 5, and 6).

It is necessary to take into account what the sensors are going to be use for. Although the dendogram (in Fig. 3) shows the medial midfoot index is the most important of the plant, we will have to make decisions about the best position for the sensors depending on the needs of the patients. Furthermore, the data from positions 7, 8, and 9 (fingers) are those with the highest values for the typical error of the mean and for the standard deviation.

At the moment we are analyzing data from a control group. We want to examine some factors that may affect the study of foot temperature, since our ultimate goal is to obtain models for diabetic patients.

Acknowledgements. The authors acknowledge support from the Fundación Memoria D. Samuel Solórzano Barruso, through the grant FS/14-2016.

References

1. Anliker, U., Ward, J.A., Lukowicz, P., Troster, G., Dolveck, F., Baer, M., Keita, F., Schenker, E.B., Catarsi, F., Coluccini, L., et al.: Amon: a wearable multiparameter medical monitoring and alert system. IEEE Trans. Inf. Technol. Biomed. **8**(4), 415–427 (2004)
2. Armstrong, D.G., Holtz-Neiderer, K., Wendel, C., Mohler, M.J., Kimbriel, H.R., Lavery, L.A.: Skin temperature monitoring reduces the risk for diabetic foot ulceration in high-risk patients. Am. J. Med. **120**(12), 1042–1046 (2007)
3. Baig, M.M., Gholamhosseini, H.: Smart health monitoring systems: an overview of design and modeling. J. Med. Syst. **37**(2), 1–14 (2013)
4. Bharara, M., Cobb, J., Claremont, D.: Thermography and thermometry in the assessment of diabetic neuropathic foot: a case for furthering the role of thermal techniques. Int. J. Lower Extremity Wounds **5**(4), 250–260 (2006)
5. Bus, S., Netten, J., Lavery, L., Monteiro-Soares, M., Rasmussen, A., Jubiz, Y., Price, P.E.: Iwgdf guidance on the prevention of foot ulcers in at-risk patients with diabetes. Diab. Metab. Res. Rev. **32**(S1), 16–24 (2016)

6. Cherenack, K., Zysset, C., Kinkeldei, T., Münzenrieder, N., Tröster, G.: Woven electronic fibers with sensing and display functions for smart textiles. Adv. Mater. **22**(45), 5178–5182 (2010)
7. Di Rienzo, M., Rizzo, F., Parati, G., Brambilla, G., Ferratini, M., Castiglioni, P.: Magic system: a new textile-based wearable device for biological signal monitoring. Applicability in daily life and clinical setting. In: 2005 IEEE 27th Annual Conference on Engineering in Medicine and Biology, pp. 7167–7169. IEEE (2005)
8. Gopalsamy, C., Park, S., Rajamanickam, R., Jayaraman, S.: The wearable motherboard^{TM}: the first generation of adaptive and responsive textile structures (arts) for medical applications. Virtual Reality **4**(3), 152–168 (1999)
9. National Institutes of Health: National institute of diabetes and digestive and kidney diseases. Diabetes in America, 2nd edn. NIH Publication (95-1468) (1995)
10. Lavery, L.A., Higgins, K.R., Lanctot, D.R., Constantinides, G.P., Zamorano, R.G., Armstrong, D.G., Athanasiou, K.A., Agrawal, C.M.: Home monitoring of foot skin temperatures to prevent ulceration. Diabetes Care **27**(11), 2642–2647 (2004)
11. Lavery, L.A., Higgins, K.R., Lanctot, D.R., Constantinides, G.P., Zamorano, R.G., Athanasiou, K.A., Armstrong, D.G., Agrawal, C.M.: Preventing diabetic foot ulcer recurrence in high-risk patients. Diabetes Care **30**(1), 14–20 (2007)
12. Mundt, C.W., Montgomery, K.N., Udoh, U.E., Barker, V.N., Thonier, G.C., Tellier, A.M., Ricks, R.D., Darling, R.B., Cagle, Y.D., Cabrol, N.A., et al.: A multiparameter wearable physiologic monitoring system for space and terrestrial applications. IEEE Trans. Inf. Technol. Biomed. **9**(3), 382–391 (2005)
13. Pandian, P., Mohanavelu, K., Safeer, K., Kotresh, T., Shakunthala, D., Gopal, P., Padaki, V.: Smart vest: wearable multi-parameter remote physiological monitoring system. Med. Eng. Phys. **30**(4), 466–477 (2008)
14. Pantelopoulos, A., Bourbakis, N.G.: A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**(1), 1–12 (2010)
15. Park, S., Jayaraman, S.: Enhancing the quality of life through wearable technology. IEEE Eng. Med. Biol. Mag. **22**(3), 41–48 (2003)
16. Reddy, P.N., Cooper, G., Weightman, A., Hodson-Tole, E., Reeves, N.: An in-shoe temperature measurement system for studying diabetic foot ulceration etiology: preliminary results with healthy participants. Procedia CIRP **49**, 153–156 (2016)
17. Tse, J., Rand, C., Carroll, M., Charnay, A., Gordon, S., Morales, B., Vitez, S., Le, M., Weese-Mayer, D.: Determining peripheral skin temperature: subjective versus objective measurements. Acta Paediatr. **105**(3), e126–e131 (2016)
18. Van Netten, J.J., Prijs, M., van Baal, J.G., Liu, C., van Der Heijden, F., Bus, S.A.: Diagnostic values for skin temperature assessment to detect diabetes-related foot complications. Diab. Technol. Ther. **16**(11), 714–721 (2014)

Multidimensional Feature Selection and Interaction Mining with Decision Tree Based Ensemble Methods

Lukasz Krol(✉) and Jonna Polanska

Faculty of Automatic Control, Electronics and Computer Science,
Data Mining Group, Silesian University of Technology, Gliwice, Poland
{lukasz.krol, joanna.polanska}@polsl.pl

Abstract. This paper demonstrates capability of detecting strong synthetic benchmark feature interactions in a set of mixed categorical and continuous variables using a modified version of Monte Carlo Feature Selection algorithm. MCFS's original way of detecting feature interactions relying on the analysis of structure of trained decision trees is compared with our modified approach consisting of a series of variable permutations combined with a decomposition of feature total effect to main effect and interaction effects. A comparison with unmodified MCFS, which by default handles only classification problems using C4.5 decision trees, shows that the new approach is slightly more robust. Furthermore, the decomposition approach is flexible by allowing to plug in different types of models to MCFS. This opens a way to handle high-throughput supervised feature selection and interaction mining problems for classification, regression and censored survival decision vector.

Keywords: Feature selection · Feature interaction · Dimensionality reduction · Classification · Decision trees · RandomForests · Random forests · Extremely randomized trees · Monte carlo feature selection

1 Background

In data analysis, the effort is often put on the sheer volume of analyzed data, which is usually caused by the number of collected observations. On the other hand, the data provided by high throughput biological experiments is characterized by feature to observation imbalance. For classical RNA microarrays, the number of features is usually several thousand. By increasing the resolution and measuring the expression of individual transcripts through RNA-sequencing, it is possible to increase the number of features to hundreds of thousands. The benefit of this increase of resolution is assessed by projects like [1]. Analyzing genomic data like Single Nucleotide Polymorphisms (SNP), Copy Number Variations or Methylation Sites can provide a number of features in the magnitude of several million. Unfortunately, this increase of number of variables is not followed by an increase of the number of observations, which usually does not exceed few thousand, even for the largest initiatives [2].

High dimensionality of the collected data combined with low number of observations is a cause of large amount of False Discoveries (FD) and low repeatability of

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_15

findings. One approach to assuring reliable results would be to perform statistical tests for each variable, and correct the results for multiple testing. More liberal procedures [3, 4] than the conservative Bonferroni correction [5] would be necessary to provide enough power for high-dimensional scenarios. This approach would however suffer from ignoring feature interactions. When dealing with mixed categorical, discrete and continuous variables, comparing p-values from different types of tests could also lead to introducing a bias in favor of certain classes of variables.

An alternative is to use machine learning based methods. In the simplest scenario, Sequential Feature Selection methods may be used. Its usefulness for high dimensional data is however questionable. One solution for this problem are ensemble methods providing a level of randomization for selecting feature groups in order to test multiple feature combinations, limit overfitting, and allow weaker features to work without the influence of stronger ones. A well-known example is the RandomForests [6] algorithm. Although primarily a classification algorithm, it provides a set of feature importance metrics. It has been successfully used for SNP data, and is able to capture the added value of feature interactions, however the ability of detecting the interactions is decreasing when the number of features increases [7]. Attempts to extend these metrics to report on pairs of features were made by Bureau et al. [8]. These metrics capture the combined total effects of two features rather than the interaction effect however.

Monte Carlo Feature Selection (MCFS) [9] is a decision tree based supervised feature selection algorithm designed to provide a human-readable list of features. Its subsequent versions [10, 11] have been enhanced with the ability to provide an explicit list of feature interactions for the purpose of visualizing them in the form of 'Interaction Networks'.

MCFS is available as the 'rmcfs' R package. As MCFS can be easily parallelized, the package allows for running the computations using multiple threads on a single machine. For larger datasets, or when running permutation tests in order to assess statistical significance, it is however worth to extend the level of parallelism beyond a single machine. Such implementation was recently successfully created in our team [12]. It is characterized by almost linear speedup when increasing the number of processors and has been successfully tested in systems as large as 192 cores. It is not ready for official release yet, however an evaluation version can be obtained from us upon request.

2 Research Goal

MCFS is a powerful and scalable algorithm for feature selection and interaction mining. It has however some fields for extension, as well as potential drawbacks. It is relying on decision trees, or more specifically on feature importance and interaction strength metrics calculated using C4.5 trees. This limits the range of feature selection problems to classification ones, like finding networks of features impacting binary survivability (death/life). Interested in the possibility of performing feature selection and interaction mining in high throughput datasets with continuous or censored survival time decision vector, we have developed a feature importance decomposition methodology that can be used with multiple model metrics like Weighted Accuracy,

Mean Absolute Error or Concordance Index. Our implementation of MCFS [12] has been modified to include these metrics. As a preliminary comparison study, it has been compared with original MCFS in an area which both approaches can handle, that is a categorical decision vector. Achieving at least as good results as the original MCFS opens the way to generalization.

While feature selection in high-dimensional datasets is an extensively investigated area, there are – to our best knowledge – no state of the art methods aimed at providing an explicit list of interactions allowing to create a human readable interaction graph. This makes MCFS generalization attempt a potentially fruitful endeavor.

3 Materials

As the objective of the study was to examine the impact of the chosen interaction detecting approach, the data must have been known and predictable. All the tests have been performed using synthetic data - six types of interacting as well as independent features have been prepared. They are presented in Fig. 1. All of the features are designed for a two-level classification problem with equal number of observations for each class, so that they can be analyzed together in a single dataset. The features are generated randomly using either normal distribution or uniform distribution. Features generated using uniform distribution are also present in categorical variants. Categorical versions of the features do not assume ordering of levels.

In addition to the strong benchmarking features, there are also two types of noise modeled. Continuous noise features are generated from standardized normal distribution, while discrete noise features from uniform discrete distribution with 3 levels.

The features were utilized to populate two datasets – **A** and **B**. **A** represents an unlikely scenario of a dataset composed entirely of interacting features. It is composed of 14 variables: 4 numerical and 3 categorical pairs of variables over 1000 observations. It is designed to test how each one of the feature-pairs scores in comparison to others in an easy scenario. Dataset **B** is more realistic. It contains all the interacting features (14), features from univariate distributions (2) and noise variables (9984). The presence of noise combined with the competition of strong features makes the discovery of interdependent features much more difficult.

4 Methods

Through this section, we will be referring to feature **total effects**, **main effects** and **interaction effects**. **Total effect** represents the overall usefulness of a feature in the presence of all the other features in a dataset. **Main effect** represents the strength of a feature alone. **Interaction effect** represents the added value of using two features together. A dataset of n features contains n different total effects, n main effects and $\binom{n}{2} = \frac{n^2-n}{2}$ interaction effects. Only second order interactions are considered.

Features and feature interactions are mined using the methodology of MCFS [9, 11]. The feature space is sampled **s** times for **m** features. Each feature subspace is then

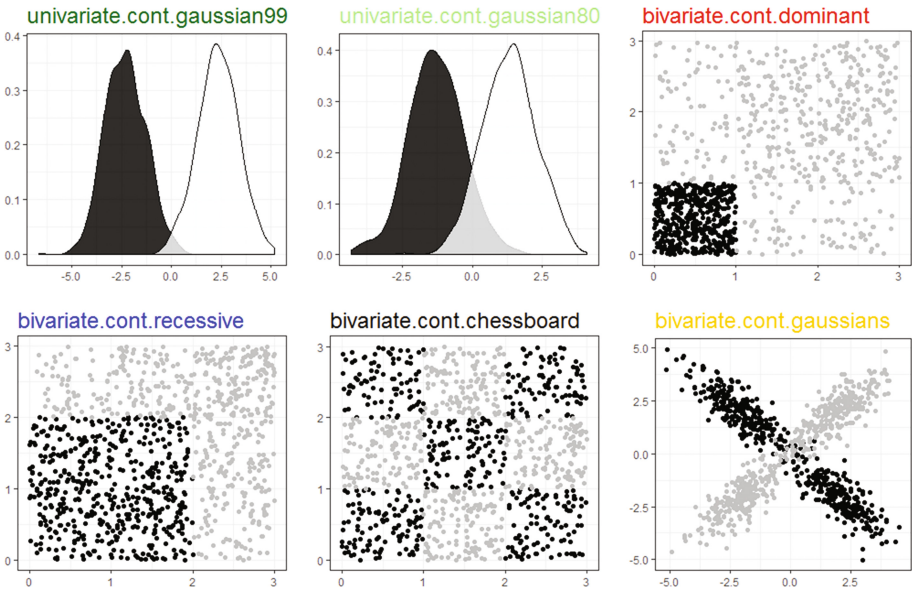


Fig. 1. The classes of features used in the study. Name of each feature consists of three parts. The first part informs whether the feature is designed as a single variable or a pair of interacting variables. Single features are plotted using probability density plots, pairs of interacting variables are visualized with scatter plots. The second part of feature name represents the nature of the feature – whether it is a numerical feature, or categorical one. The third part is the actual name of the feature. The 'dominant', 'recessive' and 'chessboard' features are also present in categorical variants.

split into a training and test set **t** times. A classifier is trained on the training set. The classifier, test set and training set can then all three serve to calculate partial scores of **total effect** and **interaction effect**. This is where our modifications are introduced.

Our way of calculating the total effect is almost the same as the basic method for calculating feature importance in RandomForests. Feature i is being permuted in the test set a number of times. The difference between the model score (weighted accuracy for classification) for the unpermuted dataset and the average weighted accuracy for permuted sets is being reported as the total effect of feature i. Our decomposition idea is that we consider the total effect of feature i to be the sum of the feature's main effect and all the interactions with other features.

$$total\ effect_i = main\ effect_i + \sum_j^{j=1..m;j\neq i} interaction\ effect_{i,j} \tag{1}$$

For m features, m equations of type (1) can be drawn. There are however m unknown main effects and $\binom{m}{2}$ unknown interaction effects, so the system is obviously lacking information. A solution for providing the missing information is to systematically

perform permutations of all $\binom{m}{2}$ pairs of features together, thus providing the missing $\binom{m}{2}$ Eq. (2). Permuting two features together impacts their both main effects, interaction between them, and all the interactions with all the other features.

$$\begin{aligned} &total\ effect_{i,k} = main\ effect_i + main\ effect_k + interaction\ effect_{i,k} \\ &+ \sum\nolimits_j^{j=1..m;j\neq i,j\neq k} interaction\ effect_{i,j} + \sum\nolimits_j^{j=1..m;j\neq i,j\neq k} interaction\ effect_{k,j} \end{aligned} \tag{2}$$

The number of equations can be reduced by considering only the feature pairs with both features having measured total effects significantly greater than 0. This method would work ideally for classifiers exploiting all the interactions of all the features in a dataset. In our scenario, each classifier works on a subset of feature space and doesn't exploit all the possible interactions, so the collected interaction effects are underestimated. They are a valuable source of information however.

Final total effect and interaction effect scores are obtained by summing partial results obtained from all s*t classifiers.

5 Results and Discussion

The original metrics of Draminski et al. obtained through decision tree structure analysis are called Relative Importance (RI) and Inter-Dependency (ID). For result comparison, we interpret RI as total effect, and ID as interaction effect. In the underlying section they are collectively referred to as white-box metrics (wb), contrary to our black-box metrics (bb) not requiring an analysis of the tree structure. While wb is tied to C4.5 decision trees, bb can work with any model.

For each configuration from Table 1, metrics were normalized by dividing by the highest scoring feature or feature pair. Results from dataset A are shown in Fig. 2 through a parallel coordinates plot. Upper subplot shows total effects, while the lower one the value of interaction effect between the features.

For the new metrics (bb-*), score for categorical feature interaction effects (dashed line) is always higher than that for continuous feature interaction effects (full line). This is reasonable, as splits on categorical features leave less space for error than binary splits on continuous ones (for the tested features). However, for the old metrics (wb-*), the continuous features and feature interactions have much higher scores than the categorical versions, especially when tree pruning is disabled (wb-C45-none). This is because the continuous features allow for growing larger decision trees, which biases the original structure dependent metrics. For all wb and most bb, numerical total effects have higher scores than categorical analogs. This may signal a bias in both approaches, however the difference is smaller for bb.

Last two axes show how the feature interaction detectability is impacted when further randomization is introduced through RandomForests and Extremely Randomized trees [13]. The 'bivariate.cont.gaussians' interaction is now on the top. An explanation is that an optimal decision tree classifier created using this feature (Fig. 1)

Table 1. Configurations of the experiments. In each case, 20000 feature samples were drawn, and 5 training-test splits were made for each one of them.

Dataset	Features	f. smp. size (m)	Classifier	Tree pruning	Metrics
A	14	2	C4.5	0.25	white-box
A	14	2	C4.5	training set	white-box
A	14	2	C4.5	none	white-box
A	14	2	C4.5	0.25	black-box
A	14	2	C4.5	training set	black-box
A	14	2	C4.5	none	black-box
A	14	2	RandomForests	none	black-box
A	14	2	ERT [15]	none	black-box
B	10000	500	C4.5	0.25	white-box
B	10000	500	C4.5	none	black-box
B	10000	500	RandomForests	none	black-box

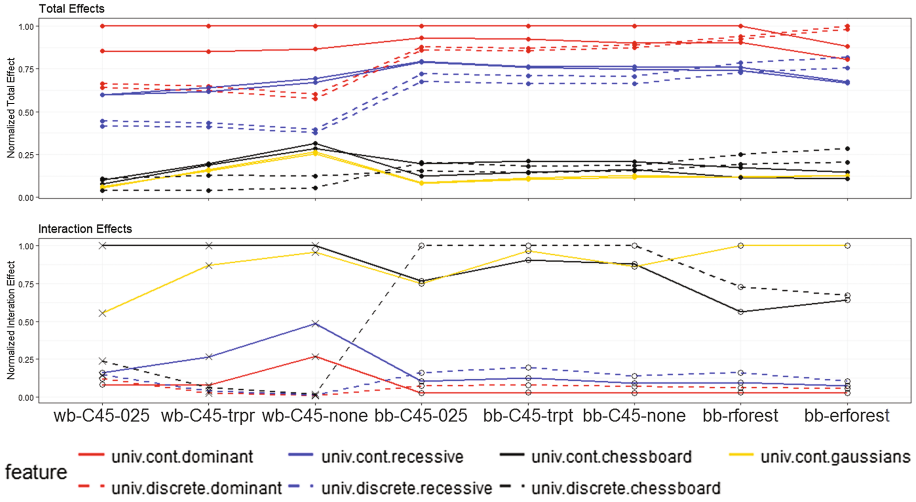


Fig. 2. Dataset A: Comparison of normalized feature total effect and interaction effect scores.

requires the first split to have little information gain. As result, it has to be forced through limited split options or randomization.

Three configurations were chosen for dataset B – the white-box metrics with C4.5 pruning, black-box metrics with unpruned C4.5 trees, and black-box metrics with RandomForests. Results are presented in Fig. 3.

Similarly as with dataset A, the discrete feature interactions are scored above their continuous analogs by the black-box metrics. There are however two major differences. First of all, the 'bivariate.cont.gaussians' interaction was not detected at all by any of the approaches. An explanation is the size of the feature sub-space – with 500 features to choose from, there is little chance that a random split unlocking the correct split sequence would be made. Secondly, the detectability of strong continuous feature

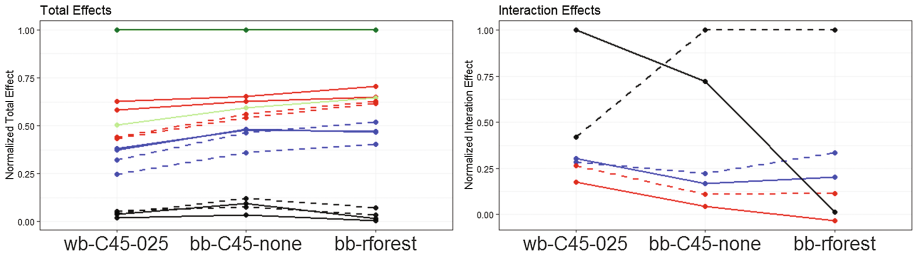


Fig. 3. Dataset B: Comparison of normalized feature total effect and interaction effect scores. Legend analogous to Fig. 2, green lines represent features from univariate distributions.

interaction (bivariate.cont.chessboard) has dramatically dropped with respect to its categorical analog for RandomForests classifier, suggesting that introducing a second level of feature space randomization may be harmful.

One final improvement over the original metrics – not shown on the plots – is the number of unexpected interactions reported by the algorithm. For dataset B analyzed with wb-C45-025, only 43% of the sum of interaction effects of all significant interactions (alpha = 0.05, validated through permutation testing) is coming from the designed interactions. This proportion rises up to 68% for wb-C45-none.

6 Conclusion

Moving the insights from artificial benchmark data to real-life problems should be made with caution. Synthetic results can however provide valuable hints about the behavior of algorithm for predictable data. The study confirms the ability of original MCFS to detect features and feature interactions in high-dimensional datasets. The original method is however – at least under certain configurations – vulnerable to bias created by overfitted decision trees. Our proposed solution is to introduce feature total effect and interaction effect metrics relying purely on classification accuracy shifts. Initial results seem promising, as the new approach is less vulnerable to overgrown decision trees, provides fewer false interactions and may allow to extend the algorithm and software to regression and censored survival time modeling.

All in all, the detectability of feature interactions both for the 'black-box' and 'white-box' approach strongly depends on the main effects of the features. Strongly interacting features without main effect component have little chance of being discovered – partially because of the greedy nature of decision tree training, but mainly because of the speed at which the interaction space grows.

Acknowledgements. We would like to thank prof. Jacek Koronacki (Polish Academy of Sciences) as well as Anonymous Reviewers for helping to increase quality of the paper.

The work was financially supported by internal grant BK/213/Rau1/2016/10. Calculations were carried out using the computer cluster Ziemowit (http://www.ziemowit.hpc.polsl.pl) funded by the Silesian BIO-FARMA project No. POIG.02.01.00-00-166/08 in the Computational Biology and Bioinformatics Laboratory of the Biotechnology Centre in the Silesian University of Technology.

References

1. Zhang, W., et al.: Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol. **16**, 133 (2015)
2. The 1000 Genomes Project Consortium: An integrated map of genetic variation from 1,092 human genomes. Nature **491**, 56–65 (2012)
3. Sidak, Z.: Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. **62**, 626–633 (1967)
4. Storey, J.: A direct approach to false discovery rates. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **64**, 499–518 (2002)
5. Perneger, T.: Whats wrong with Bonferroni adjustments. BMJ **316**, 1236–1238 (1998)
6. Breiman, L.: Random forests. Mach. Learn. **45**, 157–176 (2001)
7. Winham, S., et al.: SNP interaction detection with random forests in high-dimensional genetic data. BMC Bioinform. **13**, 164 (2012)
8. Bureau, A., et al.: Identifying SNPs predictive of phenotype using random forests. Genet. Epidemiol. **28**, 171–182 (2005)
9. Draminski, M., et al.: Monte carlo feature selection for supervised classification. Bioinform. **24**, 110–117 (2008)
10. Draminski, M., et al.: Monte carlo feature selection and interdependency discovery in supervised classification. Adv. Mach. Learn. **II** (2010)
11. Draminski, M., et al.: Discovering networks of interdependent features in high-dimensional problems. Big Data Analysis: New Algorithms for a New Society (2016)
12. Krol, L.: Distributed monte carlo feature selection: extracting informative features out of multidimensional problems with linear speedup. Beyond Databases, Architectures Struct. **12** (2016)
13. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**, 161–182 (2006)

A Normalisation Strategy to Optimally Design Experiments in Computational Biology

Míriam R. García(✉), Antonio A. Alonso, and Eva Balsa-Canto

Bioprocess Engineering Group, IIM-CSIC, Vigo, Spain
miriamr@iim.csic.es

Abstract. In this work we describe a new methodology to improve predictive capabilities of dynamic models when parameters differ in orders of magnitude. The main idea is to normalise the model unknown parameters before solving the classical problem of optimal experimental design based on the Fisher information matrix. The normalisation improves the relative confidence intervals of the estimated parameters and the conditioning of the Fisher matrix, especially for those criteria aiming to decorrelate the model parameters. Using the so-called core predictions, we show how the new approach improves the final model predictive capabilities in two terms: predictions are closer to the real dynamics and with better confidence intervals.

We illustrate the concepts using two toy examples linear and nonlinear in their parameters. Finally we test the performance of the normalisation in a model simulating the bacterial SOS response. This pathway remains of main relevance to work towards a predictive model of antimicrobial resistance.

Keywords: Normalisation · Fisher Information Matrix (FIM) · Relative parameter confidence intervals · Core predictions · Optimal Experimental Design (OED) · Bacterial SOS response

1 Introduction

Predictive capabilities of models in computational biology largely depend on the confidence we have on their parameters. Usually there is a large number of non-measurable parameters that have to be estimated fitting the model to experimental data. In most cases only a limited number of components in the network can be measured, the system may only be stimulated in very specific ways, the number of sampling times is usually limited and the experimental data are subject to substantial experimental noise [2,3]. As a consequence the confidence intervals of the parameters are too large to make useful predictions or even infinite.

Optimal Experimental design (OED) methodologies therefore become essential to find which experiments are more informative. They are currently being exploited in different areas in computational biology, mainly in systems biology

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_16

[14,18] and in pharmacokinetics [8]. The idea is to formulate an optimisation problem to find the best decision variables (such as sampling times or stimuli profiles) to maximise the quantity and quality of information using the Fisher Information Matrix (FIM).

The main challenge is that parameters in computational biology usually differ in several orders of magnitude. For example protein degradation rates are typically several orders of magnitude lower than Michaelis-Menten constants. As a consequence the FIM is ill-conditioned [1,6] and parameters with small values have too wide confidence intervals to make useful predictions.

To address this challenge we propose to normalise the parameter models before optimising the experiments. The new normalised FIM has a better condition number and the optimal experiments will focus on decreasing the relative confidence intervals of the parameters instead of their absolute confidence intervals. The improvement in the predictive capabilities of the model is analysed using the so-called model core predictions [5]. They are used to assess how the parameter uncertainty is translated to the model predictions with and without the proposed normalisation.

This work starts describing the theory of optimal experimental design based on the fisher information matrix in Sect. 2 and the proposed normalisation in Sect. 3. We will use a simple toy example (linear in the parameters) to illustrate the ideas behind these two sections. Section 4 describes the concept of core prediction and shows the improvement in predictive capabilities for another toy example (non-linear in the parameters) when using the normalisation. Finally in Sect. 5 we explore the performance of the normalisation in the context of a regulatory network describing main features of bacteria SOS response. For the sake of simplicity, elements of vectors and matrices are denoted with subindexes and vectors and matrices are denoted with same symbol as their elements and with all vectors being column vectors.

2 Classical Theory for Optimal Design of Experiments Based on the Fisher Information Matrix

The objective is to minimize the uncertainty of the estimated parameters by designing the most informative experiments. The Optimal Experimental Design (OED) problem may be mathematically formulated as a general dynamic optimisation problem searching for the manipulable variables (such as time-dependent stimuli, initial experimental conditions, experiment durations, sensor locations, sampling times and type of measurements) that maximize information. In this way the OED problem is formulated with maximum generality allowing for the sequential or parallel design of several experiments.

2.1 The Fisher Information Matrix and the Hyperellipsoid of Information

The Fisher Information Matrix (FIM) is the standard measure for the amount of information that an observable carries about an unknown parameter.

In computational biology the usual observables are discrete dynamic variables that depend on the unknown vector of parameters $\theta = [\theta_1, \ldots, \theta_{n_\theta}] \in \mathbb{R}^{n_\theta}$. Assume that each measurement is a random variable with normal distribution $x_k \sim N(\bar{x_k}, \sigma^2)$ and mean obtained with a deterministic model $\bar{x}_k = M_k(\theta, u, v)$ that depends on the stimuli $u(t) \in \mathbb{R}^{n_u}$ and other decision variables $v \in \mathbb{R}^{n_v}$ such as sampling times, sensor locations, initial conditions and experiment duration.

The FIM is defined as the variance of the score where J_{ml} is the negative log-likelihood function [12] with n_k being the total number of sampling times:

$$\mathscr{F}(\theta, u, v) = E\left\{\left(\frac{\partial J_{ml}}{\partial \theta}\right)\left(\frac{\partial J_{ml}}{\partial \theta}\right)^{\mathrm{T}}\right\}$$

$$J_{ml} = -\ln p(x; \theta, u, v) = \frac{n_k}{2}\ln 2\pi + \frac{n_k}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{n_k}(x_k - M_k(\theta, u, v))^2.$$

The FIM determines a quadric, typically a hyperellipsoid in the parameter space. This hyperellipsoid represents the quantity and quality of information of the selected experiments. The largest and the more spherical the information hyperellipsoid defined by the FIM, the better the experimental design. Different scalar functions of the FIM are formulated (J_{OED}) being the following the most common [13]:

D criterion ($J_D = \max Det[\mathscr{F}]$) maximises the volume of the information hyperellipsoid no matter its shape. The higher its value the smaller the expected parameter uncertainty for the parameter estimates. **A criterion** ($J_A = \max trace[\mathscr{F}]$) maximises the arithmetic mean of the hyperellipsoid semi-axes. **E criterion** ($J_E = \max \lambda_{min}[\mathscr{F}]$)maximises the minimum semi-axis of the information hyperellipsoid, therefore offering a compromise between D and *Emod*. **Modified E criterion** ($J_{Emod} = \min \frac{\lambda_{max}[\mathscr{F}]}{\lambda_{min}[\mathscr{F}]}$) minimises the relationship between the longest and shortest semi-axes of the information hyperellipsoid, i.e., improves the eccentricity of the hyperellipsoid. This criterion is quite appealing as the global optimal solution corresponds with $J_{Emod} = 1$, meaning that the uncertainty of the parameter estimates is equally distributed.

In general D and A are good criteria to improve overall information and *Emod* to decorrelate parameters. If the objective is to optimise a compromise between improving information and parameter decorrelation, E criterion is the best option. We should stress that this criterion is non-differentiable and requires the use of appropriate global optimisers [20].

The optimal experimental design (OED) problem may be formulated as a general dynamic optimisation problem as follows:

Calculate the time-variable stimuli $u(t)$ and other decision variables v (such as experiment duration, type of measure, initial conditions, sampling times and sensor positions) so as to optimise a scalar measure of the FIM $J_{OED} = \phi(\mathscr{F})$.

The experimental design may be subject to algebraic constraints related to experimental limitations in the manipulable dynamic variables $u^L(t) \leq u(t) \leq u^U(t)$ and in the rest of decision variables $v^L \leq v \leq v^U$ [10], where superscripts L and U represent lower and upper bound respectively.

Parameter estimation and optimal experimental design problems in computational biology require advanced numerical techniques. In this work we used AMIGO2 (Advanced Model Identification using Global Optimization), a multiplatform toolbox implemented in Matlab which covers parameter estimation but also sensitivity analysis and experimental design [4]. From the set of numerical methods offered in the toolbox, we selected the global optimizer based on scatter search (eSS, Enhanced Scatter Search) method [7]. It can optimise non-differentiable functions and it is very efficient and robust in finding the best parameter values and experimental designs. In addition, the model simulator CVODES [11] was selected to solve the model and calculate the Fisher Information Matrix. The optimisation of the FIM is approached using the so called control vector parametrisation approach (CVP), which transforms the original infinite dimension optimisation problem into a non-linear programming problem (NLP) whose solution requires the use of adequate optimisation methods.

2.2 The Covariance Matrix and the Hyperellipsoid of Uncertainty

If the Fisher information matrix represents the hyperellipsoid of information, its inverse is a hyperellipsoid that gives a sense of the confidence or uncertainty region. The Crammér-Rao inequality [21] establishes that the covariance matrix C is greater or equal than the inverse of the Fisher Information Matrix for the case that the estimator is asymptotically unbiased. Therefore the FIM is used to calculate a lower bound of the covariance matrix $C \geq \mathscr{F}^{-1}$.

The confidence intervals of a parameter may be calculated based also on this Crammér-Rao bound. The confidence intervals are $\theta_i^* \pm t_{\alpha/2}^{\gamma}\sqrt{\widetilde{C}_{i,i}}$ considering the student's t-distribution $t_{\alpha/2}^{\gamma}$ with γ being the number of degrees of freedom and $(1-\alpha)100\%$ the selected confidence interval. The correlation matrix can be also calculated from the covariance matrix $\left(Cr_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{ij}}}\right)$.

The likelihood function of non-linear models in their parameters depends on the value of the parameters (see Example 2), and the necessary conditions for the Crammér-Rao inequality are only satisfied (see [21] and [12] for details) around the optimum: $C >= \mathscr{F}^{-1}(\theta^*)$. Nevertheless the objective of the optimal experimental design is to estimate the optimum set of parameters itself. Therefore the following iterative procedure is proposed in the literature:

1. Estimate the parameters θ^0 using the available information (literature or non-optimal experiments)
2. Find the optimal set of experiments using $\mathscr{F}(\theta^0)$
3. Re-estimate the parameters θ^1
4. Find the new set of optimal experiments using $\mathscr{F}(\theta^1)$
5. Repeat steps 3 and 4 until $\theta^{i-1} \simeq \theta^i$

If the starting point θ^0 is sufficiently close to the optimum (θ^*) or $\mathscr{F}(\theta)$ is sufficiently smooth, this iterative procedure will find the optimum set of parameters ($\theta^i = \theta^*$).

Example 1: A system of two uncoupled dynamics with degradation. We illustrate the ideas behind the classical optimal experimental design using the following simple example:

$$\emptyset \xrightarrow{k_1^*=10} x_1 \quad \emptyset \xrightarrow{k_2^*=1} x_2$$

being x_1 and x_2 the dynamic variables and k_1 and k_2 the unknown reaction velocities. Contrary to common models in computational biology, this model is linear in their parameters [21] when considering mass action. It has the following simple analytical solution $x_1(t) = -k_1 t_1 + x_1^0 \quad x_2(t) = -k_2 t_2 + x_2^0$, where x_1^0 and x_2^0 are the initial conditions at $t = 0$.

We use optimal experimental design to determine the best sampling times assuming we only take one measurement per dynamic variable. For a deeper discussion on the relevance of optimising the sampling times, we advise the reading of Kutalik work [15]. We assume Gaussian noise with same standard deviation σ for all measurements to calculate the FIM:

$$\bar{x} = M(\theta, t_1, t_2) = [-k_1 t_1 + x_1^0, -k_2 t_2 + x_2^0], \qquad \mathscr{F} = \begin{pmatrix} \frac{t1^2}{\sigma^2} & 0 \\ 0 & \frac{t2^2}{\sigma^2} \end{pmatrix}$$

If we impose a maximum bound on the sampling times: $t_1 \leq 5 \in \mathbb{R}^+$, $t_2 \leq 10 \in \mathbb{R}^+$, it is trivial to see how optimal samplings $[t_1^*, t_2^*]$ in the sense of criteria D and A are these maximum bounds ($[5, 10]$). The optimal sampling times for the remaining criteria have several solutions with best value of the cost function J_{OED}. For all $t_1^* = 5, t_2^* >= 5$ sampling times are optimal in the sense of E and for all $t_1^* = t_2^*$ in the sense of *Emod*.

Finally we show in Table 1 the confidence intervals and condition number (equivalent to J_{Emod} criterion) for some optimal sampling times assuming $\sigma = 1$ and $t_{\alpha/2}^{\gamma} = 1.96$. To stress that these results refer to absolute values of the confidence intervals we also calculate the relative confidence intervals.

Table 1. Optimal sampling times, FIM condition number and uncertainty ellipsoid semi-axes in Example 1 for the different criteria

Criteria	Optimum	FIM condition number		Confidence intervals	
	$[t_1^*, t_2^*]$	Absolute	Relative	Absolute	Relative
D	[5,10]	2.0	5.0	[0.39,0.196]	[0.039,0.196]
A	[5,10]	2.0	5.0	[0.39,0.196]	[0.039,0.196]
Emod	[4,4]	1.0	10	[0.49,0.49]	[0.049,0.49]
E	[5,6]	1.2	8.33	[0.39,0.33]	[0.039,0.33]

3 Optimal Experimental Design Based on the Normalised FIM

FIM confidence intervals refers to the absolute value of the parameters. The classical approach, described in previous section, minimizes the confidence hyperellipsoid considering that all parameters have the same relevance, even if their values have different orders of magnitude. However, in general terms, same confidence intervals are considered best for parameters with larger values than for smaller parameters. See for example confidence intervals for *Emod* in Table 1 ($k_1 = 10 \pm 0.49$ and $k_2 = 1 \pm 0.49$). If we can only improve the confidence of one parameter, it is natural to focus on k_2, but with classical OED both have the same confidence and will be treated equally.

Moreover the optimisation of experiments has numerical problems because the FIM is ill-conditioned. Parameters in systems biology usually differ in orders of magnitude and therefore also the score functions ($\frac{\partial J_{ml}}{\partial \theta}$) that define the FIM.

In order to both, focus on relative confidence intervals and scaling the FIM to avoid numerical problems, we propose the following normalization: *Consider the available best estimation of the parameters θ^i, the Fisher information matrix is calculated using the following reparametrised model $M_k^{norm}(\theta/\theta^i, u, v)$. In the literature procedure described in Sect. 2.2, as i increases, the value of the unknown parameters θ^* will tend to a all-ones vector.*

We should note that for complex biological systems we need an estimation of the unknown parameters even if we do not use the normalisation. As discussed previously, the FIM depends on the parameters and has to be evaluated close to the optimum. The only case where that is not a requirement is for models linear in their parameters (such as Example 1) that are not common in biology. We use Example 1 to illustrate the effect of the normalisation, for a more detailed discussion see [10].

Example 1: A system of two uncoupled dynamics with degradation. Let us calculate for Example 1 the optimal experimental design that minimised the relative confidence intervals and compare the results with those obtained in Table 1. The normalised FIM reads $\mathscr{F}^{norm} = \begin{pmatrix} \frac{100t_1^2}{\sigma^2} & 0 \\ 0 & \frac{t_2^2}{\sigma^2} \end{pmatrix}$.

In general for uncoupled models linear in their parameters $\mathscr{F}^{norm}$ $(\text{diag}\theta^i)^2 = \mathscr{F}$. For this class of simple systems D is not affected by the normalisation and A is affected only if there are bounds or penalisation on the decision variables. We should stress that this is only a tendency for nonlinear models, but not a rule.

Contrary, the criteria focusing on decorrelating parameters (E and *Emod*) are affected by the normalisation. Assuming same bounds for the sampling times and $\sigma = 1$, best sampling times for E are now $t_2 = 10, \forall t_1 >= 1$ and for *Emod* are all sampling times satisfying $10t_1 = t_2$. Table 2 shows the results obtained with the normalised FIM.

Table 2. Optimal sampling times, FIM condition number and uncertainty ellipsoid semi-axes in Example 1 for the different criteria using the normalisation. With the new approach we improve the confidence intervals of the smallest parameter k_2 at the expenses of the largest one k_1 in *Emod* and E.

Criteria	Optimum	FIM condition number		Confidence intervals of $[k_1, k_2]$	
	$[t_1^*, t_2^*]$	Absolute	Relative	Absolute	Relative
D	[5,10]	2.0	5.0	[0.39,0.196]	[0.039,0.196]
A	[5,10]	2.0	5.0	[0.39,0.196]	[0.039,0.196]
Emod	[1,10]	10	1	[1.96,0.196]	[0.196,0.196]
E	[2,10]	5.0	2.0	[0.98,0.196]	[0.098,0.196]

4 Model Predictive Capabilities with and Without the Normalised FIM

We propose the use of the so-called core prediction to asses the predictive capabilities of the model fitted using the optimal experiments with and without the normalisation. The normalisation prioritises the minimisation of the relative confidence hyperellipsoid, resulting also in FIM with better condition number. However, it is not trivial for non-linear models how this affects to the confidence of the model predictions.

Core predictions is a standard method in systems biology to test predictive capabilities of complex models subject to uncertainty [5]. Detailed description of the method used here can be found in [9] where we explored the performance of a microbiological model after optimal experimental design using the D criterion. The idea is to compute the range of possible solutions corresponding to different realizations of the parameter statistics given by the confidence intervals before the OED and model fitting.

In the context of this work we will consider a uniform distribution between the bounds of the confidence intervals $\theta_i^* \pm t_{\alpha/2}^{\gamma}\sqrt{\widetilde{C}_{i,i}}$ and calculate 300 different realizations. Figures with core predictions will show the mean of the predictions and a coloured area defined by $\bar{x} \pm \sigma_x$ where σ_x is now the standard deviation of the prediction at each time calculated from the different realizations.

Example 2: simple metabolic pathway using mass action. To assess the performance of the normalisation, we calculate the core predictions in a simple model non-linear in its parameters. Assume that we can only measure two dynamic variables in a pathway of three compounds with the following reactions:

$$x_1 \xrightarrow{k_1^*=0.1} x_2 \xrightarrow{k_2^*=1} x_3 \xrightarrow{k_3^*=1} \emptyset$$

the objective is to find which dynamic variable to measure to maximise the information of the experiment. For the formulation of the OED problem we will define the model measurements as $\bar{x} = wM(\theta, t_1, t_2) = w[x_1(t), x_2(t), x_3(t)]$ with standard deviation 1% of the maximum value of the observable and where w is

a 3-dimensional vector that can take value 1 if the variable is observable or 0 otherwise. Sampling times here are considered fixed ($t = [0.0, 2.5, 5.0, 7.5, 10.0]$) and therefore there are only three possible solutions: $w = [1, 1, 0]$, $w = [1, 0, 1]$ and $w = [0, 1, 1]$. All criteria, except A that may give solutions with singular FIM [10], will tend to measure x_3 as it is the only possible observable with information about k_3. Otherwise the problem is structurally no identifiable.

D criterion is not affected by the normalisation and selects to measure $[x_1, x_3]$. Remaining criteria select to measure $[x_1, x_2]$. When using the normalisation all criteria coincides with D and considers that the optimum is to measure $[x_1, x_3]$. First two columns in Fig. 1 show the core predictions for *Emod* with and without normalisation. As expected, x_1 confidence is better for *Emod* and x_2 for normalised *Emod*, while x_3 is good in both cases. Attending to the overall performance, normalised *Emod* produced the best results in terms of smaller core predictions. Last columns in Fig. 1 show how the correlation matrix for *Emod* also improves with the normalisation.

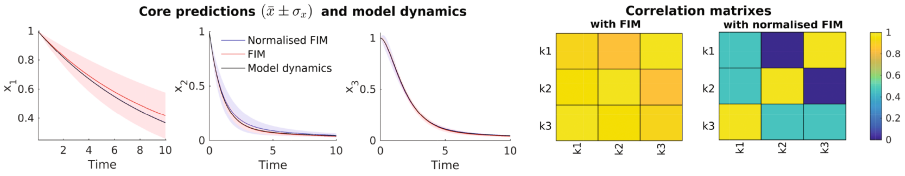


Fig. 1. First three subfigures show model dynamics and core predictions obtained with *Emod* with and without normalisation. Last subfigures show the improvement in the correlation matrix with the normalisation

5 Case Study: SOS Response in *Escherichia Coli*

Antimicrobial resistance is a threat to our health and economy that is expected to scale much faster than our ability to design new drugs. Martinez's work [17] stresses the importance of designing quantitative models to predict antimicrobial resistance using systems biology tools. The SOS response is one of the main mechanisms related with antimicrobial resistance. This pathway modulates the acquisition of bacterial mutations under DNA damage by different stressors. The response to this damage is to upregulate the production of protein recA that inactivates the transcriptional repressor LexA. Under normal conditions LexA represses the transcription of several genes involved in DNA damage repair, including recA. Therefore lexA and recA are the centre of the SOS response connected in a double-negative mixed feedback loop [19].

For our study we have used the deterministic version of that feedback loop (MODEL2937159804) in the biomodels repository [16]. Shimoni observed [19] that the model presents practical identifiability problems for the considered experiments, with different values of the parameters reproducing same results. To reproduce basal state in Fig. 2 [19], we will set the stimuli that increase the

production of recA and lexA to $f_R = 0.5$ and $f_L = 1$ obtaining the following parameters (cR = 0.5500, cL = 0.0099, gR = 0.0332, gL = 0.0196, gmL = 0.0160, gmR = 0.0011, deR = 0.0864, demR = 0.0007, sL = 0.0309, demL = 0.0002, deL = 0.2965, sR = 0.0113, cp = 0.0006). For reproducing the plateau of strongest damage the stimulus increases the production of recA changes to $f_R = 1.5$.

The objective in this work is to find the best experiment in terms of best experimental duration ($t_f \in [10000, 15000]$), best four observables and best 5-steps profile for $f_R \in [0.5, 1.5]$. We consider 20 equidistant sampling times with Gaussian noise and relative standard deviation 0.5% with respect to the maximum value of the observable.

Optimal experimental designs (data not shown here) are different for each criteria and are affected considerable by the normalisation except for the D-criterion, where differences are slight modifications on the switching times of the optimal stimulus. Figure 2 shows the core predictions of all dynamic variables for the remaining criteria with and without normalisation. For these predictions we simulate the model at different levels of the stimuli f_R in a experiment of 125000 s. Results show how predictions improve in most cases when using the normalised FIM in two senses: the mean of the predictions is closer to the model dynamics and the uncertainty (coloured areas) are smaller. Note however that some exceptions may occur. This is the case for example in the prediction of state R using the A criterion, where predictions are similar but the confidence is best using the non-normalised FIM.

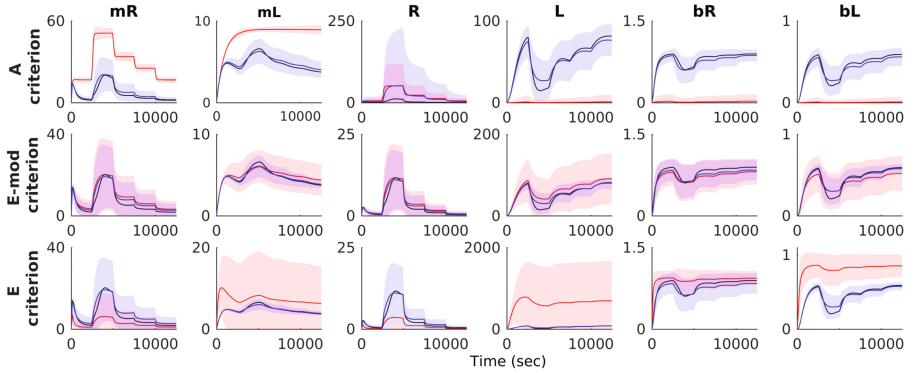


Fig. 2. Core predictions for A, $Emod$ and E criteria with conventional FIM (red) and normalised FIM (blue). Black lines represent the model dynamics with θ^* and red and blue lines represent the mean of the core predictions ($\bar{x}$) with coloured areas being $\bar{x} \pm \sigma_x$. Core predictions of the D-criterion, not depicted here, are similar to E with the normalisation

6 Conclusions

We propose a new formulation of the Optimal Experimental Design where the model parameters are normalised before defining the FIM. The new approach

searches for the minimum relative confidence intervals and improves the conditioning of the FIM avoiding numerical problems in the optimisation. Obtained optimal experimental designs change when using the normalisation, especially for those criteria focused on decorrelating parameters.

To test how the new approach improves the predictive capabilities of the models we use the core-predictions. They are a measure of the confidence in the predictions given the confidence in the parameters. We tested the normalisation in two examples, including the SOS pathway relevant to work towards a predictive model of antimicrobial resistance.

Results show how the normalisation provides more informative experiments and better model predictions than the classical approach for optimal experimental design.

Acknowledgements. This work has been funded by the Spanish Ministry of Science and Innovation throughout project RESISTANCE (DPI2014-54085-JIN).

References

1. Apgar, J.F., Witmer, D.K., White, F.M., Tidor, B.: Sloppy models, parameter uncertainty, and the role of experimental design. Mol. Biosyst. **6**(10), 1890–1900 (2010)
2. Balsa-Canto, E., Alonso, A.A., Banga, J.R.: Computational procedures for optimal experimental design in biological systems. ET Syst. Biol. **2**(4), 163–172 (2008)
3. Balsa-Canto, E., Alonso, A.A., Banga, J.R.: An iterative identification procedure for dynamic modeling of biochemical networks. BMC Syst. Biol. **4**(1), 1 (2010)
4. Balsa-Canto, E., Henriques, D., Gabor, A., Banga, J.R.: Amigo2, a toolbox for dynamic modeling, optimization and control in systems biology. Bioinformatics **32**(21), 3357 (2016)
5. Brännmark, C., Palmér, R., Glad, S.T., Cedersund, G., Strålfors, P.: Mass and information feedbacks through receptor endocytosis govern insulin signaling as revealed using a parameter-free modeling framework. J. Biol. Chem. **285**(26), 20171–20179 (2010)
6. Chis, O.T., Villaverde, A.F., Banga, J.R., Balsa-Canto, E.: On the relationship between sloppiness and identifiability. Math. Biosci **282**, 147–161 (2016)
7. Egea, J.A., Martí, R., Banga, J.R.: An evolutionary method for complex-process optimization. Comput. Oper. Res **37**(2), 315–324 (2010)
8. Galvanin, F., Ballan, C.C., Barolo, M., Bezzo, F.: A general model-based design of experiments approach to achieve practical identifiability of pharmacokinetic and pharmacodynamic models. J. Pharmacokinet. Biopharm. **40**(4), 451–467 (2013)
9. García, M.R., Vilas, C., Herrera, J.R., Bernárdez, M., Balsa-Canto, E., Alonso, A.A.: Quality and shelf-life prediction for retail fresh hake (Merluccius merluccius). Int. J. Food Microbiol. **208**, 65–74 (2015)
10. García, M.R.: Identification and real time optimisation in the food processing and biotechnology industries. Ph.D. dissertation. University of Vigo (2008)
11. Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E., Woodward, C.S.: Sundials: suite of nonlinear and differential/algebraic equation solvers. ACM Trans. Math. Softw. **31**(3), 363–396 (2005)

12. Kay, S.M.: Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory. Prentice Hall, Upper Saddle River (1993)
13. Kremling, A., Saez-Rodriguez, J.: Systems biology—an engineering perspective. J. Biotechnol. **129**(2), 329–351 (2007)
14. Kreutz, C., Timmer, J.: Systems biology: experimental design. FEBS J. **276**(4), 923–942 (2009)
15. Kutalik, Z., Cho, K.H., Wolkenhauer, O.: Optimal sampling time selection for parameter estimation in dynamic pathway modeling. Biosystems **75**(1), 43–55 (2004)
16. Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M.I., Snoep, J.L., Hucka, M., Le Novère, N., Laibe, C.: BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst. Biol. **4**, 92 (2010)
17. Martínez, J.L., Baquero, F., Andersson, D.I.: Predicting antibiotic resistance. Nat. Rev. Microbiol. **5**(12), 958–965 (2007)
18. van Riel, N.A.: Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. Brief. Bioinform. **7**(4), 364–374 (2006)
19. Shimoni, Y., Altuvia, S., Margalit, H., Biham, O.: Stochastic analysis of the SOS response in Escherichia coli. PLoS One **4**(5), e5363 (2009)
20. Telen, D., Van Riet, N., Logist, F., Van Impe, J.: A differentiable reformulation for e-optimal design of experiments in nonlinear dynamic biosystems. Math. Biosci. **264**, 1–7 (2015)
21. Walter, E., Pronzato, L.: Identification of Parametric Models from Experimental Data. Springer, London (1997)

Mitosis Detection in Breast Cancer Using Superpixels and Ensemble Classifiers

César A. Ortiz Toro[1(✉)], Consuelo Gonzalo Martín[2], Angel García Pedrero[1], Alejandro Rodriguez Gonzalez[2], and Ernestina Menasalvas[2]

[1] Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain
ca.ortiz@upm.es

[2] Departamento de Arquitectura y Tecnología de Sistemas Informáticos, Universidad Politécnica de Madrid, Madrid, Spain

Abstract. Determining the severity and potential aggressiveness of breast cancer is an important step in the determination of the treatment options for a patient. Mitosis activity is one of the main components in breast cancer severity grading. Currently, mitosis counting is a laborious, prone to processing errors, done manually by a pathologist.

This paper presents a novel approach for automatic mitosis detection, where promising candidates are selected from a superpixel segmentation of the image and classified using an ensemble classifier created from a selection from a pool of different color spaces, different features vector.

Keywords: Medical image · Mitosis detection · Ensemble classifier · Superpixels

1 Introduction

One of the main biomarkers of breast cancer patients' prognosis is the visual assessment of the tumour tissue; how closely it resembles to normal tissue in microscope images. On this assessment, the mitosis count, the number of dividing cells in the image, appears as a strong prognosticator for tumour aggressiveness and severity [1]. However, manual mitosis counting is a tedious, time consuming process burdened by human bias directly related to the massive size of the histopathology images and the high variability of the mitotic occurrences. Automatic detection of mitotic nuclei could reduce pathologist labour and provide a more consistent result, improving the quality of the diagnosis.

Currently, automatic mitosis detection methods can be divided into hand-crafted features based methods and deep convolutional neural networks (CNN).

Hand crafted features based methods relays on specific morphological, statistical or textural characteristics of the mitosis for automatic detection, both pixel-wise [2] or area-wise [3]. Hand-crafted features based methods, in the same way as non-automatic mitosis detection, suffer from the variability in texture and morphology associated to mitotic appearances.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_17

Deep convolutional neural networks (CNN), multi-layer neural networks that learn a bank of convolutional filters at each layer can find feature patterns that are difficult to describe using hand-crafted features, but training and testing CNNs is computationally demanding and time-consuming. Even thougth, there are multiples CNN approaches to mitosis detection using a pixel-wise classifier as [4] or the hybrid ensemble in [5]. In any case mitosis detection in breast cancer histology images remains as an open problem.

In this study, carried using the TUPAC16 challenge mitoses auxiliary dataset [6], we propose a method based on the use of a superpixels segmentation as a tool to isolate and select mitotic candidates, and the capabilities of an ensemble created from a pool of classifiers trained with different features vector defined in four color spaces for the creation of a simple ensemble for mitosis detection.

The rest of the paper is organized as follows: Sect. 2 describes the proposed framework for mitosis detection. Experimental results are presented in Sect. 3. Finally, the concluding remarks with future work are given in Sect. 4.

2 Materials and Methods

We approach the problem of mitosis detection in a general image segment classification sense, where the candidate object resulting from an image segmentation are tagged as mitosis/non-mitosis. The classification is performed by a fixed rule combination of a number of classifiers selected from a pool of twelve, each of them trained from three different vector sets of features (color statistics, histogram and histon) obtained from differents color space representations of the image (RGB, LAB, HSV and HE). Our method follows the typical structure of hand crafted features approaches: image pre-processing, candidates selection, feature extraction, classifier construction and detection; as can be seen in Fig. 1.

2.1 Image Pre-processing

One of the main difficulties in histopathology image analysis is variability. Tissue structures in breast histopathology images are commonly highlighted using a combination of hematoxylin (H) and eosing (E) stains. In the stained images, nuclear and cytoplasm regions appear as hues of blue and purple, while extracellular material tends towards gradations of pink. Nevertheless, inconsistencies in staining preparation of histology slides could make it difficult to perform any analysis, such as mitosis detection. In order normalize its appearance and to separate hematoxylin and eosing stains, images are preprocessed using the principal component analisys based method presented in [7].

2.2 Candidate Selection Through Superpixel Image Segmentation

Mitosis candidate selection begins with the segmentation (over-segmentation) of the normalized histopathology RGB image using superpixels.

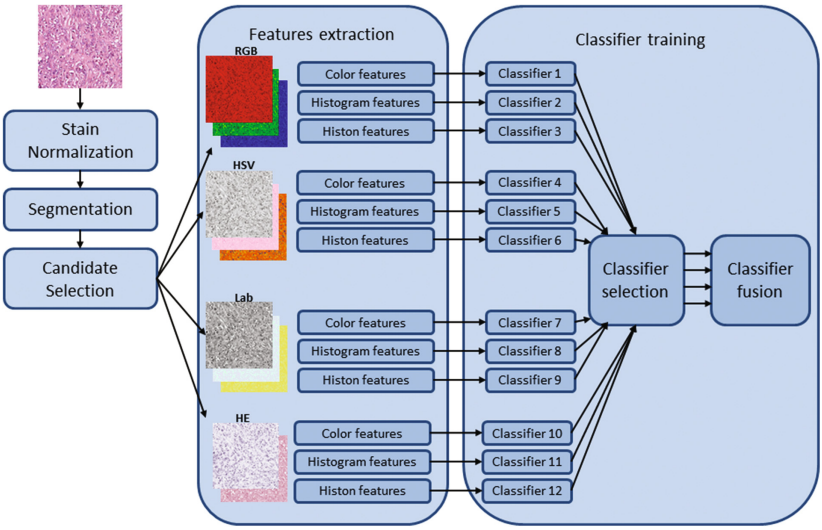


Fig. 1. Work-flow for the proposed mitosis detection method

A superpixel, first exposed in the works on binary classifiers by Ren and Malik in [8], is defined as a perceptually uniform region in the image. The idea behind the superpixels arises from the fact that the division of an image in pixels is not really a natural division, but simply an artefact of the device that captures the images. Superpixels segmentation results in an image over-segmentation composed by small, closely spectral related areas, a set of convenient primitives from which local image features are computed. There are many different techniques to generate superpixels, but we will focus on the use of SLIC (see [9]), an adaptation of a k-means clustering approach that takes in account not only spatial proximity but also local intensity similitude in the image.

Mitosis counting is usually performed in image regions corresponding to an area of 2 mm^2 (5657×5657 pixels). Depending on the features present in the region and its complexity, a superpixel segmentation (for a sampling interval of 25 pixels, chosen to isolate most of the mitosis in a unique superpixel) could produce approximately 80.000 segments. In order to select promising candidates, we take advantage on the appearance of the mitotics figures as a darker or hyperchromatic objects. To accentuate differences between possible mitotic pre-classification candidates and the background, the normalized RGB images are transformed into an image called blue ratio image representation [10], in which a pixel with a high blue intensity relative to its red and green components is given a higher value. For each superpixel, its mean blue ratio value is computed, and those within the higher 5 percentile are selected as pre-classification mitotic candidates. Figure 2 illustrates the stages of candidate selection process.

Fig. 2. Segment candidates selection process. Detail of the original normalized image (left) with its corresponding superpixel segmentation overlayed, associate blue-ratio image (center) and mitosis candidates (right)

2.3 Feature Extraction

Different color spaces express different properties relevant to color texture analysis. Despite the multiple studies related to this area no single color space can be defined as well suited for characterization of all textures. In order to obtain vector features as relevant as possible, we will train classifiers using a set of different color spaces. From the normalized RGB image, we obtain the hue saturation value, (HSV) and Lab color spaces representation. These images (RGB, HSV and Lab) and the composite created from the grayscaled eosing and hematoxylin stains (HE) are the components from which features are extracted.

For each candidate, features are calculated at superpixel level; additionally, as the surroundings of a mitosis are also relevant in mitosis detection (a mitotic cell usually presents features as an ill-defined nuclear membrane, protrusions around the edges and different staining on the cytoplasm), the feature vector of each candidate contains the mean of the features of the segments surrounding the candidate, as show in Fig. 3. Three sets of features are proposed:

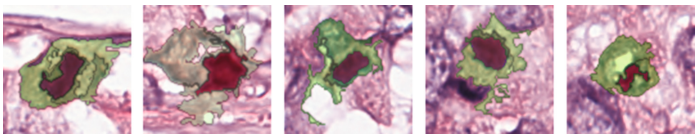


Fig. 3. Mitotic occurrences with the mitotic segment (reddish hues) and the associate surrounding segments highlighted (greenish hues)

Segment Color Statistic and Morphological Features. From the pixel intensity information of the color channels of an image, the mean, variance, skewness, kurtosis, maximum, minimum and its quartiles are extracted for each candidate segment. From the segment binary representation, we compute four morphological features, area, roundness, elongation and perimeter. This results in a vector of 60 statistical for the tree channels image (30 for the segment and 30 for the surroundings) and 42 for the HE composite.

Segment Histogram. Despite its simplicity, histograms perform adequately as input for classifiers in tasks as image indexing and retrieval as shown in [11]. In this work, the histogram of each color band in a candidate segment is calculated and used as part of a feature vector. In order to reduce inter-image variability, the number of bins is reduced to 32. The result histogram feature vector is composed by 192 characteristics in the tree channels image (96 for the segment and 96 for the surroundings) and 128 in the case of the HE composite image.

Segment Histon. A histon (see [12]) is a contour plotted on the top of any of the existing histograms of the components of the image. It exploits the correlation among the neighbouring pixels in the same spectral plane as well as the other spectral planes. In an histon, the collection of all points falling under a similar colour sphere of a predefined radius, the similarity threshold or expanse, E, belongs to a single bin in a histogram. For every intensity value g in the base histogram, the number of points encapsulated in the similar colour sphere is evaluated and added to the value in the histogram. The definition of an histon, in an image $I(x, y, s)$ of size $M \times N$, where s represent the spectral planes in the image, is given as:

$$H_{s_i} = \sum_{x=1}^{N} \sum_{y=1}^{M} (1 + S(x, y))\, \delta(I(x, y, s_i) - g) \quad 0 \leq g \leq L - 1 \text{ and } s_i \in S_p \quad (1)$$

where $\delta(\cdot)$ is the Kronecker delta, L is the total number of intensity levels in each of the spectral components (therefore, $\delta(I(x, y, s_i) - g)$ is a definition of a histogram) and $S(x, y)$ is a similarity function based on the distance measure as the sum of spectral distances of the planes $s_1, \cdots s_i, \in S_p$ that compose the image in any pixel (x, y) of a neighbourhood of sizes.

As we are extracting not the general histon of an image, but a set of segment-based local histons, the global similarity threshold, E, is substituted by a set of local similarity thresholds, corresponding to the mean standard deviation of the spectral planes in the image for each segment. The histon feature vectors, calculated from the previous histograms, encore 192 characteristics (128 in the case of the HE composite).

2.4 Classifier Training

Extracted features result in large dimensional spaces. Taking that into account, Random Forests has been selected to train the classifiers, as it can handle high dimensional data by building a large number of trees using only a subset of features, and is considered one of the most accurate classifiers. In order to reduce the classification bias produced as a result of class imbalance between non-mitotic and mitotic segment, first we down-sample the number of non-mitotic segments in the training dataset, then oversample the mitotic candidates applying the Synthetic Minority Oversampling Technique (SMOTE) [13]. Should be noted that using SMOTE for oversampling represents a trade-off between precision

and recall, the increase in true positives associated to a better balanced training set represents an increment in false positives too.

The combination of feature vectors and diferent color space images produces a pool of 12 classifiers. Unfortunately, just combining the outputs of those classifiers could result in a degradation in the final accuracy of the ensemble, as is discussed in [14]. Its theoretical framework shows that, hypothetically, using the same number of classifiers as the class labels gives the highest accuracy, proved that those classifiers are independent component classifiers, but this condition could not be easily achieved in real cases, making determining the ideal number of classifiers a complex matter. Classifier selection is based on the conclusions in [15]; classifiers should make uncorrelated errors with respect to one another. As we expect errors mainly in the form of false positives, the least correlated classifiers, using the Kendall rank correlation coefficient ([16]), from the pool of classifiers are chosen to form the ensemble. Classifiers are combinned using the maximum estimated confidence criteria (see [17]).

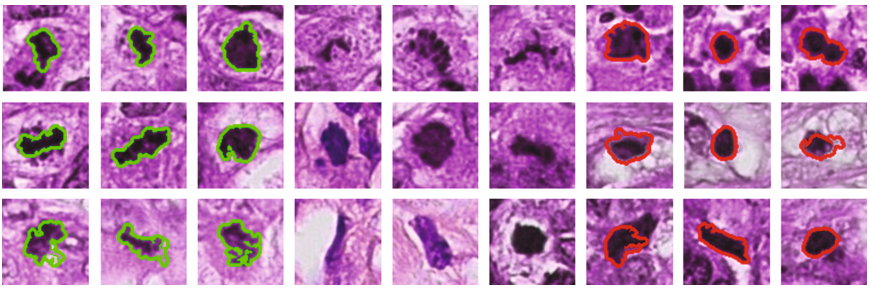


Fig. 4. Some examples of mitosis classification. True positives outlined in green, false positives outlined in red and false negatives with no outline.

3 Results and Discussion

This method is trained and evaluated using the TUPAC16 dataset, released for the MICCAI'16 Grand Challenge on Mitosis Assessment. The training dataset, consists of images from 73 breast cancer cases from different pathology centers, 23 taken by an Aperio ScanScope XT, and 50 form a Leica SCN400 ($\times$40 magnification, spatial resolution of 0.25 μm/pixel). Two expert pathologists annotated the locations of mitotic figures independently. The coincidences between pathologists were taken as ground truth objects and discrepancies were presented to a panel of an additional two observers, who made the final decision. The testing dataset consists of images from 34 breast cancer cases taken by a Leica SCN400. Segmentation in done with our own implementation of SLIC on MATLAB, class imbalance mitigation and classification is performed using WEKA.

Evaluation follows the guidelines used in the challenge. A detection will be considered a true positive only if its distance to a ground truth location is less

than 7.5 μm (30 pixels). If multiple detections are within the 30 pixels radios of a single ground truth, they are counted as a one true positive.

Empirical testing suggest an ensemble of four classifiers as the combination that produce better results. Should be noted that a small percentage (globally, between 5%-8% in our tests) of mitosis, usually ill defined mitosis in its late stages on over-stained images, could be lost in the candidate selection process (Fig. 4).

Table 1. Evaluation results for a validation subset of 10 cases (C.) of the TUPAC16 mitoses detection challenge, showing the real number of mitoses in the case (NM.) the number of true positives (TP), false positives(FP), false negatives (FN), and the prediction (Pr), recall (Re) and F-measures (Fm) values.

C.	NM	TP	FP	FN	Pr	Re	Fm
27	3	1	9	2	0.1	0.33	0.15
28	1	0	25	1	0	0	-
29	2	1	4	1	0.2	0.5	0.28
31	6	2	32	4	0.05	0.33	0.1
49	7	6	12	1	0.33	0.85	0.48

C.	NM	TP	FP	FN	Pr	Re	Fm
52	27	12	15	7	0.63	0.44	0.52
56	2	1	9	1	0.1	0.5	0.16
64	19	7	17	12	0.29	0.36	0.32
67	63	41	28	22	0.61	0.65	0.63
73	5	3	10	2	0.23	0.6	0.33

Table 1 shows the result of a validation subset of ten cases (mean F-measure 0.29) selected from the TUPAC16 training set (case numbers reference the corresponding patient in the TUPAC16 training set). The results include the effect of losing mitosis in the candidate selection process. As can be seen, the performance of the proposed method largely vary between cases. It should be noted that low performance in the classifier appears associated with a high number of false positives rather than low recall values. Results of the TUPAC16 can be seen in [6] (Task 3). (Evaluation with the testing TUPAC16 dataset is currently ongoing).

4 Conclusion and Future Work

In this paper, an automated mitosis detection method based on a multi-color space, multi-feature ensemble of classifiers has been proposed. Promising candidates are selected from a superpixel segmentation of the image. The classification is performed by an ensemble created from a selection from a pool of classifiers trained from three different feature vector extracted from four color spaces.

Despite its straightforward nature the proposed method shows promising results as the modular structure of an ensemble classifier means that any improvement in the individual classifiers, the classifier selection or the classifier combination step will affect positively to the final outcome. We expect to continue this line of research, both to improve results in general and to understand the significant differences in performance between cases.

Acknowledgment. This research has been supported by the Platform for advanced prescriptive health operational system (PAPHOS) project funded by EIT Health.

The authors would like to thank the organizers of the TUPAC16 Tumor Proliferation Assessment Challenge 2016 as well as all parties involved in preparing and providing the data.

References

1. Elston, C., Ellis, I.: pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology **19**(5) 403–410 (1991)
2. Tashk, A., Helfroush, M.S., Danyali, H., Akbarzadeh, M.: An automatic mitosis detection method for breast cancer histopathology slide images based on objective and pixel-wise textural features classification. In: The 5th Conference on Information and Knowledge Technology, pp. 406–410, May 2013
3. Khan, A.M., El-Daly, H., Rajpoot, N.M.: A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR 2012), pp. 149–152, November 2012
4. Chen, H., Dou, Q., Wang, X., Qin, J., Heng, P.A.: Mitosis detection in breast cancer histology images via deep cascaded networks. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI 2016, pp. 1160–1166. AAAI Press (2016)
5. Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., Madabhushi, A.: Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. In: SPIE Medical Imaging, vol. 9041, pp. 90410B–90410B-10, March 2014
6. MICCAI Grand Challenge: Tumor proliferation assessment challenge (tupac16). http://tupac.tue-image.nl
7. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110, June 2009
8. Ren, X., Malik, J.: Learning a classification model for segmentation. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, vol. 1, pp. 10–17, October 2003
9. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
10. Chang, H., Loss, L.A., Parvin, B.: Nuclear segmentation in h & e sections via multi-reference graph cut. In: International Symposium Biomedical Imaging (2012)
11. Chapelle, O., Haffner, P., Vapnik, V.: Svms for histogram-based image classification (1999)
12. Mohabey, A., Ray, A.: Rough set theory based segmentation of color images. In: 19th International Conference of the North American, Fuzzy Information Processing Society, NAFIPS, pp. 338–342(2000)
13. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Int. Res. **16**(1), 321–357 (2002)

14. Bonab, H.R., Can, F.: A theoretical framework on the ideal number of classifiers for online ensembles in data streams. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, pp. 2053–2056. ACM, New York (2016)
15. Ali, K.M., Pazzani, M.J.: Error reduction through learning multiple descriptions. Mach. Learn. **24**(3), 173–202 (1996)
16. Abdi, H.: The kendall rank correlation coefficient. In: Salkind, N.J. (ed.) Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks (2007)
17. Duin, R., Tax, D.: Experiments with Classifier Combining Rules, pp. 16–29 (2000)

Reproducibility of Finding Enriched Gene Sets in Biological Data Analysis

Joanna Zyla(✉), Michal Marczyk, and Joanna Polanska

Data Mining Group, Institute of Automatic Control, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland
{joanna.zyla,michal.marczyk,joanna.polanska}@polsl.pl

Abstract. Introducing the high-throughput measurement methods into molecular biology was a trigger to develop the algorithms for searching disorders in complex signalling systems, like pathways or gene ontologies. In recent years, there appeared many new solutions, but the results obtained with these techniques are ambiguous. In this work, five different algorithms for pathway enrichment analysis were compared using six microarray datasets covering cases with the same disease. The number of enriched pathways at different significance level and false positive rate of finding enrichment pathways was estimated, and reproducibility of obtained results between datasets was checked. The best performance was obtained for PLAGE method. However, taking into consideration the biological knowledge about analyzed disease condition, many findings may be false positives. Out of the other methods GSVA algorithm gave the most reproducible results across tested datasets, which was also validated in biological repositories. Similarly, good outcomes were given by GSEA method. ORA and PADOG gave poor sensitivity and reproducibility, which stand in contrary to previous research.

Keywords: Functional enrichment · Gene set analysis · Pathway analysis · Reproducibility

1 Introduction

Since the high-throughput methods were introduced into molecular biology, differentially expressed genes (DEGs) for various traits and diseases were massively detected and investigated. Some of the most commonly used measurement techniques are microarrays and next-generation sequencing (NGS) methods which can show hundreds or thousands of DEGs. Given the large number of detected genes it is hard to individually interpret and validate them, so methods which can show more complex relation between genes were developed. In the literature, they are known as enrichment or overrepresentation methods, and their main goal is to find specific collections of genes (gene sets, GSs) in which disorders caused by DEGs are observed and analyzed by statistical investigation. The GSs are defined mostly as the KEGG pathways [1], Gene Ontologies [2]

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_18

or MSigDB [3] collections. Nowadays, almost every molecular biology study based on high-throughput data looks for significant gene sets, starting from the investigation of microRNAs [4] to complex system biology approach [5]. The methods of enrichment/overrepresentation allow researchers to explain the observed processes or to perform validation of computationally derived results.

Through the years many algorithms of enrichment/overrepresentation analysis were proposed. They can be divided into three categories known as generations. First one is known as Over-Representation Analysis (ORA) generation [6]. The idea of ORA is simple and based on contingency tables constructed on proportions between DEGs and non-DEGs in given gene set. Even though the ORA method is simple there are two serious drawbacks. First, the information about the strength of phenotypes differentiation is lost by gene binarization (features in gene sets are represented only as DEGs or non-DEGs). Secondly, the assumption of signal independence in the enrichment test is not satisfied in most of the cases. To overcome these problems, the second generation of enrichment methods was proposed, known as Functional Class Sorting (FCS). These methods use an information about all investigated genes and sort them according to some metric. Further, the information from gene level is transformed to gene set level by process specific to each algorithm and statistical significance of each gene set is established. The newest, third generation is known as Pathway Topology (PT)-based approach. The idea of methods assigned to the last generation is comparable to FCS methods, however, they use the pathway structure to compute gene set enrichment statistics.

Several comparison studies of gene set enrichment methods were performed. In [7] they compared methods of first and second generation by assessment of sensitivity, prioritization and specificity. In [8] they concentrated on third generation methods and proposed an assessment score which combines pathway detection level and false positive estimation. In [9] they compared influence of ranking metrics in Gene Set Enrichment Analysis (GSEA) method to the outcome. In the presented work selected first and second generation methods are compared in terms of the number of enriched pathways at different significance level, false positive rate estimation and reproducibility of the results.

2 Material

In the presented study six microarray datasets were used. The datasets were downloaded from GEO database [10] (GEO IDs: GSE14762, GSE781, GSE6344, GSE15641, GSE14994, GSE11024). All of them are based on the investigation of gene expression for healthy controls and patients with the same type of cancer - clear cell Renal Cell Carcinoma (ccRCC). The following datasets were previously used in [8] as large collection of one specific disease among other available. The sample size of each dataset is as follows: 21, 17, 20, 55, 30 and 22. All data were normalized using RMA algorithm, and the gene duplicates were removed by keeping the probeset with the smallest p-value. The normality of expression distribution and homogeneity of expression variance were checked using Lilliefors and F test. To define gene sets the

KEGG pathway list was used. It was obtained via the KEGGREST Bioconductor package giving 299 different gene sets [1].

3 Methods

3.1 Gene Set Enrichment Algorithms

Five different algorithms for detection of overrepresented/enriched pathways were intensely analyzed. First one is the Overrepresentation Analysis (ORA) [6], in which for each gene set the contingency table is constructed from the number of DEGs and non-DEGs. Then, hypergeometric test with chi-square distribution is used to establish significance of the gene set. Next algorithm is the Gene Set Enrichment Analysis (GSEA) [3], which is the most commonly used method. The GSEA method tests if the distribution of the gene ranks in the gene set differs from a uniform distribution by weighted Kolmogorov-Smirnov test. As a ranking metric, the Baumgartner-Weiss-Schindler statistic was used [11], which was highlighted as one of the best metric [9]. The next two methods at first summarize the expression of genes within the same pathway for every sample into single value. The Gene Set Variation Analysis (GSVA) [12] estimates the expression distribution over the sample by non-parametric kernel distribution, which allows to get common scale of expression profiles. Next, it calculates the Kolmogorov-Smirnov-like statistic to get a summary score. Pathway Level Analysis of Gene Expression (PLAGE) [13] method standardizes expressions by z-score calculation and then performs the singular value decomposition (SVD). Next, the first right-singular vector of coefficients (similar to the first component in PCA) is taken as the summary score. The final p-value of pathway enrichment for both methods is calculated by performing two-sample t-test with unequal variances (Welch approximation) on summarized data. The last method is Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) [14]. The main idea of PADOG is to calculate weights for each gene to separate the genes appearing in a few gene sets, versus genes that appear in many gene sets. Further, the gene set score is calculated as the mean of absolute values of weighted moderated gene t-scores.

Table 1. Initial settings for all algorithms used in the study.

Method	Permutations	Gene set size filtration	Gene ranking metric
GSEA	1000	<15 genes & >500 genes	BWS
GSVA	NA	<15 genes & >500 genes	NA
ORA	NA	NA	Welch-test p-values
PADOG	1000	<15 genes	internal metric
PLAGE	NA	<15 genes & >500 genes	NA

Those five algorithms were chosen due to the following reasons. GSEA and ORA are the most commonly used methods. PADOG and PLAGE were shown to be the one of the most sensitive and specific algorithms in [7] and GSVA was selected as an alternative to PLAGE method. Except ORA, all algorithms are classified as self-contained

methods and perform sample permutation to obtain a score for each gene set [15]. The starting parameters for each algorithm are presented in Table 1.

3.2 Computational Experiment Scheme

The number of analyzed KEGG pathways was reduced due to internal filtration of each algorithm. The number of significantly enriched KEGG pathways was established based on the gene set score given from individual methods and Bonferroni multiple testing correction (to give the most conservative estimation). As the same disease was investigated in each dataset, the algorithm that detects similar pathways in every dataset is thought to give reproducible results. The pathways detected across five or six datasets were further validated by the literature study. To estimate the false positive rate, the original phenotypes for each dataset were permuted, creating 50 independent data collections. Using the new datasets, the level of detected pathways was checked and compared to the expected one. Finally, to compare all methods the score proposed by Jaakkola and Elo (JE score) was calculated [8]. In general, the JE measure checks the ratio between weighted number of pathways detected for multiple datasets and the average number of estimated false positives. JE score is calculated as:

$$JE\ score\ (method) = \frac{1000}{W(n - \left\lceil \frac{n}{2} \right\rceil + 1)^2} \left[\frac{\sum_{h=\left\lceil \frac{n}{2} \right\rceil}^{n} \beta(method, h) * (h - \left\lceil \frac{n}{2} \right\rceil + 1)^2}{(\alpha(method) + 1)^2} \right] \quad (1)$$

where n is the total number of analyzed datasets, β is the number of significant gene sets under h datasets, α represents the average estimated false positive rate and W is the total number of analyzed gene sets. The higher value of the JE score means that the more reproducible results and low false positive rate are provided by the algorithm.

4 Results and Discussion

At first stage, all analyzed datasets were checked for expression signal normality and homogeneity of variance. For all six, the non-normal distributions and non-homogenous variances were observed for majority of genes. It implies the usage of non-parametric method for gene ranking in GSEA and for finding DEGs in ORA. The number of investigated KEGG pathways was reduced to common level that was set by each algorithm (from 299 to 192). The statistically significant pathways were estimated at 0.05 significance level using gene set score from each method and Bonferroni correction for multiple testing was performed. Results of finding enriched pathways are presented in Table 2. Since all datasets analyze the same type of cancer (ccRCC) it is expected that the algorithms should give similar results for each investigated dataset. PLAGE algorithm detected the largest number of KEGGs common across five and six datasets (52.61% of analyzed pathways). GSVA and GSEA algorithms gave the smaller number of common pathways: 7.81% and 3.64%, respectively. ORA and PADOG did not find any significant pathways shared by at least 5 datasets. It shows their weakness in

reproducibility of the results. To support those findings and include false positive estimation, the Jaakkola and Elo score was calculated and the results are presented in Table 3.

Table 2. Results of finding common enriched pathways across tested datasets.

Method	GSEA	PADOG	PLAGE	GSVA	ORA
Total	192				
not significant	116	**185**	0	79	175
%	*60.42%*	***96.35%***	*0.00%*	*41.15%*	*91.15%*
in 1 dataset	**38**	5	3	27	5
%	***19.79%***	*2.60%*	*1.56%*	*14.06%*	*2.60%*
in 2 datasets	14	1	6	**23**	4
%	*7.29%*	*0.52%*	*3.13%*	***11.98%***	*2.08%*
in 3 datasets	9	1	21	**24**	6
%	*4.69%*	*0.52%*	*10.94%*	***12.50%***	*3.13%*
in 4 datasets	8	0	**61**	24	2
%	*4.17%*	*0.00%*	***31.77%***	*12.50%*	*1.04%*
in 5 datasets	3	0	**78**	11	0
%	*1.56%*	*0.00%*	***40.63%***	*5.73%*	*0.00%*
in 6 datasets	4	0	**23**	4	0
%	*2.08%*	*0.00%*	***11.98%***	*2.08%*	*0.00%*

Table 3. Results of calculating Jaakkola and Elo score (the higher the better).

	GSEA	PADOG	PLAGE	GSVA	ORA
JE score	42.20	0.32	434.57	92.12	4.56

The best JE score was obtained by PLAGE, which gave the largest number of significant pathways and low false positive ratio. As in previous comparison, next top algorithms are GSVA and GSEA. They gave comparable results, however, the GSVA algorithm showed higher JE score due to the larger number of KEGGs common across several datasets. Second to last was ORA algorithm and the weakest one was PADOG method. These results are partially in opposite to the one obtained in [7], where PLAGE and PADOG showed high sensitivity and specificity of finding target pathways and ORA was also distinguished. From this group, in this study only PLAGE gave reproducible results across several datasets. Furthermore, GSVA and GSEA algorithms seem to be more reproducible compared to PADOG and ORA.

The PLAGE method gave much better results in comparison to other algorithms, but the number of enriched pathways looks overestimated. In Fig. 1 the average percent of significantly enriched/overrepresented gene sets is presented for different significance level. PLAGE gave very low p-values for almost all gene sets. It indicated nearly 80% of the gene sets as relevant, even for the significance level established by the most conservative Bonferroni multiple testing correction method. Such huge percent of enriched pathways suggest that most of the analyzed gene sets are related to the

investigated disease. From the biological point of view the difference in gene expression between healthy and cancer patients occurs in at most 10% of genes [16], so the number of enriched pathways cannot be as high as PLAGE showed. Thus, it can be concluded that results of PLAGE method are overestimated. In contrary, PLAGE has an acceptable level of false positives (Fig. 1 – second row).

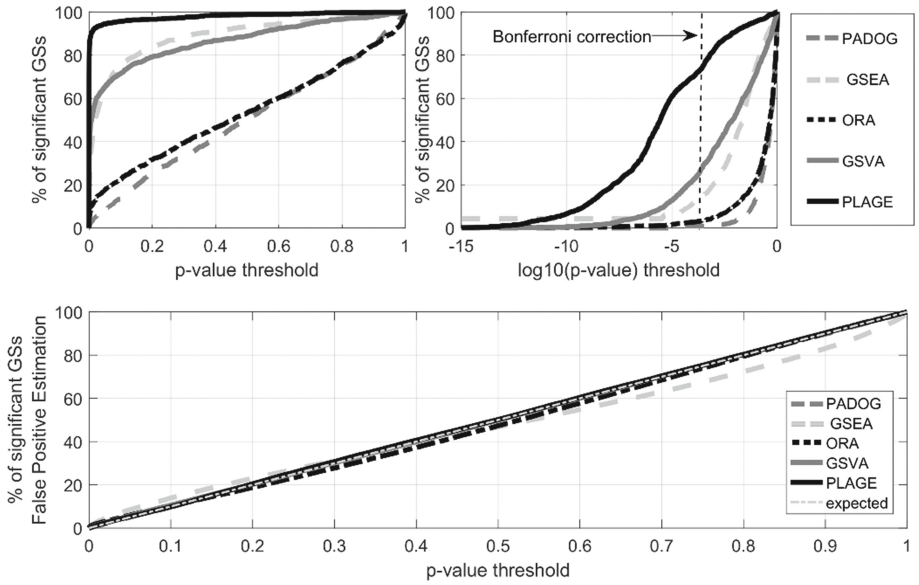


Fig. 1. Results for detecting significant gene sets across various thresholds. First row represents percent of significant pathway by average for each algorithm in linear (left figure) and logarithmic scale (right figure). Second row present false positive estimation.

The common outcomes for GSVA and GSEA can be the result of similar methodology: both use the Kolmogorov-Smirnov-like statistics. The idea of PLAGE and GSVA are similar, however, the singular value decomposition used in PLAGE leads to overestimation of results. The poor results given by PADOG can be caused by non-normal distribution of the expression data. The algorithm itself uses the moderated t-test to evaluate gene level statistic. Similar reason for weaker outcomes can be stated for ORA, where Welch approximation of t-test was used.

If we discard PLAGE due to a possible overestimation of enrichment p-values, GSVA and GSEA algorithms can be pointed as those, which give the most reproducible results across all datasets. Also, they gave an acceptable level of false positives. The KEGG pathways detected across five and six datasets by GSVA (15 gene stets) and GSEA (7 gene sets) were further investigated by literature search (Fig. 2).

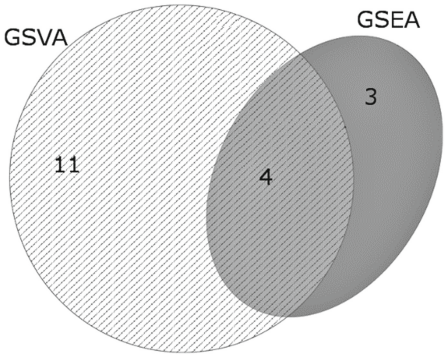


Fig. 2. Venn diagram presenting the number of detected gene sets across five and six datasets by GSVA and GSEA algorithms (created in eulerAPE software [17]).

First, four pathways detected by both algorithms were removed, and only pathways specific to given method were investigated. In case of results given by GSEA algorithm, only one pathway (Propanoate metabolism pathway) was previously reported as related to ccRCC [18–20]. The two remaining pathways were not associated with ccRCC (Staphylococcus aureus infection, Phagosome), which gave 33% validity level. For GSVA algorithm, there were two pathways evidently connected with cancer disease (DNA replication and p53 signaling pathway). Three other pathways were previously associated with ccRCC: Systemic lupus erythematosus [21], Type I diabetes mellitus [19] and Toll-like receptor signaling pathway [22]. This shows that GSVA algorithm can give not only reproducible results but also appropriate in a biological sense (45% of pathways previously reported to ccRCC). Nevertheless, all presented findings are based on one collection of datasets for ccRCC disease. This fact can affect the obtained results, thus study on other large collection of datasets devoted to different trait should be performed.

5 Conclusions

The comprehensive comparison of five algorithms for detection of pathway disorders was performed. It was shown that PLAGE method detects most pathways as associated with the investigated trait, which may indicate an overestimation. So, there is still a need to find a new measure for gene set analysis results, that favours reproducibility but gives penalty for overestimation of the results. GSVA algorithm can be highlighted for very reproducible results and keeping an acceptable level of false positives. The weakest results were obtained by ORA and PADOG. Those findings show that not only high sensitivity and specificity of enrichment algorithms should be taking into consideration like in [7] but also reproducibility of the results.

Acknowledgements. This work was financed by SUT grant no. BKM/506/RAU1/2016/t.26 (JZ), 02/010/BK_16/3015 (MM) and NCN grant no. 2015/19/B/ST6/01736 (JP). All calculations

were carried out using GeCONiI infrastructure funded by NCBiR project no. POIG. 02.03.01-24-099/13.

References

1. Kanehisa, M., et al.: KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. **44**(D1), D457–D462 (2016)
2. Ashburner, M., et al.: Gene ontology: tool for the unification of biology. Nat. Genet. **25**(1), 25–29 (2000)
3. Subramanian, A., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS **102**(43), 15545–15550 (2005)
4. Van Dongen, S., Abreu-Goodger, C., Enright, A.J.: Detecting microRNA binding and siRNA off-target effects from expression data. Nat. Methods **5**(12), 1023–1025 (2008)
5. Laaksonen, R., et al.: A systems biology strategy reveals biological pathways and plasma biomarker candidates for potentially toxic statin-induced changes in muscle. PLoS ONE **1**(1), e97 (2006)
6. Beißbarth, T., Speed, T.P.: GOstat: find statistically overrepresented Gene ontologies within a group of genes. Bioinformatics **20**(9), 1464–1465 (2004)
7. Tarca, A.L., Bhatti, G., Romero, R.: A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS ONE **8**(11), e79217 (2013)
8. Jaakkola, M.K., Elo, L.L.: Empirical comparison of structure-based pathway methods. Brief. Bioinform. **17**(2), 336–345 (2016)
9. Zyla, J., Marczyk, M., Weiner, J., Polanska, J.: Ranking metrics in gene set enrichment analysis: do they matter?. BMC Bioinform. **18**(1), 256 (2017)
10. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**(1), 207–210 (2002)
11. Baumgartner, W., Weiß, P., Schindler, H.: A nonparametric test for the general two-sample problem. Biometrics **54**, 1129–1135 (1998)
12. Hänzelmann, S., Castelo, R., Guinney, J.: GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinform. **14**(1), 7 (2013)
13. Tomfohr, J., Lu, J., Kepler, T.B.: Pathway level analysis of gene expression using singular value decomposition. BMC Bioinform. **6**(1), 225 (2005)
14. Tarca, A.L., Draghici, S., Bhatti, G., Romero, R.: Down-weighting overlapping genes improves gene set analysis. BMC Bioinform. **13**, 136 (2012)
15. Maciejewski, H.: Gene set analysis methods: statistical models and methodological differences. Brief. Bioinform. **15**(4), 504–518 (2014)
16. Anand, P., et al.: Cancer is a preventable disease that requires major lifestyle changes. Pharm. Res. **25**(9), 2097–2116 (2008)
17. Micallef, L., Rodgers, P.: euler APE: drawing area-proportional 3-Venn diagrams using ellipses. PLoS ONE **9**(7), e101717 (2014)
18. Zaravinos, A., et al.: Altered metabolic pathways in clear cell renal cell carcinoma: a meta-analysis and validation study focused on the deregulated genes and their associated networks. Oncoscience **1**(2), 117 (2014)
19. Huang, H., et al.: Key pathways and genes controlling the development and progression of clear cell renal cell carcinoma (ccRCC) based on gene set enrichment analysis. Int. Urol. Nephrol. **46**(3), 539–553 (2014)
20. Tun, H.W., et al.: Pathway signature and cellular differentiation in clear cell renal cell carcinoma. PLoS ONE **5**(5), e10696 (2010)

21. Zheng, H., Guo, X., Tian, Q., Li, H., Zhu, Y.: Distinct role of Tim-3 in systemic lupus erythematosus and clear cell renal cell carcinoma. Int. J. Clin. Exp. Med. **8**(5), 7029 (2015)
22. Morikawa, T., et al.: Identification of Toll-like receptor 3 as a potential therapeutic target in clear cell renal cell carcinoma. Clin. Cancer Res. **13**(19), 5703–5709 (2007)

Towards Trustworthy Predictions of Conversion from Mild Cognitive Impairment to Dementia: A Conformal Prediction Approach

Telma Pereira[1(✉)], Sandra Cardoso[2], Dina Silva[2], Alexandre de Mendonça[2], Manuela Guerreiro[2], and Sara C. Madeira[3]

[1] INESC-ID and Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal
telma.pereira@ist.utl.pt
[2] Laboratory of Neurosciences, Faculty of Medicine, Institute of Molecular Medicine, University of Lisbon, Lisbon, Portugal
[3] LASIGE and Faculty of Sciences, University of Lisbon, Lisbon, Portugal

Abstract. Predicting progression from a stage of Mild Cognitive Impairment to Alzheimer's disease is a major pursuit in current dementia research. As a result, many prognostic models have emerged with the goal of supporting clinical decisions. Despite the efforts, the lack of a reliable assessment of the uncertainty of each prediction has hampered its application in practise. It is paramount for clinicians to know how much they can rely upon the prediction made for a given patient, in order to adjust treatments to the patient based on that information. In this exploratory study, we evaluated the Conformal Prediction approach on the task of making predictions with precise levels of confidence. Conformal prediction showed promising results. Using high confidence levels have the drawback of leaving a large number of MCI patients without prognostic (the classifier is not confident enough to give a single class). When using forced predictions, conformal predictors achieved classification performances as good as standard classifiers, with the advantage of complementing each prediction with a confidence value.

Keywords: Conformal predictors · Confidence estimation · Mild cognitive impairment · Alzheimer's disease · Prognostic prediction

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disease, causing cognitive impairment, with devastating effect on patients and their families, and a huge socio-economic impact in modern societies. Nowadays, more than 30 million people suffer from AD worldwide and its prevalence is expected to triple by 2050 [1]. Mild Cognitive Impairment (MCI) is considered as a transitive stage between healthy aging and dementia [1], suggesting these patients as a group of singular interest to follow-up studies and interventions. In this context, studying the predictive value of MCI for the progression to dementia is a major challenge in current dementia research [2, 3].

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_19

Machine learning is at the core of many recent advances in dementia-related research [4]. By following different approaches and using different types of data, researchers have sought for robust prognostic models, to guide clinical decisions, by means of a medical decision support system to be used in clinical settings. This system would predict the most likely prognostic for a new MCI patient based on the past history of a cohort of patients with known diagnostics. If the prediction is trustworthy, clinicians then use it to timely adjust the treatment and medical appointments, and administer more effective treatments [3, 5]. Despite the advances made in prognostic prediction of MCI patients, the lack of an indication about the confidence of each prediction have hampered its practical applicability. For clinicians, it is paramount to know how much they can trust on the prognostic predicted for a new patient, in order to pursue with treatments relied on that information [6, 7]. Assessing the classifier's performance in single examples (patients) is also useful in ensemble applications [7]. The standard assessment metrics used to evaluate the average performance on an independent dataset (as the accuracy) are not suitable to these problems, as we want to assess the reliability in the classification of each individual patient.

Conformal Prediction (CP) has been proposed to tackle this problem [8, 9]. It predicts the class that makes the new example (patient) more "conform" to the training set, with precise confidence levels. A confidence level of 0.9, for instance, means that the conformal predictors commit a maximum of 10% of errors. This approach has been used in disease-related problems [10, 11]. In this study, we apply the conformal prediction framework to the prognostic problem of MCI-to-AD conversion. To our knowledge, this was not explored to date. In this exploratory study, we aimed to evaluate how accurate and reliable are the predictions given by the CP framework.

2 Conformal Prediction

We introduce the idea behind the conformal prediction framework. For a more formal description we refer to [8, 9]. Let us assume that we are given a training set $\{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$, where $x_i \in X$ is a vector of attributes and $y_i \in Y$ is the class label (assuming a binary classification problem). Given a new test example (x_n) we aim to predict its class. Intuitively, we assign each class $y_n \in Y$ to x_n, at a time, and then we evaluate how "strange" or "non-conform" the example (x_n, y_n) is in comparison with the training set. We assume that the most likely class label conforms better with the training set. A non-conformity measure, to assess the strangeness of the test example, must be extracted from the underlying classifier. To evaluate how different x_n is from the training set, we compare its non-conformity score with those of the remaining training examples $x_j, j = 1, \ldots, n-1$, using the p-value function (distinct from the p-value from statistics):

$$p(\alpha_n) = \frac{|\{j = 1, \ldots, n : \alpha_j \geq \alpha_n\}|}{n} \tag{1}$$

where α_n is the non-conformity score of x_n, assuming it is assigned to the class label y_n. If the p-value is small, then the test example (x_n, y_n) is non-conforming, since few

examples (x_i, y_i) had a higher non-conformity score when compared with α_n. On the other hand, if the p-value is large, x_n is very conforming, since most of the examples (x_i, y_i) had a higher non-conformity score when compared with α_n.

For a given significance level ε, Conformal Predictors (CPs) output a prediction region, T^{ϵ}: set of all classes with $p(\alpha_n) > \varepsilon$, contrarily to the single predictions given by standard classifiers. These prediction regions have a guaranteed error rate. This means that the frequency of errors (fraction of true values outside T^{ϵ}) does not exceed ε, at a confidence level $1 - \varepsilon$. The error rate is guaranteed under the randomness assumption, which states that the examples are independently drawn from the same distribution (this property is called validity) [8]. Prediction regions may therefore comprise more than one class (uncertain prediction), any class (empty prediction) or a single class (certain prediction). Multiple predictions are not mistakes but a reflection that the classifier was not confident enough to predict a certain class. The smaller the prediction region, the more efficient the conformal predictor [8].

Alternatively, we may force the conformal predictors to output a single prediction, predicting the class with the highest p-value (forced prediction), at the cost of losing the guaranteed confidence level. The highest p-value is the credibility of the prediction while its confidence is given by the complement to 1 of the second highest p-value.

Conformal prediction may be used in the transductive or in the inductive setting. When transductive framework is used, the training set is enriched with the test example, and the underlying classifier is updated. Non-conformity scores are then computed, for all the training examples. This process is repeated for all class labels $y \in Y$. A new prediction is therefore based on all the training examples. For large datasets, this is computationally very demanding. This led to the emergence of inductive learning [12, 13]. When inductive is used, the training set $\{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$ is divided into the proper training set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ and the calibration set $\{(x_{m+1}, y_{m+1}), \ldots, (x_{n-1}, y_{n-1})\}$, where $m < n - 1$. The proper training set is used to derive the prediction rule, by training the underlying classifier. This prediction rule is then used to classify the examples of the calibration set and the test example. Non-conformity scores are only computed with the examples of the calibration set.

Mondrian conformal prediction is a variant of CP that deals with imbalanced datasets [8]. When the number of examples of a given class is significantly larger than those of the other class, most errors are putatively from the minority class, limiting the applicability of these predictions. Mondrian conformal prediction applies CPs separately to each label class. The p-value is thus computed by comparing the non-conformity score of the test example against only training examples of the same class as the current hypothesis y_n:

$$p(\alpha_n) = \frac{|\{j = 1, \ldots, n : y_j = y_n \text{ and } \alpha_j \geq \alpha_n\}|}{|\{j = 1, \ldots, n : y_j = y_n\}|} \tag{2}$$

3 Methods

3.1 Data

Participants were selected from a revised version of the Cognitive Complaints Cohort (CCC) [2]. This is a prospective study conducted at the Faculty of Medicine of Lisbon to investigate the progression to dementia in subjects with cognitive complaints. It is based on an extensive neuropsychological evaluation at one of the participating institutions (Laboratory of Language Studies, Santa Maria Hospital, and a Memory Clinic, both in Lisbon, and the Neurology Department, University Hospital in Coimbra). The neuropsychological battery was validated in the Portuguese population and assesses different cognitive domains, such as memory and reasoning (BLAD [14]). In total, 90 variables covering clinical, demographic and neuropsychological data were used, whose description may be found in [2]. In this study, we selected patients diagnosed with MCI at baseline, who had at least one follow-up appointment and were followed for at least 3 years. The dataset comprised 160 (57%) patients who converted to dementia (positive class: converter MCI, denoted cMCI) while 122 (43%) did not convert throughout the study (negative class: stable MCI, denoted sMCI).

3.2 Conformal Prediction Framework

Given that the dataset under study does not have high dimensionality, we decided to use the Transductive Conformal Prediction framework. In addition, despite the minor imbalance of classes, we decided to study how Mondrian CP would perform in our case study. We tested four significance levels ($\varepsilon = \{0.10, 0.15, 0.20, 0.30\}$). The dataset was randomly split (keeping class proportions) in training set (60%) and test set (40%). Correlation-based feature selection was run on the training set, in order to select relevant features. The classification approach was implemented in Java using WEKA's functionalities (version 3.8.0).

Nonconformity measures. We used k-Nearest Neighbors (kNN, k set to 3) and Naïve Bayes as underlying classifiers along with the non-conformity measures described in Table 1. We decided to use Naïve Bayes since, in previous experiments, it outperformed other commonly used classifiers (such as SVMs, Decision Trees and Random Forests) in the MCI-to-dementia conversion problem [15]. We also used kNN since it is widely used in conformal prediction studies [12, 16].

Table 1. Non-conformity measures for the classifiers used in this study.

Classifier	Non-conformity measure	Comment
kNN	$\frac{\sum_{j\neq i:y_j=y_i}^{k} d(x_j,x_i)}{\sum_{j\neq i:y_j\neq y_i}^{k} d(x_j,x_i)}$	Sum of the distances to the k nearest neighbors
Naïve Bayes	$-\log p(y_i = c\|x_i)$	p is the posterior probability estimated by Naïve Bayes

4 Results and Discussion

For each significance level ε , conformal predictors may produce multiple prediction regions: (1) a single class (the p-value is smaller than ε for one of the classes, Certain prediction), (2) two classes (the p-value of both classes is larger than ε, Uncertain prediction), (3) no class (the p-value of both classes is smaller than ε, Empty prediction). One patient is thus classified as cMCI (or sMCI) if and only if cMCI (or sMCI) is the only label in the prediction region.

According to the validity property of CPs the prediction error rate (when the prediction region does not contain the real class) is not larger than the predefined significance level $\varepsilon > 0$. The CPs used in this study proved to be valid, as illustrated in Table 2 (rightmost column). When there are no Empty predictions, the error rate is the ratio of the number of certain but wrong predictions ("cMCI predicted sMCI" and "sMCI predicted cMCI") to the total number of cMCI and sMCI examples. The validity was verified across all the experiments. However, due to space limitations, we reported only the results obtained with the Transductive CP.

Table 2. Predictions obtained with Transductive CP framework using Naive Bayes (*NB*) and *k*-Nearest Neighbors (*kNN*, *k* set to 3) with different significance levels.

Significance		cMCI pred cMCI	cMCI pred sMCI	sMCI pred sMCI	sMCI pred cMCI	Empty	Uncertain	Error Rate
0.10	*NB*	5	0	7	0	0	100	0.018
	kNN	4	0	0	1	0	109	0.009
0.15	*NB*	13	1	11	2	0	87	0.026
	kNN	9	1	1	1	0	102	0.018
0.20	*NB*	17	1	15	2	0	79	0.026
	kNN	13	1	5	1	0	92	0.035
0.30	*NB*	23	3	20	3	0	65	0.052
	kNN	21	1	14	7	0	71	0.070

The implemented conformal predictors have been proven to be always valid [8, 17], theoretically or empirically (depending on the CPs settings). As such, researchers' attempts have been focused on improving their efficiency (prediction regions' size) [17]. Regarding the prognostic problem under study, we aim to train a CP to output a certain prognostic prediction (the patient will or will not evolve to dementia), with a known confidence level, in order to support the clinician's decision. A conformal predictor that outputs mostly uncertain predictions hampers its clinical applicability.

The efficiency varies with the significance level as illustrated in Table 2 and in Fig. 1. The size of the prediction region increases (number of multiple predictions increases) as the significance level decreases. This means that for a smaller significance level a certain and putatively correct prediction (sMCI or cMCI) might be then predicted as uncertain, since both p-values are then larger than the actual ε. Contrarily, for a higher significance level, it may happen that a certain prediction becomes an Empty prediction

(both p-values $< \varepsilon$). As illustrated in Table 2, when we reduce the significance level from 0.3 to 0.1 (using the CP Naïve Bayes) the number of uncertain prediction raises from 65 to 100, although the number of wrong certain predictions drops to 0. In Fig. 1 we may also observe that the number of certain predictions increases with the significance level. Being more confident has therefore the cost of having a less efficient CP, in the sense that it outputs a larger number of uncertain predictions. Depending on the problem where the model will be applied, we should find a trade-off between the number of error allowed (significance level) and the number of certain predictions obtained.

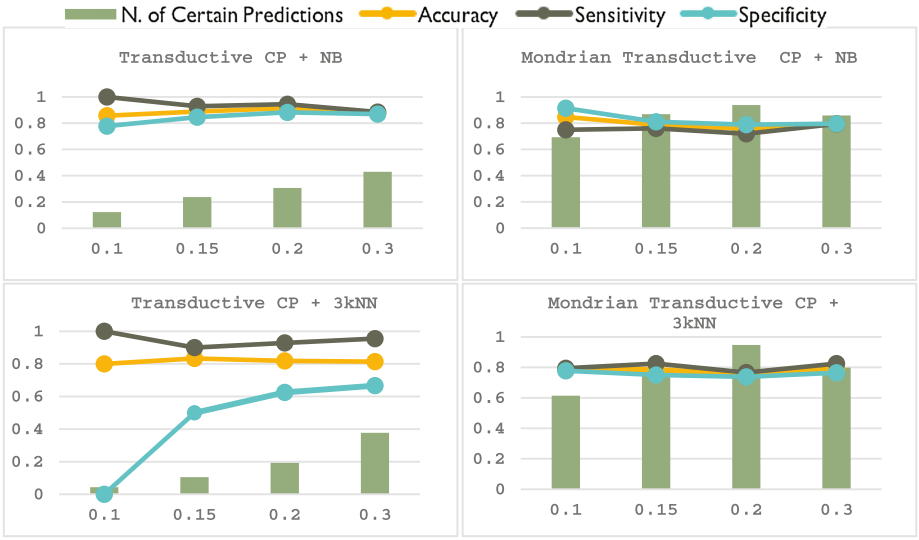


Fig. 1. Classification performance and proportion of certain predictions obtained with conformal prediction framework using Naïve Bayes and k-Nearest Neighbors (kNN, $k = 3$) as underlying classifiers and different settings (Transductive and Mondrian Transductive) in function of the significance level ($\varepsilon = \{0.10, 0.15, 0.20, 0.30\}$). Only certain predictions were used to compute the evaluation metrics (accuracy, sensitivity and specificity).

Since we were interested in evaluating how CPs performed on making predictions for individual examples, we calculated the accuracy, sensitivity and specificity of each prediction obtained with the test set (considering only certain predictions). The results are presented in Fig. 1. The accuracy is sometimes inferior to what was expected given the percentage of errors allowed by conformal predictors (validity property). This happens because we are not considering the "correct predictions" when both classes belong to the predictions region (uncertain predictions). The number of certain predictions obtained with the transductive CP (NB and kNN) was low, despite having good performances (accuracy around 0.80). Moreover, these classifiers seem to have more difficulty to predict non-converting patients (sMCI), evidenced by the smaller specificity values. This was attenuated by using Mondrian learning. Besides enhancing the specificity, Mondrian CPs also outputted a larger number of certain predictions, even for small significance level values. As an example, Mondrian CP with Naïve Bayes achieved good

performances for a high level of confidence ($\varepsilon = 0.1$, Accuracy = 0.84, Sensitivity = 0.75, Specificity = 0.91, 70% of certain predictions).

Regarding the prognostic problem, since we want to minimize classification errors while providing prognosis for as many patients as possible, a good trade-off is to use mondrian transductive conformal prediction along with low values of significance ($\varepsilon = 0.1$ or $\varepsilon = 0.15$).

Although CPs have been designed with the purpose of providing prediction regions with guaranteed error rate, we can force them to always give a single forced prediction. After the p-value assignment, we predict the class with the highest p-value. However, we should bear in mind that by doing this, we lose the guaranteed validity. We can, however, compute the credibility (highest p-value) and confidence (complement to 1 of the second highest p-value) of each prediction, as mentioned in Sect. 2. We did this experiment in order to compare CPs with standard classifiers. According to the results (Table 3), conformal predictors performs as good, or even better, as the standard classifiers, especially when using the kNN classifier.

Table 3. Results obtained with standard classifiers (NB and kNN) and with Conformal Predictors with forced predictions.

NB		CP-NB		kNN	CP-kNN	
		Transductive	Mondrian transductive		Transductive	Mondrian transductive
Acc.	0.737	0.728	0.746	0.675	0.737	0.746
Sens.	0.740	0.740	0.740	0.360	0.800	0.760
Spec.	0.737	0.719	0.737	0.922	0.688	0.734

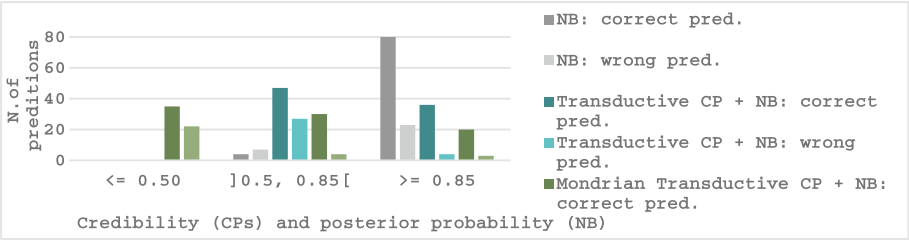


Fig. 2. Proportion of correct and wrong predictions predicted within three intervals of credibility (CP framework) or posterior probability (Naïve Bayes).

Since the posterior probabilities given by the Naïve Bayes classifier may be indicative of the quality of each prediction, we compared this measure with its equivalent on CP, the credibility (Fig. 2). More specifically, we evaluated how many correct and wrong predictions had low, moderate or high values of credibility together with the posterior probabilities. Most posterior probabilities were higher than 0.85, for both correctly and incorrectly classified instances. NB thus lacks of discriminative power to assess the trustworthiness of individual predictions, since both correct and wrong predictions have, indiscriminately, high values of posterior probabilities. On the other side, CPs produce

only 12% of errors with credibility superior to 0.85. Credibility values of correct predictions may be used to stratify patients regarding the "certainty" on their conversion risk, allowing the clinicians to adjust the treatments accordingly.

5 Conclusions

This paper presents an application of Conformal Prediction to the prognostic problem of conversion from MCI to AD, in a real-world dataset. The main purpose was to inspect whether CPs produced trustworthy predictions for a given patient. This information is paramount in the clinical practice.

Conformal predictors output prediction regions, containing the correct class within a precise level of confidence. High confidence levels have the advantage of guaranteeing a minor number of errors, but they limit considerably the number of certain predictions, and so, the number of patients with prognostic. Although this effect was attenuated by using Mondrian learning, further work should be carried out to improve their efficiency. When using forced predictions, conformal prediction proved to be a valuable framework, performing as good as standard classifiers and, complementing each prediction with credibility and confidence values.

In the clinical practice, clinicians may use CPs in a first step to reveal the set of patients with prognostic prediction with guaranteed error (prediction regions). Forced predictions may then be made for those patients with uncertain predictions. Despite not having a guaranteed validity, it gives clinicians insight of their prediction's reliability. Clinicians can then prescribe more specific exams for those patients to whom the model produced less confident predictions.

Acknowledgments. This work was partially supported by FCT under the Neuroclinomics2 (PTDC/EEI-SII/1937/2014) and UID/CEC/S0021/2013, and an individual doctoral grant to TP (SFRH/BD/95846/2013).

References

1. Prince, M., et al.: World Alzheimer Report 2015: The Global Impact of Dementia - An Analysis of Prevalence, Incidence, Cost and Trends, London (2015)
2. Silva, D., et al.: Prediction of long-term (5 years) conversion to dementia using neuropsychological tests in a memory clinic setting. J. Alzheimers Dis. **34**, 681–689 (2013)
3. Barnes, D.E., et al.: A point-based tool to predict conversion from mild cognitive impairment to probable Alzheimer's disease. Alzheimers. Dement. **10**, 646–655 (2014)
4. Moradi, E., et al.: Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage **104**, 398–412 (2014)
5. Lee, S.J., et al.: A clinical index to predict progression from mild cognitive impairment to dementia due to Alzheimer's disease. PLoS ONE **9**, e113535 (2014)
6. Ribeiro, M.T., et al.: Why should I trust you? Explaining the predictions of any classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), p. 4503 (2016)

7. Papadopoulos, H.: Reliable probabilistic prediction for medical decision support. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) AIAI/EANN -2011. IAICT, vol. 364, pp. 265–274. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23960-1_32
8. Vovk, V., et al.: Algorithmic Learning in a Random World. Springer, New York (2005)
9. Shafer, G., et al.: A tutorial on conformal prediction. J. Mach. Learn. Res. **9**, 371–421 (2008)
10. Devetyarov, D., et al.: Conformal predictors in early diagnostics of ovarian and breast cancers. Prog. Artif. Intell. **1**, 245–257 (2012)
11. Lambrou, A., Papadopoulos, H., Kyriacou, E., Pattichis, C.S., Pattichis, M.S., Gammerman, A., Nicolaides, A.: Assessment of stroke risk based on morphological ultrasound image analysis with conformal prediction. In: Papadopoulos, H., Andreou, A.S., Bramer, M. (eds.) AIAI 2010. IAICT, vol. 339, pp. 146–153. Springer, Heidelberg (2010). doi: 10.1007/978-3-642-16239-8_21
12. Toccaceli, P., Nouretdinov, I., Gammerman, A.: Conformal predictors for compound activity prediction. In: Gammerman, A., Luo, Z., Vega, J., Vovk, V. (eds.) COPA 2016. LNCS, vol. 9653, pp. 51–66. Springer, Cham (2016). doi:10.1007/978-3-319-33395-3_4
13. Norinder, U., et al.: Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. J. Chem. Inf. Model. **54**, 1596–1603 (2014)
14. Guerreiro, M.: Contributo da Neuropsicologia para o Estudo das Demências, Ph.D. thesis, Faculdade de Medicina de Lisboa (1998)
15. Pereira, T., et al.: Predicting conversion of Mild Cognitive Impairment to Alzheimer's disease: a time windows approach. In: INForum Simpósio de Informática, Lisbon (2016)
16. Balasubramanian, V.N., Baker, A., Yanez, M., Chakraborty, S., Panchanathan, S.: PyCP: an open-source conformal predictions toolkit. In: Papadopoulos, H., Andreou, A.S., Iliadis, L., Maglogiannis, I. (eds.) AIAI 2013. IAICT, vol. 412, pp. 361–370. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41142-7_37
17. Devetyarov, D., Nouretdinov, I.: Prediction with confidence based on a random forest classifier. In: Papadopoulos, H., Andreou, A.S., Bramer, M. (eds.) AIAI 2010. IAICT, vol. 339, pp. 37–44. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16239-8_8

Topological Sequence Segments Discriminate Between Class C GPCR Subtypes

Caroline König[1(✉)], René Alquézar[1], Alfredo Vellido[1,2], and Jesús Giraldo[3,4]

[1] UPC BarcelonaTech, Univ. Politècnica de Catalunya, 08034 Barcelona, Spain
{ckonig,alquezar,avellido}@lsi.upc.edu
[2] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Cerdanyola del Vallès, Spain
[3] Institut de Neurociències - Unitat de Bioestadìstica, Univ. Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain
jesus.giraldo@uab.es
[4] Network Biomedical Research Center on Mental Health (CIBERSAM), Madrid, Spain

Abstract. G protein-coupled receptors are eukaryotic cell membrane proteins with a key role as extracellular signal transmitters. While GPCRs embrace a wide and heterogeneous super-family of proteins, our interest in this study is in its Class C, of great relevance to pharmacology. The scarcity of knowledge about their full 3-D crystal structure makes the use of their primary amino acid sequences important for analysis. In this paper, we systematically analyze whether segments of the receptor sequences are able to discriminate between the different class C GPCR subtypes according to their topological location on the extracellular, transmembrane or intracellular domain. For this, we build on previous research that showed that the use of the extracellular N-terminus domain on its own for this classification task did only entail a minor decrease in subtype discrimination when compared to the complete sequence. We use Support Vector Machine-based classification models to assess the subtype discriminating power of the topological segments.

Keywords: G-protein coupled receptors · Pharmaco-proteomics · Segmentation · Support vector machines

1 Introduction

G protein-coupled receptors (GPCRs) are proteins located in the eukaryotic cell membrane. This location determines their role in transmitting extracellular signals to the interior of the cell. Such key physiological role makes them a prevalent drug target in pharmacological research [1].

The current study does not cover the whole GPCR super-family, but specifically its class C [2] (defined according to the IUPHAR[1] convention). Members

[1] http://www.iuphar.org.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_20

of this class are relevant to the investigation of therapies for neurological diseases [3]. Despite recent impressive advances in the discovery of GPCR crystal structures [4], the information about tertiary and quaternary structure is very limited in the case of class C GPCRs [5,6]. In consequence, the information of the primary amino acid sequences of class C GPCRs (in this case well known and available from publicly accessible databases) is often analyzed as of the investigation of receptor functionality.

In previous research, the discrimination between the seven defined subtypes of class C GPCRs was investigated using supervised classification approaches. Experiments revealed a relatively high level of differentiation between the subtypes, but also a clear upper threshold to classification accuracy. This research was carried out both using transformations based on the physicochemical properties of the amino acids [7] and on short n-gram features [8]. It is important to note that these prior investigations used the complete and unaligned amino acid sequence of the receptors.

The GPCRs have different structural domains, including, amongst others, a seven-helix transmembrane (7TM) domain and an extracellular domain. In the case of class C, they include a large domain in the extracellular part of the receptor (N-terminus), which is built by the Venus Flytrap (VFT) and a cystein rich domain (CRD) connecting both in many of their subtypes [9]. Recently, we investigated whether the extracellular N-terminus domain of the sequences sufficed to discriminate between class C GPCR subtypes [10], and concluded that, even if the use of the N-terminus did not suffice to completely retain the subtype discrimination capabilities of the whole sequence, the decrease in classification performance was rather small.

In the current paper, we build on these preliminary results and provide a systematic analysis of the subtype discrimination capabilities of the complete set of different topological locations in the class C sequences (in extracellular, transmembrane and intracellular domains), including their combinations. We compare this with the performance of the complete sequence.

The remainder of the paper is organized as follows: the data analyzed in this study are briefly described in Sect. 2. This is followed in Sect. 3 by the description of the Support Vector Machine (SVM)-based classification strategy, the methods of sequential data transformation, the criteria for partition of the sequence in domains and sub-domains and, finally, the metrics used for performance evaluation. Experimental results are next presented and discussed and a few final conclusions of the study are outlined.

2 Materials

The data analyzed in our study was extracted from the GPCRdb [11] database system for GPCRs. This is part of the GPCR Consortium[2], which is an industry-academia partnership. GPCRdb divides the GPCR superfamily into

[2] URL: http://gpcrconsortium.org.

five major families according to IUPHAR and, as explained in the introduction, we only focus here on class C [2,12], for their relevance in pharmacoproteomics. Class C of GPCRs is in turn subdivided into seven main subtypes: Metabotropic Glutamate (MG) receptors, Calcium sensing (CS), GABA-B (GB), Vomeronasal (VN), Pheromone (Ph), Odorant (Od) and Taste (Ta). The analyzed data set from version 11.3.4, as of March 2011, comprises a total of 1,510 sequences from the seven subtypes. The current work restricts the analysis to the subset of 1,252 sequences that have information of the complete 7-TM domain. Table 1 shows the distribution of sequences per subtype both for the original data set and for the data set comprising only sequences with complete 7-TM structure.

Table 1. Number of sequences in the original data set and in the subset with complete 7-TM structure.

Class C subtype	♯ sequ. original dataset	♯ sequ. complete 7-TM structure
MG	351	282
CS	48	45
GB	208	156
VN	344	293
Ph	392	323
Od	102	90
Ta	65	62
	1510	1252

3 Methods

3.1 Supervised Classification Techniques

In the reported experiments, we first used several supervised models for the classification of the alignment-free amino acid sequences into the seven class C GPCR subtypes. The results obtained with the complete sequences were first used to decide which classifier to select for the remaining analyses. The comparison was carried out using similar classifiers to those already used in previous research [7], namely SVM [13], Random Forest (RF) [14] and Naïve Bayes (NB) [15].

The results for all classifiers were obtained applying 5-fold cross validation (5-CV) using stratification for folder creation. In the case of the SVM, the svmLib implementation [16] was used with a one-vs-one classification approach and applying a nonlinear kernel, namely the radial basis function (RBF) kernel: $K(x_i, x_j) = e^{(-\gamma||x_i - x_j||)}$. The use of the RBF kernel requires the setting of two parameters, the error penalty parameter C and the γ parameter of the kernel, through a grid search.

3.2 Alignment Free Transformations

Prior to the creation of classification models, the class C sequences of varying length had to be transformed into fixed size representations. In proteomics research, transformations based on the physicochemical properties of the amino acids are often used [17,18], but also transformations that draw inspiration from the field of symbolic language analysis, which treat sequences as text from a 20 amino acid alphabet. In the latter, the occurrence of short "words" also known as n-grams is usually investigated [19]. In this study, we followed this approach and calculated the relative frequency of occurrence of n-grams of sizes 1 and 2, which we call the AA and Digram transformations, respectively. This n-gram-based transformations yielded good classification results in previous research when the complete sequence of the original data set was analyzed [8]. Here, we calculated not only the frequencies of AA and Digram for all sequence segments under study (called *appended frequencies*), but also the *accumulated frequencies*, which are calculated as the occurrence of AA or Digram in all the segments under study divided by the sum of the lengths of these segments (Fig. 1).

3.3 Topological Segmentation

As explained in the introduction, class C GPCRs, being transmembrane proteins, have a common complex structure: An extracellular domain comprising the N-terminus and 3 extracellular loops (EL), the 7TM and an intracellular domain built by three intracellular loops (IL) and the C-terminus. According to this segmentation, the entire sequences consists of 15 segments, which were detected using the transmembrane detection tool Phobius [20]. Table 2 shows some statistics for the lengths (in number of amino acids) of these segments.

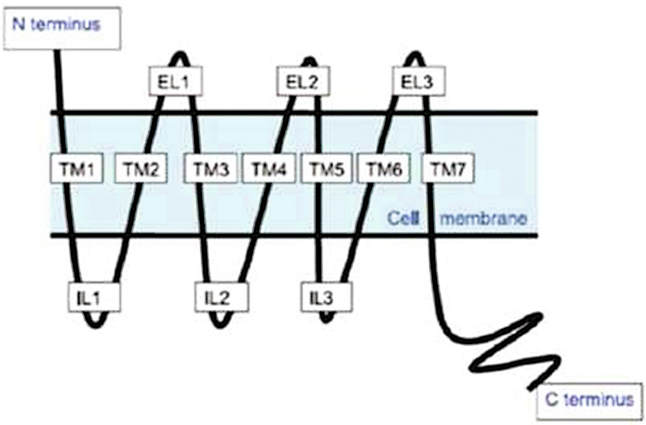


Fig. 1. Graphical representation of the common structure of GPCRs.

Table 2. Statistical information concerning the length of the segments.

Segment	Max	Min	Mean	StDev
Complete sequence	1,768	250	861.7	181
N-terminus	1,502	6	532.2	148.3
EL1	329	5	11.6	10.4
EL2	70	5	27	10.4
EL3	31	5	9	3.9
TM1	34	16	24.7	1.9
TM2	31	17	21.8	1.7
TM3	34	17	23.5	2.3
TM4	33	18	22.3	2.9
TM5	34	17	23.5	2.3
TM6	27	17	21.3	1.3
TM7	31	16	23.6	1.6
IL1	567	6	17	39.9
IL2	69	11	18.9	4.2
IL3	85	6	11.9	3.3
C-terminus	1,044	0	73	113

3.4 Metrics

The quality of the multi-class models was evaluated using the classification accuracy, which is the proportion of correctly classified receptors, and the Matthews correlation coefficient (MCC), which is, in principle, more robust when experiments involve unbalanced classes.

4 Experiments

4.1 Comparison of Classifiers

As explained in Sect. 3.1, a first batch of experiments with complete sequences was performed to select a classifier. The results in Table 3 reveal that the SVM outperforms RF and NB both for the AA and Digram transformations. In consequence, SVM was used in the remaining experiments.

4.2 Experiments with Topological Sequence Segments

The SVM experiments with the different topological segments and their combinations are reported next. Table 4 corresponds to segments in the extracellular domain; Table 5 to the 7TM and Table 6 to the four intracellular regions IL1, IL2 and IL3 and the C-terminus; Table 7, in turn, shows the classification results for

Table 3. Classification results for the entire sequence. Best results in bold.

Classifier	AA			Digram		
	Size	Accuracy	MCC	Size	Accuracy	MCC
SVM	20	**0.873**	**0.838**	400	**0.934**	**0.917**
RF	20	0.726	0.657	400	0.724	0.656
NB	20	0.703	0.625	400	0.834	0.792

Table 4. Classification results for the extracellular segments. Best results in bold.

Segments	AA			Digram		
	Size	Accuracy	MCC	Size	Accuracy	MCC
N-terminus	20	0.835	0.792	400	0.920	0.901
EL1	20	0.842	0.802	390	0.831	0.786
EL2	20	0.839	0.798	386	0.861	0.825
EL3	20	0.825	0.779	327	0.816	0.769
All EL appended freq	60	0.873	0.839	1103	0.880	0.873
All EL accum. freq	20	0.845	0.804	398	0.875	0.844
(Nterm + EL) appended freq	80	**0.904**	**0.878**	1502	0.912	0.889
(Nterm + EL) accum. freq	20	0.849	0.808	400	**0.921**	**0.901**

Table 5. Classification results for the transmembrane segments. Best results in bold.

Segments	AA			Digram		
	Size	Accuracy	MCC	Size	Accuracy	MCC
TM1	20	0.794	0.741	321	0.823	0.778
TM2	20	0.850	0.809	298	0.847	0.806
TM3	20	0.866	0.829	290	0.878	0.846
TM4	20	0.822	0.776	320	0.860	0.822
TM5	20	0.859	0.818	293	0.856	0.817
TM6	20	0.836	0.794	262	0.848	0.810
TM7	20	0.808	0.755	281	0.843	0.801
TM appended frequency	140	**0.900**	**0.873**	2066	**0.900**	**0.873**
TM accumulated frequency	20	0.879	0.847	384	0.894	0.864

the N-terminus combined with the 7TM region. For each experiment the table shows the name of the segments under study, the size of the feature set and the classification performance as measured by accuracy and MCC.

Table 6. Classification results for the intracellular segments. Best results in bold.

Segments	AA			Digram		
	Size	Accuracy	MCC	Size	Accuracy	MCC
IL1	20	0.825	0.777	398	0.795	0.739
IL2	20	0.853	0.815	388	0.872	0.837
IL3	20	0.857	0.817	304	0.834	0.789
C-terminus	20	0.793	0.740	400	0.805	0.753
(IL+ C-terminus) append. freq	80	**0.906**	**0.880**	1490	**0.895**	**0.874**
(IL + C-terminus) accum. freq	20	0.837	0.795	400	0.885	0.854

Table 7. Classification results for the N-terminus concatenated with the 7TM regions. Best results in bold.

Segments	AA			Digram		
	Size	Accuracy	MCC	Size	Accuracy	MCC
Appended frequency	160	**0.919**	**0.897**	2467	0.915	0.889
Accumulated frequency	20	0.866	0.830	400	**0.928**	**0.909**

4.3 Discussion

The experimental results reported in the previous section show, as we might come to expect, an increasing deterioration of classification as we remove more parts of the sequence. Note though that this performance never drops below 0.75 (neither in accuracy nor in MCC), even for very small segments, and rarely below 0.8. This indicates a remarkable preservation of the discriminability throughout the sequence.

The best classification in our experiments using the entire sequences was found for the Digram representation with an accuracy of 0.934 and MCC of 0.917. The N-terminus by itself or in combination with the extracellular loops (see Table 4) drops little more than a percentage point, both in accuracy and MCC, when compared with the entire sequence for the Digram transformation. Note that the combination of the N-terminus with the 7TM also yields similar results (see Table 7). The classification results of the extracellular loops, transmembrane and intracellular segments are less accurate than those of the complete sequence or the N-terminus. In general, the combination of topologically-alike segments outperforms the classification results of single segments (with the aforementioned exception of the N-terminus). It is noteworthy that some very small segments such as IL2, EL2, TM3 and TM4 (some of them including on average no more than 2.2% of the sequence) barely drop more than 6% in classification performance as compared with the best results.

The Digram transformation provided overall the best results but with interesting exceptions: the 7TM regions, and the IL + C-terminus for the appended frequencies. Also interestingly, the appended frequencies yielded their better

results with the AA transformation, whereas the accumulated frequencies did it for Digram.

5 Conclusions

Preliminary research hinted the potential use of separated domains of complete class C GPCR sequences as the basis for subtype classification. In this study, we have carried out a systematic analysis of the performance of each of the individual sequence segments and some of their combinations. None of them yields better classification than the complete sequence, but the extracellular domain, the combination of the N-terminus and 7TM and, to some extent, the intracellular domain have all performed almost as well as the entire sequence. This, by itself, allows us to focus our work on the most discriminative segments. Future research should involve feature selection starting from these separate regions as a way to discover specific motifs with subtype discriminative capabilities.

Acknowledgments. This research was partially funded by the Spanish MINECO TIN2016-79576-R and SAF2014-58396-R project.

References

1. Santos, R., et al.: A comprehensive map of molecular drug targets. Nature Rev. Drug Discov. **16**, 19–34 (2017)
2. Leach, K., Gregory, K.J.: Molecular insights into allosteric modulation of Class C G protein-coupled receptors. Pharmacol. Res. **116**, 105–118 (2017)
3. Kniazeff, J., et al.: Dimers and beyond: the functional puzzles of class C GPCRs. Pharmacol. Ther. **130**, 9–25 (2011)
4. Cooke, R.M., Brown, A.J., Marshall, F.H., Mason, J.S.: Structures of G protein-coupled receptors reveal new opportunities for drug discovery. Drug Discov. Today **11**, 1355–1364 (2015)
5. Wu, H., et al.: Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science **344**(6179), 58–64 (2014)
6. Doré, A.S., et al.: Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain. Nature **551**, 557–562 (2014)
7. König, C., Cruz-Barbosa, R., Alquézar, R., Vellido, A.: SVM-based classification of class C GPCRs from alignment-free physicochemical transformations of their sequences. Lecture Notes in Computer Science, vol. 8158, pp. 336–343 (2013)
8. König, C., Alquézar, R., Vellido, A., Giraldo, J.: Finding class C GPCR subtype-discriminating n-grams through feature selection. J. Integr. Bioinform. **11**(3), 254 (2014)
9. Pin, J.P., Galvez, T., Prezeau, L.: Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. Pharmacol. Ther. **98**(3), 325–354 (2003)
10. König, C., Alquézar, R., Vellido, A., Giraldo, J.: The extracellular N-terminal domain suffices to discriminate class C G protein-coupled receptor subtypes. In: Proceedings of the Joint-Conference on Artificial Neural Networks (IJCNN 2015), pp. 1–7 (2015)

11. Isberg, V., et al.: GPCRdb: an information system for G protein-coupled receptors. Nucleic Acids Res. **44**, 356–364 (2016)
12. Pin, J.P., Bettler, B.: Organization and functions of mGlu and GABAB receptor complexes. Nature **540**, 60–68 (2016)
13. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience, New York (1998)
14. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
15. John, G., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pp. 338–345 (1995)
16. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011)
17. Opiyo, S.O., Moriyama, E.N.: Protein family classification with partial least squares. J. Proteome Res. **6**(2), 846–853 (2007)
18. Liu, B., et al.: Using amino acid physicochemical distance transformation for fast protein remote homology detection. PLoS ONE **7**(9), 1–7 (2012)
19. Cheng, B., et al.: Protein classification based on text document classification techniques. Proteins: Struct. Funct. Bioinf. **58**(4), 955–970 (2005)
20. Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S.: Improved prediction of signal peptides: signalP 3.0. J. Mol. Biol. **340**(4), 783–795 (2004)

QmihR: Pipeline for Quantification of Microbiome in Human RNA-seq

Bruno Cavadas[1,2,3], Joana Ferreira[1,2], Rui Camacho[4,5], Nuno A. Fonseca[6], and Luisa Pereira[1,2,7(✉)]

[1] Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal
luisap@ipatimup.pt
[2] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal
[3] Instituto de Ciências Biomédicas Abel Salazar (ICBAS), Universidade do Porto, Porto, Portugal
[4] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
[5] LIAAD/INESC TEC, Porto, Portugal
[6] European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK
[7] Faculdade de Medicina da Universidade do Porto, Porto, Portugal

Abstract. The huge amount of genomic and transcriptomic data obtained to characterize human diversity can also be exploited to indirectly gather information on the human microbiome. Here we present the pipeline QmihR designed to identify and quantify the abundance of known microbiome communities and to search for new/rare pathogenic species in RNA-seq datasets. We applied QmihR to 36 RNA-seq tumor tissue samples from Ukrainian gastric carcinoma patients available in TCGA, in order to characterize their microbiome and check for efficiency of the pipeline. The microbes present in the samples were in accordance to published data in other European datasets, and the independent BLAST evaluation of microbiome-aligned reads confirmed that the assigned species presented the highest BLAST match-hits. QmihR is available at GitHub (https://github.com/Pereira-lab/QmihR).

Keywords: Microbiome · RNA-seq data · Identification · Quantification

1 Introduction

A mutualist symbiotic relationship between microbes and their animal hosts has been estimated to occur for at least the last 500 million years [1]. A big impulse on our knowledge on the 'normal' human microbiome is being contributed by large scale studies such as the Human Microbiome Project (HMP) [2] and MetaHIT [3]. Major findings of HMP [4] indicated an overall high diversity of community members, heterogeneous in terms of within host versus between host ratio diversities, and ethnicity was amongst one of the strongest associations with microbiome. An intact microbial community is essential for a healthy development of the host [5], and several changes

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_21

to the microbiome are beginning to be described as associated with complex diseases, such as cancer [6, 7].

Initially, most studies of microbial communities depended on the sequencing of the gene coding the bacterial and archaea 16s rRNA, but the paradigm shift in sequencing technologies is also changing this analyses. Efforts have been applied to complete sequence the microbiome directly [8], and the huge amount of human-focused omics data (for e.g., international consortia such The Cancer Genome Atlas (TCGA) [9] and Genotype-Tissue Expression project (GTEx) [10]) has the potential to indirectly contribute information on the human microbiome [11]. In fact, it has been already shown [11] that human whole genome/exome (WGS/WES) and transcriptome sequences (RNA-seq) contain human-unmapped reads that match bacteria, viruses and fungi that colonize/infect the individuals. However, a technical challenge is that a large number of short reads cannot be uniquely mapped to a specific location at one genome, mapping instead to multiple locations at one or related genomes, influencing the bacterial abundance classification. This issue must be taken into account in the development of efficient pipelines, which can incorporate probabilistic methods that attribute these reads to the most abundant species already identified through unique-location mapping reads (such as RSEM [12]).

In this paper, we describe a pipeline to characterize the microbiome inferred from human-focused RNA-seq data, designed to perform a reliable classification of bacterial abundance. We assess its efficiency through a real TCGA RNA-seq dataset collected in 36 Ukrainian patients from gastric carcinoma. This dataset was selected as it can be compared with published information of the gut microbiome in European individuals, inferred from traditional techniques of 16s rRNA sequencing [13].

2 Description of the Pipeline

We designed a pipeline (Fig. 1) aiming to best characterize known microbiome communities, despite also allowing to collect reads that can be processed in BLAST for identification of new/uncommon pathogenic species. Currently, the most common microbiome species occurring in various human habitats are well characterize, rendering more efficient to design pipelines that search first for a reference panel of microbial species, and allow identification of the subset of unmapped non-human reads. HMP is a good departing database to construct these reference panels per location in the human body.

QmihR begins by trimming of reads using Trimmomatic [14]. It checks if: (1) the mean of two consecutive bases is below 20 Phred; and (2) the resulting read is smaller than 40 bases. This pre-processing step removes adapters and low quality reads, following the best practices for accurate RNA-seq expression estimates [15]. Even when using the pipeline in already indexed non-human mapped reads, we advise to perform this trimming as in our experience there are still low-quality reads classified as unmapped.

Then the global alignment of the reads against the bacterial reference database is made with Bowtie2 [16] and quantification of bacterial genera is performed through RSEM [17]. This tool takes a probabilistic approach to the quantification of reads in

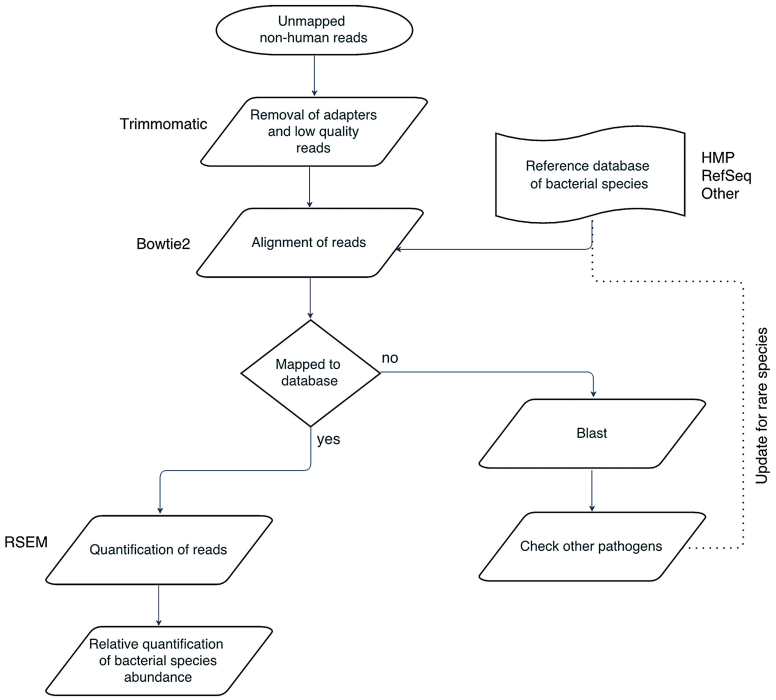


Fig. 1. Scheme of the QmihR pipeline.

cases of multi-mapping, and avoids discarding all reads that would multi-map in diverse species, conducting to a more real solution. A previous publication [18] has shown that RSEM presents the higher accuracy amongst probabilistic algorithms, guiding our choice. RSEM produces as output counts of mapped reads per gene belonging to a species (giving an indication of the most expressed genes). The pipeline takes the counts of the various genes within a species and aggregates them to produce counts of reads aligned per species, which are then normalized by the library size for the mapped reads against the bacterial reference database, as indicated in the Eq. (1).

$$normalizedcounts = \frac{counts\ gene}{\sum all\ reads\ mapped\ to\ database} \times 10^6 \tag{1}$$

3 Application to a TCGA RNA-seq Dataset

The original human-unmapped raw RNA-seq reads obtained in tumor tissue from 36 Ukrainians patients of gastric carcinomas were found in the TCGA Genomic Data Commons repository (https://gdc.cancer.gov/). The microbe reference panel used contains 194 bacterial whole genomes (one representative strain per species) collected from NCBI following the species identified by the HMP [2] in the gastrointestinal tract.

QmihR reported that the microbiome in the cohort (Fig. 2) is dominated by the genera *Bacillus* and *Pseudomonas* (around 21% and 17%, respectively), then *Escherichia* and *Enterobacter* (10–15%). The class I carcinogen *Helicobacter* reaches 3% overall frequency in Ukraine. This microbiome diversity is in accordance to published data in other European cohorts [3].

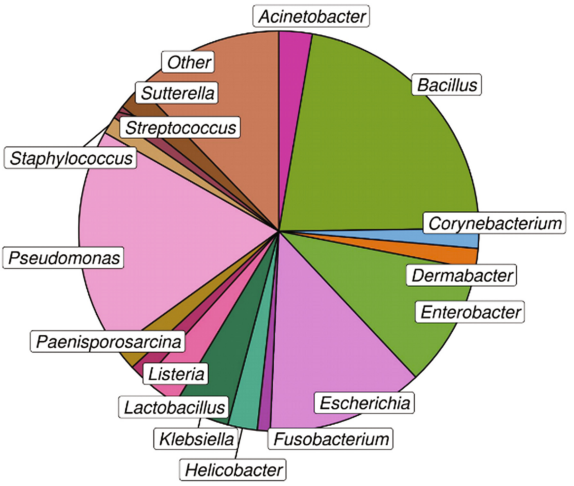


Fig. 2. Overall microbiome abundance in gastric tumor samples from Ukraine (n = 36). Only genera that passed a threshold of 1% of mean abundance are displayed in the graph, otherwise they are summed together in a class denominated as "other".

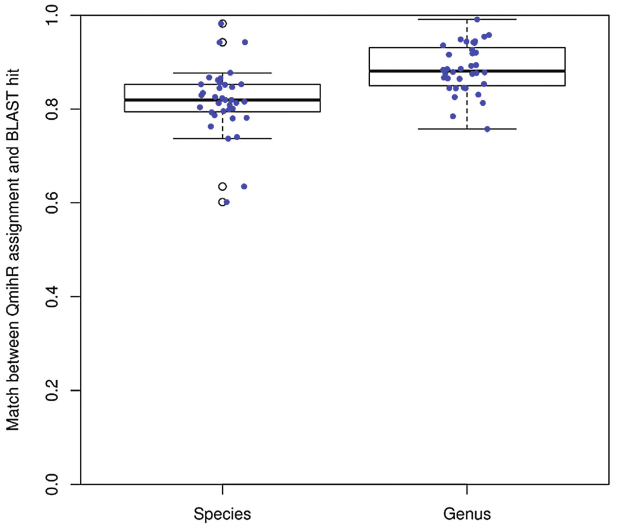


Fig. 3. Comparison of hit-species/genera matches between QmihR and BLAST for all microbe-aligned reads in the 36 gastric tumor samples from Ukraine.

In order to double-check the assignment of microbe species, we run the total amount of QmihR-assigned reads in BLAST (database downloaded on 3th February 2017, and curated for excluding sequences from uncultured species). In the Ukrainian dataset (Fig. 3), in around 82% of the reads the species identified in QmihR would also be on the list of top hits provided by BLAST, and the value raises to 88% when limiting to the genus level. We also took a closer look into the two samples with poorer results, and confirmed in BLAST that some read-pairs would align with an identity of 97–100% in the forward and 93–100% identity in the reverse in the QmihR-assigned species.

4 Benchmarking

QmihR took in average 30 min per sample to calculate the microbiome abundance, based on the reference microbe panel provided (mean 14 Gb of raw un-mapped reads in fastq format), when using an Intel Core i7-4700 2.4 GHz with 8 cores and 16 Gb of RAM. It is a fast and efficient tool that may be used in human microbiome inference from RNA-seq, in health and disease conditions.

To run the full set of unmapped reads in BLAST tool would take weeks. Even the test of running the QmihR-mapped reads in bacteria took between 2 and 8 h per sample

5 Conclusions

QmihR is a fast and efficient tool that may be used in human microbiome inference from RNA-seq, both in health and disease conditions. To our best knowledge, this is the first pipeline for quantification of the microbiome (bacterial) from RNA-seq data. A similar pipeline was developed to infer viral infection in RNA-seq TCGA samples [19], a case-study that presents, nevertheless, some differences to the situation analyzed here. Viral genomes are smaller than bacterial ones and the genes detected in the RNA-seq are the ones important for the infection and display low homology between species. In the bacteria, the reads detected in RNA-seq are mostly from rRNA genes (higher than 90%; similarly to the human genes), which display certain similarity between species, generating the multi-location read problem.

Acknowledgements. We wish to thank TCGA for the access provided to the protected data used in this work. Funds were guaranteed by the project "Advancing cancer research: from basic knowledge to application"; NORTE-01-0145-FEDER-000029; "Projetos Estruturados de I&D&I", funded by Norte 2020 – Programa Operacional Regional do Norte. I3S is financed by FEDER - Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Competitiveness and Internationalization Operational Programme (POCI), Portugal 2020, and by Portuguese funds through FCT/Ministério da Ciência, Tecnologia e Inovação in the framework of the project "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274).

References

1. Cho, I., Blaser, M.J.: The human microbiome: at the interface of health and disease. Nat. Rev. Genet. **13**, 260–270 (2012)
2. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C., Knight, R., Gordon, J.I.: The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature **449**, 804–810 (2007)
3. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D., Wang, J.: A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464**, 59–65 (2010)
4. Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature **486**, 207–214 (2012)
5. Bäckhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., Gordon, J.I.: Host-bacterial mutualism in the human intestine. Science **307**, 1915–1920 (2005)
6. Thomas, R.M., Jobin, C.: The microbiome and cancer: is the 'Oncobiome' mirage real? Trends Cancer **1**, 24–35 (2015)
7. Brawner, K.M., Morrow, C.D., Smith, P.D.: Gastric microbiome and gastric cancer. Cancer J. **20**, 211–216 (2014). (Sudbury, Mass)
8. Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., Wang, J., Imhann, F., Brandsma, E., Jankipersadsing, S.A., Joossens, M., Cenit, M.C., Deelen, P., Swertz, M.A., Weersma, R.K., Feskens, E.J., Netea, M.G., Gevers, D., Jonkers, D., Franke, L., Aulchenko, Y.S., Huttenhower, C., Raes, J., Hofker, M.H., Xavier, R.J., Wijmenga, C., Fu, J.: Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science **352**, 565–569 (2016)
9. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Cancer Genome Atlas Research Network.: The cancer genome atlas pan-cancer analysis project. Nat. Genet. **45**(10), 1113–1120 (2013)
10. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B.: The genotype-tissue expression (GTEx) project. Nat. Genet. **45**(6), 580–585 (2013)
11. Samuels, D.C., Han, L., Li, J., Quanghu, S., Clark, T.A., Shyr, Y., Guo, Y.: Finding the lost treasures in exome sequencing data. Trends Genet. **29**, 593–599 (2013)
12. Chandramohan, R., Wu, P.Y., Phan, J.H., Wang, M.D.: Benchmarking RNA-seq quantification tools. In: 35th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBC), pp. 647–650 (2013)
13. Dicksved, J., Lindberg, M., Rosenquist, M., Enroth, H., Jansson, J.K., Engstrand, L.: Molecular characterization of the stomach microbiota in patients with gastric cancer and in controls. J. Med. Microbiol. **58**(4), 509–516 (2009)
14. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114–2120 (2014)
15. Williams, C.R., Baccarella, A., Parrish, J.Z., Kim, C.C.: Trimming of sequence reads alters RNA-seq gene expression estimates. BMC Bioinform. **17**, 103 (2016)

16. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat Meth **9**, 357–359 (2012)
17. Li, B., Dewey, C.N.: RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinform. **12**, 323 (2011)
18. Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. **34**(5), 525–527 (2016)
19. Tang, K.W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M., Larsson, E.: The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat. Commun. **4**, 2513 (2013)

Improving Prognostic Prediction from Mild Cognitive Impairment to Alzheimer's Disease Using Genetic Algorithms

Francisco L. Ferreira[1(✉)], Sandra Cardoso[2], Dina Silva[2], Manuela Guerreiro[2], Alexandre de Mendonça[2], and Sara C. Madeira[3(✉)]

[1] INESC-ID and Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal
francisco.lourenco.ferreira@tecnico.ulisboa.pt
[2] Laboratory of Neurosciences, Faculty of Medicine, Institute of Molecular Medicine, University of Lisbon, Lisbon, Portugal
sandradcardoso@gmail.com, dlsilva@ualg.pt, mmgguerreiro@gmail.com, mendonca@medicina.ulisboa.pt
[3] LASIGE and Faculty of Sciences, University of Lisbon, Lisbon, Portugal
sacmadeira@fc.ul.pt

Abstract. Alzheimer's disease is becoming a global epidemic. Its impact is devastating for patients', their families and the economy. As such, it is important to build good prognostic models that can predict conversion to dementia so that treatment measures could be taken. In this work, we applied a genetic algorithm to choose the most relevant neuropsychological and demographic features for prognostic prediction. The results show improvements over other feature selection methods, with our model being able to predict conversion to dementia with AUC and sensitivity of 88% . Moreover, we found that with only 7 features it is possible to achieve good classification results. These results could help physicians to adjust treatment and select which exams should be performed regularly to increase efficiency in clinical practice.

Keywords: Mild cognitive impairment · Alzheimer's disease · Prognostic prediction · Neuropsychological tests · Genetic algorithm

1 Introduction

Alzheimer's disease (AD) accounts for 60 to 80% of all cases of dementia [1]. It affects 5.4 million Americans nowadays and an expected 13.8 million by 2050 [2], mainly due to population shifting to older ages. These numbers represent not only a true global epidemic, but also a huge socio-economic burden [3]. The problem is even greater considering the fact that low and middle income countries will have the most increase in numbers of these patients [3].

Mild Cognitive Impairment (MCI) is a condition in which patients have cognitive complaints not affecting their ability to perform daily tasks [4]. It is a common disorder, affecting 15 to 20% of people older than 65 [4]. These patients are more likely to develop

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_22

AD. Reliably predicting MCI to AD conversion could help physicians taking decisions about their patients' treatment or selecting, among them, those who could be included in clinical trials for new possible treatments.

Since no treatment is available to revert or reduce brain damage caused by AD, it is critical to understand AD and its progression, not only to guide clinical decisions and managing patients and families' expectations, but to develop new effective treatments. Thus, understanding AD's biomarkers, correctly identifying patients in the disease spectrum and predicting patients' decline is of maximum importance.

In this work, we used GA for feature selection in MCI to AD prognosis. Our goal is to (1) improve the classification results of current prognostic models and (2) find a small feature set highly predictive of conversion. This will be accomplished using data from a large national database. We also compare our approach with other FS methods and discuss the most chosen features by the GA. Our future goal is to apply this knowledge to choose which biomarkers should be gathered regularly to save time, resources and to better predict converting patients.

The paper is organized as follows: Sect. 2 presents related work; Sect. 3 presents the database used as well as the main technique employed to the classification task; Sect. 4 discusses the results and provides a comparison with alternative procedures; and Sect. 5 concludes and presents future work.

2 Related Work

2.1 Predicting Conversion to AD Using Neuropsychological Tests

Neuropsychological tests (NPTs) are commonly used to classify dementia patients [5]. They test patients in multiple cognitive domains such as episodic memory, learning and language. Common NPTs include the California Verbal Learning Test (CVLT) [6], the Alzheimer's Disease Assessment Scale – cognitive subscale (ADAS-Cog) [7] and the Mini-Mental State Examination (MMSE) [8].

Among AD's biomarkers, NPTs have revealed the best results in predicting converting patients. Silva et al. [5] showed that a Linear Discriminant Analysis (LDA) model constituted by Digit Span backward, Semantic Fluency, Logical Memory (immediate recall), and Forgetting Index is able to predict conversion from MCI to AD, in a 5 years' time period, with high values of accuracy, specificity and sensitivity (around 80%). In another study, Chapman et al. [9] predicted conversion using 17 common NPTs with Principal Component Analysis (PCA) followed by discriminant analysis of those components. Good accuracy (84%), sensitivity (86%) and specificity (83%) were reported, although their dataset was fairly small (43 patients). Lee et al. [10] also showed the prognostic power of these inexpensive and non-invasive tests that can be conducted in any environment.

2.2 Genetic Algorithms in Alzheimer's Disease Prognostic Tasks

When building prognostic models for AD, feature selection (FS) methods are usually applied, searching for a subset of features highly predictive of conversion. Such task is

important not only to increase the predictive power of these models, but also to gather insight about which NPTs are more important as biomarkers. These results can then guide physicians in choosing the best test battery to apply to their patients.

In this context, genetic algorithms (GAs) can be useful since they are particularly suitable for large, complex or poorly understood feature sets, as the GAs demonstrates fast convergence to nearly-optimal solutions [11]. Moreover, their optimization procedure can be directed to any measure of interest, such as higher accuracy or lower number of features used to train the classification model, leading to better predictive models using a fraction of the initial feature set.

In AD, GAs have been used to select the most relevant features in diagnostic and prognostic tasks using blood-based biomarkers [12] and MRI data [13]. To our knowledge, this analysis was not sufficiently applied to NPTs. We found only one study [14], with promising results in prognosis but using data of only 77 patients.

3 Methods

3.1 The Cognitive Complaints Cohort

The Cognitive Complaints Cohort (CCC) [5] is a study conducted at the Faculty of Medicine of Lisbon in partnership with Laboratory of Language Studies (Santa Maria Hospital), Memoclínica and the Neurology Department of Coimbra's University Hospital. Its goal is to investigate AD progression in patients with cognitive complaints. All participants are evaluated through a neuropsychological battery validated for the Portuguese population (BLAD [15]). This includes the MMSE, CVLT, Logical Memory and Clock Drawing tests. Age and years of formal education were also gathered from these patients. Z-Scores, data corrected by age and education, were used when available. In total, 58 features are associated with this dataset.

To predict MCI to AD conversion, we chose patients diagnosed with MCI at their baseline assessments, and asked whether they converted to dementia within 3 years. We excluded reverting patients, that is patients that reverted from MCI to normal cognition and from AD to MCI. This is usually the method employed, as reversion is clearly unexpected [16] and can be related to errors in diagnosis such as the presence, at the time of evaluation, of diseases other than AD. The final dataset includes 282 patients, including 122 converting, and 160 non-converting MCI patients.

3.2 Genetic Algorithm and Naïve Bayes

We implemented a GA using MATLAB's Global Optimization Toolbox aimed at selecting a subset of features to be trained with Naïve Bayes from Weka software [17]. This classifier was chosen as it was the one presenting the best results. Figure 1 depicts the learning architecture defined in this work. The parameters presented were defined through experimentation, maximizing the predicting ability of the trained models, as well as accounting for solutions with low number of features.

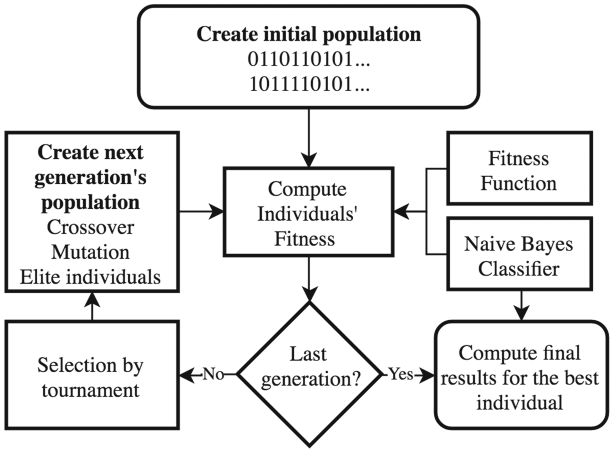


Fig. 1. Architecture of the genetic algorithm implemented for feature selection.

The GA starts by generating (through a uniform distribution) an initial population of 50 binary individuals coded as strings. Each of these is a solution to the FS problem. They are composed of a sequence of 1s and 0s with a length of 58 (the total number of features in the dataset) where each 1 indicates that a given feature should be used in the classification task. Thus, the total number of possible solutions is 2^{58}.

These individuals are then tested for their fitness: the solutions are trained with Naïve Bayes (through a 5-fold cross-validation learning process) and tested for their predictive ability. We tested the GA with two different fitness functions. First, we defined the individuals' fitness as their area under the ROC curve (AUC). Such experiment is labeled "Test 1" (T1). Then, to obtain solutions with low number of features, we altered the fitness function to reward small solutions (T2). This fitness function (minimized by MATLAB toolbox) is presented in (1).

$$fitness = (1 - AUC) + \rho\left(n_{ind}/n_{total}\right) \tag{1}$$

Where n_{total} and n_{ind} define the total and the individual's selected features, respectively. The value ρ determines the importance being attributed to small solutions. We defined ρ as 0.35, leading to a good compromise between classification results and a low number of features. This value was chosen empirically.

A selection procedure is then used to determine the next generation's parents. We chose selection by tournament (of size 2), to avoid early convergence of the GA that leads to poor solutions [14]. This procedure repeatedly selects 2 random individuals from the population and chooses the one with the highest fitness.

Next, reproduction by single-point crossover combines the selected parents to create new individuals. It first selects one of the binary sites, creating a new individual from the first parent's features up to that site and the rest from the other parent. Through experimentation, we defined that 60% and 80%, for T1 and T2 respectively, of the next generation's population should be created using crossover. Also, we selected the 5%

best solutions (elite individuals) to continue to the next generation, ensuring the continuation of the best fitted individuals. The remaining 15% for T1, or 35% for T2, of the next generation's population is then created by mutating the parents. We defined a mutation rate of 3% for T1 and 6% for T2.

After creating the next generation, the algorithm starts another cycle by re-computing the fitness of each individual. When the total number of 150 generations is reached (value defined through experimentation, maximizing classification results while minimizing classification training time), the algorithm stops and chooses the best fitted individual. It then computes its final classification results through a 5-fold cross validation procedure averaged 100 times.

4 Results and Discussion

4.1 Feature Selection Using a Genetic Algorithm

The accuracy of the baseline model (without using FS) to predict MCI to AD conversion within 3 years was 76.47%. This result is accompanied by an AUC of 0.84, sensitivity of 0.79 and specificity of 0.74. Using the GA, the results improved as shown in Tables 2 and 3 for the first (T1) and second (T2) tests, respectively.

We tested several methods implemented in Weka: *CfsSubsetEval* is an implementation of Hall's work [18], measuring the predictive power of each feature while minimizing the redundancy between them; *CorrelationAttributeEval* and *InfoGain* measure the Person's correlation between features and the class, and the worth of an attribute by its information gain with respect to the class, respectively.

Table 1 shows that the FS methods implemented in Weka had modest improvements over the baseline model. A small increase in sensitivity and specificity is achieved, using less than half the features of the original set.

Table 1. MCI to AD prognostic prediction using various FS methods implemented in Weka. Baseline results are presented for reference.

Method	AUC	Accuracy	Sensitivity	Specificity	# features
Baseline	0.84	76.47%	0.79	0.74	58
CfsSubsetEval	0.84	77.57%	0.82	0.74	19
CorrelationAttributeEval	0.85	77.46%	0.81	0.75	16
InfoGain	0.85	77.84%	0.81	0.76	20

On the other hand, GA solutions show substantial improvements over the baseline and other FS methods. Some simulations present improvements between 6 to 8 percent points in accuracy, with high values of specificity and sensitivity. They also improve previous results using the same database [5]. Moreover, the results presented have a higher value of AUC and are tested in a bigger set of patients than another work [14] using GAs and NPTs for MCI to AD prognostic prediction.

Some heterogeneity is present in the GA solutions, which can be advantageous depending on the classification objective. In particular, high values of sensitivity are preferred in prognostic tasks, as the cost of misclassifying a converter is usually higher than predicting conversion in a non-converting patient. This solution can be found in simulation 36 (Table 2). Simulation 20 also presents high values of sensitivity, but a more balanced specificity. Solutions with balanced measures (simulation 3) and higher values of specificity (simulation 23) are also presented.

Table 2. MCI to AD prognostic prediction for T1 using GA as feature selection. These results are computed through 50 simulations. The last row presents the average of such results.

Simulation	AUC	Accuracy	Sensitivity	Specificity	# features
#3	0.87	82.94%	0.84	0.83	23
#20	0.88	82.95%	0.86	0.81	25
#23	0.87	83.81%	0.83	0.85	24
#36	0.87	81.95%	0.88	0.77	21
Average	0.87	81.95%	0.84	0.80	25

Table 3 shows T2's results, when a bonus for solutions using less features is used in the fitness function. These models use a third of the features of the earlier GA solutions, with minor compromises in terms of their predictive ability. We show that it is possible to build useful models with only a small subset of the original features, while increasing the prognostic classification results when compared with the baseline and other FS methods. Simulations 3, 8, 16, 29 and 30 reached the same solution after 150 generations, with high sensitivity and using only 7 features.

Table 3. GA results (50 simulations) for MCI to AD prognostic prediction using AUC as the fitness function and a bonus for small feature set solutions.

Simulation	AUC	Accuracy	Sensitivity	Specificity	# features
#3,8,16,29, 30	0.87	81.56%	0.85	0.79	7
#39	0.87	82.15%	0.86	0.79	8
#4	0.88	82.26%	0.80	0.84	8
#36	0.87	80.99%	0.81	0.81	6
Average	0.86	80.83%	0.82	0.80	8

An exhaustive search performed on the 8 most selected features by the GA showed that it was possible to obtain similar results to the ones presented by other FS methods (Table 1) using a minimum of 4 features, although such compromise in classification power does not have advantages compared to similar solutions presented in Table 3.

4.2 Features Selected by the Genetic Algorithm

Both T1 and T2 showed a similar pattern regarding the most and least chosen features. The most selected features are related to conceptual thinking (proverbs exercise and

abstract reasoning), short-term memory (word recall exercise), verbal semantic fluency and non-associated learning cognitive domains.

In T1, age and a proverb exercise were selected in all simulations. Interestingly, age and education level were not amongst the top 10 features in T2, probably because most NPTs are z-scores of the original features corrected by these measures. Some of the most selected features, as those related to semantic fluency and short-term memory, are in accordance with previous work [5]. But this study may have revealed an important role of conceptual thinking degradation in predicting converting patients.

On the least selected features, the memory domain is also present, but related to episodic memory. This set is also dominated with depression scales and measures of sustained and divided attention and visuo-motor processing speed. Interestingly, MMSE scores were not selected as the most predictive features as happened in other studies [14]. In fact, MMSE [19] has been shown to be a poor predictor of conversion in later stages of AD and was not selected in any simulation for T2.

5 Conclusions and Future Work

Prognostic prediction from MCI to AD is key to tackle its global epidemic. Good classification models, predicting MCI to AD conversion within a time-window of interest should help physicians managing their patients' progression and prescribing new possible treatments. Moreover, more efficiency could be achieved if only a handful of measures are necessary to perform such prediction.

In this work we showed how to use a genetic algorithm to optimize the features used by the model. The proposed approach was able to find a very small subset of features, highly predictive of conversion, and superior to other common FS methods using only 7 features. It was able to predict converting patients with an accuracy of 81.56%, specificity of 0.79 and sensitivity of 0.85. The sensitivity is particularly important due to the importance of predicting conversion in true converting patients. Our analysis, while reinforcing the roles of semantic fluency and short-term memory degradation in AD, shows a possible role of conceptual thinking deterioration.

The results are promising, but further work should be done to explore the way in which the most selected features by the GA should be combined, in practice, to achieve the best prognostic and efficiency results in the clinic. Moreover, these results should be validated with a different database, such as ADNI [20]. The difference between the sets of NPTs data available in both datasets will increase the challenge of comparing the models. Finally, prognostic tasks in earlier phases of the disease, such as predicting conversion from normal cognition to MCI should also be explored. Further insight could be gained by comparing the most predictive features between such task and the one presented in this work.

Acknowledgments. This work was partially supported by FCT under the projects NEUROCLINOMICS2 (PTDC/EEI-SII/1937/2014) and UID/CEC/50021/2013, and an individual doctoral grant to FF (SFRH/BD/118872/2016).

References

1. Barker, W.W., et al.: Relative frequencies of Alzheimer disease, Lewy body, vascular and Frontotemporal dementia, and Hippocampal sclerosis in the State of Florida Brain Bank. Alzheimer Dis. Assoc. Disord. **16**, 203–212 (2002)
2. Hebert, L.E., Scherr, P.A., Bienias, J.L., Bennett, D.A., Evans, D.A.: Alzheimer disease in the US population: prevalence estimates using the 2000 census. Arch. Neurol. **60**, 1119 (2003)
3. Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M., Karagiannidou, M.: World Alzheimer Report 2016 Improving healthcare for people living with dementia (2016)
4. Roberts, R., Knopman, D.S.: Classification and epidemiology of MCI. Clin. Geriatr. Med. **29**, 753–772 (2013)
5. Silva, D., Guerreiro, M., Santana, I., Rodrigues, A., Cardoso, S., Maroco, J., De Mendonça, A.: Prediction of long-term (5 years) conversion to dementia using neuropsychological tests in a memory clinic setting. J. Alzheimer's Dis. **34**, 681–689 (2013)
6. Silva, D., Guerreiro, M., Maroco, J., Santana, I., Rodrigues, A., Bravo Marques, J., de Mendonça, A.: Comparison of four verbal memory tests for the diagnosis and predictive value of mild cognitive impairment. Dement. Geriatr. Cogn. Dis. Extra **2**, 120–131 (2012)
7. Kolibas, E., Korinkova, V., Novotny, V., Vajdickova, K., Hunakova, D.: ADAS-cog (Alzheimer's Disease Assessment Scale-cognitive subscale)–validation of the Slovak version. Bratisl. Lek. Listy **101**, 598–602 (2000)
8. Folstein, M.F., Folstein, S.E., McHugh, P.R.: Mini-mental state. J. Psychiatr. Res. **12**, 189–198 (1975)
9. Chapman, R.M., Mapstone, M., Mccrary, J.W., Gardner, M.N., Porsteinsson, A., Sandoval, T.C., Reilly, L.A.: Predicting conversion from mild cognitive impairment to Alzheimer´s disease using neuropsychological test and multivariate methods. J. Clin. Exp. Neuropsychol. **33**, 187–199 (2012)
10. Lee, S.J., Ritchie, C.S., Yaffe, K., Cenzer, I.S., Barnes, D.E.: A clinical index to predict progression from mild cognitive impairment to dementia due to Alzheimer's disease. PLoS ONE **9**, 1–15 (2014)
11. Yang, J., Honavar, V.: Feature subset selection using a genetic algorithm. In: Feature Extraction, Construction and Selection, pp. 117–136. Springer, Boston (1998)
12. Vandewater, L., Brusic, V., Wilson, W., Macaulay, L., Zhang, P.: An adaptive genetic algorithm for selection of blood-based biomarkers for prediction of Alzheimer's disease progression. BMC Bioinf. **16**, S1 (2015)
13. Spedding, A.L., Di Fatta, G., Cannataro, M.: A genetic algorithm for the selection of structural MRI features for classification of mild cognitive impairment and Alzheimer's disease. In: Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine, pp. 1566–1571 (2015)
14. Johnson, P., et al.: Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. BMC Bioinf. **15**(Suppl. 1), S11 (2014)
15. Guerreiro, M.: Contributo da Neuropsicologia para o Estudo das Demências, Ph.D. thesis, University of Lisbon (1998)
16. Grande, G., et al.: Reversible mild cognitive impairment: the role of comorbidities at baseline evaluation. J. Alzheimer's Dis. **51**, 57–67 (2016)
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. ACM SIGKDD Explor. Newsl. **11**, 10 (2009)
18. Hall, M.: Correlation-based feature selection for machine learning. Methodology, pp. 1–5 (1999)

19. Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Collins, D.L.: Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. Neurobiol. Aging **36**, S23–S31 (2015)
20. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.: Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers. Dement. **1**, 55–66 (2005)

Novel Method of Identifying DNA Methylation Fingerprint of Acute Myeloid Leukaemia

Agnieszka Cecotka(✉) and Joanna Polanska

Data Mining Group, Faculty of Automatic Control,
Electronics and Computer Science, Institute of Automatic Control,
Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland
{agnieszka.cecotka, joanna.polanska}@polsl.pl

Abstract. Finding new statistical approaches to high throughput data analysis is a very hot topic nowadays. Such a data needs dedicated methods and algorithms of analysis due to huge number of features, but often also due to a small number of samples. Methylation data are also very special, because of dependencies between features and their neighbourhood. There is a need to find a novel, data driven algorithm for these data owing to big variety of distributions data sets. Purpose of this method is detection of regions with different levels of demethylation. From the biological point of view, the most important genome regions are TSS (transcription start site) regions. Hypermethylation of these part of a gene leads to repression and thus stop the gene expression. This phenomenon often happens in cancer disease and impairs a number of molecular processes in the cell. The proposed algorithm is performed for AML patients data in comparison to healthy control. By combination of statistics methods and mathematical modelling together, it enables detection of demethylated regions or DNA and their classification as low, medium or high demethylated.

Keywords: DNA methylation · Epigenetics · Acute Myeloid Leukaemia · Gaussian mixture model · Mathematical modelling · Robust estimator

1 Background

Methylation is an epigenetic process which controls the mechanism of transcription. It is based on changing cytosine to 5-methylcytosine in CpG sites of genome. Cytosine must be followed by guanine in a DNA strand [1]. In cancer diseases, because of alterations in DNA methylation, expression of tumour suppressor genes can be stopped and protooncogenes transcription can be increased [2].

There are several analysis methods which can be used to compare two sets of methylation data (case-control study). The simplest bases of parametric statistical tests for mean equality as t-Student test. They detect demethylated CpG sites of genome. Demethylated region is defined by the amount of demethylated CpG sites in the region. An example can be methyAnalysis algorithm [3]. More advanced methods take into account the neighbourhood of CpG sites. One of such algorithms is an A-clustering algorithm [4]. Another interesting algorithm, Bump-hunting method [5] aims at doing peak detection across a genome. The last mentioned method is Probe Lasso [6], which

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_23

defines kernels with specific size, determined by density of CpG sites in the examined region. Presented work aims at proposing a new method, which does not use predefined cut-off levels of demethylation but is data-driven due to big variety of distributions data sets. It enables not demethylated regions to be checked but also to be categorised as low, medium or high demethylated.

2 Materials

The data set used in the presented study is free accessible and was downloaded from GEO database (GSE63409) [7]. The data were normalized with *minfi* package [8]. The data set consisted of 14 samples of CD34 + 38- cells from AML patients and 5 samples of hematopoietic stem cells from healthy donors, so in total 19 samples. The data were collected in the Illumina Infinium450 k microarray experiment [9]. As a result, they got methylation level for 485 512 CpG sites of human genome. Methylation level is defined as methylated signal to sum of methylated and unmethylated signals ratio and it is called Beta value. Beta value must range from 0 to 1, where 0 means no methylation and 1 means full methylation [10] (Fig. 1).

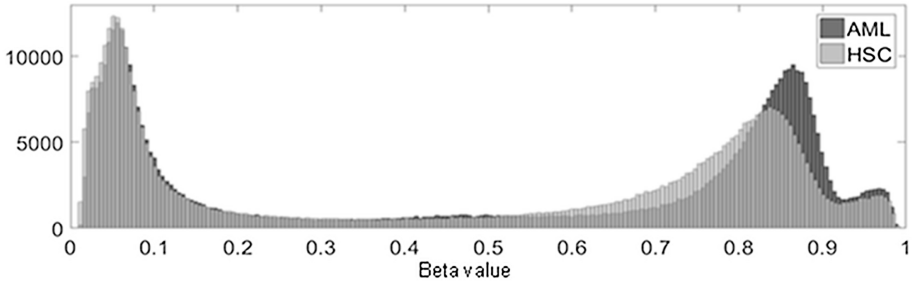


Fig. 1. Histogram of Beta-value in HSC and AML cells.

Using Illumina annotation system, each CpG site is assigned to i.a. chromosome number, locus, its sequence, RefGene Name and RefGene Accession (if it belongs to the gene region), RefGene Group, Relation to CpG Island and Regulatory Feature Group. The whole genome is divided into several regions according to the gene structure. It consists of intergenic, TSS, 5'UTR, 1stExon, Body and 3'UTR regions, which form RefGene Groups. The whole genome is also divided into groups according to the density of CpG sites. The regions with the highest density are called CpG Islands, then are Shore, Shelf and Open sea [9].

3 Methods

3.1 Set Difference Estimation

The first step of analysis is to estimate the difference between the AMLs and control. Because of non-normality of data distribution, robust estimator of shift is

Hodges-Lehmann statistic (HL) [11]. HL statistic is calculated for distance between AML patients and control. For two sets of data with N_1 and N_2 elements, a new set, containing $N_1 \times N_2$ elements, is created. One element comes from each pair from set 1 and set 2 and equals difference of pair of values. The applied estimator is the median of $N_1 \times N_2$ differences.

$$\begin{aligned} d_{ij} &= x_i - y_j,\ i \in (1..N_1),\ j \in (1..N_2) \\ HL &= median(d) \end{aligned} \tag{1}$$

where:

- d is set of distances between each pair of set 1 and set 2
- x_i is i-th element of set 1
- y_j is j-th element of set 2
- N_1 and N_2 is set 1 size (number of AML samples) and set 2 size (number of healthy samples) respectively

The above procedure is carried out for each CpG site.

3.2 Gaussian Mixture Modelling

The second step of described method is decomposition of distribution of values of Hodges Lehmann statistic into Gaussian components. Let $f(x)$ denote the probability density function corresponding to the analyzed signal x. The Gaussian mixture decomposition model (GMM) of $f(x)$ is:

$$f(x) = \sum_{k=1}^{K} \alpha_k f_k(x, \mu_k, \sigma_k), \sum_{k=1}^{K} \alpha_k = 1. \tag{2}$$

where:

- K is the number of Gaussian components,
- α_k are non-negative component weights,
- f_k is the probability density function of a normal distribution ($N(\mu_k, \sigma_k)$) of the k-th component,
- μ_k, σ_k are k-th Gaussian component mean and standard deviation, respectively.

The Gaussian mixture model is fitted to HL statistic's distribution by using the method of maximization of the log-likelihood function (3).

$$\log L = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \alpha_k f_k(x_n, \mu_k, \sigma_k) \tag{3}$$

where:

- N is the total number of elements in modeled vector.

The expectation maximization (EM) algorithm [12] for recursive maximization of the likelihood function was applied. The initial values of decomposition parameters are randomly generated.

For finding the best number of Gaussian components, the algorithm was performed for different values of K (K = 2..12). Bayesian information criterion (4) (*BIC*) [13] was used for each examined K:

$$BIC = -2\log L + (3K - 1)\log N \tag{4}$$

Minimal value of *BIC* indicates what number of Gaussian components is the best to create the model.

Finding Gaussian components marks off subpopulations of HL statistic, hence enables categorization of CpG sites as low, medium or high demethylated.

According to the maximum probability rule [14], cut-off levels based on Gaussian components were detected. Cut-off levels are determined by intersection points of probability density functions of components.

3.3 Statistical Tests

For each CpG site a statistical comparison between AML patients and healthy donors was performed. Because data do not come from normal distribution, the applied test was Mann-Whitney U-test [15]. In a basic approach, the tested null hypothesis said that the difference between AMLs and control equals 0. In the modified situation the null hypothesis was defined by the found cut-offs. Such a method enables statistically significant detection of low, medium and high demethylated CpG sites of genome.

3.4 P-value Integration

Array annotation provided by Illumina let us indicate CpG sites belonging to regions which play a crucial role in gene expression control. Hypermethylation of TSS regions led to repression. In order to learn if TSS region of a particular gene was statistically demethylated, p-value integration was performed. Stouffer's method for p-value integration [16] was applied for CpG sites belonging to particular TSS regions. Each gene's TSS region was consider as demethylated by comparison integrated p-value with integrated significance level. Integrated significance level is based on number of CpG sites in particular TSS region. Z-value, which is basis to compute integrated p-value can be computed according to (5).

$$Z \sim \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{5}$$

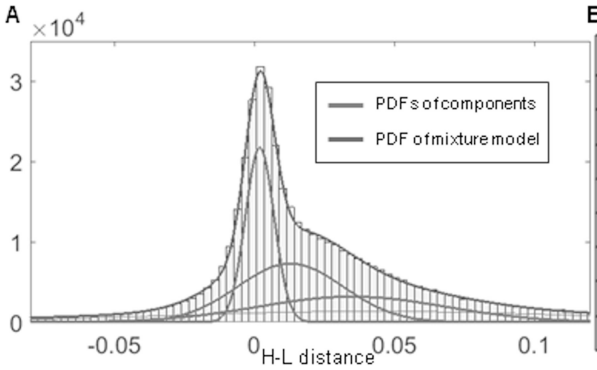
where:

- $Z_i = \Phi^{-1}(1 - p_i)$,
- p_i is the p-value for the i-th hypothesis test,

- Φ is the standard normal cumulative distribution function,
- k is number of integrated p-values.

4 Results and Discussion

After Gaussian decomposition of HL statistic's distribution, several Gaussian components were found (Fig. 2):



B

μ	σ	α
0.0128	0.0189	0.2645
0.0019	0.0051	0.2148
0.0348	0.0334	0.2045
0.0427	0.0748	0.1942
0.1792	0.1269	0.0597
-0.1248	0.1107	0.0358
-0.3006	0.1940	0.0158
0.3818	0.1775	0.0107

Fig. 2. (A) Histogram of HL distance with PDFs of components from GMM decomposition, (B) parameters of model components

Most of differences between AML cells and control cells happens in the "right side" of distribution, so for the examined situation, hypermethylation processes are more common than hypomethylation processes. It is possible to distinguish several levels of upmethylation, but it is impossible for downmethylation. Levels of demethylation can be called (just) demethylated, medium or high demethylated and high demethylated. Three the most important components (with highest weight and lowest standard deviations) describe three subpopulations of HL distance, hence three levels of demethylation (Fig. 3).

The first step of testing consisted of checking if CpG sites were just up- or downmethylated. For this purpose null hypothesis was equal to 0. For each CpG site, it was examined whether HL statistic was significantly lower or greater than 0. The next step was to check if particular CpG site was medium or high upmethylated. For these approaches cut-off levels found according to Gaussian decomposition were used as null hypotheses (Table 1).

Because of one-tailed test, significance level equals 2.5%. In basic approach, with null hypothesis equal to 0, much more CpG sites were detected as upmethylated than as downmethylated. It happens in whole genome as well as in TSS regions. About a half of upmethylated CpG sites are medium or high upmethylated. Part of them are only high upmethylated (Table 2).

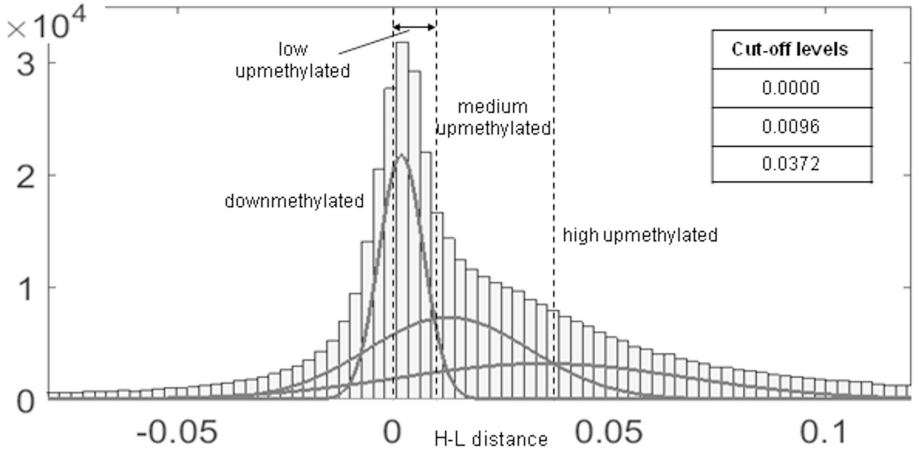


Fig. 3. Cut-off levels found according to maximum probability rule

Table 1. Number of demethylated probes for each tasted case

	AML downmethylated	AML upmethylated	AML medium or high upmethylated	AML high upmethylated
Test type	Left-tailed	Right-tailed		
Threshold	0	0	0.0096	0.0372
Probes within whole genome				
# significantly demethylated probes	15 260	84 073	47 659	17 317
% of 485 512 probes	3.14%	17.32%	9.82%	3.57%
Probes within TSS region only				
# significantly demethylated probes	3 772	20 354	10 034	3 982
% of 140 003 probes	2.69%	14.54%	7.17%	2.84%

Table 2. Number of demethylated TSS regions according to p-value integration

	AMLs downmethylated	AML upmethylated	AML medium or high upmethylated	AML high upmethylated
Test type	Left-tailed	Right-tailed		
Threshold	0	0	0.0096	0.0372
# significantly demethylated gene's TSS regions	106	1 088	474	122
% of 21 227	0.50%	5.13%	2.23%	0.57%

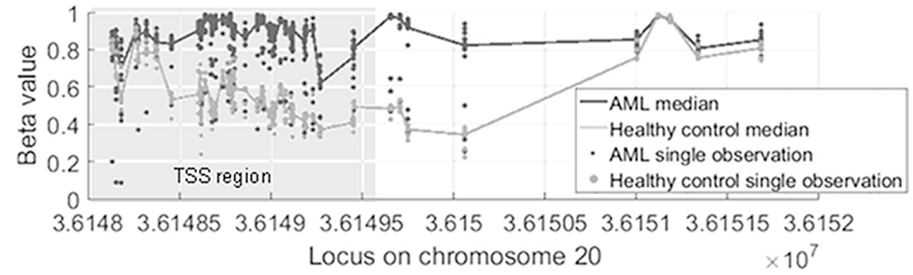


Fig. 4. Methylation level along NNAT gene in AML cells and healthy cells

More gene's TSS regions were high upmethylated than just downmethylated. TSS region with highest demethylation comes from NNAT gene. NNAT gene was described as trancsriptionally silenced because of hypermethylation in pedriatric AML [17]. It confirms the thesis that hypermethylation of TSS gene regions causes repression (Fig. 4).

5 Conclusions

Novel methylation data analysis method for efficient detection of demethylated DNA regions was proposed. In contrary to standard approaches, the developed algorithm is data driven and does not use a priori assumed cut-off thresholds. Such approach enables detecting of demethylated CpG sites of genome independently of initial methylation level in examined data and their distribution. It trades on Gaussian components which characterize subpopulation of demethylation level. Modified null hypotheses in U Mann-Whitney-Wilcoxon test enables to check whether particular CpG site is greater or lower not only than 0, but another thresholds. Hence, proposed method gives a possibility to classify CpG sites as low, medium or high demethylated. Due to p-value integration enables to conclude about particular gene TSS regions demethylation. Found upmethylated genes were successfully confirmed at literature and thereby validated the algorithm. Evaluation of proposed method by comparison with existing approaches needs extended study which will be part of further work.

Acknowledgements. This work was financed by SUT grant no. BKM/506/RAU1/2016/t.29 (AC) and SUT grant no. 02/010/BK_16/3015 (JP). All the calculations were carried out using infrastructure funded by GeCONiI project (POIG.02.03.01-24-099/13).

References

1. Zemach, A., McDaniel, I.E., Silva, P., Zilberman, D.: Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science **328**(5980), 916–919 (2010)
2. Gonzalo, S.: Epigenetic alterations in aging. J. Appl. Physiol. **109**(2), 586–597 (2010)

3. Du, P., Bourgon, R.: MethyAnalysis: DNA methylation data analysis and visualization, R package version 1.10.0 (2014)
4. Sofer, T., Schifano, E.D., Hoppin, J.A., Hou, L., Baccarelli, A.A.: A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. Bioinformatics **29**, 2884–2891 (2013)
5. Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P., Irizarry, R.A.: Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int. J. Epidemiol. **41**, 200–209 (2012)
6. Butcher, L.M., Beck, S.: Probe Lasso: a novel method to rope in differentially methylated regions with 450 K DNA methylation data. Methods **72**, 21–28 (2015)
7. Jung, N., Dai, B., Gentles, A.J., Majeti, R., Feinberg, A.P.: An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. Nat. Commun. **6**, 8489 (2015)
8. Aryee, M.J., et al.: Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics **30**(10), 1363–1369 (2014)
9. Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M., Esteller, M.: Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics **6**(6), 692–702 (2011)
10. Houseman, E.A., et al.: Model-based clustering of DNA methylation array data: a re-cursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distri-butions. BMC Bioinform. **9**(1), 365 (2008)
11. Hodges Jr., J.L., Lehmann, E.L.: Estimates of location based on rank tests. Ann. Math. Stat. **34**, 598–611 (1963)
12. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2004)
13. Claeskens, G., Hjort, N.L.: Model Selection and Model Averaging, vol. 330. Cambridge University Press, Cambridge (2008)
14. Huberty, C.J.: Applied Discriminant Analysis, vol. 297. Wiley, New York (1994)
15. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**(1), 50–60 (1947)
16. Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams Jr., R.M.: The American soldier: adjustment during army life. (Studies in social psychology in World War II), vol. 1 (1949)
17. Kuerbitz, S.J., Pahys, J., Wilson, A., Compitello, N., Gray, T.A.: Hypermethylation of the imprinted NNAT locus occurs frequently in pediatric acute leukemia. Carcinogenesis **23**(4), 559–564 (2002)

Metadata Analyser: Measuring Metadata Quality

Bruno Inácio, João D. Ferreira(✉), and Francisco M. Couto

LaSIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
jdferreira@fc.ul.pt

Abstract. Scientific research is increasingly dependent on publicly available information and data sharing. So far, the best practices to ensure that data is accessible and shareable has been to deposit it in public repositories. However, these repositories often fail to implement mechanisms that measure data quality, which could lead to improving the discoverability of existing data, and contribute to its future integration. In light of this, we present Metadata Analyser, a tool that measures metadata quality. It assesses the quality of metadata by considering the proportion of terms actually linked to ontology concepts, as well as the specificity of the terms used in the metadata. Metadata Analyser applied to Metabolights, a real-world repository of metabolomics data, and results show that the tool successfully implements the proposed measures, that there is indeed a lack of effort in the annotation task, and that our tool can be used to improve this situation. Metadata Analyser's frontend is available at http://masterweb-metadataanalyser.rhcloud.com.

Keywords: Metadata quality · Data sharing · Ontologies · Specificity · Coverage

1 Introduction

A significant portion of scientific research has recently become producer and consumer of large volumes of data, from multiple sources and in various formats [3]. In this scenario, data sharing takes an important role in the success of any scientific endeavour, as it allows scientific advances to "stand on the shoulders" of previous works, either performed by the authors themselves or by other teams [5]. This can only happen if data is properly integrated (i.e. categorized and organized in meaningful groups that reflect the data's similarities and differences), which enables information to be retrieved automatically [1]. However, ensuring data integration is a non-trivial task, sometimes regarded as non-scientific, and costly both in terms of human and time resources. Thus, it tends to be postponed, or even neglected.

The goals of this work are thus threefold: (*i*) to propose two measures of metadata quality, (*ii*) to implement a tool that is able to evaluate these measures in a public repository, and (*iii*) to show that these measures are valid and significant in a real-world scientific repository.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_24

2 Materials and Methods

We propose two measures of metadata quality: (*i*) the proportion of annotations in the metadata file that link to an ontology concept, and (*ii*) the average specificity of those ontology concepts.

The dependence on a notion of ontology is justified because ontologies are regarded by the biomedical community as standard representations of knowledge [6]. An ontology can be thought of as a graph that connects nodes (the relevant concepts) with edges (the relations between the concepts). For example, CHEBI contains statements about small molecules such as "carbon dioxide is-a greenhouse gas" and "glucose is-a carbohydrate". On the one hand, relying on ontologies allows us to base our measures in community-approved knowledge; on the other hand, an ontology concept is unambiguous, traceable, and represents a quantum of information that can be shared between the scientific community without potential for misinterpretation, enabling and enhancing data sharing.

2.1 Term Coverage

Usually, metadata files contain a mixture of ontology concepts and natural language terms. Since data sharing relies on the ability to find and retrieve information with automatic tools, ensuring that metadata is expressed as reference to ontology concepts improves its potential for being found in the future.

The first measure of metadata quality, therefore, is **term coverage**. It is the ratio between the number of annotations that refer to ontology concepts and the total number of annotations in the metadata file.

2.2 Semantic Specificity

Each ontology concept contains a certain amount of information, which can be measured by its specificity. More specific concepts have a higher information content and thus contribute with more specific knowledge to the metadata file. As such, we propose **semantic specificity**, a measure that reflects the average specificity of the concepts in the metadata file.

For a given concept, we consider the path from itself up to the root of the ontology and all the paths from itself down to the leaves of the tree. Let $T = \{t_1, t_2, \cdots, t_n\}$ be the set of ontology concepts found in a metadata file. For each t in T, its specificity $S_{\text{concept}}(t)$ is computed as

$$S_{\text{concept}}(t) = \frac{A(t)}{A(t) + D(t)} \tag{1}$$

where $A(t)$ is the number of ascendant concepts up from t and $D(t)$ is the average distance between t and all its leaf descendants, measured in number of edges. Concepts with low specificity are located at the top of the tree (near the root). A non-specific annotation contains small amounts of knowledge and is a weak descriptor of the contents of the resource: a more specific descendant concept

would be a better descriptor, since it would provide a more specific semantics to the resource and thus increase its potential for future integration. Concepts with high specificity are located near the leaves of the ontology, and correspond to informative annotations.

In order to determine the semantic specificity of an annotated resource, we average the specificity of the concepts in its metadata.

2.3 Motivation for the Measures

The two measures presented reflect the quality of the metadata associated with a resource. On the one hand, high coverage by ontology concepts in a resource's metadata file reflects a greater amount of computationally meaningful knowledge provided about that resource. On the other hand, as demonstrated above, the highest the specificity value of a concept, the better it is in describing the content of the resource. Therefore, high values for these measures enhance the meaning and discoverability of the data to those who wish to use it.

2.4 Metadata Analyser Architecture

To automatically evaluate the quality of a metadata file based on the quality measures described previously, we designed an architecture to analyse and evaluate the metadata file contents, Metadata Analyser. This is a modular architecture that can be adapted to other domains. For example, one module is responsible for reading Metabolights metadata files (see "Case Study" below), and another for computing the quality measures. Both modules can be exchanged by other ones, specific to other repositories or designed to compute other measures.

The tool is composed of the following layers:

1. An **interface layer** that interacts with the user by requesting a metadata file, informing the user on the analysis progress, and outputting the result.
2. An **application layer** that analyses the metadata file and evaluates the annotations found therein.
3. A **data layer** that holds the ontologies in local databases.
4. A web **API layer** that connects the interface layer to the application layer, coded in commonly used web technologies.

Source code is available at https://github.com/lasigeBioTM/MetadataAnalyser.

3 Case Study

To evaluate our work, we applied Metadata Analyser to Metabolights, a database of metabolomics experiments [4,9]. Metabolomics is the study of the chemical processes that occur in life-related contexts, usually within a cell or in its surroundings. This data often refers to a large number of scientific domains, as it can be cross-species and cross-technique, while covering metabolite structures, biological roles, locations and concentrations, as well as experimental factors.

Metabolights stores metadata associated with the actual data describing the information in each resource. For example, the metadata of the resource called "LCMS analysis of seven apple varieties with a leaking chromatographic column" claims that the data was collected through "liquid chromatography" and "mass spectroscopy", and that the study factors include "Sample type", "Apple number", etc.[1]. These pieces of metadata are collected (by the researcher or the curator) using the ISA-tools software suite [8]; in particular, metadata is saved in the ISA-TAB format, which has the built-in ability to refer to ontology concepts. At the moment of this study, the repository had 161 resources.

This repository has been developed and maintained by the EBI since 2012, and is therefore a relatively recent addition to the panorama of knowledge stores in the biomedical domain. Its use of ontology concepts in the metadata files has been advocated since the beginning, since Metabolights has always recommended its users to prepare and submit the data with that possibility in mind.

Our evaluation consisted of three steps: we first evaluated the measures on all the resources of the Metabolights database, then manually evaluated the results obtained in a selection of resources, and finally performed an evaluation of the metadata quality before and after a curation step performed by a team of metabolomics experts.

3.1 Metadata Quality in Metabolights

From the 161 resources, 6 did not contain any ontology annotation, i.e. the semantic specificity was 0.0. The average coverage was 0.25. Only 9 resources show a coverage of 0.40 or higher. The average semantic specificity was 0.81. Histograms of distributions are shown in Fig. 1. From these distributions, we can see that more effort is put into the semantic characterization of the resources than into making sure the terms are actually from a reference ontology.

Discarding the 6 resources with no ontology annotation, a small negative correlation was found between the two measures (see Fig. 2), with a Pearson correlation coefficient of -0.28 (corresponding to a p-value of 5.1×10^{-4}). This trend is only slightly negative, at best, even if statistically significant. Nonetheless, we argue that this may be related to the fact that the tasks of (*i*) looking for the most specific concept to use in the annotation and (*ii*) finding all the locations in the metadata file where an ontology concept can be used take time and thus cannot both be performed perfectly given time constraints.

The most relevant conclusion is that the semantic annotation of metadata describing the Metabolights resources is still far from the desired state of affairs.

3.2 Manual Evaluation of the Measures

To validate the correctness of the implementation, we randomly selected 6 resources and calculated the two quality measures manually. Results were compared both with a manual verification as well as with a previous work [7].

[1] See www.ebi.ac.uk/metabolights/MTBLS99.

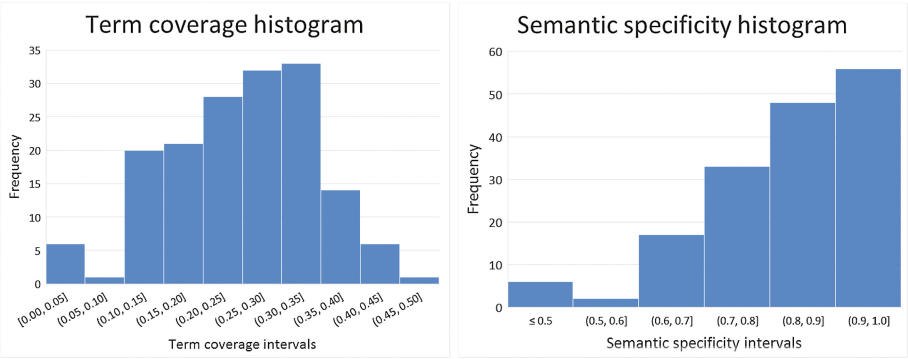


Fig. 1. The histograms of the distribution for the two measures of metadata quality in Metabolights. On the left, the distribution for term coverage; on the right, the distribution for semantic specificity.

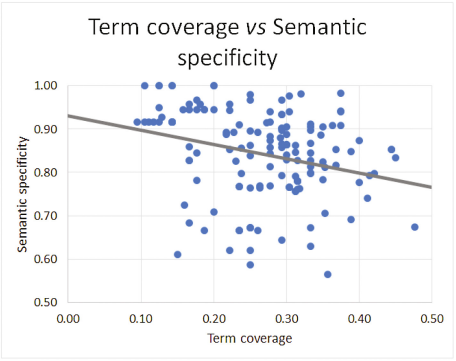


Fig. 2. The correlation between the term coverage and semantic specificity of all the metadata files for all resources in Metabolights.

The previous work computes the same results based on a web API that can be used to query biomedical ontologies (BioPortal). The metadata quality measures for the present work, the previous work and the manual validation are presented for the selected resources in Table 1.

These results show that values from the Metadata Analyser are close to the ones obtained from a manual computation. The only significant difference is that in two of the resources the term coverage is lower for our tool. This reduced amount of ontology concepts found in the metadata file leads to an artificial increase in the semantic specificity since the concepts that were exclusively found in the manual validation are non-specific. This limitation in our methodology is due to the fact that not all ontologies used to annotate the resources were included in the local database (e.g. one of the concepts used in MTLBS166 is from MeSH, but since Metadata Analyser did not include it in the database, it failed to compute a semantic specificity for the concept).

Table 1. Results from the manual validation

Resource	Manual results		This work		Previous work	
	Semantic specificity	Term coverage	Semantic specificity	Term coverage	Semantic specificity	Term coverage
MTBLS1	0.89	0.30	0.88	0.30	0.00	0.00
MTBLS36	0.96	0.17	0.96	0.17	0.00	0.00
MTBLS88	0.75	0.31	0.75	0.31	0.69	0.75
MTBLS110	0.84	0.28	0.91	0.14	0.87	0.50
MTBLS137	0.94	0.20	0.94	0.20	0.87	0.37
MTBLS166	0.60	0.23	1.00	0.14	0.00	0.54

The results from the previous study show a small semantic specificity value compared with the manual validation. They also present higher values of term coverage because that study uses a different algorithm to compute it. Finally, given that the previous work relies on a service over the web and that our methodology uses a local knowledge base, it is unsurprising to notice that Metadata Analyser is faster. In fact, it computes the results, on average, more than 10,000 times faster than the previous work (results not shown).

3.3 Evolution of Metadata Quality

To study the effect of an expert-driven curation process, we applied our measures of metadata quality to consecutive versions of three resources in the repository. The development team of Metabolights provided the pre- and post-curation versions of the resources MTBLS286, MTBLS287 and MTBLS288. The numbers for these three resources are presented in Table 2.

There are three general differences between the pre- and post-curation process. First, we notice an increase in the number of annotations, from 9 to 16 in each of the three resources. Furthermore, even though there are more annotations, we observe an increase in the amount of annotations that make use of ontology concepts, since the term coverage measure increases from an average

Table 2. Results from the pre- and post-curation analysis. **N** is the number of annotations that refer to ontology concepts.

Resource	Pre-curation			Post-curation		
	Semantic specificity	Term coverage	N	Semantic specificity	Term coverage	N
MTBLS286	0.00	0.00	9	0.96	0.25	16
MTBLS287	0.92	0.22	9	0.96	0.25	16
MTBLS288	0.92	0.22	9	0.87	0.25	16

of 0.15 to an average of 0.25. Finally, we also observe a mild increase in specificity. These three facts suggest that curators are able to increase the amount of machine-readable metadata that is available for each resource as well as its information content, measured by the semantic specificity. This experiment suggests that our measures do indeed capture a notion of metadata quality, since both experienced an increase after being handled by curation experts.

It is interesting to notice that the resources MTBLS287 and MTBLS288 already presented high values of semantic specificity prior to curation (higher than the full repository average), which means expert-driven curation could not improve then by much. That the curation process did not significantly alter them suggests that the annotation from the authors was already of high quality.

4 Discussion

There is an increasing usage of linked data techniques in Life and Health Sciences and many of them using biomedical ontologies, however to enhance their impact and value they need to produce high quality semantic descriptions of the data [2].

This work proposes two measures of metadata quality: (*i*) semantic specificity, which measures the average specificity of the ontology concepts referred to in the metadata and (*ii*) term coverage, which measures the proportion of annotations associated with actual ontology concepts. Based on them, we developed Metadata Analyser, an application that assesses metadata quality. It was evaluated by comparing its results both with a manual evaluation and a previous tool: results suggest that our measure corresponds to the expectations for metadata quality, as they increase after an expert-driven curation process. The tool is also significantly faster than the previously presented one and more accurate.

The major conclusion is that the two proposed measures can effectively measure the effort put into the semantic annotation of digital resources. This includes the annotation of a resource's metadata with explicit references to concepts from ontologies accepted by the community as machine-readable, standard representations of a domain of knowledge.

The results obtained from the Metabolights case study confirm the problem that motivated the creation of this tool, as we observe a weak term coverage (average of 0.25) and we hope it can be applied in existing repositories as a way to provide users feedback on their metadata quality, as well as motivating the general scientific community to increase their annotation efforts, so that we can, as a whole, spend more effort in ensuring proper data integration.

4.1 Helping with Scarce Semantic Integration

One possible cause behind the poor state of affairs in semantic annotation is of social nature, rather than technical [3]: metadata files are usually compiled by the authors of the data, who (*i*) may not know the ontologies that contain the concepts they need, (*ii*) do not fully know the structure of the ontologies in order to perform annotation with the appropriate specific terms, (*iii*) lack the

proper skills to carry on the annotation process because of the technical difficulties associated with this task, (*iv*) do not consider data sharing to be relevant, or (*v*) consider that the cost of ensuring proper semantic integration outweighs the benefits. The apparent correlation between specificity and coverage shown in Fig. 2 shows that a general effort exists to ensure specific concepts are used in the metadata, but not to ensure that ontology concepts are used throughout the metadata files, which suggests that indeed the perceived benefits may not significantly counterbalance the time costs of doing so. Without mandatory high quality metadata publication, it becomes difficult or even impossible to create automatic information retrieval mechanisms that can handle these author-created metadata files. While the short-term solution is to leverage on curators to help increase metadata quality, in a long-term scenario we wish to empower data creators with a means to measure the quality of their metadata, who would then use this feedback to improve metadata quality and thus the integration potential of the data.

Acknowledgments. This work was supported by FCT through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013. We thank the EBI team in charge of the development and maintenance of Metabolights for their support in this study.

References

1. Baker, M.: Quantitative data: learning to share. Nat. Methods **9**(1), 39 (2012)
2. Barros, M., Couto, F.M., et al.: Knowledge representation and management: a linked data perspective. IMIA Yearbook, pp. 178–183 (2016)
3. Couto, F.M.: Rating, recognizing and rewarding metadata integration and sharing on the semantic web. In: Proceedings of the 10th International Conference on Uncertainty Reasoning for the Semantic Web, vol. 1259, pp. 67–72 (2014)
4. Haug, K., Salek, R.M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González Beltrán, A., Sansone, S., Griffin, J.L., Steinbeck, C.: Metabolights - an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res. **41**(Database-Issue), 781–786 (2013). http://dx.doi.org/10.1093/nar/gks1004
5. Innovative Medicine Initiatives: IMI2: 9th Call for proposals. http://www.imi.europa.eu/sites/default/files/uploads/documents/IMI2Call9/IMI2_Call9_TopicsText.pdf. Accessed Apr 2016
6. Noy, N.F., McGuinness, D.L.: Ontology development 101: a guide to creating your first ontology (2001)
7. Ramos, C., Louro, M., Santos, M., Couto, F.M.: Knowledge ratings in metabolights. arXiv preprint arXiv:1604.07997 (2016)
8. Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., et al.: ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics **26**(18), 2354–2356 (2010)
9. Salek, R.M., Haug, K., Conesa, P., Hastings, J., Williams, M., Mahendraker, T., Maguire, E., González Beltrán, A., Rocca-Serra, P., Sansone, S., Steinbeck, C.: The metabolights repository: curation challenges in metabolomics. In: Database 2013 (2013). http://dx.doi.org/10.1093/database/bat029

Vascular Contraction Model Based on Multi-agent Systems

J.A. Rincon[1(✉)], Guerra-Ojeda Sol[2(✉)], V. Julian[1(✉)], and C. Carrascosa[1(✉)]

[1] D. Sistemas Informáticos y Computación, Universitat Politècnica de València, València, Spain
{jrincon,vinglada,carrasco}@dsic.upv.es
[2] Departamento de Fisiologìa, Universitat de València, València, Spain
solanye.guerra@uv.es

Abstract. This paper presents a first approximation to the simulation of vascular smooth muscle cell following an agent-based simulation approach. This simulation incorporates mathematical models that describe the behaviour of these cells, which are used by the agents in order to emulate vascular contraction. A first tool, implemented in Netlogo, is provided to allow the performance of the proposed simulation.

Keywords: Multi-agent system · Agent-based simulation · Vascular simulation

1 Introduction

Vascular tone is the vessel's property of increasing (vasocontriction) or decreasing (vasorelaxation) the tension of its walls in response to a given stimulus. Vascular tone is regulated by the simultaneous influence of intravascular vasoactive substances (hormones and platelet derivatives), neurotransmitters and the production of vasoactive substances released by the endothelium [1,2]. The walls of the blood vessels are arranged in three concentric layers: intima, media, and adventitia. The intimal layer, also called endothelium, is the layer located in the lumen of the vessel and is composed of a monolayer of endothelial cells (EC). The ECs are flat and elongated and release numerous vasoactive compounds such as superoxide anion (O^{2-}), thromboxane A2 (TXA2) and endothelin-1 (ET-1) [3]. The media consists of vascular smooth muscle cells (VSMC) and are organized into fiber bundles concentrically layered. VSMC are responsible for maintaining vasomotor tone [4]. The adventitial is the outer layer of the vascular wall and is formed by dense fibroelastic tissue, without smooth muscle cells, surrounded by connective tissue with fibroblasts and macrophages. The adventitia grants the vascular wall stability and transport nutrients to SMC [5].

The study of vascular tone modulation is important to understand and predict the response of the vascular system in a physiological and pathological environment. Traditionally, in the study of vascular reactivity the tissue-organ

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_25

bath methodology is used. Tissue-organ baths are used for in vitro dose-response experiments to investigate the physiology and pharmacology of tissue preparations. For these experiments vascular tissue is extracted from animal models and it can be easily subject to controlled changes in oxygen availability or drug administration.

In biomedical research the use of animal-based experimental processes is very common. However, the use of experimental animals has involved some ethical issues in the scientific community. For this reason, the new paradigm proposed in biomedical research aims to change animal-based experimental processes to a combination of in vitro cell-based experimental processes and computational model (in silico models).

The relationship between chemical activation and mechanical response in the vascular network is a complex system and the simulation with computational models requires a multi-scale simulation [6] that is able to reproduce this behaviour. Multi-scale simulation allows dynamic interaction simulation at the molecular, cellular and tissue level of biological systems. This type of simulation can use a continuous or discrete approach. In biological systems there are microscopic elements such as molecules or cells that interact with each other in a cooperative or competitive way. To model the behaviour of these elements a discrete approach, such as that offered by agent-based modelling (ABM) is the best option. ABM is based on intelligent autonomous entities or agents. Agents are able to perceive, act and communicate behaviours. These characteristics make ABM the most used technique to perform this type of multi-scale simulations.

Thus, the purpose of this work is to present a first approach of a tool for a biological simulation following a multi-agent system approach. The proposed tool will focus on the simulation of a vascular contraction model that is able to predict the generation of contraction force in response to a chemical stimulus in a VSMC.

The rest of this paper is structured as follows: Sect. 2 presents the problem description; Sect. 3 shows the previous works regarding vascular cell models and ABM simulations; Sect. 4 presents the model for the simulation of VSMC simulation using multi agent-systems; and, finally, Sect. 5 presents the conclusions and future work.

2 Problem Description

The increase and decrease of calcium concentration $[Ca^{2+}]$ are the main mechanisms that cause contraction and relaxation of vascular smooth muscle and consequent regulation of vascular tone. Contraction in vascular smooth muscle cells occurs by increasing the concentration of intracellular calcium $[Ca^{2+}]_i$ by Ca^{2+} influx through channels of the cellular plasma membrane and/or its release from its intracellular deposits (e.g., from the sarcoplasmic reticulum (SR)). After this increase, Ca^{2+} binds to the 4 binding sites of calmodulin (CM) producing the calcium-calmodulin complex (CaCM) and actives the myosin light chain kinase (MLCK). Once activated, MLCK induces phosphorylation of the serine at position 19 of the light chain (20 kD) of myosin (MLC20). This phosphorylation

allows the actin binding and the activation of myosin ATPase, which leads to cross-bridge formation between both proteins and the development of the active force necessary to produce the contraction.

Under physiological conditions an increase in $[Ca^{2+}]_i$ may lead to coupling excitation-contraction in smooth muscle, and this can occur by two mechanisms: Electro-mechanical coupling or pharmaco-mechanical coupling. Both mechanisms can occur simultaneously in a cell, acting individually or both at the same time. In fact, both mechanisms are not independent and there is usually a two-way interaction between them. During electro-mechanical coupling the membrane potential of the VSMC change. The resting potential of these cells ranges from -50 to -60 mV. When these values become less negative (depolarization) the voltage operated calcium channels (VOCC) are activated leading to Ca^{2+} influx and $[Ca^{2+}]_i$ increase. On the other hand, pharmaco-mechanical coupling is based on the binding of a contractile agonist to a receptor causing an increase in $[Ca^{2+}]_i$ without a previous change in membrane potential. Although, these changes in membrane potential may occur secondary. Several mechanisms of pharmaco-mechanical coupling have been proposed. Agonist binding to a G-protein coupled receptor is the most important. This mechanism involves phosphatidylinositol cascade activation producing increase of Inositol-(1,4,5)-triphosphate (IP_3) levels. The IP_3 releases Ca^{2+} from the SR producing an increase in $[Ca^{2+}]_i$.

3 Related Work

Over the last few years, we have seen different approaches related to biological simulation. These approaches is focused on how to model systems, tissue or cell behaviours and the most common tool used are the multi-agent system. Multi-agent systems are an artificial intelligent tool which define a group of entities with the capability of perception, cognition and actuation. These special features make agents the most adequate tool in this type of simulations. The biological simulation using multi-agent systems is used to understand the complexity of biological systems. The integration of biomedical and computational research has facilitated the modelling and simulation of biological complex units as cancer cells [7], heart cells [8] and others [9]. Based on these models it is possible, for example, to simulate the cell permeability and assess the pharmacological action of some drugs in the gastrointestinal tract. [10,11]. Similarly, while we find models that simulate intracellular absorption process, we also find models that simulate drug clearance. Drug metabolism and clearance determines the drug effectivity and it can be simulated since a low clearance in some circumstance can compensate low intestinal absorption [12].

On the other hand, in vascular biology, we can find models focused on simulating the mechanical behaviour of vascular walls and how some drugs can affect it. These models incorporate a mathematical approach describing the behaviour of vascular smooth muscle cell it includes the interaction. Furthermore, we can find that these models include the interaction to vasoactive compounds or ion

channels, as calcium (Ca^{2+}) and potassium ($K+$), involved in cellular homeostasis.

In all in all theses models is notorious the use of computational models. However, artificial intelligent (AI) techniques are also used in some cases. For example, Stephen Johnson [13] used artificial neural networks (ANN) for describing plasma protein binding (ppb) models based on literature data. In another work, Nathan McElroy (Penn State) presented an aqueous solubility model using genetic algorithms and simulated annealing to select the most useful subset of descriptors [14].

Nevertheless, the use of multi-agent systems in this kind of simulation is not very common, since designers do not have adequate tools to model and simulate biological systems. In our opinion, and after analyzing previous works, the use of multi-agent systems in these simulations provides to the designer the capacity to define very interesting concepts such as behaviours, norms, roles and also to be the container of other needed AI techniques. According to this, next section presents our proposal of a biological simulation based on multi-agent systems.

4 Simulation of Vascular Cells Based on Multi Agent-Systems

This sections presents an agent-based simulation model which incorporates AI tools and biological mathematical models. The presented model aims to give a first approximation to a vascular cell simulation tool, using the A&A [15] methodology. In recent years, the use of multi-agent systems has increased as a simulation tool. This is mainly due to the ability of this tool to incorporate behaviors, as well as to serve as a container for other AI tools useful in this type of simulations. On the other hand, the incorporation of other useful concepts such as the artifact, presented in the meta-model of agents and artifacts (A&A), allows the designer to easily differentiate between intelligent agents and objects of the environment. This model allows to perform a clear separation of entities in this type of simulations, since, not all the entities that interact with each other must have an intelligent behavior. In some cases, these artifacts are simply tools used by the agents to interact with the environment.

Our model is centred in the cells located in the arterial smooth muscle, which is responsible for vasoconstriction and vasodilation. Due to the complexity of these simulations, this paper will focus especially on vasocontriction through Electrochemical and chemomechanical interaction. These interactions have been studied in depth by Murtada et al. [16] and Yang J. et al. [17]. They introduced electrical, chemical, and mechanical phenomena in their models, driven by calcium, in order to predict the force generation in a VSMC.

The Fig. 1 shows a general view of the agent simulating a cell. This agent is based on BDI architecture (desires, beliefs, and intentions) incorporating two types of behaviours. The first one is related to the selection of the mechanism that will excite the cell and vasoconstriction occurs. Therfore two sub-behaviours are established: electro-mechanical coupling and pharmaco-mechanical coupling.

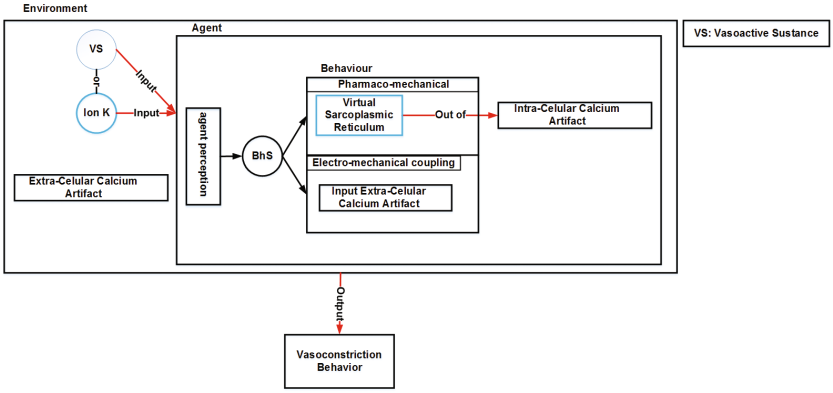


Fig. 1. General view of agent cell components

- **Pharmaco-mechanical coupling.** This behavior is activated when the agent perceives a vasoactive substance (VS). As a consequence, the agent activates the virtual sarcoplasmic reticulum (RSV). This activation creates the $Artifact\_Ca^{2+}$ + within the agent triggering the behavior of vasoconstriction.
- **Electro-mechanical coupling.** This behavior is triggered when the $Artifact\_Ca^{2+}$ + entry is required inside the *Cell_Agent*. This activation occurs when the artifact *VOCC_Artifact* is used inside the environment (extra-cellular medium). This artifact makes the environment less negative than the interior of the agent producing a depolarization that allows the $Artifact\_Ca^{2+}$ + to enter the agent and produce a vasoconstriction. To determine the depolarization levels, the agent perceives the levels of *Artifact_K* by using the equation of *Nernst* (Eq. 1) and takes the decision to use the environment $Artifact\_Ca^{2+}$ + and performs the vasoconstriction. Equation 1 is described as:

$$V_{eq} = \frac{RT}{zF} \cdot \frac{[X]_o}{[X]_i} \tag{1}$$

Where, **Veq** is the equilibrium potential (Nernst potential) for a given ion. It is common to use the ion symbol as a subscript to denote the equilibrium potential for that ion (e.g., VK, VNa, VCl, VCa, etc.). If only one ionic species is present in the system and channels for only the ionic species are present (and open), Veq will also be the membrane potential (Vm). Units for Veq. are Volts. However, the equilibrium potential is typically reported as millivolts (mV). **R** is the universal gas constant and is equal to $8.314 J.K - 1.mol - 1$ (Joules per Kelvin per mole). **T** is the temperature in Kelvin (K = °C + 273.15). **z** is the valence of the ionic species. **F** is the Faraday's constant and is equal to 96.485 C.mol-1 (Coulombs per mole). **[X]out** is the concentration of the ionic species X in the extra-cellular fluid.Concentration unit must match

with **[X]in**. **[X]in** is the concentration of the ionic species X in the intracellular fluid. In this case the concentration unit must match with **[X]out**. Concentrations units are milimolar (mM).

The second behavior performs the mechanical response. The activation is done through a Behaviour Selector (BhS). This selector allows the agent to determine what kind of sub-behaviors activate in each situation.

The proposed model has been divided into the representation of the environment (the extra-cellular medium) and a set of entities (including agents and artifacts) that interact with the environment and among them. Specifically, the components of the proposed model are the following:

- **Medium or Environment:** Is the space where our agents perform their interactions. It simulates the extra-cellular conditions where the cell is subjected to chemical stimuli like vasoactive substances modelled as artifacts.
- **VS Artifact:** This artifact will simulate the vasoactive substances that are coupled to receptor-dependent calcium channels to perform the pharmacological activation. To simulate vasoconstriction, this artifact will mimic the coupling of phenilefrine (Phe) to an alpha1 adrenergic receptor. Activation of adrenergic alpha 1 receptors results in the release of intra-cellular calcium from the SR and the opening of receptor-dependent calcium channels thus increasing [Ca2+]i, and generating mechanical response.
- **Ion-Calcium Artifact:** This artifact models the Ion Calcium which is the primary signal responsible for activation of vasocontriction mechanism. This artifact can be found within the environment (in the middle) or can be released by the VSMC through the SR. Activation of VSMC contractile response requires that $[Ca_i^{2+}]$ increase.
- **VOCC_Artifact:** This artifact will simulate changes in membrane potential. This artifact makes environment less negative as occur when potassium chloride blocks $K+$ channels and opens VOCC by producing a membrane depolarization and causing $Ca^{2+}+$ influx.
- **Agent Cell:** This agent is in charge of modelling the VSMC, allowing to simulate the interaction of the VS Artifact, VOCC_Artifact and Ion-Calcium artifact. This interaction will produce a mechanical response, that in this case will be vasoconstriction, by incorporating the previously described behaviors.

Figure 2 shows a first approximation of the simulation interface. This was done using the NetLogo tool[1]. In the interface the user can modify the number of agents (VSMC) to build the blood vessel. These agents are *red* color circles. In turn, the user can determine the number of the different artifacts. e.g. the yellow circles are *VOCC_Artifacts*, the green ones are *Ion-Calcium Artifacts* and the blue ones are *VS Artifacts*. The result of the simulation is the force generated by *VSMC* and is represented as tension-time relationship. MiliNewtons (mN) are tension units and minutes (min) are time units.

[1] https://ccl.northwestern.edu/netlogo/.

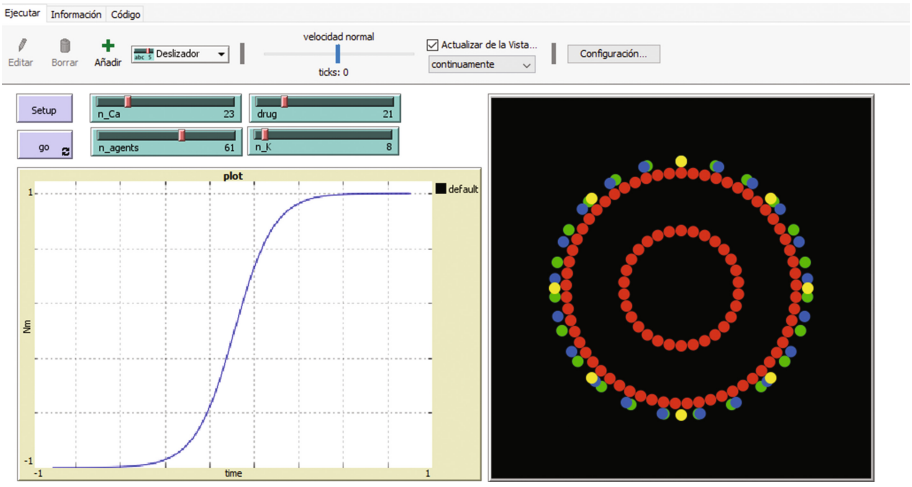


Fig. 2. Example of the Netlogo simulation.

5 Conclusions and Future Work

This paper presents a new tool for the simulation of vascular smooth muscle cells modulating vasoconstriction and vasodilation mechanisms. The proposed tool has been designed as an agent-based simulation and it is especially focused on simulating vasoconstriction mechanism. VSMC have been modelled as agents which interact with the environment through different elements modelled as artifacts. The designed agents incorporate some mathematical models taken from other studies exploring how smooth muscle responds to electrical or mechanical forcing. Moreover, we have implemented a prototype of the proposed system using the Netlogo tool. At this moment, we are evaluating the implementation of the model against real data.

References

1. Behrendt, D., Ganz, P.: Endothelial function: from vascular biology to clinical applications. Am. J. Cardiol. **90**(10, Supplement 3), L40–L48 (2002)
2. Lüscher, T.F., Richard, V., Tschudi, M., Yang, Z., Boulanger, C.: Endothelial control of vascular tone in large and small coronary arteries. J. Am. Coll. Cardiol. **15**(3), 519–527 (1990)
3. Galley, H.F., Webster, N.R.: Physiology of the endothelium. BJA: Br. J. Anaesth. **93**(1), 105 (2004)
4. Lacolley, P., Regnault, V., Nicoletti, A., Li, Z., Michel, J.-B.: The vascular smooth muscle cell in arterial pathology: a cell that can take on multiple roles. Cardiovasc. Res. **95**(2), 194–204 (2012)
5. Majesky, M.W., Dong, X.R., Hoglund, V., Mahoney, W.M., Daum, G.: The adventitia. Arterioscler. Thromb. Vasc. Biol. **31**(7), 1530–1539 (2011)

6. Thorne, B.C., Hayenga, H.N., Humphrey, J.D., Peirce, S.M.: Toward a multi-scale computational model of arterial adaptation in hypertension: verification of a multicell agent-based model. Front. Physiol. **2**, 1–12 (2011)
7. Wang, Z., Butner, J.D., Kerketta, R., Cristini, V., Deisboeck, T.S.: Simulating cancer growth with multiscale agent-based modeling. Semin. Cancer Biol. **30**, 70–78 (2015). Elsevier
8. Dada, J.O., Mendes, P.: Multi-scale modelling and simulation in systems biology. Integr. Biol. **3**(2), 86–96 (2011)
9. Qu, Z., Garfinkel, A., Weiss, J.N., Nivala, M.: Multi-scale modeling in biology: how to bridge the gaps between scales? Prog. Biophys. Mol. Biol. **107**(1), 21–31 (2011)
10. Wessel, M.D., Jurs, P.C., Tolan, J.W., Muskal, S.M.: Prediction of human intestinal absorption of drug compounds from molecular structure. J. Chem. Inf. Comput. Sci. **38**(4), 726–735 (1998)
11. Stenberg, P., Luthman, K., Artursson, P.: Virtual screening of intestinal drug permeability. J. Control. Release **65**(1), 231–243 (2000)
12. Gao, H., Lajiness, M.S., Van Drie, J.: Enhancement of binary QSAR analysis by a GA-based variable selection method. J. Mol. Graph. Model. **20**(4), 259–268 (2002)
13. Eyer, C.L.: Goodman & gilman's: the pharmacological basis of therapeutics. Am. J. Pharm. Educ. **66**(1), 95 (2002)
14. Shen, M., LeTiran, A., Xiao, Y., Golbraikh, A., Kohn, H., Tropsha, A.: Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. J. Med. Chem. **45**(13), 2811–2823 (2002)
15. Omicini, A., Ricci, A., Viroli, M.: Artifacts in the A& A meta-model for multi-agent systems. Auton. Agents Multi-Agent Syst. **17**(3), 432–456 (2008)
16. Murtada, S.I., Kroon, M., Holzapfel, G.A.: A calcium-driven mechanochemical model for prediction of force generation in smooth muscle. Biomech. Model. Mechanobiol. **9**(6), 749–762 (2010)
17. Yang, Jin, Clark, John W., Bryan, Robert M., Robertson, Claudia S.: The myogenic response in isolated rat cerebrovascular arteries: vessel model. Med. Eng. Phys. **25**(8), 711–717 (2003)

Study of the Epigenetic Signals in the Human Genome

Susana Ferreira[3(✉)], Vera Afreixo[1,2], Gabriela Moura[2], and Ana Tavares[1]

[1] Department of Mathematics & CIDMA, University of Aveiro, Aveiro, Portugal
[2] Department of Medical Sciences & iBiMED, University of Aveiro, Aveiro, Portugal
[3] Department of Medical Sciences, University of Aveiro, Aveiro, Portugal
catarina.fer@hotmail.com

Abstract. Epigenetics can be defined as changes in the genome that are inherited during cell division, but without direct modification of the DNA sequence. These genomic changes are supported by three major epigenetic mechanisms: DNA methylation, histone modification and small RNAs. Different epigenetic marks function regulate gene transcription, some of them when altered can trigger various diseases such as cancer. This work is focus on the epigenetic signals in the human genome, studding the dependency between the nucleotide word context and the occurrence of epigenomic marking. We based our study on histone epigenomes available in the NIH Roadmap Epigenomics Mapping Consortium database that contains various types of cells and various types of tissues. We compared genomic contexts of epigenetic marking among chromosomes and among epigenomes. We included a control scenario, the DNA sequence regions without epigenetic marking. We identified significant differences between context occurrence of control and epigenetic regions. The genomic words in epigenetic marking regions present significant association with chromosome and histone modification type.

Keywords: Epigenome · Histone modification · Epigenetic marking · Genome context · Data analysis

1 Introduction

Epigenetics is one of the most promising and intriguing areas of genetics. It is the science that studies the interaction between gene regulation, i.e. how genes are expressed, and its surrounding environment without involving changes in the DNA sequence level, which may still persist in future generations [1–3]. The inheritance of epigenetic marks from mother to daughter cells is crucial for the maintenance of a cell differentiation state and could be propagated by various epigenetic mechanisms, such as, DNA methylation, histone modifications and replacement of histone variants. The cell differentiation is a natural event in every organism, which involves no alteration of DNA sequence. However, all cells in an organism share the same genome (except B lymphocytes), each cell type has different kinds of epigenetic signatures, and each has a cell-type specific epigenome [1–3]. In other words, epigenetic has to do with changing the whole genome regulatory activity and this can be resumed in the epigenome, which is a kind of map that overlays the map of the genome, with epigenetic means that turn on or off genes,

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_26

increasing or reducing its activity. The epigenome can be studied through genomics and an important note is that epigenome is not static as the genome, it can be dynamic, influenced by environmental factors and extracellular stimuli, and change rapidly in response to these factors [4, 5].

Epigenetics can be regulated by three mechanisms: DNA methylation, histone modification and small RNAs. In this essay, we have studied epigenetic signals presented in the human genome, focusing mainly on the epigenetic regulation related to histone modification. Its importance will be described as follows.

Histone modifications and chromatin structure: The DNA is wrapped around two copies of each of the four core histone proteins H3, H4, H2B, and H2A, to form the nucleosome which is the fundamental repeating unit of chromatin [6–8]. The chromatin will be necessary for efficient packaging of the DNA into the nucleus of the cell. However, when DNA is compacted into the chromatin, its accessibility becomes greatly limited, it serves as a mechanism by which the cell protects DNA from external damage but it also regulates DNA mediated processes, such as transcription, DNA replication, DNA repair and chromosome segregation [6–8].

So, these histone proteins can influence chromatin organization and regulate many DNA-templated processes, through their chemical modification patterns (acetylation, methylation, sumoylation, and ubiquitylation) [6, 9]. Changes on the histone modification status may be associated with active or inactive chromatin. In addition, the combinatorial nature of various histone modifications occurring at different times during development, and at specific sites within histones, provides additional levels of regulation and complexity to the epigenome [8]. However, from these modifications can exert some biological effects, and how the addition or removal of many of these modifications is regulated, is still unclear.

The work presented in this paper studied the epigenetic signals of the human genome: the dependence between the context and the occurrence of epigenetic marking, chromosome and histone type; the identification of specific contexts related to epigenomic modification of a specific chromosome or histone.

2 Materials and Methods

We used the NIH Roadmap Epigenomics Mapping Consortium database that was designed, in 2008, to store human epigenomic data in order to encourage research [6, 10]. This database contains data for 31 histones modifications H2AZ, H2AK5ac, H2AK9ac, H2BK5ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK120ac, H3H4ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me2, H3K14ac, H3K18ac, H3K23ac, H3K23me2, H3K27ac, H3K27me3, H3K36me3, H3K56ac, H3K79me1, H3K79me2, H3T11ph, H4K5ac, H4K8ac, H4K12ac, H4K20me1, H4K91ac.

The epigenome files contain the sites of epigenomic marking (start and end positions) relative to the reference genome (GRCh37), defining the epigenetic regions. The word (k-mers) counts are obtained by DNA segment regions. The sequences are classified in two subgroups:

Control regions. All regions without epigenetic marking were used as control, consisting of 43916 fragments. The control regions are represented by 19369407 nucleotides, has 5619417 A nucleotides; 5625109 T nucleotides; 4063926 C nucleotides; and 4060955 G nucleotides.

Epigenomic regions. All regions with at least one epigenetic marking. If two fragments present intersection then we join them into one. Epigenomic regions are represented by 11325056856 nucleotides, has 1672399545 A nucleotides; 1674863703 T nucleotides; 1157288962 C nucleotides; and 1157976218 G nucleotides.

The word context analysis was subdivided essentially in three subanalysis: a global analysis comparing the control and epigenomic regions; a chromosome comparison; and a histone comparison.

We use standard statistical procedures: t-test, chi-square test, Cohen's d, Cramer's V, residual analysis and hierarchical clustering methods. To obtain k-mer ($k = 1, 2, 3, 4$) counts and perform the statistical analysis we use R software.

3 Results

Control and epigenomic regions analysis. In word context, the control and epigenomic regions present significant differences with low effect size difference, for the word lengths under analysis ($k = 1,\ldots,4$). The comparison was performed with chi-square test (p-values < 0.001) and complemented with the Cramer's V ($0.001 < V < 0.01$).

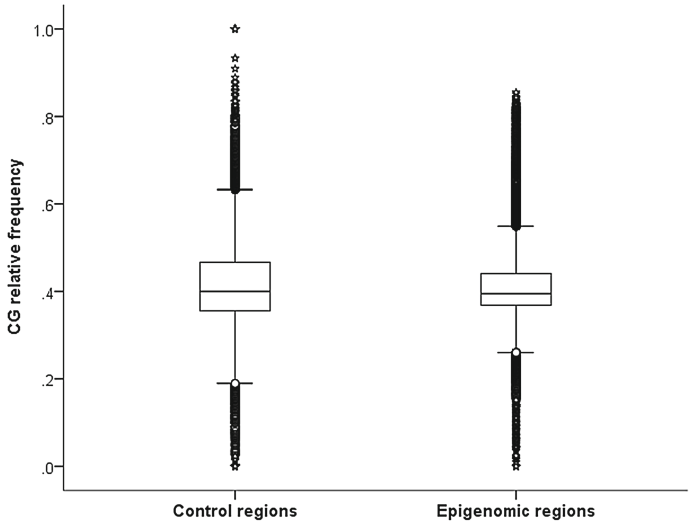


Fig. 1. Boxplot of C + G content for control and epigenomic samples. Both regions have several outlier fragments with similar median values, but the set of control regions presents high values dispersion.

It is know that the human genome has regions of high C + G content, alternating with regions of low C + G content. To rule out the hypothesis that the C + G contents could be marking the occurrence of epigenetic marking, we explore the differences between CG relative frequencies of control and epigenomic regions. Figure 1, show the CG relative frequency for the epigenomic and the control subset, where the differences between the two groups are globally low. We also applied the t-test and we concluded the C + G content in the two groups of sequences presents significant differences (p-value < 0.001). Through the Cohen's d, we concluded that the size effect of C + G content of our analysis is very small ($d = 0.039$). Thus, we classified the C + G bias between the two groups as negligible.

Chromosomes analysis. In this analysis, we wanted to evaluate if the genomic context associated with the occurrence of epigenetic marking is homogeneous, among chromosomes. For this, we applied the chi-square test and the Cramer's V value (Table 1).

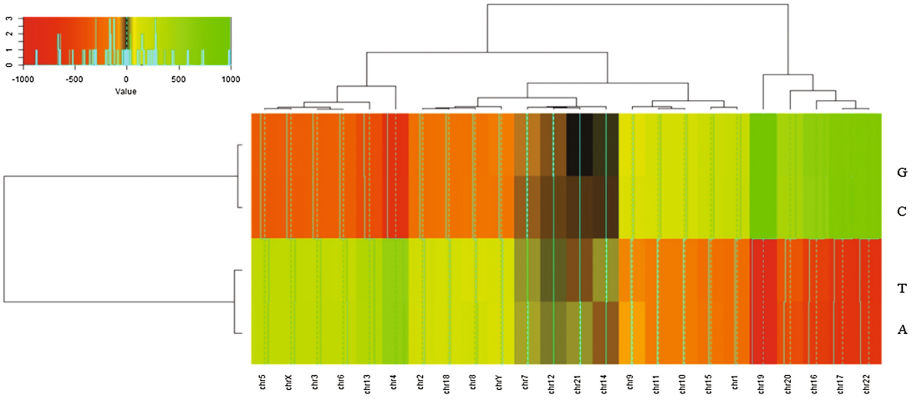


Fig. 2. Heatmap of chromosomes vs nucleotides, for epigenomic regions. Three chromosomes clusters were formed, two of which have strong nucleotides preferences (for A/T or G/C, respectively) and another cluster with more similar nucleotide preferences.

Table 1. Chi-square test to evaluate the homogeneity between chromosomes for epigenomic regions word context. X^2 - chi-square test; df - degrees of freedom; V - Cramer's V association measure; N - the sample size. *p-value is < 0.001.

Parameters	X^2	df	p-value	V	N
Nucleotide	10063000	69	*	0.0243	11325056856
Dinucleotide	22290000	345	*	0.0161	11324685040
Trinucleotide	34156000	1449	*	0.0161	11324313240
Tetranucleotide	46218000	5865	*	0.0188	11323941448

For nucleotide, dinucleotide, trinucleotide and tetranucleotide contexts, we concluded that there was a significant heterogeneity between chromosomes. Taking into

account the residuals values and the hierarchical analysis, we identified specific k-mers that were able to differentiate the human chromosomes taking into account the epigenomic regions context.

For example, in the nucleotide context, through a residual analysis we can observe that there are identical profiles in various chromosomes, with similar nucleotides preference (see Fig. 2). Table 2 presents in simultaneous the trinucleotide words and chromosomes with the highest residual values (>20) identifying specific preferred genomic contexts.

Table 2. Identification most favored genomic contexts in some chromosomes.

Chromosome with favored genomic contexts	Identify specific genomic contexts in some chromosomes
Chr3; Chr4; Chr5; Chr6; Chr13; ChrX	TAT; ATA
Chr16; Chr17; Chr19; Chr20; Chr22	GGG; CCC; GCC; GGC

Histone modifications analysis. This analysis was performed in order to compare the word context between different histone modification types. Specific contexts are associated with specific histone modification (p-value < 0.005, qui-square test).

For example, in the trinucleotide context, we concluded from the analysis of residues that each modifications has specific preferences. Table 3 presents in simultaneous the trinucleotide words and histones with the highest residual values (>20).

Table 3. Identify specific genomic contexts in histone modifications.

Histone modifications	Identify specific genomic contexts
H2AZ; H2AK5ac; H2BK5ac; H2Bk15ac; H3K4ac: H3K4me1; H3K4me2: K3K4me3; H3K9ac; H3K9me3; H3K23ac; H3K27ac; H3K27me3; H3K36me3; H3K79me1; H3K79me2	TTA and TAA
H2AK9ac	GGG; GCC; GGC
H2BK12ac; H2BK20ac; H3K14ac; H3K18ac; H3K56ac; H4K8ac	GCT; CTC
H2BK120ac	TTA; TAA; CTC; GCT
H3K9me1	AGG; GCT; CTG; CCT
H3K23me2	AGG; CTC; GCT
H3T11ph	GCT
H4K5ac	GCT; CTC;TTA
H4K12ac	GGG; GGC; CCC; GCC
H4K20me1	CAG; CAC; AGG; CTC; GCT
H4K91ac	GCT; CTT; AGG

4 Discussion

In this study we globally study the human epigenome, and the main objective was to identify motifs that could be associated with histone modification to further understand the relationship between DNA sequences and the occurrence of epigenetic marking.

Through heatmaps and the hierarchical clustering analysis, we could identify specific genomic contexts associated to each histone modification. One of the strongest contexts was TTA and TAA trinucleotides that are present mainly in regions of H2 and H3 histone modification, for both acetylation and methylation. However, there are other histone modifications that have other enriched motifs, as shown in the results. So, with these results, it may be possible to predict the occurrence of a modification from the nucleotide context of the region. Our epigenomic data is obtained from healthy cells, so with these profiles and the identification of the words with the greatest effect on the modifications, a comparison should be made between healthy and unhealthy cells and evaluate what differentiates them. It was also possible to create groups according to the type of histones (H2, H3 and H4) and the type of modification (acetylation or methylation) [6, 8, 9]. Curiously, it was observed two distinct groups: one including transcription-activating histone modifications (normally acetylations) and other including transcription-inactivating ones (normally methylation).

The trinucleotide and tetranucleotides contexts were the most informative ones differentiate chromosomes and histones. From the results, we can speculate that increasing the word size of contexts, more information and conclusions could be addressed. Because the computational complexity, we did not study higher word length, which is a limitation of this analysis.

Acknowledgment. This work was partially supported by Portuguese funds through the FCT (Portuguese Foundation for Science and Technology), CIDMA and iBiMED, within projects: UID/MAT/04106/2013, UID/BIM/04501/2013.

References

1. Ng, R.K., Gurdon, J.B.: Epigenetic inheritance of cell differentiation status. Cell Cycle **7**(9), 1173–11797 (2008)
2. Roloff, T.C., Nuber, U.A.: Chromatin, epigenetics and stem cells. Eur. J. Cell Biol. **84**(2–3), 123–135 (2005)
3. Probst, A.V., Dunleavy, E., Almouzni, G.: Epigenetic inheritance during the cell cycle. Nat. Rev. Mol. Cell Biol. **10**(3), 192–206 (2009)
4. Bernstein, B.E., Meissner, A., Lander, E.S.: The Mammalian Epigenome. Cell **128**(4), 669–681 (2007)
5. WHO. Genetics, genomics and the patenting of DNA: review of potential implications for health in developing countries. World Heal Organ (2005)
6. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., et al.: Analysis of dynamic changes in posttranslational modifications of human histones during cell cycle by mass spectrometry. NIH Public Access. **28**(10), 1045–1048 (2013)

7. Scholz, B., Marschalek, R.: Epigenetics and blood disorders. Br. J. Haematol. **158**(3), 307–322 (2012)
8. Chen, T., En, L.: Structure and function of eukaryotic DNA methyltransferases. Curr. Top. Dev. Biol. **60**, 55–89 (2004)
9. Bird, A.: DNA methylation patterns and epigenetic memory. Genes Dev. **16**, 6–21 (2002)
10. Consortium, R.E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al.: Integrative analysis of 111 reference human epigenomes. Nature **518**(7539), 317–330 (2015)

Cloud-Assisted Read Alignment and Privacy

Maria Fernandes[1(✉)], Jérémie Decouchant[1], Francisco M. Couto[2], and Paulo Esteves-Verissimo[1]

[1] SnT – Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, Luxembourg
maria.fernandes@uni.lu

[2] LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

Abstract. Thanks to the rapid advances in sequencing technologies, genomic data is now being produced at an unprecedented rate. To adapt to this growth, several algorithms and paradigm shifts have been proposed to increase the throughput of the classical DNA workflow, e.g. by relying on the cloud to perform CPU intensive operations. However, the scientific community raised an alarm due to the possible privacy-related attacks that can be executed on genomic data. In this paper we review the state of the art in cloud-based alignment algorithms that have been developed for performance. We then present several privacy-preserving mechanisms that have been, or could be, used to align reads at an incremental performance cost. We finally argue for the use of risk analysis throughout the DNA workflow, to strike a balance between performance and protection of data.

Keywords: Read alignment · Cloud computing · Genomic data privacy

1 Introduction

Genome sequencing evolved at an unprecedented rate with the advances of Next-Generation Sequencing (NGS) technologies. These new technologies allowed the sequencing costs to fall down to less than $1000 per genome, the machines throughput to increase from MB to TB of raw data produced per day, and the development of optimized parallelized procedures [19]. Medicine and biomedical research are benefiting from this evolution and started including sequenced data in their workflows [5]. However, to produce more comprehensive analysis using the large amount of NGS data generated, clinical and research entities faced new technical challenges. Indeed, they now have to share data and collaborate to improve the quality of their studies and the development of larger datasets [13].

Going further than traditional sharing schemes, domain experts established the e-biobanking vision [4], which calls for multi-research environment models and architectures facilitating the sharing of data. However, biomedical data (e.g.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_27

genomic sequences, medical reports, diseases information) is sensitive, as it is unique for each person and reveals information about herself and her relatives (e.g., predispositions to diseases). Therefore, a collaborative environment needs not only to enable the storage, the access and the analysis of biomedical data, but also be secure and reliable. Developing such an environment still remains a challenge.

As this integrated environment does not yet exist, scientists mostly relied on clouds to store and analyse sequencing data, due to their data sharing platform and improved computing schemes. However, the question remains on their ability to store and exploit genomes without breaking privacy policies. Despite the best efforts of cloud providers, the challenge is now set to accurately determine a threshold between the privacy and the openness of genomic data [23].

In this paper, we focus on the first step of the DNA analysis workflow — read alignment — which finds the location of a sequenced portion of DNA or RNA in a reference sequence. Section 2 summarizes the privacy-related features of genomic data, and describes the privacy attacks that have been presented in the literature, highlighting the importance of protecting genomic data. Section 3 describes cloud-based alignment algorithms which first emerged in response to the fast growth of sequenced data, highlighting their lack of consideration for privacy. Section 4 introduces the more costly algorithms that have been developed with privacy in mind. Finally, Sect. 5 gives some final remarks for the development of genomic data protective cloud environments, and argues for a risk-scale analysis that would be both practical and efficient. Section 6 concludes this paper.

2 Privacy Attacks on Genomic Data

Protecting genomic data is a non-trivial task, due to its many specificities which have been exploited in recent attacks. The attacks performed in order to obtain private information from genomic data all rely on one or several of the following characteristics.

Long-lived and static data. Genomic data stays sensitive at least as long as her owner lives, and contains particular properties, which make standard encryption mechanisms insufficient to protect it on the long term. Furthermore, once the privacy of genomic data has been compromised, there is no way to recover it, as the genome of a subject evolves very slightly during her life.

Hereditary information. Genomic information is transmitted from generation to generation. Thus, privacy leaks also affect the relatives of a victim.

Revealing diseases risk. Hereditary diseases are embedded in genes. The possession of even parts of a person's genome makes it possible to infer about her/his risk to develop certain disease. This information can lead to discrimination, for example, an employer might not offer a job to someone suffering from a chronic disease, or a health insurance could be denied to a person whose genome revealed a high risk to develop a disease. The same can occur with mortgage, if a person has a disease which decreases her lifespan.

Revealing personal response to medicines and risks to diseases. Prospects of personalized medicine show the benefits of using genomic information to adapt a patient's treatment to his particular expected reactions to it. However, this information could also be used for less glorious goals, since knowing the patient's reaction to a set of medicines can expose potential weaknesses.

Prone to manipulation. Ongoing research led to the belief that in the near future it will be possible to artificially recreate the DNA of any sequenced subject. As DNA samples are now used in forensics investigation to study crime scenes, artificial DNA samples could be introduced to influence investigations. This practice would compromise the ongoing investigations, by obfuscating any potential result or worse, lead to a wrong accusation.

In practice, the approach that has been followed by the existing platforms or services that work on genomic data until now has been a reactive one: data is made available and once a new attack is discovered, sensitive data is removed from public access. Several privacy-related attacks have been studied and described in the literature, we summarize them here.

Identification attacks are performed to determine the relation between the DNA profile of an individual and a data set. Taking as example a disease study case, an identification attack would reveal if a person is in the case or control data set, therefore breaking the privacy policies. Such an attack would typically reveal that a subject has a given disease [11].

Identity tracing attacks use records of genetic information and personal published information, which is available, for example, on genealogical databases (Ancestry[1]), diseases studies databases (DisGenet[2]), and surnames databases (e.g. Surname Navigator[3]). In the past, those databases reacted to reported attacks — such as the one determining Dr. Watson's APOE gene status [18] or the one using identification by surname inference [10] — by removing the detailed information used for the concerned attack from the database.

Recovery attacks determine a subject's sensitive genomic sequences using statistics and frequency information combined with released sensitive data (e.g. single nucleotide polymorphisms). Once the sequence is known, the attacker can use this information to launch the two previously mentioned attacks [25].

These attacks alerted the research community and the databases administrators of the possible data privacy threats. However, they cannot protect genomic data against future unknown attacks, as an attacker could collect and save data, and run an attack on it once it has been made public. Therefore, several privacy-preserving approaches to handle genomic data have appeared, which propose to protect data preventively.

In the next section, we discuss the existing cloud-based alignment solutions that the scientific community has adopted in order to leverage their high throughput, and we study them from a privacy-related point of view.

[1] Ancestry – https://www.ancestry.com.
[2] DisGeNet – http://www.disgenet.org.
[3] Surname Navigator – http://www.surnamenavigator.org.

3 Alignment in the Cloud

Aligning reads to a reference genome is one of the most important steps, and the first, of the sequencing analysis workflow that ultimately leads to genomic insights. Due to the throughput of NGS technologies and computational resources of research centers being unable to follow it, reads alignment is now often a bottleneck [21] and traditional algorithms, like BLAST [2] cannot be used as is. Hence, researchers started to offload the alignment of reads to cloud providers. Clouds are scalable computing infrastructures that allow users to adapt the resources they use to each of their analysis. These infrastructures allow users to benefit from their important computational power and storage space provided on demand through a simple internet connection, and at a manageable cost.

Several popular alignment tools have been adapted to run in clouds using Hadoop's MapReduce to execute code in parallel (Cloud-MAQ [22], Cloud-BLAST [15]). MapReduce's performance can be affected by the large amounts of data that has to be uploaded in the cloud, before executing the processing step. In addition, this data transfer increases cloud storage costs and causes increased latencies. This main limitation of MapReduce algorithms can be partly addressed using stream processing engines, which have also been explored in combination with alignment algorithms. Kienzler et al. [14] proposed a stream-based sequence analysis approach where the transfer step is replaced by data streaming, thus avoiding the huge amount of data transfer. Even though streaming approaches improved performance, since they apply data compression and decompression, read alignment remains computationally intensive and time consuming.

Although cloud processing improves performance and provides more storage space, it poses security concerns. A cloud infrastructure is controlled by a Cloud Service Provider (CSP), which does not provide the users full control over their own data. Additionally, CSPs can copy, transfer and store the data into multiple-locations (for fault-tolerance or economic reasons), and do not guarantee that the data cannot be accessed by the CSP or an intruder [20]. Thus, researchers need to consider a cloud as an untrusted, and possibly insecure, environment. To deal with genomic data on clouds, researchers and CSPs should discuss and adapt the privacy policies (eg. data control, security, confidentiality, transferring) to guarantee data protection [8].

Cloud computing offers the best solution in terms of modularity concerning computational power and costs to analyse large quantities of data. However, the algorithms described in this section require the client to upload his data into the cloud, where it is treated in plain text (i.e., without using any encryption mechanism). Considering that the user-cloud communications are made via internet, where communications could be intercepted and genomic data decrypted, given enough time, and the trust we give to the cloud provider, using such an infrastructure presents privacy issues which need to be addressed.

4 Privacy-Preserving Alignment in the Cloud

Privacy-preserving methods for execution in the cloud can involve cryptographic or non-cryptographic mechanisms and client-CSP agreements that must be followed. Both, client and CSP need to be aware of the sensitivity of biomedical data to ensure the adequate privacy protection [1]. In this section, we introduce the current privacy-preserving methods that could be applied to biomedical data, and then present real-life applications of these mechanisms.

The non-cryptographic techniques include data anonymization, the control of accesses, and privacy agreements.

Data anonymization consists in removing the personal information (e.g. name, surname, birth data, address, age) to avoid direct associations between genomes and their donors. Some portions of genomes have been considered privacy-critical information as well [10], which raises the challenge of identifying such genomic portions.

Access control consists in specifying who is allowed to access the data, often with different access levels, to limit and track its usages. For example, a medical center may have access to the disease genes of a patient and another research unit would only have access to the genes related to a particular disease under investigation [9].

Privacy agreements are signed documents specifying that a donor grants access to his data. All the entities (e.g. donor, researcher, medical institution) that can access the data sign the agreement, and it is assumed that all the concerned parties are trustworthy. Historically, privacy agreements were the first privacy-preserving technique developed around genomic data. However, the necessary uses of untrusted machines and communications links render privacy agreements unable to fully protect data.

These three methods are considered insufficient to protect genomic data alone, however it is believed that when combined with cryptographic privacy-preserving techniques they increase the protection of sensitive data [9]. Cryptographic techniques provide high privacy guarantees to very specific scenarios. However, the scientific community has been working towards extending their range of applicability to study genomic data.

Keyed-hash functions convert clear-text to hashes and combine them with a secret key. This technique however relies on the assumption that the key is never stolen, since in that case all the data would be accessible [6]. In addition, this approach does not allow direct collaboration between multiple entities.

Differential privacy introduces randomness to the input of a function in order to protect its privacy-sensitive features. Intuitively, the output of a function must not vary much whether an individual is part of the study or not. The main issue of this technique is to control the amount of randomness introduced in human genomes, so that studies can produce meaningful results [17,23].

Garbled circuits are a cryptographic technique for two-party secure computation. This technique allows a user to send his data to a receiver (e.g. cloud

service) to make some computations and receive back the final output. During this process, neither the input nor intermediate values are revealed [3].

Lastly, **homomorphic encryption** schemes have been explored as a security method for genomic data. These schemes allow a computation to be executed on encrypted data, and its result to be decrypted, therefore providing insight on the plaintext data. However, their performance is currently unsatisfactory and it only allows a limited number of operations [3].

Several privacy-preserving cloud alignment solutions have been recently published. Those solutions rely on hybrid clouds environments where the most sensitive data computations are performed on a private cloud and the less sensitive is processed on a public cloud [24]. Some solutions apply keyed-hash functions on the sensitive data and then send the hash-values to the cloud [6]. Homomorphic encryption has also been applied on other steps of the analysis of genomic data, e.g. for disease susceptibility tests [16]. However, these examples still present some limitations: the most CPU intensive task (i.e., the extend step) has to be performed in the private cloud; the need of an efficient and reliable sensitive data classifier; the use of hash algorithms that may be broken before the expiration of the genomic data they protect.

5 Towards a Differentiated Protection of Genomic Data

Several privacy-preserving methods have been developed, however their limited usability stills cannot address all the different issues found in the workflow analyses steps. In this section, we describe how classifying the sensitivity of genomic data would contribute to a thorough use of the potential of existing algorithms, at the best possible cost.

Enabling technologies. A filtering approach that classifies reads as embedding sensitive or non-sensitive information has been described in [7]. Adding this filtering step would allow the reduction of data encryption costs by encrypting only the critical information and improve the data usability, while ensuring the protection of genomic data. In addition, the level of sensitivity of reads could be determined according to the attack power it provides to an attacker through a risk-analysis study. Doing so, however, requires further work. We are convinced that such approaches will be developed in the future, and now present the benefits they would bring to different stages of the DNA workflow.

Privacy-preserving alignment can be obtained in mainly two ways: rely on plaintext conventional algorithms in a secure environment (e.g. local computer, private cloud) [14,22], or protect data through cryptographic methods. In the former there is always a risk for an adversary to get access to the machines, and therefore to the sensitive data. The second solution can be too costly or even unpractical since encryption makes data unavailable for some operations [3,12]. Classifying data into sensitivity levels would allow both approaches to be combined, globally improving performance, as the more-costly algorithms would be applied only to the most sensitive data, while improving the performance of the low-sensitivity reads.

Storage security requires long-term protection techniques. The most sensitive data could be stored in highly restricted and protected areas, while less sensitive data could be stored encrypted on the cloud. Splitting data and differentiating the way it is stored based on its sensitivity would reduce the storage costs as the most secure environments are usually more costly.

Release of and access to sensitive data require an extensive understanding of genomic data privacy breaches. For a privacy protective data release it is, of course, necessary to hide all the unique individual information (e.g. names, address, genes) [25]. Differentiating the sensitivity of genomic data would allow more data to be released to scientists, while the most sensitive one would still be protected. Data aggregation was also purposed as a secure solution for data release, however it remains in a early stage of understanding and application. For example, a human genome contains around 10 million single nucleotide polymorphisms (SNPs), and therefore a secure aggregate of full genomes would have to involve more than 80 millions of subjects [25] (≈1.15% of the world population).

6 Conclusion

The migration of read alignments to the clouds and the parallelization of the process using MapReduce, have greatly improved the performance of this essential step of the DNA workflow. However, these solutions require data to be manipulated in plain-text in the cloud, which poses privacy concerns, which were highlighted by the genomic privacy attacks reported in the last years. As researchers became more aware of those data vulnerabilities, the last years saw the development of privacy-preserving solutions to replace the typical alignment algorithms, which are deprived of privacy measures. Until now, it seems that privacy protection and performance are inversely related, since the improvement of one leads to the decrease of the other. Thus, the golden question is how to provide data privacy protection while taking advantage of the storage and computational power that cloud environments provide. Accurately determining the level of sensitivity of genomic information seems to be a way to go to benefit entirely for the broad range of algorithmic, storage and access solutions that have been developed. Such a secure cloud environment for biomedical data analysis is still an open challenge.

Acknowledgements. This work was supported by the Fonds National de la Recherche Luxembourg (FNR) through PEARL grant FNR/P14/8149128, and by the Fundação para a Ciência e para a Tecnologia (FCT) through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013.

References

1. Akgün, M., Bayrak, A.O., Ozer, B., et al.: Privacy preserving processing of genomic data: a survey. J. Biomed. Inf. **56**, 103–111 (2015)
2. Altschul, S.F., Gish, W., Miller, W., et al.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 403–410 (1990)

3. Baron, J., El Defrawy, K., Minkovich, K., et al.: 5pm: secure pattern matching. In: SCN, pp. 222–240 (2012)
4. Bessani, A., Brandt, J., Bux, M., et al.: Biobankcloud: a platform for the secure storage, sharing, and processing of large biomedical data sets. In: DMAH (2015)
5. Chan, I.S., Ginsburg, G.S.: Personalized medicine: progress and promise. Ann. Rev. Genomics Hum. Genet. **12**(1), 217–244 (2011)
6. Chen, Y., Peng, B., Wang, X., et al.: Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. In: NDSS (2012)
7. Cogo, V.V., Bessani, A., Couto, F.M., et al.: A high-throughput method to detect privacy-sensitive human genomic data. In: ACM WPES, pp. 101–110 (2015)
8. Dove, E.S., Joly, Y., Tasse, A.M., et al.: Genomic cloud computing: legal and ethical points to consider. Eur. J. Hum. Genet. **23**, 1271–1278 (2015)
9. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**, 409–421 (2014)
10. Gymrek, M., McGuire, A.L., Golan, D., et al.: Identifying personal genomes by surname inference. Science **339**(6117), 321–324 (2013)
11. Homer, N., Szelinger, S., Redman, M., et al.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genet. **4**(8), e1000167 (2008)
12. Huang, Y., Evans, D., Katz, J., et al.: Faster secure two-party computation using garbled circuits. In: USENIX Security Symposium, vol. 201(1) (2011)
13. Kaye, J., Heeney, C., Hawkins, N., et al.: Data sharing in genomics re-shaping scientific practice. Nat. Rev. Genet. **10**(5), 331–335 (2009)
14. Kienzler, R., Bruggmann, R., Ranganathan, A., et al.: Large-scale DNA sequence analysis in the cloud: a stream-based approach. In: ICPP, vol. 2, pp. 467–476 (2012)
15. Matsunaga, A., Tsugawa, M., Fortes, J.: Cloudblast: combining mapreduce and virtualization on distributed resources for bioinformatics applications. In: ESCIENCE 2008, pp. 222–229 (2008)
16. Namazi, M., Troncoso-Pastoriza, J.R., Pérez-González, F.: Dynamic privacy-preserving genomic susceptibility testing. In: ACM MMSec, pp. 45–50 (2016)
17. Naveed, M., Ayday, E., Clayton, E.W., et al.: Privacy in the genomic era. ACM CSUR **48**(1), 1–44 (2015)
18. Nyholt, D.R., Yu, C.E., Visscher, P.M.: On Jim Watsons apoe status: genetic information is hard to hide. Eur. J. Hum. Genet. **17**, 147–149 (2009)
19. O'Driscoll, A., Daugelaite, J., Sleator, R.D.: "Big data", hadoop and cloud computing in genomics. J. Biomed. Inf. **46**(5), 774–781 (2013)
20. Rocha, F., Correia, M.: Lucy in the sky without diamonds: stealing confidential data in the cloud. In: DSNW, pp. 129–134 (2011)
21. Stein, L.D.: The case for cloud computing in genome informatics. Genome Biol. **11**(5), 207 (2010)
22. Talukder, A., Gandham, S., Prahalad, H., et al.: Cloud-maq: the cloud-enabled scalable whole genome reference assembly application. In: WOCN, pp. 1–5 (2010)
23. Vayena, E., Gasser, U.: Between openness and privacy in genomics. PLoS Med. **13**(1), 1–7 (2016)
24. Zhang, K., Zhou, X., Chen, Y., et al.: Sedic: Privacy-aware data intensive computing on hybrid clouds. In: ACM CCS, pp. 515–526 (2011)
25. Zhou, X., Peng, B., Li, Y.F., et al.: To release or not to release: Evaluating information leaks in aggregate human-genome data. In: ESORICS, pp. 607–627 (2011)

On the Role of Inverted Repeats in DNA Sequence Similarity

Morteza Hosseini(✉), Diogo Pratas, and Armando J. Pinho

IEETA/DETI, University of Aveiro, Aveiro, Portugal
{seyedmorteza,pratas,ap}@ua.pt

Abstract. In this paper, we propose a computational approach to quantify inverted repeats. This is important, because it is known that the presence of inverted repeats in genomic data may be associated to certain chromosomal rearrangements. First, we present a reference-based relative compression method, which employs statistical characteristics of the genomic data. Then, for determining the similarity between genomic sequences, we use the normalized relative compression measure, which is light-weight regarding computational time and memory. Testing this approach on various species, including human, chimpanzee, gorilla, chicken, turkey and archaea genomes, we unveil unreported results that may support several evolution insights.

Keywords: Inverted repeats · Relative compression · Finite-context model · Reference-based compression · Chromosomal rearrangement

1 Introduction

With increasing rise in the production of genomic data, our scientific knowledge of genome sequence information is continuously being updated. Along with this, there are challenges regarding storage, processing and transmission of this data deluge, as well. Compression is a solution to address these challenges. Heretofore, several methods have been proposed for this purpose [1–4].

Genomic sequences have specific properties, such as the presence of inverted repeats (IR) [5]. IRs are sub-sequences of genomic sequences which are reversed and complemented copies of some other sub-sequences [6]. They may play an important role in chromosomal rearrangements [7]. The compression methods, aside from providing more efficient storage, processing and transmission, can also help investigating the properties of IRs.

To investigate the properties of the IRs in DNA data, a measure is required. For this purpose, different measures have been proposed, such as normalized compression distance [8], normalized conditional compression distance [9,10] and normalized relative compression [11]. These measures rely on the notion of Kolmogorov complexity [12], which is the length of a shortest binary program that computes a binary string of finite length in a universal Turing machine and

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_28

halts [13]. In this paper, we use the normalized relative compression, given by $\mathrm{NRC}(x,y) = \frac{C(x\|y)}{|x|\,\log_2|\mathcal{B}|}$, where $C(x\|y)$ denotes the relative compression of x based on y, which is the size of the compressed version of a string x given *exclusively* the information contained in a string y, $|x|$ denotes the size of x, and $|\mathcal{B}|$ denotes the alphabet size [11].

The rest of this paper is organized as follows: In Sect. 2, we present a reference-based relative compression method for quantifying IRs. The key idea of this reference-based compression is to build a model based on a reference sequence, then freeze the model and compress a target sequence, based on that model. In Sect. 3, we test this method on various datasets, both by considering and not considering IRs. Then, we use compact heatmaps to compare the results. Finally, in Sect. 4, we draw some conclusions.

2 Method

A finite-context model (FCM) relies on the Markov property, i.e., it employs the $k > 0$ most recent symbols (context-order size k) of the information source, to estimate the probability of the next symbol [14]. Denoting the k most recent symbols as $x_{n-k+1}^n = x_{n-k+1}\cdots x_n$, the probability estimates $P(x_{n+1}|x_{n-k+1}^n)$ are calculated based on the number of symbols which are accumulated while the information source is being processed. Therefore, we have

$$P(s|x_{n-k+1}^n) = \frac{\mathcal{C}(s|x_{n-k+1}^n) + \alpha}{\mathcal{C}(x_{n-k+1}^n) + \alpha\,|\mathcal{B}|}, \tag{1}$$

in which $|\mathcal{B}|$ denotes the size of alphabet $\mathcal{B} = \{s_1, s_2, \ldots, s_{|\mathcal{B}|}\}$, containing the objects of interest. Heretofore, the alphabet $\{A, C, G, T\}$ has been considered for DNA data [15]; We consider the 'N' symbol, as well. Therefore, we have $\mathcal{B} = \{A, C, G, T, N\}$ and $|\mathcal{B}| = 5$. Also, in Eq. (1), $\mathcal{C}(s|x_{n-k+1}^n)$ represents the number of times that symbol $s \in \mathcal{B}$ has been found in the past, considering x_{n-k+1}^n as the conditioning context. We have $\mathcal{C}(x_{n-k+1}^n) = \sum_{a\in\mathcal{B}} \mathcal{C}(a|x_{n-k+1}^n)$, which denotes the total number of events occurred within context x_{n-k+1}^n. Parameter α allows balancing between the maximum likelihood estimator and a uniform distribution. For large number of events, n, the estimator behaves as a maximum likelihood estimator. Also, when $\alpha = 1$, Eq. (1) turns out to be the Laplace estimator [16].

After the first n symbols of x are processed, the per symbol information content average, which is provided by an order-k FCM, is given by

$$H_{k,n} = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P_R(x_i|x_{i-k}^{i-1}) \quad \text{bpb}, \tag{2}$$

in which P_R is the probability regarding the reference sequence, and also "bpb" stands for bits per base. The upper bound of the bpb for sequences of five symbols (A, C, G, T and N) is $\log_2 5 = 2.32$. This value is obtained in a situation where the symbols are independent and equally likely. Note that the smaller the value of

bpb is, the better the model is [17]. The values of NRC are obtained by dividing $H_{k,n}$ by $\log_2$ of the alphabet size (in this case, 2.32). Note that $0 < \text{NRC} \leq 1$ and that the closer its value is to zero, the better the sequence can be compressed using the reference.

We have implemented the method, using C++ language, and provided a command-line tool, `Phoenix` [18], which can be applied to any genomic sequence, in FASTA or bare sequence ($ACGTN$) format. Along with `Phoenix`, a set of bash scripts were written for downloading, installing and running `Phoenix`, as well as plotting the results in an automatic way. It is worth mentioning that `Phoenix` is a complementary tool to `SMASH` [16], since it allows to indicate, in a compact map, which possible rearrangements, of IR type, need to be looked by `SMASH`. Besides, it offers the possibility to have a metric to quantify the overall inversions.

3 Results

In order to test the reference-based relative FCM method, we employed genomes with different species origins and lengths, in FASTA format (see Table 1). The tests were carried out on a server with 16-core 2.13 GHz Intel® Xeon® CPU E7320 and with 256 GB of RAM. For all datasets, the parameters α and context-order size associated with our method, were set to 0.01 and 20, respectively. The results can be replicated using the `Phoenix` [18] and `GOOSE` [19] softwares.

Table 1. Target and reference genomic sequence datasets [18].

Target genome	Scientific name	Abbr	File size (GB)	Reference genome	File size (GB)
Human	*Homo sapiens*	HS	3.3	Chimpanzee	3.3
Chimpanzee	*Pan troglodytes*	PT	3.3	Human	3.3
Gorilla	*Gorilla gorilla*	GG	4.4	Human	3.3
Chicken	*Gallus gallus*	GGA	1.3	Turkey	1.2
Turkey	*Meleagris gallopavo*	MGA	1.2	Chicken	1.3
Archaea	*Archaea*	A	0.5	Archaea	0.5

Hereinafter, the notation X.i refers to chromosome i of sequence X. Moreover, notations MT, UL, UP, AL and LG refer to mitochondrial DNA, unlocalized sequence, unplaced sequence, alternate locus and a linkage group, which is not assigned to a chromosome, associated with different sequences.

In Fig. 1, the heatmaps of the NRC values, regarding the compression of human and chimpanzee chromosomes, are plotted, in an all to all scheme. Squares show how much similar the reference and target sequences are. The less the NRC value is, the more similar the corresponding chromosomes are, since less bits per base are used for their relative compression. Figure 1a and b show

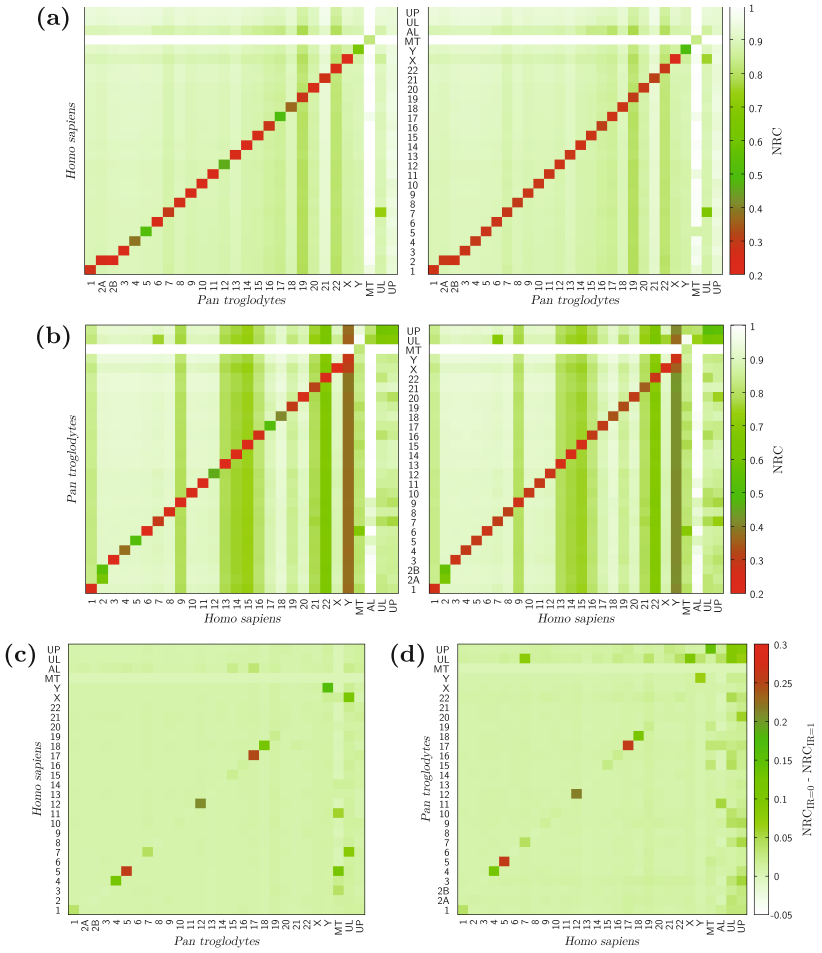


Fig. 1. NRC values obtained by compression of human and chimpanzee chromosomes. (a) left: IRs not applied (IR = 0), right: IRs applied (IR = 1) [reference: HS, target: PT], (b) left: IR = 0, right: IR = 1 [ref.: PT, tar.: HS], (c) the difference in NRCs between applying and not applying IRs ($NRC_{IR=0} - NRC_{IR=1}$) [ref.: HS, tar.: PT] and (d) the difference between NRCs [ref.: PT, tar.: HS].

NRC values with and without considering IRs. As can be seen, there is higher similarity between reference and target chromosomes with the same numbers, rather than others, except for HS.2, HS.MT, HS.UL and HS.UP related to PT.2A & PT.2B, PT.MT, PT.UL and PT.UP, respectively. Also, HS.AL is not similar to any PT chromosome. Note HS.2 (related to PT.2A & PT.2B) is presumed to contain an ancestral chromosome fusion [20]. HS.Y (as a target) is highly correlated to PT.X, since they had possibly exchanged information in recombination processes [21].

Figure 1c and d show the difference in NRC between considering and not considering IRs in the compression. The larger the difference of NRC values for two sequences is, the more similar those sequences are when considering IRs, and, consequently, possibly the higher the probability of chromosomal rearrangement would be. As can be seen, chromosomes 4, 5, 12, 17 and 18 of HS and PT have more similarity than others, when considering IRs. This conforms with the results reported in [22,23], in which pericentric inversions were detected by fluorescence in situ hybridization (FISH) analysis. Also, HS.Y and PT.Y are correlated, which conforms to [24]. Additionally, we have found a high correlation between HS.MT and PT.UP (as the reference), when considering IRs.

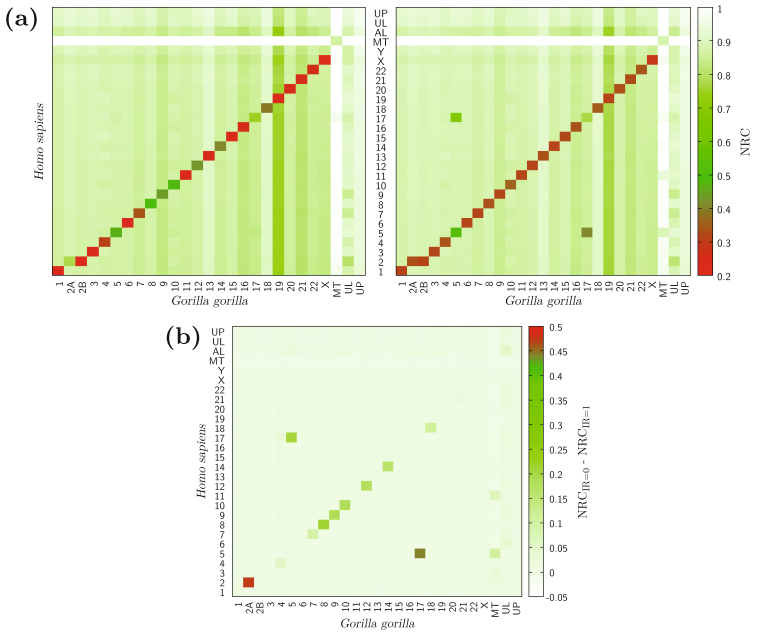


Fig. 2. NRC results concerned with compression of gorilla using human chromosomes as the reference. (a) left: IR = 0, right: IR = 1 and (b) the difference between NRCs.

The NRC results regarding compression of gorilla chromosomes, using human as the reference, are shown in Fig. 2. Considering IRs, a similarity between GG.17 and HS.5 is seen, which is justified by a chromosomal translocation [9,16,25]. Moreover, GG.2B and HS.2 are similar, either with or without considering IRs. Surprisingly, however, there is a remarkable difference between considering and not considering IRs in the compression of GG.2A using HS.2 as reference. Thus, these two chromosomes are similar, only when IRs are considered.

In Fig. 3, the NRC results regarding the compression of chicken and turkey chromosomes are plotted. Many different similarities, such as in chromosomes 3 & 2, 6 & 8, 8 & 10, 11 & 13 and 18 & 20 of GGA and MGA is seen, which were

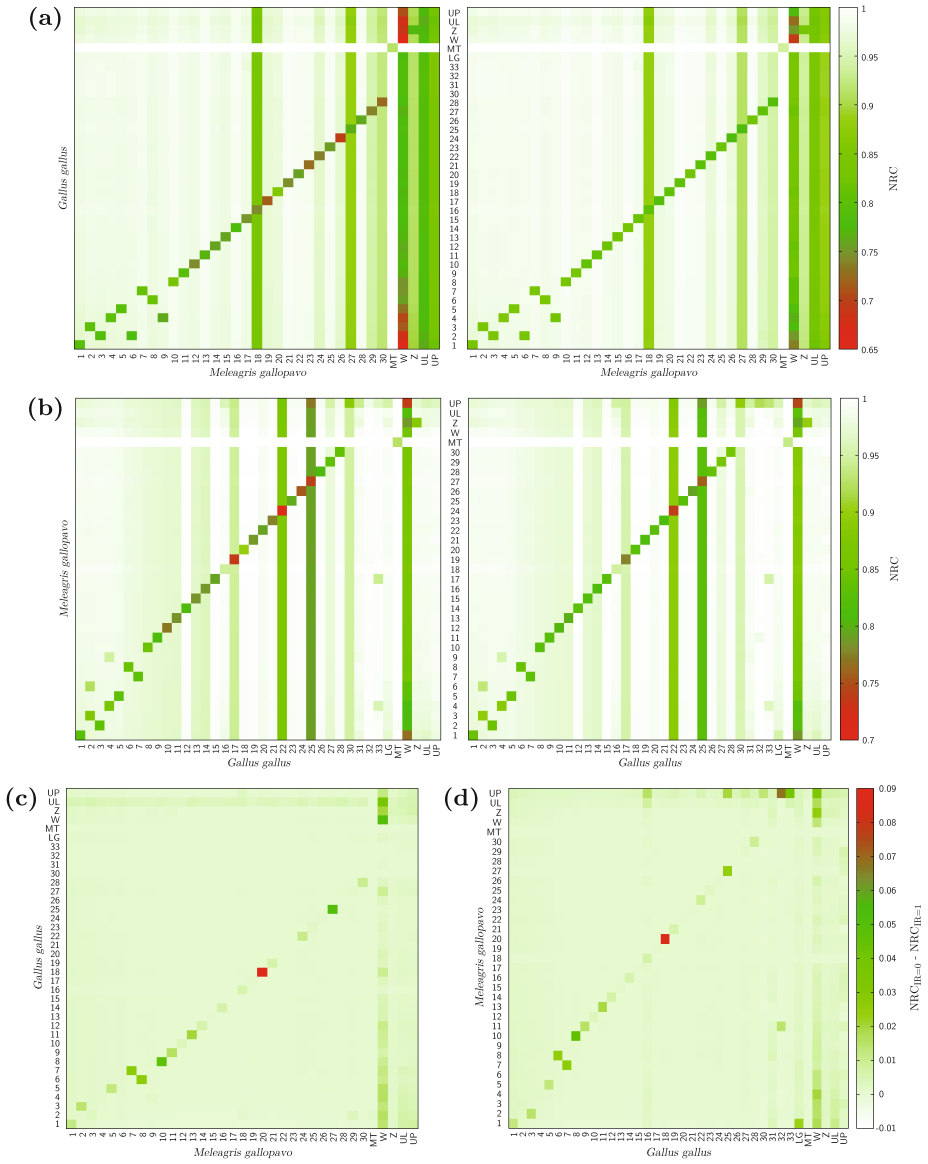


Fig. 3. NRC values associated with the compression of chicken and turkey chromosomes. (a) left: IR = 0, right: IR = 1 [ref.: GGA, tar.: MGA], (b) left: IR = 0, right: IR = 1 [ref.: MGA, tar.: GGA], (c) the difference between NRCs [ref.: GGA, tar.: MGA] and (d) the difference between NRCs [ref.: MGA, tar.: GGA].

reported in [26] as chromosomal rearrangements. Moreover, we found similarities between MGA.W and GGA.1 & GGA.UL as well as GGA.W and MGA.1 & MGA.Z & MGA.UP. Additionally, we found that MGA.27 and GGA.25, as well as GGA.32 and MGA.UP, are highly similar, when IRs are considered.

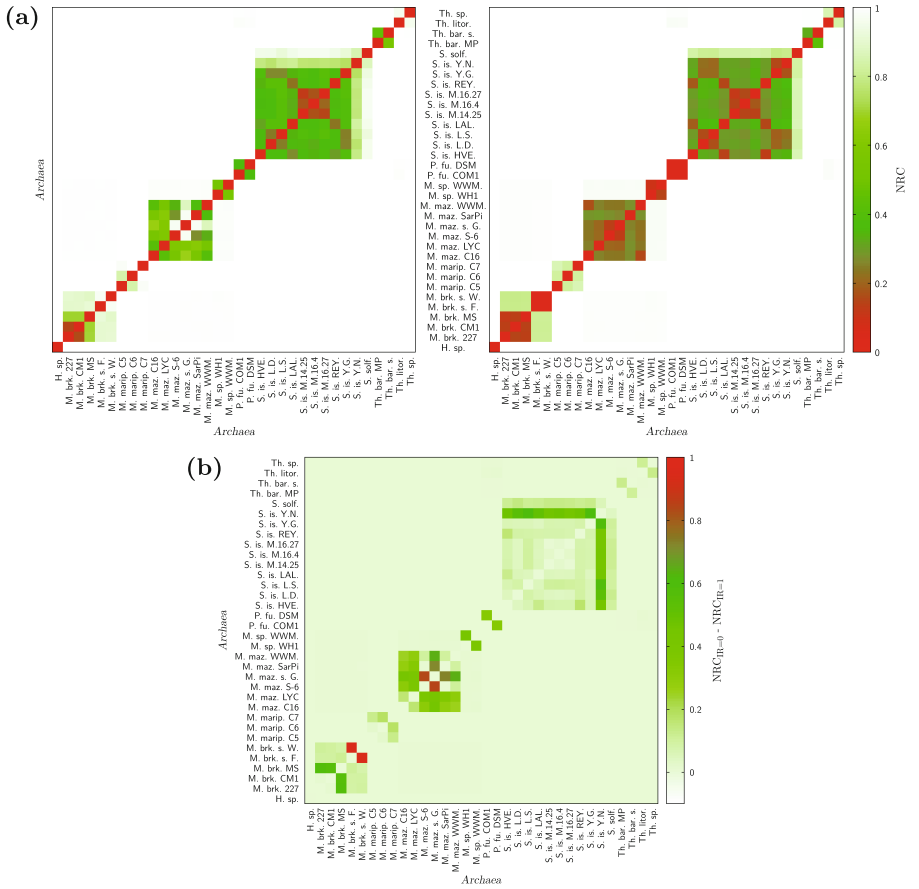


Fig. 4. NRC values obtained by compression of archaea using archaea chromosomes as the reference. (a) left: IR = 0, right: IR = 1 and (b) the difference between NRCs.

Figure 4 shows the results associated with the compression of archaea chromosomes, using archaea as the reference. The highest similarities are seen between different strains of the same archaeon, for example, *227*, *CM1* and *MS* strains regarding *Methanosarcina barkeri* archaeon. We have found remarkable differences between considering and not considering IRs in the compression of the followings: for *Methanosarcina barkeri* archaeon, *Fusaro* & *Wiesmoor*, *MS* & *227* and *MS* & *CM1*, for *Methanosarcina mazei* archaeon, *S-6* & *Goe1*, *Goe1* & *SarPi* and *Goe1* & *WWM610*, for *Sulfolobus islandicus* archaeon, *L.D.8.5* & *Y.N.15.51*, *L.S.2.15* & *Y.N.15.51* and *Y.G.57.14* & *Y.N.15.51*. It is worth mentioning that because of high similarities between different strains of the same archaeon, the plots in Fig. 4 are approximately symmetric.

4 Conclusions

One of properties of genomic sequences is the presence of inverted repeats, which are reversed and complemented copies of some sub-sequences of the genomic sequences. It is known that they may play an important role in certain genomic rearrangements. To quantify IRs, the similarity between genomic sequences needs to be determined. A common biological approach, FISH, is a time-consuming and expensive method. In this paper, we employed an affordable computational approach, relying on reference-based relative compression, to quantify IRs.

We tested the proposed approach on several genomic datasets from various species (14 GB of data). We found some sequences are more similar to each other when IRs are considered in the compression process. This raises the possibility to detect chromosomal rearrangements. The results obtained conform to the results that were attained in previous works, using the expensive FISH approach as well as computational approaches, but we also unveiled undocumented ones.

Acknowledgments. We would like to thank the FCT—Foundation for Science and Technology in Portugal, for their support of this research, within the Doctoral Programme FCT MAP-i in Computer Science, and also acknowledge european funds through FEDER, under the COMPETE 2020 and Portugal 2020 programs, in the context of the projects UID/CEC/00127/2013 and PTDC/EEI-SII/6608/2014.

References

1. Kahn, S.: On the future of genomic data. Science **331**, 728–729 (2011)
2. Alberti, C., et al.: Investigation on genomic information compression and storage. ISO/IEC JTC 1/SC 29/WG 11 N15346, pp. 1–28 (2015)
3. Giancarlo, R., et al.: Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies. Briefings Bioinform. **15**, 390–406 (2014)
4. Hosseini, M., et al.: A survey on data compression methods for biological sequences. Information **7**, 56 (2016)
5. Lesk, A.: Introduction to Bioinformatics. Oxford University Press, Oxford (2013)
6. Pinho, A.J., et al.: Inverted-repeats-aware finite-context models for DNA coding. In: 2008 16th European Signal Processing Conference, pp. 1–5 (2008)
7. Lee, J., et al.: Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. PLoS One **3**(12), e4047 (2008)
8. Cilibrasi, R., Vitányi, P.M.B.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005)
9. Pratas, D., Pinho, A.J.: A conditional compression distance that unveils insights of the genomic evolution. In: Data Compression Conference, p. 421 (2014)
10. Nikvand, N., Wang, Z.: Generic image similarity based on Kolmogorov complexity. In: IEEE International Conference on Image Processing, pp. 309–312 (2010)
11. Pinho, A.J., et al.: Authorship attribution using relative compression. In: Data Compression Conference, pp. 329–338 (2016)
12. Kolmogorov, A.: Three approaches to the quantitative definition of information. Probl. Inf. Transm. **1**(1), 1–7 (1965)

13. Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and Its Applications, 3rd edn. Springer, New York (2009)
14. Sayood, K.: Introduction to Data Compression, 4th edn. Morgan Kaufmann, Waltham (2012)
15. Pinho, A.J., et al.: Information profiles for DNA pattern discovery. In: Data Compression Conference, p. 420 (2014)
16. Pratas, D., et al.: An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Sci. Rep. **5**, 10203 (2015)
17. Pinho, A.J., et al.: On the representability of complete genomes by multiple competing finite-context (Markov) models. PloS One **6**, e21588 (2011)
18. Hosseini, M.: 21 March 2017. github.com/smortezah/Phoenix
19. Pratas, D.: 21 March 2017. github.com/pratas/goose
20. Ijdo, J., et al.: Origin of human chromosome 2: an ancestral telomere-telomere fusion. PNAS **88**, 9051–9055 (1991)
21. Hughes, J., et al.: Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature **463**(7280), 536–539 (2010)
22. Kehrer-Sawatzki, H., et al.: Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (Pan troglodytes). Hum. Mutat. **25**(1), 45–55 (2005)
23. Mikkelsen, T.S.: Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**, 69–87 (2005)
24. Bachtrog, D.: Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. **14**(2), 113–124 (2013)
25. Samonte, R.V., Eichler, E.E.: Segmental duplications and the evolution of the primate genome. Nat. Rev. Genet. **3**(1), 65–72 (2002)
26. Dalloul, R.A., et al.: Multi-platform next-generation sequencing of the domestic Turkey (Meleagris gallopavo): genome assembly and analysis. PLoS Biol. **8**(9), e1000475 (2010)

An Ensemble Approach for Gene Selection in Gene Expression Data

José A. Castellanos-Garzón[1,2(✉)], Juan Ramos[1], Daniel López-Sánchez[1], and Juan F. de Paz[1]

[1] IBSAL/BISITE Research Group, Edificio I+D+i USAL, University of Salamanca, C/Espejo s/n, 37007 Salamanca, Spain
{jantonio,juanrg,lope,fcofds}@usal.es
[2] CISUC, ECOS Research Group, Pólo II - Pinhal de Marrocos, University of Coimbra, 3030-290 Coimbra, Portugal

Abstract. Feature/Gene selection is a major research area in the study of gene expression data, generally dealing with classification tasks of diseases or subtype of diseases and identification of biomarkers related to a type of disease. In such a context, this paper proposes an ensemble approach of gene selection for classification tasks from gene expression datasets. This proposal provides a four-staged approach of gene filtering. Each stage performs a different gene filtering task, such as: data processing, noise removing, gene selection ensemble and application of wrapper methods to reach the end result, a small subset of informative genes. Our proposal has been assessed on two different datasets of the same disease (Pancreatic ductal adenocarcinoma) for which, good results have been achieved in comparison with other gene selection methods. Hence, the proposed strategy has proven its reliability with respect to other approaches.

Keywords: DNA-microarray · Gene expression data · Feature/Gene selection · Ensemble method · Wrapper method · Filter method

1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most aggressive types of cancer [1], with a five-year survival rate of 8% [2]. It is usually asymptomatic in early stages which makes early detection difficult and contributes to low survival. In addition, the chemotherapeutic drugs available are not very effective in PDAC. The latter has been associated with the dynamic relation between the tumor cells and the stroma [3]. PDAC originates as a consequence of the successive accumulation of mutations which affect different oncogenes and tumor suppressors. These genes usually play an important role in key signaling pathways. Among them are RAS, AKT, CDKN2A, TP53, DPC4, among others affected by punctual mutations or allelic loss [4,5].

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_29

Another particular feature of PDAC is its resistance to drugs, this is because of the desmoplastic stroma which constitutes a protective barrier against drugs. Today, many therapies focus on targeting the stroma, but there is little improvement in overall survival and high toxicity associated. Moreover, the complete ablation of the stroma enhances tumor progression, so current research aims to develop a stromal targeted therapy which aims to maintain an equilibrium between stroma abundance and complete depletion [2]. Regarding diagnosis, most common methods such as tomography, magnetic resonance image (MRI) and endoscopic ultrasonography (EUS) are not able to detect injuries produced by PDAC in early stages. Due to this limitation, research on diagnosis and PDAC treatment is mainly focused on the identification of new biomarkers based on differential expression analysis. Thus, the search for biomarkers is critical for the early and accurate diagnosis of PDAC, aiming to increase lifespan.

Feature/Gene selection has received a lot of attention in Bioinformatics, and many approaches for reducing dimensionality and selecting biomarkers have been proposed [6–8]. However, the wide range of existing techniques has resulted in different results, making it difficult to apply the gained knowledge to clinical practice. Gene selection methods have been divided into four categories: *filters, wrappers, embedded and ensemble* [6,7,9]. Filter methods determine the relevance of features by ranking them on the basis of statistical criteria whereas wrappers use a classifier to determine feature sets with high discrimination power. Similar to wrappers, embedded methods are based on learning methods but allowing to interact with them, which decreases the runtime taken by wrappers. Meanwhile, ensembles are the most recent among feature selection methods and merge different strategies to face instability problems presented by other methods due to data perturbations.

In consequence with the above, this paper proposes a hybrid technique operating as an ensemble approach for gene selection. The goal is to select biomarkers for PDAC diagnosis (classification tasks) by combining results from different gene selection methods to face information loss and provide a unified and coherent biomarker subset. This challenge is justified since obtaining a set of informative genes from PDAC is a complex task because PDAC is a particularly unstable and variable cancer from the genomic point of view [10].

2 An Ensemble Approach for Gene Selection

This section explains the main features of our gene selection approach, which consists of four linked stages, each developing a different gene filtering process until reaching an informative gene subset. In general sense, the first stage (Stage-I) prepares the data for the following stages, while Stage-II is responsible for removing noise presented in the data (genes considered noise). Meanwhile, Stage-III represents an ensemble of different gene selection methods applied to the input dataset. The result of this stage, a gene set, is passed to Stage-IV, which carries out a wrapper-based gene filtering to achieve an informative gene subset as an end result. The four stages above have been displayed in the flowchart given in

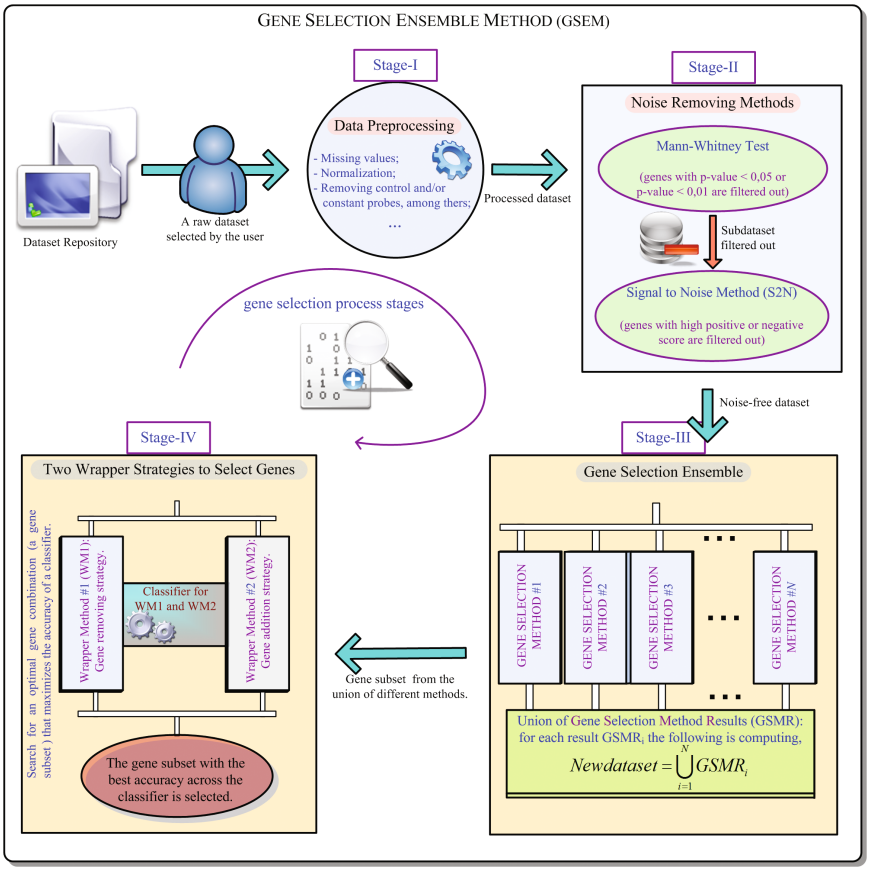


Fig. 1. Flowchart representing different stages of the gene selection process of the GSEM method: data preprocessing, noise removing, gene selection ensemble and finally, gene filtering with two wrapper strategies.

Fig. 1. In the following subsections, we are going to describe in detail each stage shown in this figure, which builds the proposed method.

2.1 Data Preprocessing: Stage-I

In this stage, a raw dataset is given as input by the user for its processing. This implies that several processes such as, data transformation, missing value estimation and data cleaning will be run if needed. Thus, this stage is in charge of preparing data for the next stages, which are that actually perform the filtering process. At the end of this stage, a new dataset is returned to Stage-II.

2.2 Noise Removing Methods: Stage-II

As its name suggests, this stage is responsible for removing noise in the data. This process involves two gene filter methods to reach such a propose. By applying the Mann-Whitney test to the input dataset as the first filter method, we will have a gene significance test, relating genes to the studied disease. Mann-Whitney is a widely used test in differential gene expression analysis [11]. Besides that, this test is nonparametric and states a null hypothesis by relating samples to the same population whereas the alternative hypothesis relates samples to different populations [12]. Thus, once this test is applied, genes with *p-value* under 0.01 are filtered out towards the next filter method, S2N. Note that such genes are who reject the null hypothesis and in consequence, they have the greatest statistical significance.

On the other hand, S2N (Signal-to-Noise, [13,14]) performs a second noise filtering from the input data and computes the statistic that determines the correlation of each gene with respect to both tissue sample classes given in the dataset. Thus, the most positive values are more correlated with the positive class whereas the most negative values are more correlated with the negative class. Hence, a determined number of genes is selected for each class based on a threshold and finally passed to the next stage. Once both methods have been applied, the resulting dataset is assumed as noise-free and the gene selection processes can be run.

2.3 Gene Selection Ensemble: Stage-III

This stage acts as an ensemble of gene selection methods by combining solutions of different methods in a single gene set. The idea consists of individually applying each gene selection method and merge their results by running the union operation (in mathematical terms) between them. Therefore, the gene set resulting from this operation (which we call *Union-set*) will have all genes found by each of different applied methods. Hence, it would be desirable to find a gene combination from such a Union-set, being representative for the remaining genes and optimizing the classification process of the study disease. Another important factor in this stage is that new gene selection methods can be included to the list of existing ones to improve the results.

Once Union-set has been achieved, it is necessary to run some strategy able to find a small gene subset from Union-set whose genes maximize the accuracy of the classifier used to identify tissue samples from the input dataset. This is the goal of the following stage.

2.4 Two Wrapper Strategies: Stage-IV

This stage is in charge of finding a small gene subset from Union-set, whose genes maximize the accuracy of a determined classifier. To deal with this problem, we have developed two greedy strategies acting as wrapper methods, which involve, on the one hand, a gene removing strategy and on the other hand, a

gene addition strategy. Both strategies share the same classifier to maximize its accuracy and the strategy whose gene subset reaches the best accuracy across the presented classifier is the winner. The gene subset of the winner strategy will be the subset of informative genes end. The operation mode developed by both strategies (which we call WM1 and WM2) is presented as follows:

- *Gene removing strategy WM1:* This strategy takes as input the Union-set set and a classifier. In each step, it deletes a gene from Union-set to evaluate the accuracy of the remaining genes. If the accuracy of the classifier is greater than or equal to the accuracy of the previous Union-set, then such a deleted gene is not significant for classification and it is permanently removed from Union-set. The new Union-set replaces the previous one. The process is repeated for the resulting Union-set by selecting (deleting) a new gene until all genes have been selected. Note that if a deleted gene decreases the accuracy of the classifier, then it is returned back to the set (because it is important for the classifier) to select another gene. As a final result, a small gene subset where no gene can be removed is returned.
- *Gene addition strategy WM2:* The Union-set set and a classifier are also the input to this method. The strategy applied in this method performs in reverse sense to WM1. It starts from choosing a single gene from Union-set in such a way that maximizes the accuracy of the input classifier. Such a gene is added to an empty set (which we call NG) and removed from Union-set. The process above is repeated by adding another gene from the remaining genes of Union-set to NG in such a way that, the accuracy of the new NG is greater than the accuracy of the previous NG across the classifier. The strategy above continues until no more genes can be added to NG, i.e., any other gene added to NG decreases the accuracy of the input classifier. Finally, NG is returned as an informative gene subset.

3 Case Study

This case study outlines two Pancreas datasets (Pancreatic ductal adenocarcinoma, PDAC), which we call PDAC#1 and PDAC#2. Our proposal is applied to both datasets and the results are compared with respect to the individual ones reached by each gene selection method used in the ensemble. The goal of this case study is to evaluate the significance of the genes found by our proposal in classification tasks from PDAC#1 and extend such results to another dataset of the same disease (in this case, PDAC#2) in order to assess the generality of the results of the current approach. The latter deals with evaluating in PDAC#2, the same genes discovered by GSEM (Fig. 1) from PDAC#1. We want to assess how the accuracy of the such a gene subset chosen in PDAC#2 is decreased, since PDAC#2 presents features very different to PDAC#1, as will be seen later. The same process above with GSEM is also applied to each gene selection method used in GSEM and the results on PDAC#2 are compared with respect to GSEM.

3.1 Datasets

As previously explained, two datasets of Pancreatic ductal adenocarcinoma (PDAC) have been used in the experiments. The first dataset, PDAC#1, comes from the NCBI public repository and available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15471. PDAC#1 is a 39-paired tumor and non-tumor tissue samples dataset, for a total of 78 samples against 54675 gene probes. The second dataset is PDAC#2, which has been selected from the same chip model as PDAC#1 but with different features since it comes from a different source. PDAC#2 consists of 25 tumor tissue samples and 7 normal tissue samples, for a total of 32 samples against 54675 gene probes. This dataset also comes from the NCBI public repository at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32676.

3.2 Results on PDAC#1

Once the case study has been described, we are going to show the results reached in each stage after applying the approach given in Fig. 1, that is:

1. *Stage-I:* After applying the data processing to the raw dataset with 54675 gene probes given as input, a new dataset is returned with 54623 probes.
2. *Stage-II:* After applying the Mann-whitney test, ranking the dataset in ascending order for the p-value of each gene probe and selecting those gene probes whose p-value < 0.01, we have achieved an output dataset with 36751 probes to be passed to the next filter method, S2N. This method assigns significance values related to the class (positive or negative class) of each gene. From the value range of each class the middle-point is computed and then, genes with both, the most positive and negative values above the middle-point in each class are chosen. In this case, a new subset with 1094 probes is taken out.
3. *Stage-III:* This stage applies a set of gene selection methods to the input data (1094 probes) and computes the union of the gene sets found by each method to form Union-set. In this case, the methods used by GSEM to select gene subsets are: kofnGA [15], Boruta [16], propOverlap [17], SDA [18], Spikeslab [19] and SubLasso [20]. The individual results of this methods on PDAC#1 have been listed in Part-A of Table 1. This table shows the name of the method applied, number of genes found and the accuracy evaluated on a Support Vector Machine (SVM) as the classifier (*stratified 10-fold cross-validation* has been applied). Then, after applying the union of results, a new dataset with 100 gene probes is obtained and identified as Union-set. Part-B in Table 1 shows the accuracy reached by Union-set through the SVM classifier.
4. *Stage-IV:* This stage uses a SVM classifier for the two defined wrapper methods in order to reduce the number of genes given in the input dataset (with 100 gene probes) and increase the accuracy of this classifier. At the end of this stage, the selected gene subset is one whose wrapper method reaches the best accuracy. Part-B in Table 1 lists the results achieved by WM1 and WM2

Table 1. Comparative table of the gene selection methods applied to PDAC#1 by the GSEM method. Part-A shows, the used methods with the number of genes found and its accuracy for a SVM classifier. Part-B shows the results reached by stages III and IV of GSEM.

Part	Method	Number of genes	SVM-Accuracy (%)
A	KofnGA	20	94.87
	Boruta	27	91.02
	propOverlap	30	92.31
	SDA	40	93.59
	Spikeslab	23	92.31
	SubLasso	11	96.15
	Union-set	100	93.59
B	WM1	6	94.87
	WM2	3	96.15

in this stage. As shown, WM2 and SubLasso reached the best accuracy. However, WM2 achieved the best accuracy with only 3 genes whereas SubLasso made it with 11 genes. Thus, our strategy is able to find a minimal gene subset from the results of other methods by increasing the accuracy of classification. Hence, GSEM improves the results of other methods. Reinforcing everything explained here, Fig. 2 presents a global view of the whole gene filtering process involved in each stage of GSEM. This figure shows the reduction progress, by stages, of the gene number from PDAC#1 until reaching the final result: 3 genes through WM2 in Stage-IV.

3.3 Assessing the Results from PDAC#1 in PDAC#2

As previously explained, one of the goals of this case study is to evaluate the decrease in accuracy of the genes discovered from the PDAC#1 dataset when they are selected from the PDAC#2 dataset, which represents the same type of cancer. This test will give us an appreciation of the universality of our results. Table 2 lists the results for this test, where Part-A and Part-B show the accuracy reached by the genes selected from PDAC#2, which have been discovered from PDAC#1 in Table 1. Part-C in Table 2 shows the results reached by WM1 and WM2 when they were applied to Union-set (with 76 probes) in this table for PDAC#2. Finally, note that the number of genes given in this table for Part-A and Part-B is not the same as the one given in Table 1. This is because PDAC#2 has genes whose expression values are constant for almost all samples. Those genes have been removed from this dataset due to they do not contribute to the classification process.

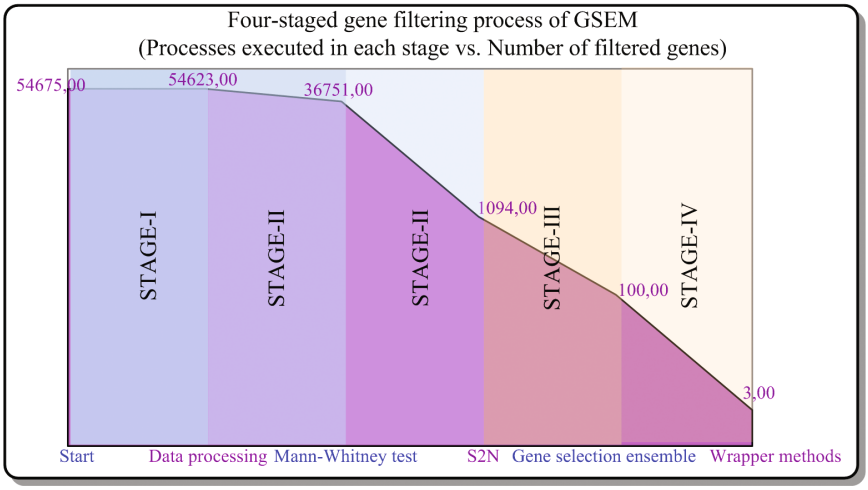


Fig. 2. Chart summarizing the gene filtering process involved in each stage and the remaining number of genes when the processes of each stage in GSEM are applied to PDAC#1.

Table 2. Comparative table of gene selection methods for PDAC#2. The genes discovered by the methods given in Table 1 for PDAC#1, have been selected from PDAC#2. Part-A and Part-B measures the accuracy of such gene sets of PDAC#2 by using a SVM classifier whereas Part-C measures the accuracy of the genes found after applying WM1 and WM2 to Union-set given in Part-B of this table.

Part	Method	Number of genes	SVM-Accuracy (%)
A	KofnGA	14	78.12
	Boruta	25	78.12
	propOverlap	28	78.12
	SDA	25	78.12
	Spikeslab	19	75
	SubLasso	6	75
B	Union-set	76	78.12
	WM1	5	75
	WM2	1	78.12
C	WM1	1	78.12
	WM2	5	96.88

3.4 Result Discussion

As for the results given from PDAC#1 in Table 1, the Union-set accuracy given in Part-B performed as an intermediate value from the accuracy given by the six methods applied. Both gene subsets given by WM1 and WM2 (with 3 and 6 genes)

showed the same accuracy as those of kofnGA and SubLasso respectively. This is because they probably use a small common subset of high relevance genes for classification. That is, only few genes are used to classify, so redundancy is successfully removed from the data by our proposal, which has found 3 genes with high accuracy.

As expected in the results listed from PDAC#2 in Table 2 (Part-A and Part-B), all methods suffered a decrease of their accuracy. This should be related to that on one hand, the classes of PDAC#2 are very unbalanced (25 tumor samples vs. 7 normal samples) and on the other hand, many of genes discovered by the methods given in Table 1 have been removed from the results given in Table 2, since such genes have constant values in their expression levels for PDAC#2. Then, taking into account the problems above, we have that the accuracy reached in Part-A and Part-B in Table 2 is not low. Moreover, Part-C of this table shows that the result of applying WM2 to Union-set given in Part-B, improved the accuracy to 96.88% only using 5 genes. This proves that Union-set given in both tables at least contains a minimal subset of significant genes that maximizes the classification task. Such minimal subsets have been found by our proposal, GSEM.

4 Conclusions

The goal of this paper has been to provide an ensemble method for gene selection from DNA-microarray data. Our proposal has been divided into four stages which have been explained throughout of this paper. To assess the proposed approach, we have used a case study with two datasets of the same disease, pancreatic ductal adenocarcinoma (PDAC). The goal of this study has been to evaluate and compare the results of our proposal with other methods in classification tasks from one of the datasets and generalize such results to the other dataset. In consequence with the above, our approach showed its reliability with respect to the other methods and that its results can be extended to other datasets of the same disease. Finally, by way of future work, our approach will be tested on RNA-seq expression data in addition to analyzing the pathway context of the selected genes.

Acknowledgments. This work has been supported by project MOVIURBAN: Máquina social para la gestión sostenible de ciudades inteligentes: movilidad urbana, datos abiertos, sensores móviles. SA070U 16. Project co-financed with Junta Castilla y León, Consejería de Educación and FEDER funds.

The research of Daniel López-Sánchez has been financed by the Ministry of Education, Culture and Sports of the Spanish Government (University Faculty Training (FPU) program, reference number FPU15/02339).

References

1. Badea, L., Herlea, V., Olimpia, S., Dumitrascu, T., Popescu, I.: Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. Hepatogastroenterology **88**, 2015–2026 (2008)
2. Kota, J., Hancock, J., Kwon, J., Korc, M.: Pancreatic cancer: stroma and its current and emerging targeted therapies. Cancer Lett. **391**, 38–49 (2017)
3. Bhaw-Luximon, A., Jhurry, D.: New avenues for improving pancreatic ductal adenocarcinoma (pdac) treatment: selective stroma depletion combined with nano drug delivery. Cancer Lett. **369**(2), 266–273 (2015)
4. Korc, M.: Pancreatic cancer-associated stroma production. Am. J. Surg. **194**(4), S84–S86 (2007). Elsevier
5. Hidalgo, M., Cascinu, S., Kleeff, J., Labianca, R., Löhr, J.M., Neoptolemos, J., Real, F.X., Van Laethem, J.L., Heinemann, V.: Addressing the challenges of pancreatic cancer: future directions for improving outcomes. Pancreatology **15**(1), 8–18 (2015). Elsevier
6. Natarajan, A., Ravi, T.: A survey on gene feature selection using microarray data for cancer classification. Int. J. Comput. Sci. Commun. (IJCSC) **5**(1), 126–129 (2014)
7. Shraddha, S., Anuradha, N., Swapnil, S.: Feature selection techniques and microarray data: a survey. Int. J. Emerg. Technol. Adv. Eng. **4**(1), 179–183 (2014)
8. Tyagi, V., Mishra, A.: A survey on different feature selection methods for microarray data analysis. Int. J. Comput. Appl. **67**(16), 36–40 (2013)
9. Castellanos-Garzón, J.A., Ramos, J.: A gene selection approach based on clustering for classification tasks in colon cancer. Adv. Distrib. Comput. Artif. Intell. J. (ADCAIJ) **4**(3), 1–10 (2015). http://dx.doi.org/10.14201/ADCAIJ201543110
10. Hezel, A., Kimmelman, A., Stanger, B., Bardeesy, N., DePinho, R.: Genetics and biology of pancreatic ductal adenocarcinoma. Genes & Dev. **20**, 1218–1249 (2006)
11. Fang, Z., Du, R., Cui, X.: Uniform approximation is more appropriate for wilcoxon rank-sum test in gene set analysis. PLoS ONE **7**(2), e31505 (2012)
12. Weiss, P.: Applications of generating functions in nonparametric tests. Math. J. **9**(4), 803–823 (2005)
13. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., deSchaetzen, V., Duque, R., Bersini, H., Nowé, A.: A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**(4) 1106–1118 (2012)
14. Berrar, D.P., Dubitzky, W., Granzow, M.: A Practical Approach to Microarray Data Analysis. Kluwer Academic Publishers, New York (2003)
15. Wolters, M.: A genetic algorithm for selection of fixed-size subsets with application to design problems. J. Stat. Softw. **68**(1), 1–18 (2015)
16. Kursa, M., Rudnicki, W.: Feature selection with the Boruta package. J. Stat. Softw. **36**(11), 1–13 (2010)
17. Mahmoud, O., Harrison, A., Perperoglou, A., Gul, A., Khan, Z., Metodiev, M., Lausen, B.: A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. BMC Bioinform. **15**(274), 1–20 (2014)

18. Ahdesmaki, A., Strimmer, K.: Feature selection in omics prediction problems using CAT scores and false non-discovery rate control. Ann. Appl. Stat. **4**, 503–519 (2010)
19. Ishwaran, H., Rao, J.: Spike and slab variable selection: frequentist and bayesian strategies. Ann. Stat. **33**(2), 730–773 (2005)
20. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear modelsvia coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2008). http://www.stanford.edu/~hastie/Papers/glmnet.pdf

Dissimilar Symmetric Word Pairs in the Human Genome

Ana Helena Tavares[1(✉)], Jakob Raymaekers[2], Peter J. Rousseeuw[2], Raquel M. Silva[3,4], Carlos A.C. Bastos[4,5], Armando Pinho[4,5], Paula Brito[6], and Vera Afreixo[1,3,4]

[1] Department of Mathematics and CIDMA, University of Aveiro, Aveiro, Portugal
ahtavares@ua.pt
[2] Department of Mathematics, KU Leuven, Leuven, Belgium
[3] Department of Medical Sciences and iBiMED, University of Aveiro, Aveiro, Portugal
[4] IEETA, University of Aveiro, Aveiro, Portugal
[5] DETI, University of Aveiro, Aveiro, Portugal
[6] FEP and LIAAD - INESC TEC, University of Porto, Porto, Portugal

Abstract. In this work we explore the dissimilarity between symmetric word pairs, by comparing the inter-word distance distribution of a word to that of its reversed complement. We propose a new measure of dissimilarity between such distributions. Since symmetric pairs with different patterns could point to evolutionary features, we search for the pairs with the most dissimilar behaviour. We focus our study on the complete human genome and its repeat-masked version.

Keywords: Inter-word distance · Reversed complements · Dissimilarity measure · Human genome

1 Introduction

Chargaff's second parity rule states that within a single strand of DNA the number of complementary nucleotides is similar [6]. An extension of this rule says that the frequency of an oligonucleotide should be similar to that of its reversed complement (the word obtained by reversing its letters and interchanging A–T and C–G). This phenomenon is known as single strand symmetry. Several authors discuss the prevalence of Chargaff's second parity rule (e.g., [1–4]). Various lines of research are being explored in an attempt to explain its cause. One approach postulates that the phenomenon would be an original feature of the primordial genome, the most primitive nucleic acid genome, and the maintenance of strand symmetry would rely on evolution mechanisms [11].

The similarity between the number of occurrences of symmetric word pairs in one strand of DNA can be verified using frequency analysis. However, two words with the same frequency in a sequence may exhibit very distinct distributions

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_30

along that sequence. This leads to the natural question how both words are distributed along the DNA sequence. Are their distributions similar?

If we constrain a random generator of sequences to respect single strand symmetry (e.g., using a high-order Markov process), it is expected that the distance distribution of a word be similar to that of its reversed complement. A reasonable hypothesis is that the distance distribution of symmetric pairs should usually be similar, and that strong deviations may have a biological origin.

As the word length increases, more unexpected patterns may be observed in the inter-word distance distributions, which may result in increased dissimilarity between symmetric pairs. The similarity between distance distributions of symmetric word pairs of length $k \leq 5$ was studied in [10]. For such short words the dissimilarity between symmetric pairs was basically negligible.

This work focuses on the dissimilarity between distance distributions of symmetric pairs of length $k = 7$ in the human genome. We propose a new dissimilarity measure between such distributions, based on the gap between the locations of their peaks and the difference between the sizes of these peaks.

The paper is organized as follows. In Sect. 2 we introduce a new dissimilarity measure between distributions based on their peaks. Section 3 then identifies and investigates the symmetric word pairs with the most dissimilar distance distributions, and Sect. 4 concludes.

2 Methods

2.1 Materials

In this study, we used the complete genome assembly (GRCh38.p2) downloaded from the website of the National Center for Biotechnology Information. We also investigate how well our results hold up in a masked sequence, which excludes major known classes of repeats [8]. We used pre-masked data, available from UCSG Genome Browser (http://genome.ucsc.edu/index.html), in which the repeats determined by Repeat Masker [9] and Tandem Repeats Finder [5] are replaced by N's. The chromosomes were processed as separate sequences and non-ACGT symbols were used as sequence separators. Distance distributions of words were generated using the C language. We programmed the new dissimilarity measures (1) to (3) in the R language used for our statistical analysis.

2.2 Inter-word Distance Distribution

A genomic word (or oligonucleotide) w is a subsequence in the nucleotide alphabet $\mathcal{A} = \{A, C, G, T\}$. Words of length k are elements of $\mathcal{A}^k$. The inter-word distance sequences are defined as the lags between the positions of the first symbol of consecutive occurrences of that word. For instance, in the DNA segment $A\underline{CG}T\underline{CG}AT\underline{CG}TG\underline{CG}\,\underline{CG}$, the inter-$CG$ distance sequence is $(3, 5, 4, 2)$.

The inter-w distance distribution (or simply distance distribution of w) gives the relative frequency of each inter-w distance and is denoted by f^w.

2.3 Dissimilarity Measure

The distance distributions may present several peaks, i.e., distances with frequencies much higher than the global tendency of the distribution. In general, the strongest peaks occur at short distances, whereas peaks at longer distances have lower frequencies. Looking only for the highest frequencies would not capture such local maxima. In what follows we will take that effect into account.

Identifying Peaks. To determine peaks we slide a window of fixed width h along the domain of the distribution. In each such interval of width h we average the absolute values of the differences between successive frequencies, and call the result the (average) *size* of the peak on that interval. The peak's *location* is defined as the midpoint of the interval. The strongest peak is then determined by the interval with the highest size. For the second strongest peak we only consider intervals that do not overlap with the first one, and so on.

Dissimilarity Between Peaks. To measure the dissimilarity between two peaks p_1 and p_2 of the same distribution we consider the difference between their sizes and between their locations. We will use the following measure:

$$d_1(p_1, p_2) = \left(\frac{|l_1 - l_2|}{R} + 1\right)\left(\frac{|v_1 - v_2|}{v} + 1\right) - 1 \tag{1}$$

where l_1 denotes the location and v_1 denotes the size of peak p_1 (and similarly for p_2). Note that we standardize $|l_1 - l_2|$ by the range R of the domain of the distribution, and $|v_1 - v_2|$ by the size v of its strongest peak. In general, the dissimilarity given by Eq. (1) increases with both the location difference and the size difference. If the peaks have the same location the dissimilarity is reduced to a relative size difference $|v_1 - v_2|/v \leq 1$, and if they have the same size it is reduced to a relative location difference $|l_1 - l_2|/R \leq 1$. When $p_1 = p_2$ Eq. (1) becomes 0, and in general it takes values between 0 and 3.

Now consider two different words w and $\bar{w}$ and let f^w and $f^{\bar{w}}$ be their distance distributions, defined on the same domain with length R. Let $p_i^w = (l_i, v_i)$ and $p_j^{\bar{w}} = (\bar{l_j}, \bar{v_j})$ be peaks in each. To measure the dissimilarity between these peaks we propose to use

$$d_2(p_i^w, p_j^{\bar{w}}) = \left(\frac{|l_i - \bar{l_j}|}{R} + 1\right)\left(\frac{|v_i - \bar{v_j}|}{\min\{v, \bar{v}\}} + 1\right) - 1 \tag{2}$$

where v and $\bar{v}$ are the highest peak sizes observed in each distribution. The denominator $\min\{v, \bar{v}\}$ yields a high dissimilarity when one distribution has strong peaks and the other doesn't.

Note that (2) satisfies the semimetric property: it reduces to zero when the two peaks have the same location and size, and is symmetric and non-negative. This makes it quite effective. When $f^w = f^{\bar{w}}$ it reduces to Eq. (1).

Dissimilarity Between Distributions. To measure the dissimilarity between two distributions we compare their n strongest peaks, for fixed n. We propose

$$d(f^w, f^{\bar{w}}) = \min_{\pi \in \mathcal{P}_n} \{ \sum_{i=1}^{n} d_2(p_i^w, p_{\pi(i)}^{\bar{w}}) \} \tag{3}$$

where π is a permutation of the indices $i = 1, \ldots, n$ meaning that $\pi(i)$ is the new position of the i-th element. The minimum is over the set $\mathcal{P}_n$ of all such permutations. The proposed measure (3) depends on n, the number of peaks considered, and on the bandwidth h used in the peak search. Note that (3) is a semimetric too.

3 Results and Discussion

There are 4^7=16384 distinct genomic words of length $k = 7$, corresponding to 8192 symmetric word pairs. We restrict our distance distributions from $k+1$ to 1000 (some distances from 1 to k may be absent due to the word structure). The dissimilarity measure (3) between distance distributions is computed with bandwidth $h = 5$ and the $n = 3$ strongest peaks (for $n = 4, \ldots, 7$ we obtained similar results in much higher computation time). Our bandwidth choice $h = 5$ is a compromise which combines peaks that lie close together without oversmoothing the distribution.

Some words w of length $k = 7$ have a distance distribution with low total absolute frequency S^w, so in our analysis we exclude symmetric pairs in which at least one word has S^w below the first quartile of $S = \{S^w : w \in \mathcal{A}^k\}$.

3.1 Complete Genome Assembly

In the complete genome this first quartile is 1498, so we exclude the symmetric pairs with $\min\{S^w, S^{\bar{w}}\} \leq 1498$ (see Table 1) and measure the dissimilarity (3) in the remaining 6054 symmetric pairs. Let D be the set formed by these 6054 dissimilarity values.

We then automatically select the symmetric pairs with dissimilarity under 0.129, the 10^{th} percentile of D, and those above 12.638, its 90^{th} percentile.

The symmetric pairs with low values of (3) have very similar distributions. For some words this dissimilarity is surprisingly low in spite of their distributions having some strong peaks, which are almost the same in the distribution of their reversed complement, as illustrated in Fig. 1(a)–(d). This also suggests that the dissimilarity measure (3) achieves its intended purpose.

The symmetric pairs with high dissimilarity are usually formed by one distribution with strong peak(s) and another displaying low variability or small peaks. Figures 1(e)–(f) show the distance distributions for two symmetric pairs discovered by our procedure. Especially the distance pattern of $w = CACAGGC$ is noteworthy. It shows several peaks whose size goes up, which is a very unusual behavior in distance distributions between words.

Table 1. Sum of distance frequencies S^w, their maximal ratio $\max\{S^w/S^{\bar{w}}\}$ in a symmetric pair, and dissimilarity $d(f^w, f^{\bar{w}})$ inside a symmetric pair, for the complete genome and the masked genome. Results are given for all 8192 symmetric pairs and for those with $\min\{S^w, S^{\bar{w}}\}$ above its first quartile.

	Complete sequence				Masked sequence			
	All pairs		6054 pairs		All pairs		6075 pairs	
	S^w	$\max\{\frac{S^w}{S^{\bar{w}}}\}$	$d(f^w, f^{\bar{w}})$	$\max\{\frac{S^w}{S^{\bar{w}}}\}$	S^w	$\max\{\frac{S^w}{S^{\bar{w}}}\}$	$d(f^w, f^{\bar{w}})$	$\max\{\frac{S^w}{S^{\bar{w}}}\}$
Min	10	1.000	0.003	1.000	3	1.000	0.032	1.000
Q1	1498	1.012	0.350	1.009	546	1.015	0.507	1.011
Med	11850	1.037	0.915	1.022	2771	1.039	0.832	1.026
Q3	28510	1.165	2.936	1.075	6265	1.112	1.471	1.055
Max	927376	86.74	178.7	83.29	277460	14.64	21.19	2.041

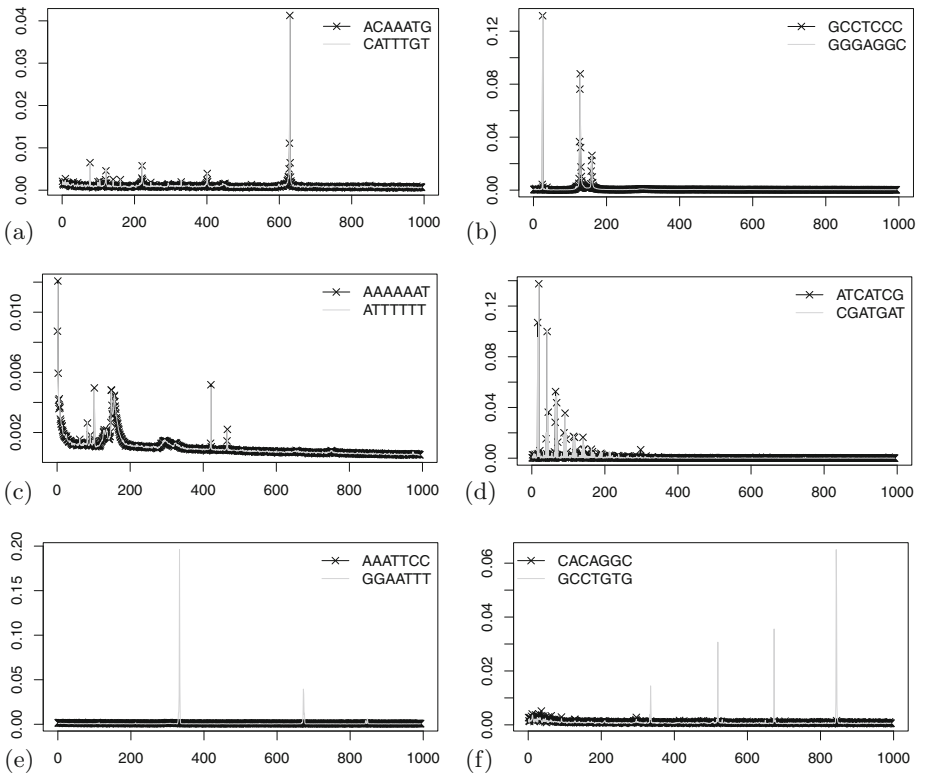


Fig. 1. Inter-word distance distributions of some reversed complements, f^w and $f^{\bar{w}}$, with low dissimilarity values: 0.036 (a), 0.003 (b), 0.058 (c), 0.116 (d); and with high dissimilarity values: 178.749 (e), 51.767 (f). Sequence: complete human genome.

3.2 Masked Genome Assembly

To reduce the effect of repetitive sequences in the original genome assembly, we also analyse a masked version of the genome. All distributions and measures in this subsection are from the masked sequence.

Masking the genome sequence affects the shape of the distance distributions. Several strong peaks observed in the complete genome are eliminated by masking. For example, the distance distribution of $w = CACAGGC$ (Fig. 1(f)) loses the four strong peaks in the masked sequence (not shown).

We repeat the previous procedure in the masked sequence, to detect symmetric pairs whose distance distributions have very similar or very dissimilar patterns. The first quartile of $S = \{S^w : w \in \mathcal{A}^k\}$ becomes 546, so we exclude the pairs for which $\min\{S^w, S^{\bar{w}}\} \leq 546$, leaving D with 6075 pairs (see Table 1).

As before we automatically select the symmetric pairs with dissimilarity below the 10^{th} percentile of D (0.328), and those with dissimilarity above the 90^{th} percentile of D (2.454). The pairs with lowest dissimilarity may be divided

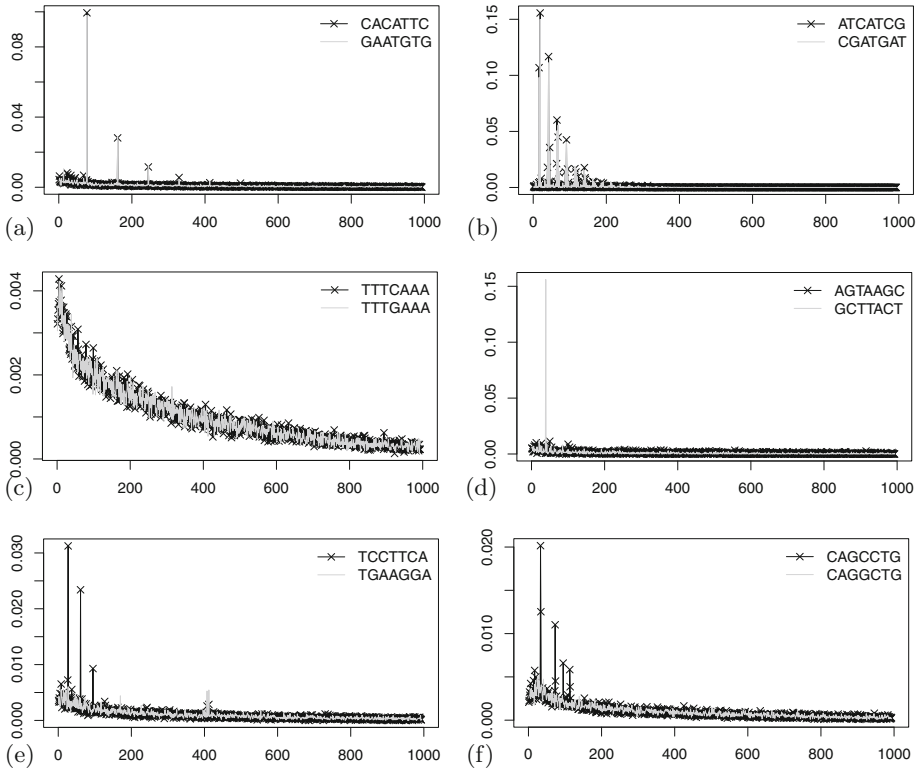


Fig. 2. Inter-word distance distributions of some reversed complements with low dissimilarity values: 0.032 (a), 0.125 (b), 0.144 (c); and with high dissimilarity values: 11.744 (d), 11.310 (e), 6.486 (f). Sequence: masked human genome.

in two groups: those for which both distributions have strong peaks at short distances, and those whose distributions look like exponential curves without strong peaks. These patterns are illustrated in Fig. 2(a)–(c). Interestingly, the unusual pattern of $w = ATCATCG$ in the complete sequence (Fig. 1(d)) remains in the masked sequence (see Fig. 2(b)).

Symmetric pairs with high dissimilarity usually have one distribution with one or more strong peaks at short distances (<200) whereas the other has low variability. Some very dissimilar pairs are shown in Fig. 2(d)–(f).

To investigate whether an association exists between dissimilar reversed complements and functional DNA elements, we perform an annotation analysis for the 15 most dissimilar symmetric pairs. For each such pair we list the word with the strongest peaks. Then we look for the 'favoured' distance(s), i.e. those where the strongest peak(s) are located. These peaks are often concentrated in one chromosome rather than being spread over the entire genome sequence. Table 2 lists the chromosome in which the favoured distances are most pronounced, for each of the 15 pairs. The positions of the words occurring at that distance from each other are recorded. Then, we retrieve annotations within these genomic coordinates from UCSC GENCODE v24 (August 2015) [7]. Interestingly, the words we obtained that are located on chromosome 13 all fall within the gene LINC01043 (long intergenic non-protein coding RNA 1043) and all of our words on chromosome 1 are contained in the gene TTC34 (tetratricopeptide repeat domain 34). These results suggest that the most dissimilar distributions may be related to repetitive regions associated with RNA or protein structure.

A deeper investigation into the biological meaning of these words is necessary to investigate whether the observed dissimilarities reflect the selective evolutionary process of the DNA sequence.

Table 2. The 15 most dissimilar symmetric pairs with $k = 7$, characterized by their word with the strongest peaks. The chromosome on which these peaks are prominent is listed. Masked sequence.

Chromosome	13		1	4	3	8
Word w	*ACCATTC*	*GGTAAGC*	*AGCATCT*	*GTTGGTA*	*TGGTATG*	*GCTTACT*
	CTTCAGG	*TAAGCAT*	*GAGCATC*	*TGGTAGA*		
	GACCATT	*TCAGGAT*	*TGAGCAT*			
	TCCTTCA	*TTCAGGA*				

4 Conclusions

We propose a new dissimilarity measure between distance distributions, based on discrepancies between their peaks. Here we use it to evaluate the dissimilarity between reversed complements.

In the complete human genome, we confirm the expected existence of many symmetric pairs with low dissimilarity, both in word frequency and in distance distribution. Even an irregular distribution with strong peaks is often very

similar to that of its reversed complement. However, our main interest lies in using the proposed dissimilarity measure to detect symmetric pairs with highly distinct distributions. In such cases, one of the distance distributions typically has well defined peaks and the other has low variability.

We also investigate how well our results hold up in the masked sequence, which excludes major known classes of repeats. Even though masking generally reduces the dissimilarity between distributions of symmetric pairs, there remain quite a few word pairs with high dissimilarity, which in our study was mainly localized on a specific chromosome and even a specific gene. A question worth investigating is to what extent the high dissimilarities may be linked to evolutionary processes, that are not the result of recent local DNA block expansions.

Acknowledgment. This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Center for Research & Development in Mathematics and Applications (CIDMA), Institute of Biomedicine (iBiMED) and Institute of Electronics and Informatics Engineering of Aveiro (IEETA), within projects UID/MAT/04106/2013, UID/BIM/04501/2013 and UID/CEC/00127/2013, and by PhD grant PD/BD/105729/2014. The research of P. Brito was financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation (COMPETE 2020) within project POCI-01-0145-FEDER-006961, and by the FCT as part of project UID/EEA/50014/2013. The research of J. Raymaekers and P. Rousseeuw was supported by projects of Internal Funds KU Leuven.

References

1. Afreixo, V., Bastos, C.A.C., Garcia, S.P., Rodrigues, J.M.O.S., Pinho, A.J., Ferreria, P.J.S.G.: The breakdown of the word symmetry in the human genome. J. Theoret. Biol. **335**, 153–159 (2013)
2. Afreixo, V., Rodrigues, J.M.O.S., Bastos, C.A.C.: Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics **16**(2), 209–221 (2015)
3. Albrecht-Buehler, G.: Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc. Natl. Acad. Sci. **103**(47), 17828–17833 (2006)
4. Baisnée, P.-F., Hampson, S., Baldi, P.: Why are complementary DNA strands symmetric? Bioinformatics **18**(8), 1021–1033 (2002)
5. Benson, G., et al.: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**(2), 573–580 (1999)
6. Forsdyke, D.R., Mortimer, J.R.: Chargaff's legacy. Gene **261**(1), 127–137 (2000)
7. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., Kent, W.J.: The UCSC Table Browser data retrieval tool. Nucleic Acids Res. **32**(suppl 1), D493–D496 (2004)
8. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.: Initial sequencing and analysis of the human genome. Nature **409**(6822), 860–921 (2001)
9. Smit, A.F.A., Hubley, R.M., Green, P.: Repeatmasker open-4.0. 2013–2015 (http://repeatmasker.org)

10. Tavares, A.H., Afreixo, V., Rodrigues, J.M.O.S., Bastos, C.A.C.: The symmetry of oligonucleotide distance distributions in the human genome. In: Proceedings of ICPRAM, vol. 2, pp. 256–263 (2015)
11. Zhang, S.-H., Huang, Y.-Z.: Strand symmetry: characteristics and origins. In: 2010 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE), pp. 1–4. IEEE (2010)

A Critical Evaluation of Automatic Atom Mapping Algorithms and Tools

Nuno Osório[1(✉)], Paulo Vilaça[1,2], and Miguel Rocha[1]

[1] Centre of Biological Engineering, University of Minho, Campus de Gualtar, Braga, Portugal
nuno.m.c.osorio@gmail.com

[2] SilicoLife-Computational Biology Solutions for the Life Sciences, Braga, Portugal

Abstract. The identification of the atoms which change their position in chemical reactions is an important knowledge within the field of Metabolic Engineering. This can lead to new advances at different levels from the reconstruction of metabolic networks to the classification of chemical reactions, through the identification of the atomic changes inside a reaction. The Atom Mapping approach was initially developed in the 1960s, but recently suffered important advances, being used in diverse biological and biotechnological studies. The main methodologies used for atom mapping are the Maximum Common Substructure and the Linear Optimization methods, which both require computational know-how and powerful resources to run the underlying tools.

In this work, we assessed a number of previously implemented atom mapping frameworks, and built a framework able of managing the different data inputs and outputs, as well as the mapping process provided by each of these third-party tools. We evaluated the admissibility of the calculated atom maps from different algorithms, also assessing if with different approaches we were capable of returning equivalent atom maps for the same chemical reaction.

Keywords: Metabolic engineering · Chemical reactions · Atom mapping algorithms · Open-source software · Maximum common structure

1 Introduction

Cell metabolism is composed of chemical reactions which are catalysed by enzymes responsible for transforming the nutrients uptaken by the cell into energy and cellular building blocks. When needed, the cell uses its anabolic pathways to produce essential macromolecules, from energy and cellular building blocks, maintaining its regular behaviour [1].

Glimpsing the cells as industrial factories, the raw materials prices persistent climbing, and the reduction of their reserves, take researchers to build models which help to understand and optimize cellular systems (such as genetically

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_31

altered microorganisms) to produce native and non-native high-value industrial compounds like biofuels, antibiotics or aminoacids [2,3]. These approaches, largely applied in industry, help Metabolic Engineering to solve problems like tracing metabolic pathways from a metabolite A to a metabolite B [4], analysing the conservation of metabolites in metabolic networks [5], calculating all possible paths inside a metabolic network, from the initial to the goal atom, classifying chemical reactions (e.g. assigning EC numbers to enzymes) [6] or identifying which atoms are preserved.

All these applications have a common approach, crucial to accomplishing their goals: in a chemical reaction, performing the matching of its reactants' and products' atoms. This correspondence, called Atom Mapping, allows a correct atom trace of the desired reaction, identifying what are the changes between the reactants and products. Atom Mapping assigns a different index (integer number) to each atom from the reactions' substrates and tries to map these atoms onto the products, thus assigning them the same index. With this information, it is possible to determine what are the changes performed by a reaction (catalysed by specific enzymes). In other words, the atom mapping procedure identifies which are the broken/formed bonds or which bond's change their order [7].

The atom mapping approach allows diverse uses and applications, for instance, in the reconstruction of metabolic networks, which represents the atom level of the pathways, it will improve understanding of the metabolic network [8]. Atom mapping can also be used to do consistency checking of pathways [4], to analyse the conservation ratios of atoms in a reaction [5] and to classify chemical reactions based on their chemical transformation [6]. Also, to optimise drug design, it is necessary to predict which atoms, from the candidate drug, change during the chemical reaction. It may also be used to deduce the relevant pathways of a certain metabolite or a particular drug [9].

With this work, we aim to study strategies to collect atom mappings from databases, by analysing reaction databases and build a framework to extract atom mapping information; analyse methods for automatic atom mapping of reactions, by automatically extract atom mappings from published atom mapping software (API's); and evaluate comparison metrics of atom mapping, namely, evaluate against atom mapping from databases and other atom mapping algorithms.

Here, the comparison of four algorithms within four different frameworks was performed to verify the differences between each other, in terms of valid and equivalent maps assignment.

2 Methods

2.1 Data

A biological database was chosen to build our set of reactions, namely MetaCyc, from where 11575 reactions were collected, in which more than 90% had an associated atom map. The set contains balanced, not balanced, incomplete and

elemental reactions, with the objective of obtaining the most complete sample possible.

2.2 Algorithms

The group of tools and algorithms selected to perform the atom mapping process will be briefly described. Note that these tools use a combination of different algorithms to obtain their results.

MetaCyc. The atom mappings collected from the MetaCyc database [10] were calculated using the Minimum Weighted Edit-Distance metric (MWED) [11]. It uses a Mixed-Integer Linear Programming (MILP) approach, that identifies which bonds have more tendency to react. MWED finds multiple optimal maps, but with the particularity of having less symmetric maps, due to the introduction of bond weights which represent the tendency of a bond to break. Within the reactions, bonds can be broken, formed or change their type (e.g. single to double). The cost of a transformation is calculated taking into account the weights assigned to the bonds involved in the bond breaking/forming/changing process. The sum of the costs of all the changes in the chemical reaction results in the weight-edit distance of the reaction. This process only handles fully balanced biochemical reactions (reactions with the same number of atoms on both sides).

AutoMapper. AutoMapper performs the atom mapping based on Maximum Common Structure (MCS) and MILP algorithms. It provides some options on the mapping style: *Complete*: where all atoms are mapped; *Changing*: as the name indicates, only maps the atoms that have their bonds modified; *Matching*: only maps the atoms which do not have any bond modified.

Reaction Decoder Tool. The Reaction Decoder Tool (RDT) [12] calculates the atom maps for balanced and unbalanced reactions using MCS and MILP algorithms. It uses the Chemistry Development Kit (CDK) [13], a cheminformatics framework which offers diverse functionalities in molecular informatics (e.g. input/output features for SMILES or RXN files, rendering chemical structures, modelling, building chemical graphs - isomorphism checker or MCS searchers, fingerprinting or Nuclear Magnetic Resonance prediction, etc.).

ICMap. ICMap maps and determines the reaction's centres based on MCS and MILP approaches. Some chemical rules are applied to help the MILP approach finding the best possible map (e.g. breaking/forming hetero-atoms bonds are preferable to carbon-carbon bonds). It has some restrictions on the mapping process: it has a limit on the number of molecules in the reaction (no more than 15 on each side), on the molecules' size (no more than 100 non-hydrogen atoms) and on single atom mapping (single atoms without non-hydrogen bonds e.g. Phosphor or Sulphur). The ICMap cannot map a reaction in which all chemical bonds were broken and remade.

2.3 AtomMapper Framework

To ensure that the four algorithms followed the same analysis pipeline, it was implemented a framework, called AtomMapper Framework (AMF). AMF is 100% developed in Java^{TM} and joins different algorithms of atom mapping into a single program. It allows users to map their chemical reactions with different approaches and verify if their atom maps are equivalent or not.

AMF is also implemented as an abstraction that provides generic functionalities, which can be specified with the addition of new code. It is an universal, reusable software environment, which facilitates the development of additional applications. AMF defines which functions the user should implement (interface classes) and releases users of thinking in low-level details. It is especially useful for users wanting to test their own tools and algorithms, once it is easy to add new methods following the existing interfaces.

Figure 1 illustrates the two main step of the atom mapping process. On A the reading process and on B the atom mapping.

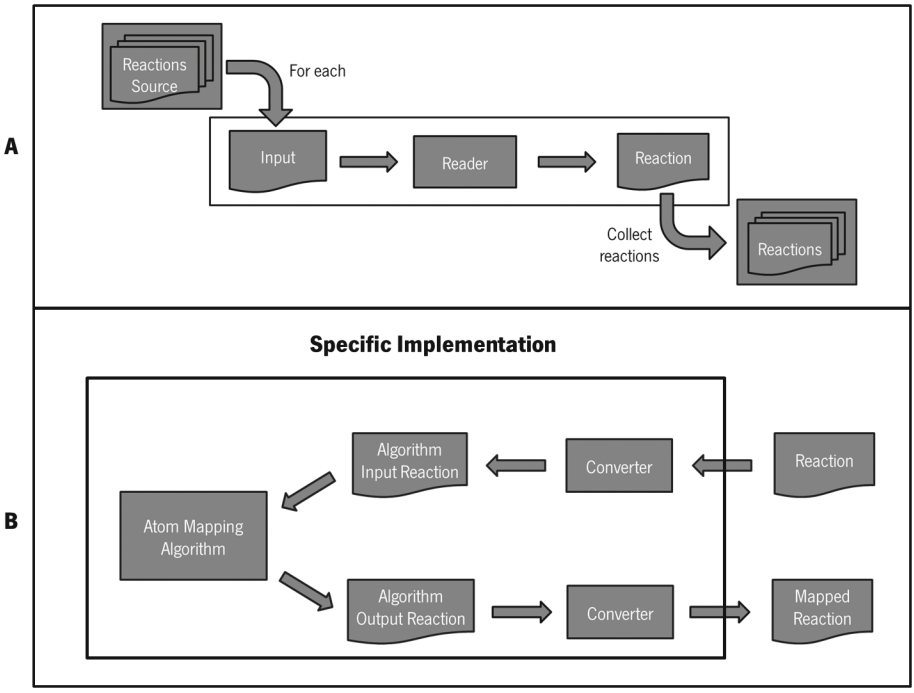


Fig. 1. Schematic representation of the AMF implementation philosophy. (A) It shows the reading of different types of input files to build a collection of reactions. (B) It represents the implementation needed to handle with each different algorithm input and output.

3 Results and Discussion

This section presents the results of the evaluation of different atom mapper algorithms. To do so, the Metacyc database was chosen as the reactions main set. It is constituted by 11575 different reactions, of which 10870 are already mapped, meaning that 705 reactions did not have a valid atom map on the Metacyc database.

It is important to differentiate a valid mapped reaction and an equivalent mapped reaction. A valid mapped reaction is a reaction where all atoms are assigned with a continuous numeration in both left and right sides, as well as both sides have the same elementary composition. An equivalent mapped reaction is a reaction for which different algorithms assigned the same atom linkage between left and right sides, i.e. all atoms in the right pair to the same atom in the left in both results, ignoring the individual numbers assigned to each atom (in one algorithm a right-left atom pair can have one label, while in the other algorithm the same pair has a different label, but they are the same pair).

The validation step will filter the reactions which have complete and plausible atom maps. This highlights the reactions for which their atom maps are comparable.

The first analysis of the mapping process was to consider the mappings provided by the four algorithms, checking the number of valid maps defined for each reaction. A total of 604 reactions were not mapped by any of the used atom mapping algorithms. This way, the number of admissible reactions decreased to 10971 valid reactions. Adding to this, the number of reactions with one or two valid maps was 1603, which is significantly lower when it is compared to the 9368 reactions with at least three valid maps. This indicates that over 80% of the reactions had three or four algorithms which were capable of assigning a valid map.

In terms of percentages, Metacyc presents 99.1%, AutoMapper 83.6%, RDT 99.8% and ICMap 40.9% of the whole set of reactions with at least one valid atom map. We can verify that the ICMap algorithm had the lowest percentage of valid maps, followed by AutoMapper algorithm, Metacyc database and RDT algorithm.

After analysing the behaviour of each individual algorithm, it was found that the MetaCyc and the RDT algorithms presented a similar number of reactions with valid maps assigned. The AutoMapper also presented a similar number, concerning the reactions with three and four valid atom maps, although, it did not have the same concordance with reactions containing one or two valid atom maps. About the ICMap, the numbers do not show very promising results, as its number of valid atom maps was less than half of the total reactions analysed and the concordance with the remaining algorithms was almost restricted to the reactions with four valid atom maps.

Figure 2A shows a Venn diagram with the intersection of the four sets of valid maps computed by each algorithm, assessing the reactions where pairs of algorithms are able to produce valid maps. Furthermore, the sum of all numbers of each oval form, gives the total number of valid reactions from each algorithm.

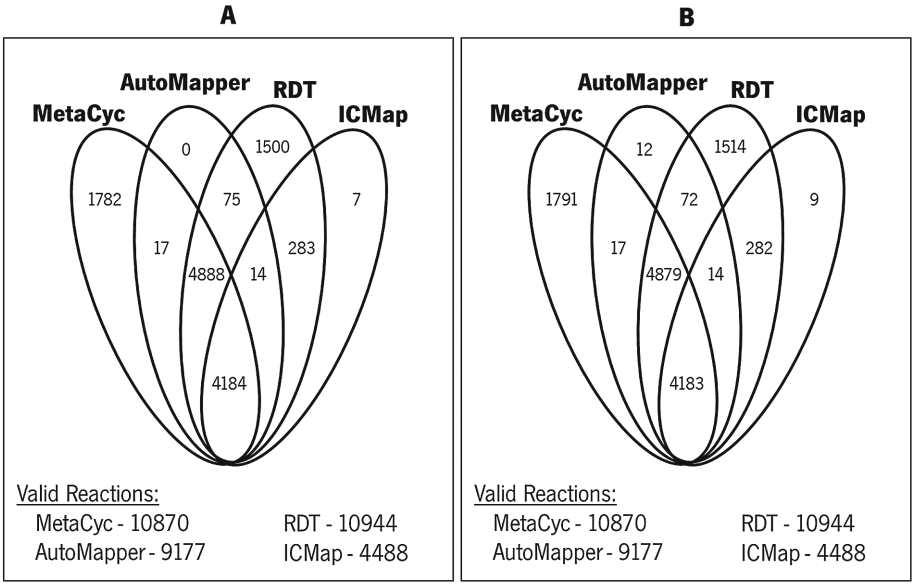


Fig. 2. Venn diagram showing the relations between the atom maps produced by the four algorithms: Metacyc, AutoMapper, RDT and ICMap sets, showing the intersection of reactions where each algorithm produced maps. (A) Counting of valid reactions, where intersections will show reactions where both algorithms produced valid maps; (B) Each intersection represents the number of reactions with equivalent atom maps assigned by the different algorithms. In both cases, if all numbers from an oval are added, it will represent the number of valid reactions on each set.

The four algorithms assign the same 4184 reactions as valid, corresponding to 38.1% of all reactions with at least one valid atom map (i.e. 10971 reactions). Nevertheless, if the ICMap algorithm is not considered in the analysis, the percentage of valid reactions raises from 38.1% to 82.7%, which represents 9072 reactions with three valid atom maps each. So, it may be admissible to say that the ICMap is pulling the number of common valid reactions down.

Having in mind that all analyses made so far do not imply that two valid maps, assigned to the same reaction, are equivalent, it is now time to check this. Considering all reactions from each set, and getting their atom maps, the comparison approach was performed to evaluate the atom maps equivalence.

Figure 2B shows the same representation from Fig. 2A, but now describing the comparison process. It represents the intersection of the four sets, and each intersection shows the number of reactions with equivalent maps between both algorithms. The intersection of the MetaCyc with the AutoMapper sets represents 9079 reactions with equivalent maps, which means 82.8%. The intersection of the AutoMapper with the RDT sets represents 9148 reactions, 83.4%, while 4479 reactions (40.8%) had equivalent atom maps calculated with the RDT and

the ICMap algorithms. Note that all percentages were calculated considering the 10971 reactions with at least one valid atom map.

When the intersection of more than two algorithms was analysed, the number of equivalent reactions tends reduce. The intersection of MetaCyc, AutoMapper and RDT represents 9062 reactions with three equivalent atom maps (82.6% of the valid reactions), still an interesting number. If it is now analysed the intersection of AutoMapper, RDT and ICMap, it joins 4197 reactions with three equivalent atom maps, with 38.3% of reactions. Finally, it was performed the intersection of all algorithms, and obtained 4183 reactions (38.1%), which were assigned with four equivalent atom maps for all analysed algorithms. Comparing the equivalence values with the ones from the validation, it is visible the high correlation between them. The ICMap was the algorithm with the lower percentage of valid atom maps. However, it was not significant in the comparison process, once it presented a similar percentage of equivalent atom maps.

Additionally, as referred before, 705 reactions did not have an atom map from the Metacyc database. Having into account that there are 604 reactions where none of the algorithms could provide a valid atom map, only 101 have the potential to have an atom map assigned by the remaining three algorithms. It was found that 14 reactions of those were assigned with four valid maps, all with four equivalent atom maps, which is a very interesting starting point to add new atom maps to the Metacyc database.

4 Conclusions

AMF enables the scientific community to explore the atom mapping process as well as, due its extensibility properties, be the base block to support additional implementation of atom mapping algorithms and comparison methods. It was shown that the studied algorithms had different behaviours: in the attribution of valid atom maps to this biological reactions set, they scaled from nearly 40% (ICMap) to almost 95% (RDT) of valid maps. However, despite this behaviour on the validation process, all algorithms, on the comparison step, had presented similar percentages of equivalent maps. Concerning the number of reactions which had four valid atom maps in the validation process, the majority had their atom maps considered equivalent, which proves the good precision of all tested algorithms. This may indicate that the atom mapping algorithms could assign different numbers to the atoms, but the matching of the left with the right reaction sides shows they are equivalent. The algorithms also had different techniques to assign the atom maps, which indicates that despite the theoretical differences, the result is somehow similar.

Acknowledgments. This study was partially supported by the Portuguese FCT under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by ERDF under the scope of Norte2020.

References

1. Heinonen, M., Lappalainen, S., Mielikäinen, T., Rousu, J.: Computing atom mappings for biochemical reactions without subgraph isomorphism. J. Comput. Biol. J. Comput. Mol. Cell Biol. **18**(1), 43–58 (2011)
2. Li, R., Townsend, C.A.: Rational strain improvement for enhanced clavulanic acid production by genetic engineering of the glycolytic pathway in Streptomyces clavuligerus. Metab. Eng. **8**, 240–252 (2006)
3. Rokem, J.S., Lantz, A.E., Nielsen, J.: Systems biology of antibiotic production by microorganisms. Nat. Prod. Rep. **24**, 1262–1287 (2007)
4. Arita, M.: Introduction to the ARM database: database on chemical transformations in metabolism for tracing pathways. In: Tomita, M., Nishioka, T. (eds.) Metabolomics, pp. 193–210. Springer, Tokyo (2005)
5. Hogiri, T., Furusawa, C., Shinfuku, Y., Ono, N., Shimizu, H.: Analysis of metabolic network based on conservation of molecular structure. Biosystems **95**, 175–178 (2009)
6. Yamanishi, Y., Hattori, M., Kotera, M., Goto, S., Kanehisa, M.: E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics **25**, i179–i186 (2009)
7. Fooshee, D., Andronico, A., Baldi, P.: ReactionMap: an efficient atom-mapping algorithm for chemical reactions. J. Chem. Inf. Model. **53**(11), 2812–2819 (2013)
8. Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., Palsson, B.O.: Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Natl. Acad. Sci. **104**, 1777–1782 (2007)
9. Blum, T., Kohlbacher, O.: Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. J. Comput. Biol. **15**, 565–576 (2008)
10. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., Karp, P.D.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. **42**, D459–471 (2014)
11. Latendresse, M., Malerich, J.P., Travers, M., Karp, P.D.: Accurate atom-mapping computation for biochemical reactions. J. Chem. Inf. Model. **52**(11), 2970–2982 (2012)
12. Rahman, S.A., Torrance, G., Baldacci, L., Cuesta, M.S., Fenninger, F., Gopal, N., Choudhary, S., May, J.W., Holliday, G.L., Steinbeck, C., Thornton, J.M.: Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics **32**, 2065–2066 (2016)
13. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E.: The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. J. Chem. Inf. Comput. Sci. **43**, 493–500 (2003)

Substitutional Tolerant Markov Models for Relative Compression of DNA Sequences

Diogo Pratas(✉), Morteza Hosseini, and Armando J. Pinho

IEETA, University of Aveiro, Aveiro, Portugal
{pratas,seyedmorteza,ap}@ua.pt

Abstract. Referential compression is one of the fundamental operations for storing and analyzing DNA data. The models that incorporate relative compression, a special case of referential compression, are being steadily improved, namely those which are based on Markov models. In this paper, we propose a new model, the substitutional tolerant Markov model (STMM), which can be used in cooperation with regular Markov models to improve compression efficiency. We assessed its impact on synthetic and real DNA sequences, showing a substantial improvement in compression, while only slightly increasing the computation time. In particular, it shows high efficiency in modeling species that have split less than 40 million years ago.

Keywords: Markov models · Tolerant Markov models · Relative compression · Genomic sequences

1 Introduction

Several applications in bioinformatics require the compression of a string, x, given other string, y. This is the case when one needs to analyze or store compactly as possible the data [1–6]. The information in y can be used together with that on x or alone. In the so called conditional approach [7,8], the compressor can explore the information that is contained in y, as well as that from x (assuming causality), according to

$$C(x|y) = \sum_{i=1}^{|x|} -\log_2 P(x_i|x_1^{i-1}, y), \tag{1}$$

where $|x|$ is the size of x and x_i is i^{th} element of x. So, for example, x_3^5 is a substring of x composed by x_3, x_4 and x_5.

The relative approach [6,9–14], $C(x\|y)$, assumes that information comes exclusively from y, according to

$$C(x\|y) = \sum_{i=1}^{|x|} -\log_2 P(x_i|x_{i-\pi}^{i-1}, y), \tag{2}$$

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_32

where $i - \pi$ is the allowed size of elements from x that can be used in order to search for regularities in y. For $i \leq \pi$ we assume a uniform distribution.

In order to calculate the probabilities of Eq. 2, we need data models that describe y efficiently. Both Ziv-Merhav dictionary-based models [9,13,15] and Markov models [5,14,16,17] have been successfully used in diverse data type applications. However, for DNA sequences, Markov models proved to be more efficient [6].

Markov models (MMs), also known as finite-context models (FCMs), are statistical models. A MM of an information source assigns probability estimates to the symbols of an alphabet, Θ, according to a conditioning context computed over a finite and fixed number, k, of past outcomes (order-k MM) [18]. At element i, these conditioning outcomes are represented by $x_{i-k+1}^{i-1} = x_{i-k+1}, \ldots, x_{i-1}$. A non relative MM can store each outcome of the past in memory, while a MM working in relative mode can only store the outcomes seen in y. The number of conditioning states of the model in DNA sequences is 4^k. The cooperation between MM of different orders has proved to be a more efficient solution for representing DNA sequences, instead of competition [19].

High order MM, typically with $k \geq 13$, proved to be one of the most important models for DNA sequence representation [20], as well as to address other applications [21–23]. However, when substitutional mutations occur between two identical sequences, high order MM fall short to represent the data. This happens because, if, for example, we use an order-20 MM and we have a probability of one random substitution for each 20 bases, the probability that the same context is seen again is low. The DNA data between close species is frequently of this nature, because they share a common ancestral. Moreover, the distinct majority of the editions in the DNA sequences are of substitutional nature.

Aware of these characteristics, we have recently proposed a preliminary approach to deal with substitutional mutations in DNA sequences [6]. In this paper, we consolidate the concept of substitutional tolerant Markov models (STMM) and we apply them to the relative compression case. After, we measure its impact on synthetic genomic data, exploring some characteristics of compressing the elements from a reverse order, as well as some combinations between both. Finally, we show some comparative results between whole genomes.

2 Substitutional Tolerant Markov Model (STMM)

A substitutional tolerant Markov model (STMM) is a probabilistic-algorithmic finite-context model. It assigns probabilities according to a conditioning context that considers the last symbol, from the sequence to occur, as the most probable, given the occurrences stored in the memory, such as those from y, instead of the true occurring symbol.

For a symbol $s \in \Theta$, the estimator of a STMM, working in relative mode, is given by

$$P(s|x'^{i-1}_{i-k}, y) = \frac{N(s|x'^{i-1}_{i-k}, y) + \alpha}{N(x'^{i-1}_{i-k}, y) + \alpha|\Theta|}, \tag{3}$$

where function N accounts for the memory counts regarding the model and x' is a copy of x, edited according to

$$x'_i = \underset{\forall s \in \Theta}{\operatorname{argmax}} P(s|x'^{i-1}_{i-k}, y). \tag{4}$$

The parameter α allows balancing between the maximum likelihood estimator and a uniform distribution. For deeper orders, α should be generally lower than one.

When a STMM (relative or non-relative model) is cooperating with any other model, besides being probabilistic, can also be algorithmic, because they can be switched on or off given its performance, according to a threshold, t, defined before the computation.

Both relative and non-relative modes work with a threshold, t, that enables or disables the model according to the number of times that the context has been seen. Listing 1.1. describes the process for enabling or disabling a STMM.

Listing 1.1. Algorithm of a STMM, described in C language, with comments.

```
int GetBestId(int *array){
  int x, best = 0, maximum = array[0];
  for(x = 1 ; x < N_SYMBOLS ; ++x)                // N_SYMBOLS = 4 (bases)
    if(array[x] > maximum){
      maximum = array[x];
      best = x;
    }
  return best;          // RETURN THE HIGHEST ELEMENT POSITION OF AN ARRAY
}

void Fail(Model *M){                                        // ACTION FOR FAIL
  int x, fails = 0;
  for(x = 0 ; x < M->k ; ++x)                            // USING HISTORY COUNT
    if(M->history[x] != 0)                               // THE NUMBER OF FAILS
      ++fails;
  if(fails > M->threshold)                    // FAILS MORE THAN THRESHOLD?
    M->on = 0;                                                 // SET STMM OFF
  else                                                         //    OTHERWISE
    ShiftBuffer(M->history, M->k, 1);                        // ADD ONE FAIL
}

void Hit(Model *M){                              // ACTION FOR HIT (SUCCESS)
  ShiftBuffer(M->history, M->k, 0);                            // ADD ONE HIT
}

void CorrectSTMM(Model *M, PModel *P, int sym){
  int best = 0;
  if(M->on == 0){                                                 // IF IS OFF
    M->on = 1;                                                 // TURNS STMM ON
    memset(M->history, 0, M->k);
  }
  else{                                                          // ELSE IF IS ON
    if((best = GetBestId(P->freqs) == sym){                 // IF BEST ID = SYM
      Hit(M);                                               // CALL HIT FUNCTION
    }
    else{                                                          //     OTHERWISE
      Fail(M);                                              // CALL FAIL FUNCTION
      M->seq->buf[M->seq->idx] = best;                      // UPDATE NEW SYMBOL
    }
  }
  UpdateCBuffer(M->seq);                                // UPDATE SEQUENCE BUFFER
}
```

The threshold, t, is set at the beginning of the computation. We also need a Boolean cache-array (history) to store the past k hits/fails. For example, consider that $k = 7$ and that $c_0 = \text{CACGTCA}$ is the current context. Also, consider that the number of past symbol occurrences following c_0 was $\text{A} = 1, \text{C} = 0, \text{G} = 0, \text{T} = 0$. If the symbol that is being compressed is G (contradicting the probabilistic model), a MM would have as next context $c_1 = \text{ACGTCAG}$. However, the STMM would use a c'_1, taking into account the most probable outcome and, hence, $c'_1 = \text{ACGTCAA}$. Therefore, the next probabilistic model would be dependent on the past context assumed to be seen and, hence, it assumes that the symbol that was compressed is A.

3 Results

For producing the results, we have used synthetic and real data. The synthetic data made available a controlled comprehension of the STMMs, while the real data shown the characteristics that are also not controlled. The materials to replicate both results on synthetic and real data are available, under GPL v3 license, at the repository https://github.com/pratas/STMM. All experiments were run on Ubuntu Linux v16.04 LTS, with gcc v5.3.1, using only one Intel Core i7-6700K 3.4 GHz CPU, 32 GB of RAM and a solid-state hard drive.

3.1 Synthetic Data

In Fig. 1 we have simulated a sequence y with 200 bases, copied y to x and inserted edits in several positions of x, specifically at positions 50, 100, 102, 150, 152 and 154. Then we have compressed x relatively to y, assuming the order of each element of x as $x_1, x_2, ..., x_{|x|}$ as right direction, $x_{|x|}, ..., x_2, x_1$ as left direction and the minimum complexities of both directions as min.

As it can be seen, the cooperation between MMs and STMMs led to a much better approximation of the data. While the MMs can not address efficiently the data after a substitution occurs, between a period of time that seems related with the k-size, the cooperation between MMs and STMMs address them efficiently, having an almost strict decay to a low complexity value.

In Fig. 2 we have simulated a sequence y. Then, we have made 12 copies, for each one applied some degree of random substitutional mutations, and concatenated all into a final sequence, called x. Then we have compressed, using $C(x_i \| y)$, and plotted it. As it can be seen, with 7.5% of substitutional mutations the cooperation of only MMs reaches the average of 1 BPB (bits per base), while the cooperation between MMs and STMMs reaches the same BPB only at 15% of substitutional mutations.

3.2 Real Data

We have used two eagle whole genomes in non-assembled mode, namely White-tailed eagle (*Haliaeetus albicilla*, 1.14 GB, 26X) and Bald eagle (*Haliaeetus leucocephalus*, 1.26 GB, 88X), from [24]. We have also used the reference genomes of

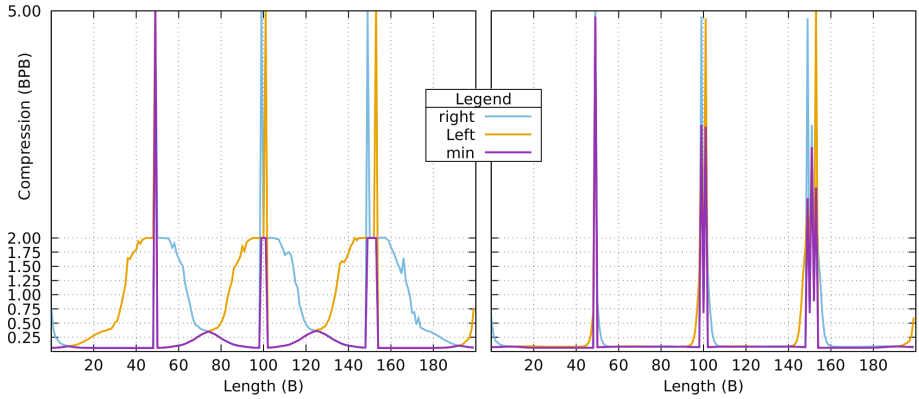


Fig. 1. Relative compression using a cooperative set of MMs (left plot) and a cooperative set of MMs and STMMs (right plot). The compression direction is included for right and left, as well as the minimum (min) between both for each elements. The data is synthetic. The length is in bytes (B). The experiment can be replicated using the script *runSmallBidirection.sh*, from the repository described in this paper.

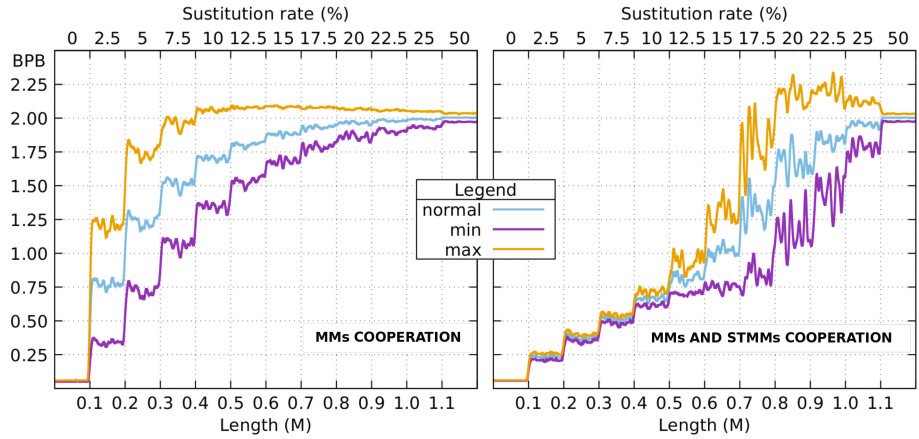


Fig. 2. Relative compression using a cooperative set of MMs (left plot) and a cooperative set of MMs and STMMs (right plot). The synthetic data has been copied from y, creating multiple concatenated x's. For each 100k of data (bottom axis), a substitution mutation rate has been applied (top axis). Besides normal, the legend shows the computation of min and max. These are the minimum (min) and maximum (max) functions of each element processed in left and right directions. The length is in mega bytes (M). The experiment can be replicated using the script *runRelativeBidirection.sh*, from the repository described in this paper.

human, chimpanzee, gorilla, orangutan, and marmoset from the NCBI. We have used a setup of 4 MMs in cooperation with order-k of $\{4, 6, 13, 20\}$ and the α of, respectively, $\{1, 1, 0.5, 0.005\}$. Only one STMM was used with $k = 20$, $\alpha = 0.5$

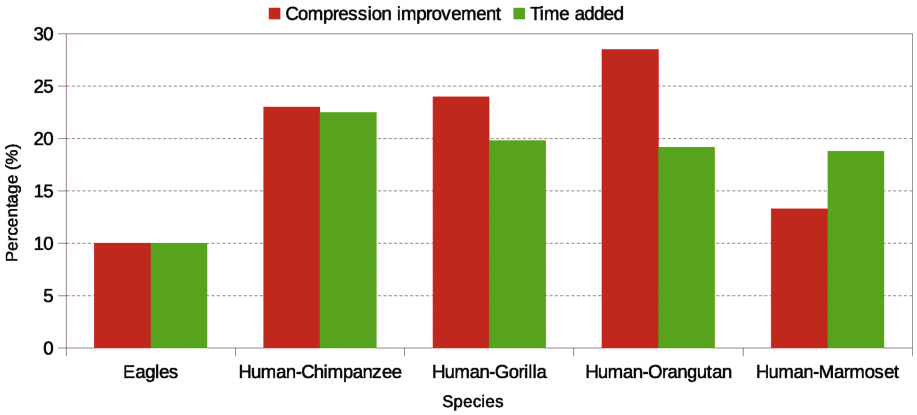


Fig. 3. Compression improvement and compression time added between the relative compression using a cooperative set of MMs and a cooperative set of MMs and STMMs. Percentages are given by $STMM_{bytes}/MM_{bytes} \times 100$ for compression improvement and $MM_{minutes}/STMM_{minutes} \times 100$ for time added.

and $t = 5$. The experiment can be replicated using the script *runBirds.sh* and *runPrimates.sh*.

As can be seen in Fig. 3, to compress the Bald eagle relatively to White-tailed eagle, using only a cooperation between MMs, we needed 31,561,247 bytes. Adding the cooperation of the STMMs, we reached 34,864,683 bytes, which is around 10% of improvement, using the same RAM memory (13.8 GB) and around 10% more computational time. These species are believed to have diverged ≈1 million years ago (mya) [25].

As can be seen in Fig. 3, to compress a chimpanzee relatively to a human genome, using only a cooperation between MMs, we needed 274,450,972 bytes and near 80 min. Adding the cooperation of the STMMs we were able to spend only 210,691,987 bytes, which is around 23% of improvement, using the same RAM memory (26.3 GB) and around more 22.5% of computational time. The human and chimpanzee lineages are believed to have diverged ≈3–4.5 mya [26].

To compress a gorilla relatively to a human genome, using only a cooperation between MMs, we needed 262,271,376 bytes. Adding the cooperation of the STMMs we were able to spend only 199,204,749 bytes, which is around 24% of improvement, using the same RAM memory (26.3 GB) and around 19.8% more computational time. The human and gorilla lineages are believed to have diverged before ≈5–9 mya [26].

To compress a orangutan relatively to a human genome, using only a cooperation between MMs, we needed 418,481,411 bytes. Adding the cooperation of the STMMs we were able to spend only 299,316,387 bytes, which is around 28.5% of improvement, using the same RAM memory (26.3 GB) and around 19.2% more computational time. The human and orangutan lineages are believed to have diverged before 10 mya [26].

Finally, to compress a marmoset relatively to a human genome, using only a cooperation between MMs, we needed 562,916,901 bytes. Adding the cooperation of the STMMs we were able to spend only 488,238,361 bytes, which is around 13.3% of improvement, using the same RAM memory (26.3 GB) and around 18.8% more computational time. The human and marmoset lineages are believed to have diverged around ≈40 mya [27].

4 Conclusions

In this paper, we have proposed a new model for relative compression of DNA sequences—the substitutional tolerant Markov model (STMM). We have shown that it addresses efficiently some degree of substitutional mutations, being a model efficient to use between species that divergence less than 40 million years ago, such as between some primates or eagles. The time added by the model to the compressor is affordable, given the compression improvement—for example, between human and orangutan is around 28.5%. This model is, therefore, a strong candidate to be used in ancient DNA analysis, namely because of the high substitutional mutation rates of the data.

Acknowledgments. This work was partially funded by FEDER (POFC-COMPETE) and by National Funds through the FCT - Foundation for Science and Technology, in the context of the projects `UID/CEC/00127/2013` and `PTCD/EEI-SII/6608/2014`.

References

1. Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., Valiente, G.: Compression-based classification of biological sequences and structures via the universal similarity metric: experimental assessment. BMC Bioinform. **8**(1), 252 (2007)
2. Pinho, A.J., Garcia, S.P., Pratas, D., Ferreira, P.J.S.G.: DNA sequences at a glance. PLoS ONE **8**(11), e79922 (2013)
3. Campagne, F., Dorff, K.C., Chambwe, N., et al.: Compression of structured high-throughput sequencing data. PLoS ONE **8**(11), e79871 (2013)
4. Benoit, G., Lemaitre, C., Lavenier, D., et al.: Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph. BMC Bioinform. **16**(1), 288 (2015)
5. Pratas, D., Silva, R.M., Pinho, A.J., Ferreira, P.J.S.G.: An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Sci. Rep. **5**, 10203 (2015)
6. Pratas, D., Pinho, A.J., Ferreira, P.: Efficient compression of genomic sequences. In: Proceedings of the Data Compression Conference on DCC-2016, Snowbird, Utah, pp. 231–240, March 2016
7. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Probl. Inf. Transm. **1**(1), 1–7 (1965)
8. Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and Its Applications, 3rd edn. Springer, New York (2008)
9. Ziv, J., Merhav, N.: A measure of relative entropy between individual sequences with application to universal classification. IEEE Trans. Inf. Theory **39**(4), 1270–1279 (1993)

10. Benedetto, D., Caglioti, E., Loreto, V.: Language trees and zipping. Phys. Rev. Lett. **88**(4), 048702-1–048702-4 (2002)
11. Cilibrasi, R.L., et al.: Statistical inference through data compression. Ph.D. thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam (2007)
12. Cerra, D., Datcu, M.: Algorithmic relative complexity. Entropy **13**, 902–914 (2011)
13. Coutinho, D.P., Figueiredo, M.: Text classification using compression-based dissimilarity measures. Int. J. Pattern Recogn. Artif. Intell. **29**(5), 1553004 (2015)
14. Pinho, A.J., Pratas, D., Ferreira, P.: Authorship attribution using relative compression. In: Proceedings of the Data Compression Conference on DCC-2016, Snowbird, Utah, March 2016
15. Coutinho, D.P., Figueiredo, M.A.: An information theoretic approach to text sentiment analysis. In: ICPRAM, pp. 577–580 (2013)
16. Fink, G.A.: Markov Models for Pattern Recognition: From Theory to Applications. Springer Science & Business Media, London (2014)
17. Brás, S., Pinho, A.J.: ECG biometric identification: a compression based approach. In: Engineering in Medicine and Biology Society (EMBC), pp. 5838–5841. IEEE (2015)
18. Sayood, K.: Introduction to Data Compression, 3rd edn. Morgan Kaufmann, Burlington (2006)
19. Pinho, A.J., Pratas, D., Ferreira, P.: Bacteria DNA sequence compression using a mixture of finite-context models. In: Proceedings of the IEEE Workshop on Statistical Signal Processing, Nice, France, June 2011
20. Pratas, D., Pinho, A.J.: Exploring deep Markov models in genomic data compression using sequence pre-analysis. In: Proceedings of the 22nd European Signal Processing Conference on EUSIPCO-2014, Lisbon, Portugal, pp. 2395–2399, September 2014
21. Zhao, W., Wang, J., Lu, H.: Combining forecasts of electricity consumption in China with time-varying weights updated by a high-order Markov chain model. Omega **45**, 80–91 (2014)
22. Kwak, J., Lee, C.H., et al.: A high-order Markov-chain-based scheduling algorithm for low delay in CSMA networks. IEEE/ACM Trans. Netw. **24**(4), 2278–2290 (2016)
23. Kárnỳ, M.: Recursive estimation of high-order Markov chains: approximation by finite mixtures. Inf. Sci. **326**, 188–201 (2016)
24. Jarvis, E.D., Mirarab, S., Aberer, A.J., et al.: Whole-genome analyses resolve early branches in the tree of life of modern birds. Science **346**(6215), 1320–1331 (2014)
25. Wink, M., Heidrich, P., Fentzloff, C.: A mtDNA phylogeny of sea eagles (genus haliaeetus) based on nucleotide sequences of the cytochrome b-gene. Biochem. Syst. Ecol. **24**(7–8), 783–791 (1996)
26. Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., et al.: Great ape genetic diversity and population history. Nature **499**(7459), 471–475 (2013)
27. Sequencing, T.M.G., Consortium, A., et al.: The common marmoset genome provides insight into primate biology and evolution. Nat. Genet. **46**(8), 850–857 (2014)

Biomedical Word Sense Disambiguation with Word Embeddings

Rui Antunes(✉) and Sérgio Matos

DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
{ruiantunes,aleixomatos}@ua.pt

Abstract. There is a growing need for automatic extraction of information and knowledge from the increasing amount of biomedical and clinical data produced, namely in textual form. Natural language processing comes in this direction, helping in tasks such as information extraction and information retrieval. Word sense disambiguation is an important part of this process, being responsible for assigning the proper concept to an ambiguous term.

In this paper, we present results from machine learning and knowledge-based algorithms applied to biomedical word sense disambiguation. For the supervised machine learning algorithms we used word embeddings, calculated from the full MEDLINE literature database, as global features and compare the results to the use of local unigram and bigram features.

For the knowledge-based method we represented the textual definitions of biomedical concepts from the UMLS database as word embedding vectors, and combined this with concept associations derived from the MeSH term co-occurrences.

Both the machine learning and the knowledge-based results indicate that word embeddings are informative and improve the biomedical word disambiguation accuracy. Applied to the reference MSH WSD data set, our knowledge-based approach achieves 85.1% disambiguation accuracy, which is higher than some previously proposed approaches that do not use machine-learning strategies.

Keywords: Biomedical word sense disambiguation · Word embeddings

1 Introduction

Large volumes of biomedical data are produced every day, and this is accompanied by an increasing amount of textual data, mostly in the form of scientific publications. In order to efficiently treat and interpret these data it is necessary to create tools that automatically do this job, reducing the human efforts. This led to the application of text mining methods for extracting information from the literature and linking that to repositories of biomedical data [1].

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_33

Word Sense Disambiguation (WSD), an important subtask of Natural Language Processing (NLP) [2], is a challenging task that consists of finding the correct sense of an ambiguous term. Usually, this is achieved using the surrounding context of the term. Currently, there are mainly two distinct approaches for WSD, those based on Machine Learning (ML) algorithms and the ones based on knowledge sources. The ML approaches can follow supervised, semi-supervised or unsupervised algorithms, with supervised classification approaches currently offering the best results, achieving macro and micro accuracy around 96% on the MSH WSD data set using a Support Vector Machine (SVM) classifier [3].

Knowledge-based approaches to WSD have also attracted large interest, as these approaches are usually less dependent on training data, which may lead to better generalization when compared to supervised learning algorithms. The use of multiple knowledge databases brings benefits to the problem of concept disambiguation [4]. WordNet [5] is a large knowledge database of the English language that has been extensively applied for word sense disambiguation [2]. In the case of biomedical texts, the largest and most relevant knowledge database is the Unified Medical Language System (UMLS) [6], which offers a rich integrated metathesaurus and semantic network for the biomedical domain. In this work we used the Medical Subject Headings (MeSH), a hierarchically-organized biomedical vocabulary resource used by the MEDLINE database to index scientific publications, and which is part of the UMLS metathesaurus.

Word embeddings [7] is a recent technique that consists in deriving vector representations of the words within an unlabelled corpus. These vectors can be used for different NLP tasks, namely for the disambiguation process. We used them as global features in the ML classification problem. In our case, these features showed to be almost as effective as local features, such as unigrams and bigrams. Also, we made use of the word embeddings in our knowledge-based approach to represent concepts, and the textual context of ambiguous words, as embedded vectors that can be directly compared. In [3], the authors present a work on supervised biomedical word sense disambiguation applied to the MSH WSD data set, exploring the combination of unigrams as local features and word embeddings as global features. Other approaches using word embeddings for word sense disambiguation have also been proposed by Wu et al. [8], and Taghipour and Ng [9].

In this work, we applied knowledge-based methods and machine learning techniques to the MSH WSD data set in order to measure the WSD accuracies. The UMLS database were used to extract textual definitions of biomedical concepts. Also, we used the co-occurrences of the MeSH descriptors[1] to derive concept-concept associations between. The ML classifiers used in this experiment were the decision tree, the k-nearest neighbours, and the linear SVM with stochastic gradient descent. Textual data from the MEDLINE database were used to generate the word embeddings, which were used in the machine learning and knowledge-based approaches.

[1] https://ii.nlm.nih.gov/MRCOC.shtml.

2 Methods

2.1 The MSH WSD Data Set

The MSH WSD data set was automatically generated using the UMLS metathesaurus and MEDLINE citations [10]. The data consist of scientific abstracts, each with one ambiguous term identified and mapped to the correct sense. It contains 203 ambiguous terms with a total of 423 distinct senses. Most terms (189) have only two different meanings, 12 terms have three different meanings, and the remaining 2 terms have four and five different meanings. The dataset contains around 37 thousand abstracts, each representing an ambiguity example for a term, therefore averaging 187 ambiguity examples per ambiguous term.

Since we extracted textual definitions and MeSH relations from the UMLS database, not all concepts of the MSH WSD data set were present. Thus, a minor part containing 12 terms[2] of the MSH WSD data set were not used for this disambiguation task. All the presented results do not include these terms.

2.2 Machine Learning

For each ambiguous term, we applied 5-fold cross-validation to subdivide the corresponding abstracts for training and testing the model. A bag-of-words model was used to represent the texts, with local features acquired from the context, namely unigrams and bigrams, with tf-idf weighting. We also applied supervised ML algorithms using word embedding vectors, calculated from the full MEDLINE, as global features. A list of 364 stopwords obtained from the UMLS repository was used to filter out very frequent words in the corpus. All these tasks were implemented using the framework Scikit-learn [11], a machine-learning library for the Python programming language. Word embedding models were obtained with the Word2Vec [7] implementation in the Gensim framework [12].

We tested three machine learning classifiers: decision tree classifier, k-nearest neighbours, and linear SVM with stochastic gradient descent. The local features used were unigrams and bigrams, and the global features used were the word embeddings from the full MEDLINE.

The word embedding models were calculated with PubMed articles, which are specific to biomedical domain, from the full MEDLINE. Around 20 million abstracts corresponding to the years 1900 to 2015 were used, containing around 800 thousand distinct words. We trained six models, with windows of five, twenty and fifty words and for feature vectors of sizes 100 and 300. Each abstract, instance of the MSH WSD data set, was represented by the weighted average of the embedding vectors of the containing words, with the tf-idf value of each word used as weight.

[2] Terms not considered: Ca; CNS; Crown; DBA; FAS; Gamma-Interferon; Hybridization; ITP; PCP; Plaque; Pneumocystis; Semen.

2.3 Knowledge-Based

We developed a knowledge-based method to choose the most related concepts from a text and which was applied in the disambiguation task. From the UMLS database we extracted all the available concept textual definitions. Additionally, we used the co-occurrence counts of MeSH terms in MEDLINE articles[3] to calculate the normalized Pointwise Mutual Information (nPMI) as an association metric between all pairs of MeSH terms. Since the MSH WSD data set uses UMLS Concept Unique Identifiers (CUIs) to identify the distinct term senses, we used the MeSH to CUI mapping in UMLS to translate these MeSH term associations to (UMLS) concept-concept associations.

We used the same word embedding models as described above for the machine learning approach. Each specific CUI was represented as an embedding vector calculated as the tf-idf weighted average of the words in the concept definition, therefore mapping each concept to an high-dimensional vector. Using the same approach we were able to calculate an embedding vector for each abstract in the MSH WSD data set. Thus, it was possible to infer the most related sense for an ambiguous term by measuring the cosine similarity between its textual context and each possible UMLS concept, selecting the most similar one.

Additionally, we extended this document-concept similarity score using the concept associations obtained from the MeSH co-occurrences, as shown in Eq. 1.

$$score(CUI) = \frac{1}{N} \sum_{j} nPMI(CUI, CUI_j) \cdot CS(\boldsymbol{t}, \boldsymbol{CUI_j}) \tag{1}$$

According to Eq. 1, for each possible CUI of an ambiguous target term is assigned a score given by the average of the cosine similarities between the term context vector $\boldsymbol{t}$ and the concept vector of all the concepts $\boldsymbol{CUI_j}$, weighted by the concept association score $nPMI(CUI, CUI_j)$. Each considered CUI has a $nPMI$ value equal to a unit in relation to himself. As before, the concept with highest score is selected as the correct sense for the ambiguous term.

3 Results

Table 1 shows that the state-of-the-art results for this problem can be almost reproduced using simple word-based features. It is also noticeable that bigram

Table 1. Accuracies using local features. Results shown are the average across five folds. U: Unigrams; B: Bigrams; DT: Decision Tree; kNN: k-Nearest Neighbour (k = 5); SVM: linear Support Vector Machine with stochastic gradient descent.

	U	B	U + B
DT	0.903	0.862	0.901
kNN	0.913	0.918	0.924
SVM	**0.947**	0.931	0.946

[3] https://ii.nlm.nih.gov/MRCOC.shtml.

features contribute only slightly to the results, and unigram features alone achieve almost as good if not better results than the combination of unigram and bigrams. Also, comparing these results with Table 2, one can observe that word embedding features alone allow obtaining results that are very close to the best results obtained with unigram features.

With the machine learning classifiers the highest accuracy, 94.7%, was obtained with unigram features alone, using the support vector machine linear classifier. On the other hand, using only global features the accuracies were similar, and the highest accuracy, 94.0%, was also obtained using the support vector machine classifier.

In Table 3 the knowledge-based results are presented. One can see that these results are only about 10% below the machine learnings results, since it is a more generalized method that do not use train data from the data set to predict the correct meanings. We applied a threshold to the concept association nPMI score, in order to filter the associated concepts that contribute to the final score for a

Table 2. Accuracies using word embedding models from the full MEDLINE as global features. S: Size; W: Window; DT: Decision Tree; kNN: k-Nearest Neighbour (k = 5); SVM: linear Support Vector Machine with stochastic gradient descent.

	S100			S300		
	W5	W20	W50	W5	W20	W50
DT	0.907	0.909	0.909	0.910	0.911	0.909
kNN	0.931	0.933	0.933	0.931	0.931	0.931
SVM	0.931	0.934	0.937	0.934	0.938	**0.940**

Table 3. Accuracies using the CUI definitions, the CUI relations from the UMLS and word embeddings from the full MEDLINE. CS: cosine similarity between term context vector and concept vector only; nPMI $\geq$ *thresh*: cosine similarity plus related concepts with a nPMI value higher than the threshold; S: Size; W: Window; nPMI: normalized Pontwise Mutual Information.

	S100			S300		
	W5	W20	W50	W5	W20	W50
CS	0.800	0.812	0.813	0.799	0.811	0.810
nPMI $\geq$ 0.9	0.799	0.812	0.813	0.799	0.811	0.809
nPMI $\geq$ 0.8	0.799	0.812	0.813	0.799	0.812	0.810
nPMI $\geq$ 0.7	0.797	0.811	0.813	0.798	0.811	0.809
nPMI $\geq$ 0.6	0.783	0.798	0.799	0.785	0.797	0.795
nPMI $\geq$ 0.5	0.789	0.803	0.805	0.790	0.802	0.798
nPMI $\geq$ 0.4	0.816	0.829	0.831	0.817	0.826	0.826
nPMI $\geq$ 0.3	0.835	0.849	**0.851**	0.835	0.846	0.844
nPMI $\geq$ 0.2	0.827	0.842	0.844	0.826	0.838	0.837

CUI (see Eq. 1). A smaller value for the nPMI threshold means that more related concepts contribute to the final score, and the results show that using more related concepts, and not only the ones with stronger association score, improves the disambiguation accuracy. The highest accuracy, 85.1%, was obtained with a nPMI threshold of 0.30 using the word embedding model with a size vector of 100 and a window of 50 words.

4 Conclusions

As has been previously shown, machine learning algorithms outperform the knowledge-based algorithms in biomedical word sense disambiguation. However, the latter have the advantage of being directly applied to any ambiguous term, since they do not rely on training data. Our approach achieves a robust disambiguation performance that is on par with the best methods that do not use annotated data in a supervised setting, and slightly above the results obtained with the Automatic Extracted Corpus (AEC) [10], which can be applied to obtain training data from MEDLINE to create the disambiguation classifiers on-the-fly, therefore reducing the need for pre-compiled training data. Tulkens [13] obtained a disambiguation accuracy of 84% with a knowledge-based method applied to the same data set using word embeddings from BioASQ corpora. In a recent work, Sabbir et al. [14] combined a knowledge-approach with neural concept embeddings and distant supervision, achieving an accuracy of 92%.

One of the limitations of our approach is that not all UMLS concepts have rich definitions. Also, some concepts of the MSH WSD data are not present in the UMLS database, leading to an incapacity of disambiguation. As future work, we will investigate ways of overcoming this by constructing concept vectors from associated MEDLINE texts.

Acknowledgments. This work was supported by Portuguese National Funds through FCT - Foundation for Science and Technology, in the context of the project IF/01694/2013. Sérgio Matos is funded under the FCT Investigator programme.

References

1. Campos, D., Matos, S., Oliveira, J.L.: A modular framework for biomedical concept recognition. BMC Bioinform. **14**(1), 281 (2013)
2. Navigli, R.: Word sense disambiguation: a survey. ACM Comput. Surv. **41**(2), 10 (2009)
3. Yepes, A.J.: Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation (2016). arXiv:1604.02506v3
4. Tsai, C.T., Roth, D.: Concept grounding to multiple knowledge bases via indirect supervision. Trans. Assoc. Comput. Linguist. **4**, 141–154 (2016)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)

6. Bodenreider, O.: The unified medical language systems (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(1), 267–270 (2004)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. **3111**, 3119 (2013)
8. Wu, Y., Xu, J., Zhang, Y., Xu, H.: Clinical abbreviation disambiguation using neural word embeddings. ACL-IJCNLP, pp. 171–176 (2015)
9. Taghipour, K., Ng, H.T.: Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In: HLT-NAACL, pp. 314–323 (2015)
10. Yepes, A.J., McInnes, B.T., Aronson, A.R.: Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinform. **12**(1), 223 (2011)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
12. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50 (2010)
13. Tulkens, S., Šuster, S., Daelemans, W.: Using distributed representations to disambiguate biomedical and clinical concepts (2016). arXiv:1608.05605v1
14. Sabbir, A.K.M., Yepes, A.J., Kavuluru, R.: Knowledge-based biomedical word sense disambiguation with neural concept embeddings and distant supervision (2017). arXiv:1610.08557v3

Classification Tools for Carotenoid Content Estimation in *Manihot esculenta* via Metabolomics and Machine Learning

Rodolfo Moresco[1(✉)], Telma Afonso[3], Virgílio G. Uarrota[1], Bruno Bachiega Navarro[1], Eduardo da C. Nunes[2], Miguel Rocha[3], and Marcelo Maraschin[1]

[1] Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianopolis, Brazil
rodolfo_moresco@yahoo.com.br
[2] Santa Catarina State Agricultural Research and Rural Extension Agency (EPAGRI), Experimental Station of Urussanga, Urussanga, Brazil
[3] Centre Biological Engineering, School of Engineering, University of Minho, Braga, Portugal

Abstract. Cassava genotypes (*Manihot esculenta* Crantz) with high pro-vitamin A activity have been identified as a strategy to reduce the prevalence of deficiency of this vitamin. The color variability of cassava roots, which can vary from white to red, is related to the presence of several carotenoid pigments. The present study has shown how CIELAB color measurement on cassava roots tissue can be used as a non-destructive and very fast technique to quantify the levels of carotenoids in cassava root samples, avoiding the use of more expensive analytical techniques for compound quantification, such as UV-visible spectrophotometry and the HPLC. For this, we used machine learning techniques, associating the colorimetric data (CIELAB) with the data obtained by UV-vis and HPLC, to obtain models of prediction of carotenoids for this type of biomass. Best values of R^2 (above 90%) were observed for the predictive variable TCC determined by UV-vis spectrophotometry. When we tested the machine learning models using the CIELAB values as inputs, for the total carotenoids contents quantified by HPLC, the Partial Least Squares (PLS), Support Vector Machines, and Elastic Net models presented the best values of R^2 (above 40%) and Root-Mean-Square Error (RMSE). For the carotenoid quantification by UV-vis spectrophotometry, R^2 (around 60%) and RMSE values (around 6.5) are more satisfactory. Ridge regression and Elastic Network showed the best results. It can be concluded that the use colorimetric technique (CIELAB) associated with UV-vis/HPLC and statistical techniques of prognostic analysis through machine learning can predict the content of total carotenoids in these samples, with good precision and accuracy.

Keywords: Chemometrics · Descriptive models · Machine learning · Cassava genotypes · Carotenoids · HPLC · UV-vis

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_34

1 Introduction

Carotenoids refer to the most important natural pigments, being found in all photosynthetic organisms, with colors varying between yellow and dark-red. One of the most important trait of carotenoids is their physiological function as vitamin A precursors to animals [1]. Vitamin A deficiency is a leading cause of morbidity and mortality, especially in young children and pregnant and lactating women. Food-based interventions focused on alleviating vitamin A deficiency in susceptible populations have advantages over supplementation and fortification programs, especially in rural areas, because they can provide a sustainable source of a variety of nutrients and other phytochemicals without the recurring transport and administration costs of these other methods [2]. It is estimated that among all known carotenoids, about 50 can act as precursors of vitamin A in mammals. However, only α-carotene, β-carotene, γ-carotene, and β-cryptoxanthin are common in fruits and vegetables [3]. Cassava genotypes with high contents of pro-vitamin A carotenoids have been identified as a strategy to reduce the prevalence of deficiency of this vitamin [4].

The cassava crops are characterized by the color variability of their roots, which can vary from white to red. The color is related to the presence of several carotenoid pigments, their associations and contents [5]. However, the possibility of adopting the color of roots as an indirect criterion for selection of higher carotene content is questionable, since color is a characteristic of difficult visual evaluation.

In order to standardize color measurements, the CIE (Commission Internationale de L'Eclairage) recommended the use of the CIE L* a* b* or CIELAB color scale. It is currently the most used system for quantitative color description of an object, due to its uniformity, ease of acquisition, and very low cost technique [6].

Chemical extraction followed by the identification and quantification of carotenoid pigments, especially by UV-vis spectrophotometry and high performance liquid chromatography (HPLC) are very accurate, but extremely expensive, also requiring a long time for the analysis. The CIELAB color measurement is a non-destructive and very fast technique, which allows to obtain a series of parameters, in a few seconds. Thereby, it facilitates performing measurement in the field, avoiding the degradation of these compounds in consequence of their chemical extraction, for instance.

The aim of this work is to validate a quantification method for carotenoid contents in roots of *M. esculenta* from colorimetric data using the CIE L* a* b* system, assuming that the statistical techniques of prognostic analysis, as well as machine learning, can correlate colorimetric data easily obtained in the field, with the contents obtained through traditional techniques, e.g., UV-vis spectrophotometry and HPLC and, from this, construct prediction models of carotenoids content for cassava roots. This study applies analytical techniques and bioinformatics tools to detect genotypes of *M. esculenta* with high levels of carotenoids. In addition, it provides tools that can support the plant-breeding program at Epagri (Agricultural Research Company and Rural Extension of the State of Santa Catarina- http://www.epagri.sc.gov.br/) that aims to obtain genotypes with high levels of pro-vitamin A carotenoids and superior nutritional traits.

2 Materials and Methods

Roots of fifty genotypes of *M. esculenta* (2015/2016 season) from the EPAGRI's germplasm bank (Urussanga Experimental Station, 28º31′18″S, 49º19′03″W, Santa Catarina, southern Brazil) were used in this study due to their economic and social importance.

Carotenoids were extracted from fresh roots as described by Rodriguez-Amaya & Kimura (2004) [7]. The absorbances of the organosolvent extracts were recorded on an UV-vis spectrophotometer (Gold Spectrum lab 53 UV-Vis spectrophotometer, BEL photonics, Brazil) over a spectral window from 200 to 700 ηm. Aliquots (10 μl) of the extracts were also injected into a liquid chromatograph (LC-10A Shimadzu) system equipped with a C18 reversed-phase column (Vydac 201TP54, 250 mm × 4.6 mm, 5 μm Ø, 35°C) coupled to a pre-column (C18 Vydac 201TP54, 30 mm × 4.6 mm, 5 μm Ø) and a spectrophotometric detector (450 nm). Methanol: acetonitrile (90: 10, v/v) was used for elution at a rate of 1 ml/min.

The color attributes of the roots samples were measured by a colorimeter (CR-400, Minolta, Japan) immediately after harvest and the results were expressed according to the CIELAB color space scale [4]. Three readings were performed at different sites in fifty samples. Data were collected, summarized, and submitted to analysis of variance (ANOVA) followed by the *post-hoc* Tukey's test ($p < 0.05$) for mean comparison. Spectrophotometric data and the amounts of the target carotenoids determined by HPLC were treated using multivariate statistical analysis and chemometrics techniques, supported by scripts written in R language (v. 3.3.1) [8]. Additionally, we used prognostic tools through machine learning techniques, associating the colorimetric data (CIELAB) with the data obtained by UV-vis and HPLC, to obtain models of prediction of carotenoids for this type of biomass and technique.

The data analysis was supported and structured using the R *specmine* package [9] developed by our research team for metabolomics studies that includes a number of machine learning methods implemented through the package *caret* [10]. In supplementary material, provided in http://darwin.di.uminho.pt/pacbb2017/cassava-carotenoids, we include the data analysis reports automatically generated from the R scripts using the features provided by R Markdown, as well as the respective data and metadata files. This allows fully understanding and reproducing the computational experiments.

3 Results and Discussion

The values of the carotenoid quantification through UV-vis spectrophotometry and HPLC are given in the metadata of the dataset. The roots white-colored pulp presented the lowest concentrations of total carotenoids (values from 0.57 $\mu g.g^{-1}$), while highest concentrations were observed in genotypes with pigmented pulp (yellow and red) roots, i.e., 54.93 $\mu g.g^{-1}$. These results are consistent with data reported in the literature that observe a positive relation between the color of the root pulp and the total content of those pigments [11, 12]. The contents of the major carotenoid compounds, *trans*-β-carotene and *cis*-β-carotene, ranged from 1.82 to 42.82 $\mu g.g^{-1}$ for *trans*-β-carotene and 1.19 to 28.86 $\mu g.g^{-1}$ for *cis*-β-carotene.

The visual interpretation of the sample's location in the CIELAB' space is enough to verify which samples have higher levels of carotenoids [13]. Figure 1 shows the samples location according to the color of roots, in the CIE L* a* b* plane. Samples 105, 119 and 125 (Fig. 1 - ellipse I) contain the highest levels of total carotenoid. The sample 74, due to its reddish color, was represented in the CIELAB space on the positive axis (Fig. 1 - ellipse IV), mostly due to its lycopene contents, which confer reddish coloration to the roots [14]. Samples with lower amounts of carotenoids (123, 27, 05, AO47) shown values of b* closer to zero (ellipse III), while those with medium contents were grouped in a* negative and b* positive (ellipse II).

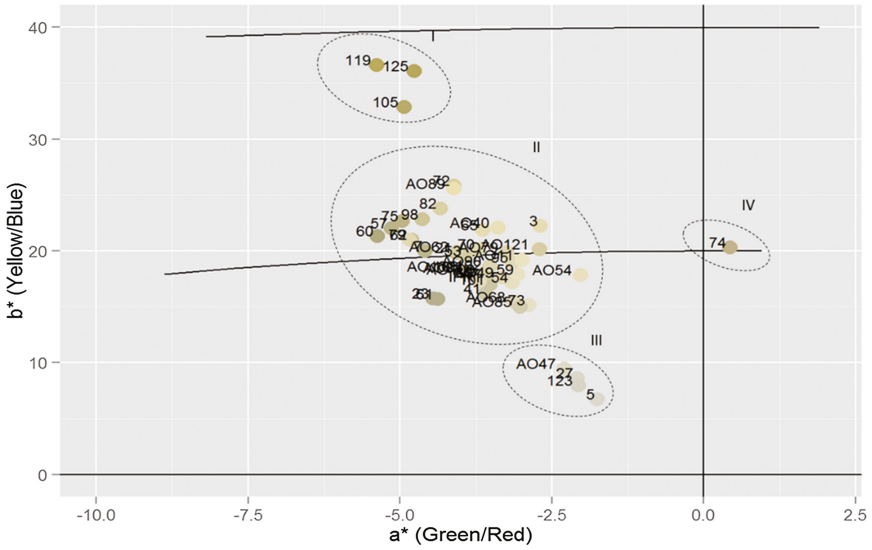


Fig. 1. Location of the cassava samples in the CIE L* a* b* plane according to their root pulp colors. The a* value characterizes the coloration in the regions of red (+a*) to green (−a*). The b*, value b* indicates coloring in the range of yellow (+b*) to blue (−b*). The L indicates the luminosity, varying from white (L = 100) to black (L = 0).

The next step of this work was to correlate the colorimetric data obtainable in the field (CIELAB) with the contents found by traditional techniques, e.g., UV-vis spectrophotometry and HPLC, through statistical techniques of prognostic analysis such as machine learning. From this, we constructed a set of carotenoid concentration predictive regression models for this type of biomass using the information from the samples' color values and the UV-vis spectra.

The *specmine* package provides a number of functions to train, use, and evaluate machine learning methods, being mostly based in the R package *caret* [10], covering both classification and regression methods. In addition, there are functions to evaluate the importance of each variable in the models. A list of possible models and tunable parameters can be seen in https://topepo.github.io/caret/available-models.html.

The implemented functions enable executing model training and can be used to predict new data posteriorly. Also, it is possible to optimize a set of model parameters testing a set of possible values and evaluating those according to the selected validation method and metric errors. The CIELAB data were considered as continuous variables. In this way, regression-derived statistical data mining models (5-fold cross-validation repeated 10 times, testing all models with feature selection with 80, 60, and 40% data filtering) were used, such as Least Absolute Shrinkage and Selection Operator (Lasso) [15], Ridge Regression [16], Elastic Net Regression (Enet) [17], Decision Trees/Randon Forest (RF) [18], Partial Least Squares (PLS), Artificial Neural Net (NNs), and Support Vector Machines (SVMs). These validation methods are available to estimate the metric errors, and usually the decision is based on simple criteria based on the residual values. The chosen evaluation metrics to compare model performance were the Root-Mean-Square Error (RMSE) and the coefficient of determination (R^2), since they explicitly show how much the model predictions deviate, on average, from the actual values in the dataset.

Table 1 shows the performance values of a set of machine learning regression models (RMSE and R^2) associating UV-vis scanning spectrophotometry in the typical region of fingerprint for carotenoids (400–500 nm) as inputs, with the total carotenoids contents determined by HPLC (TCC HPLC), total carotenoids contents determined by UV-vis spectrophotometry (Lambert-Beer formula), and the majoritarian carotenoid found in cassava roots (*trans*-β-carotene), each predicted as an output in distinct experiments using the different methods (details are given in the reports in supplementary materials).

It can be verified that the best R^2 values (>90%) were observed for the predictive variable TCC, determined by UV-vis spectrophotometry. These values were higher than the predictive variables *trans*-β-carotene (best model with R^2 47%) and total carotenoids contents determined by HPLC (with values of R^2 around 60%). This is expected, since they are methodologies that employ the same physical phenomenon of detection of compounds (absorbance). When observed the values of variable importance in this analysis (supplementary material), it can be detected that the wavelength at 450 nm (precisely the wavelength that is used for the quantification of β-carotene through the Lambert-Beer formula) was the most prevalent. This result is important because it attests to the robustness of the models in predicting the contents of these compounds in these samples.

Then we tested the machine learning models using the CIELAB values as inputs, with the same outputs as before. For the total carotenoids contents quantified by HPLC, the Partial Least Squares (PLS), Support Vector Machines (kernlab), and Elastic Net models presented the best values of R^2 and lower values of RMSE (Table 2). It can be verified that these values are smaller than when the inputs are the UV-vis (400–500 nm) data (Table 1). This is due to the fact that the colorimetric and chromatographic techniques are different in their physicochemical bases, and the UV-vis data has many more variables measured.

Table 1. Performance values (RMSE and R^2) associating UV-vis scanning spectrophotometry (400–500 nm) with the total carotenoids contents determined by HPLC (TCC HPLC), total carotenoids contents determined by Lambert-Beer formula (TCC Spectrophotometry), and the majoritarian carotenoids of cassava roots samples (*trans*-β-carotene).

	UV-vis. 400–500 nm					
	TCC Spectrophotometry		TCC HPLC		*trans*-β-carotene	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Partial Least Squares (simpls)	**3.492**	**0.920**	5.789	0.572	4.309	0.362
Support Vector Machines (e1071)	**3.709**	**0.931**	5.844	0.597	4.218	0.399
PLS (widekernelpls)	**3.732**	**0.923**	5.779	0.570	4.324	0.453
Random Forest	**3.768**	**0.948**	7.275	0.359	5.753	0.239
Elastic Net	3.793	0.918	**5.934**	**0.634**	4.191	0.412
Partial Least Squares (pls)	3.800	0.952	**5.643**	**0.597**	**4.265**	**0.470**
Ridge Regression (w/FS)	3.855	0.947	**5.880**	**0.603**	4.159	0.356
Ridge Regression	3.877	0.928	7.282	0.616	4.407	0.316
SVM (kernlab)	3.928	0.940	5.907	0.589	4.230	0.466
PLS (kernelpls)	4.096	0.896	5.878	0.566	4.211	0.422
Linear Regression (Stepwise)	4.158	0.919	8.341	0.526	6.135	0.206
Linear Regression (Forward)	4.178	0.888	8.783	0.471	5.142	0.311
Linear Regression (Backwards)	4.392	0.871	6.373	0.522	5.355	0.278
K-Nearest Neighbors	4.732	0.922	6.277	0.445	4.597	0.224
Lasso	5.207	0.817	17.508	0.249	16.145	0.189
Conditional Inference RF	6.713	0.791	6.806	0.558	4.703	0.369
Conditional Inference Tree	7.363	0.711	6.916	0.480	4.894	0.288
Decision Trees	7.582	0.683	6.795	0.473	5.189	0.053

When the CIELAB values were used to predict the values of carotenoid contents by UV-vis spectrophotometry, R^2 and RMSE values were more satisfactory. Ridge regression and Elastic Network showed the best results. Observing the importance of the variables in the prediction (supplementary material), it can be verified that the values of b* were more relevant. In the CIELAB space, the value b^* indicates coloration in the range from yellow ($+b^*$) to blue ($-b^*$), an important finding since most carotenoids confer yellowish pigmentation in foods, associating their pro-vitamin A activity.

Table 2. Performance values (RMSE and R^2) associating CIELAB colorimetric data with the total carotenoids contents determined by Lambert-Beer formula (TCC Spectrophotometry), total carotenoids contents determined by HPLC (TCC HPLC), and the content of the majoritarian carotenoid found in cassava roots samples (*trans*-β-carotene).

	CIELAB Data					
	TCC Spectrophotometry		TCC HPLC		trans-β-carotene	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Partial Least Squares (simpls)	7.043	0.543	6.789	0.414	4.781	0.194
Support Vector Machines (e1071)	7.136	0.500	6.645	0.380	4.800	0.155
PLS (widekernelpls)	6.771	0.541	6.696	0.396	4.857	0.170
Random Forest	7.280	0.448	7.571	0.293	5.393	0.149
Elastic Net	**6.515**	**0.573**	**6.534**	**0.412**	**4.690**	**0.212**
Partial Least Squares (pls)	7.085	0.538	6.622	0.394	4.859	0.164
Ridge Regression (w/FS)	**6.469**	**0.608**	6.653	0.389	4.951	0.238
Ridge Regression	**6.497**	**0.590**	**6.584**	**0.421**	4.848	0.238
SVM (kernlab)	6.919	0.528	**6.534**	**0.366**	**4.745**	**0.201**
Partial Least Squares (kernelpls)	6.865	0.540	6.756	0.431	4.815	0.162
Linear Regression	6.651	0.558	6.749	0.400	4.945	0.220
K-Nearest Neighbors	7.267	0.525	7.278	0.256	4.956	0.153
Lasso	6.757	0.575	6.669	0.411	4.793	0.182
Conditional Inference RF	8.021	0.454	6.930	0.408	4.782	0.223
Conditional Inference Tree	9.636	0.339	7.307	0.384	4.929	0.130
Decision Trees	9.737	0.316	7.641	0.353	5.000	0.297

These results are very promising because they enable CIELAB technique as an alternative for measuring carotenoids in cassava roots to the use of more expensive analytical techniques such as UV-vis spectrophotometry and HPLC. Thus, it has been shown that the concomitant use of UV-vis and color (CIELAB) techniques with statistical techniques of prognostic analysis (i.e., machine learning) can predict the content of total carotenoids in cassava roots, with good precision and accuracy and low metrical error.

4 Conclusions

The present study has shown how CIELAB color measurement can be used as a fast and non-destructive method to calibrate for the total carotenoid content of cassava genotypes roots with acceptable prediction error. In addition, the information obtained by coupling the analysis of pro-vitamin A biochemical markers to bioinformatics tools helps supporting the rational design of biochemically-assisted breeding programs of *M. esculenta*, that aims to obtain cultivars with high levels of pro-vitamin A carotenoids and superior nutritional traits.

Acknowledgements. To CNPq (National Counsel of Technological and Scientific Development) for financial support (Process no. 407323/2013-9), to CAPES (Coordination for the Improvement of Higher Education Personnel (CAPES), and EPAGRI (Agricultural Research and Rural Extension Company of Santa Catarina). The research fellowship from CNPq on behalf of M. Maraschin is acknowledged. The work is partially funded by Project PropMine, funded by the agreement between Portuguese FCT and Brazilian CNPq.

References

1. Rodriguez-Amaya, D.B.: A Guide to Carotenoid Analysis in Foods (2001)
2. Tanumihardjo, S.A., Palacios, N., Pixley, K.V.: Provitamin a carotenoid bioavailability: what really matters? Int. J. Vitam. Nutr. Res. **80**, 336–350 (2010)
3. Stahl, W., Sies, H.: Antioxidant activity of carotenoids. Mol. Aspects Med. **24**, 345–351 (2003)
4. La Frano, M.R., Woodhouse, L.R., Burnett, D.J., Burri, B.J.: Biofortified cassava increases β-carotene and vitamin A concentrations in the TAG-rich plasma layer of American women. Br. J. Nutr. **110**, 310–320 (2013)
5. Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., Zum Felde, T., Domínguez, M., Davrieux, F.: Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. Food Chem. **151**, 444–451 (2014)
6. CIE: The Evaluation of Whiteness. Color, 3rd edn., vol. 552, p. 24 (2004)
7. Rodriguez-Amaya, D., Kimura, M.: HarvestPlus handbook for carotenoid analysis. Harvest. Tech. Monogr. **59**, 525–528 (2004)
8. R Core Team: R: A Language and Environment for Statistical Computing (2014). http://www.r-project.org/
9. Costa, C., Maraschin, M., Rocha, M.: An R package for the integrated analysis of metabolomics and spectral data. Comput. Methods Programs Biomed. **129**, 117–124 (2015)
10. Max, A., Contributions, K., Weston, S., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C.: Package "caret". Max Kuhn (2016)
11. Champagne, A., Bernillon, S., Moing, A., Rolin, D., Legendre, L., Lebot, V.: Carotenoid profiling of tropical root crop chemotypes from Vanuatu. South Pacific. J. Food Compos. Anal. **23**, 763–771 (2010)
12. Chávez, A.L., Sánchez, T., Jaramillo, G., Bedoya, J.M., Echeverry, J., Bolaños, E., Ceballos, H., Iglesias, C.: Variation of quality traits in cassava roots evaluated in landraces and improved clones. Euphytica **143**, 125–133 (2005)

13. Kljak, K., Grbeša, D., Karolyi, D.: Reflectance colorimetry as a simple method for estimating carotenoid content in maize grain. J. Cereal Sci. **59**, 109–111 (2014)
14. Meléndez-Martínez, A.J., Britton, G., Vicario, I.M., Heredia, F.J.: Relationship between the colour and the chemical structure of carotenoid pigments. Food Chem. **101**, 1145–1150 (2006)
15. Tibshirani, R.: Regression Selection and Shrinkage via the Lasso (1994). http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574
16. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970)
17. Zou, H.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Series B **67**, 301–320 (2005)
18. Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R.P., Feuston, B.P.: Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. **43**, 1947–1958 (2003)

UV-Vis Spectrophotometry and Chemometrics as Tools for Recognition of the Biochemical Profiles of Organic Banana Peels (*Musa* sp.) According to the Seasonality in Southern Brazil

Susane Lopes[1(✉)], Rodolfo Moresco[1], Luiz Augusto Martins Peruch[2], Miguel Rocha[3], and Marcelo Maraschin[1]

[1] Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianopolis, Brazil
susane.lopes@ufsc.br
[2] Agricultural Research and Rural Extension Company of Santa Catarina, Criciúma, Brazil
[3] School of Engineering, Centre Biological Engineering, University of Minho, Braga, Portugal

Abstract. Banana (*Musa* sp.) has received wide interest in popular and scientific medicine because of its rich composition in bioactive metabolites, e.g., phenolic compounds, found in interesting concentrations in its peel. Banana peel is a residue that is under-exploited by the industry. Thus, with the intention to give a destination to this by-product towards health care or cosmetics industries, we evaluated its aqueous extract (AE) as a source of bioactive phenolic compounds, aiming at to apply them in future studies of biological activities. For that, in this study samples of banana peels were chemically profiled throughout the year to identify the best harvest time of those biomasses regarding their phenolic composition. In this sense, we used additional information on the chemical heterogeneity of the samples determined by the seasoning, through a set of analytical and climatic data to elaborate chemometric models, supported by bioinformatics tools. Through PCA and HCA analyzes, it was detected that low temperatures; normally observed in winter; strongly modulate the banana metabolism, leading to increased amounts of phenolic compounds, and improving the antioxidant activity of the banana peel AE. The samples collected during the months of winter showed a similar profile and a relatively high concentration of phenolic compounds with potential for future studies of biological activities.

Keywords: Banana · *Musa* sp. · Peels · Phenolic compounds · Antioxidant activity · Spectrophotometry · Chemometrics · Seasonality · Metabolic profile

1 Introduction

Banana (*Musa* sp.) is an edible fruit grown in tropical and subtropical regions with seasonal chemical variation in its pulp and peel composition due to the effect of climatic factors, e.g., rainfall and temperature [1]. In Brazil, one of the largest banana producers worldwide [2], the peel is the main by-product of the banana industrial processing,

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_35

accounting for approximately 38% of the total weight of the fruit. This residual biomass is considered a waste with low economical value [3]. Studies have shown that banana is a good source of carbohydrates, mostly starch, minerals, vitamin B6, natural antioxidants [4, 5], as well as carotenoids and biogenic amines [6].

In the last few decades, banana has been evaluated by scientific and medicinal interests as an important source of bioactive compounds, such as flavonoids, anthocyanins, condensed tannins, and biogenic amines. These compounds have been extensively documented for their actions in promoting health in the reduction of chronic diseases, e.g., cancer, cardiovascular dysfunction, and muscular degeneration [4], besides the antibacterial, antiulcerogenic, antihypertensive antidiabetic, and antioxidant activities [7].

Phenolic compounds are secondary metabolites responsible for several of these therapeutic properties, mainly due to their antioxidant potential, and obtained in interesting concentrations in banana peels [3]. In local and traditional Brazilian medicine, the banana peel has a useful history to promote the healing of wounds mainly by burns when used topically [6], assigning an interesting destination to this residual biomass. Thus, our group aims to recover the phenolic compounds from the banana peel and confer a use to this residue. In this study, we determined the chemical profiles of AEs of banana peels collected over the year to better understand their seasonal heterogeneity. For that, we used a set of analytical and climatic data to build chemometric models, supported by bioinformatics tools. By applying multivariate statistical techniques (principal component analysis - PCA and hierarchical clustering analysis - HCA), we investigated the influence of climatic variables and their relation with biosynthesis of phenolic compounds and their antioxidant activity in samples of banana peel, collected over the seasons in the Santa Catarina State, southern Brazil. This strategy aims to obtain additional information about the biochemical heterogeneity of the samples caused by the seasonality, in order to select the best samples for future studies of biological activities, also driving eventual technological usage.

2 Materials and Methods

2.1 Banana Samples and Processing of Plant Material

Samples of banana peels (*Musa* sp., cv. Prata Anã) were monthly collected (February 2015 to January 2016) from an orchard agroecologically managed, and provided by the Agricultural Research and Rural Extension Company of Santa Catarina (EPAGRI - Criciúma County, 28º40′39″ S, 49º22′11″ W, Santa Catarina State, southern Brazil). The banana samples were sanitized in running water and dried with a paper towel. Then, the fruit peels were manually removed and dried in an oven (45 °C), with air flow, until constant weight. The dry biomass was packed in polyethylene bags and stored at −20 °C. Dried banana peels samples (0.5 g) were added of 7.5 ml distilled water and incubated (water bath, 37 °C, 30 min), followed by centrifuging and recovering of the supernatant as the aqueous extract (AE) according to Pereira (2014) [8].

2.2 Spectrophotometric Analysis

The AE's chemical profiles were determined through UV-Vis scanning spectrophotometry (Gold Spectrum lab 53 UV-Vis spectrophotometer, BEL photonics, Brazil) over a spectral window of 200 to 800 nm (1 nm resolution/data point). The content of total phenolic compounds was determined according to Randhir, Shetty and Shetty (2002) [9], while the total flavonoids amounts followed the methodology proposed by Woisky and Salatino (1998) [10]. The ability of the AE's to scavenge the 1, 1-diphenyl-2-picrylhydrazyl (DPPH) free radical was determined based on the method of Ribeiro et al. (2008) [11].

2.3 Statistical and Chemometric Analysis

Data were collected, summarized, and submitted to analysis of variance (ANOVA) followed by the *post-hoc* Tukey's test ($p < 0.05$) for mean comparison. All procedures were performed in triplicate, in three independent experiments (n = 9). The processing of the spectrophotometric profile considered the definition of the spectral window of interest (200–800 nm), baseline correction, normalization, and optimization of the signal/noise ratio (smoothing). The processed data set was subjected to multivariate statistical analysis, by applying principal component analysis (PCA) and clustering methods, as well as predictive machine learning tools. All analyses were supported by scripts written in the R language using tools developed by our research group (the specmine package) [12] and some functions from the packages Chemospec [13], HyperSpec [14], and ggplot2. All R scripts, raw data, and additional chemometric analysis are available in supplementary material, in http://darwin.di.uminho.pt/pacbb2017/banana-peels, as well as the data analysis report automatically generated from the R scripts using the features provided by R Markdown. This allows anyone to fully reproduce and document the experiments.

3 Results and Discussion

Phenolic compounds are secondary metabolites found in plants with important biological activities. We propose to recover these compounds by producing AE's of banana peels collected over the seasons, to select the best sample for future biological assays. Initially, biochemical (total contents of phenolic compounds and flavonoids, and the inhibition activity of the DPPH radical) and spectrophotometric (absorbance at $\lambda = 200$–800 nm) assays were done, followed by chemometric analysis.

Initial exploratory analysis, with simple descriptive statistics and boxplots of the main variables, indicated differences, namely in the total phenolics and flavonoids concentrations, which appear to be higher in the winter samples. By performing a linear regression relating phenolic and flavonoid contents and climatic factors, the results indicated that lower temperatures increased the amounts of total phenolics and flavonoids in banana peels. This trend was clearer for the total phenolics (mg gallic acid equivalent/g dry peels; i.e., mg GAEq/g), where for the minimum temperature (°C) (Fig. 1) the variance explained ($R^2 = 0{,}702$) is over 70%, with a quite low p-value (4×10^{-4}).

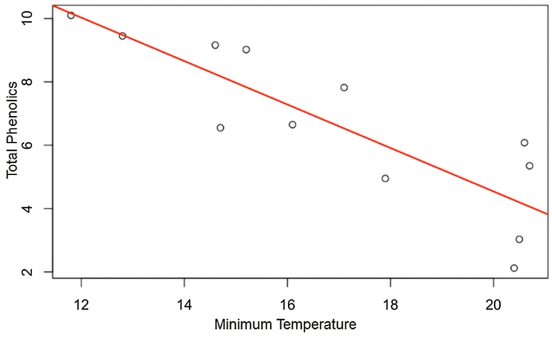


Fig. 1. Linear regression overlapping the data points (sample/month) of the variables total phenolic compounds (mg GAEq/g) *vs*. minimum temperature (°C). R^2 (0,702) is above 70%.

Further, the Pearson correlations for the dataset were calculated, corroborating the previous findings, where a significant relationship ($r = -0.854$) between the minimum temperature and the total phenolic compounds was detected (Fig. 2). In a similar statistical approach, the variables air moisture, total rainfall, and insolation (amount of solar energy/cm^2/min) did not show significant relationship with the biochemical variables.

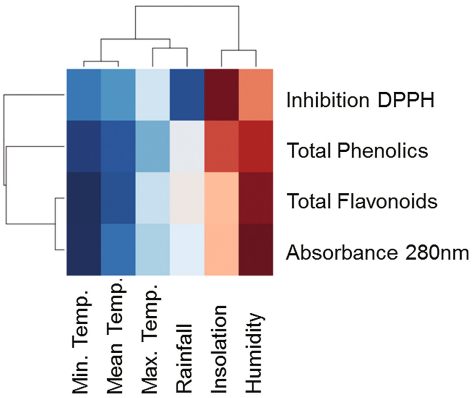


Fig. 2. Pearson correlation between biochemical and climatic variables. The correlation increases in blue and decreases in red.

In a second stage, the UV-Vis dataset (λ = 200–800 nm) of the AE's was pre-processed, where an offset correction and smoothing were applied. All the spectral profiles (λ = 200–800 nm) of AE's showed absorbances unity (Au) in the spectral window typical of phenolics (λ = 280–320 nm), indicating that the extraction system was able to recover the secondary metabolites from the residual biomass. In addition, the spectral profiles were similar, suggesting a homogeneous chemical composition among samples over the seasons (Fig. 3).

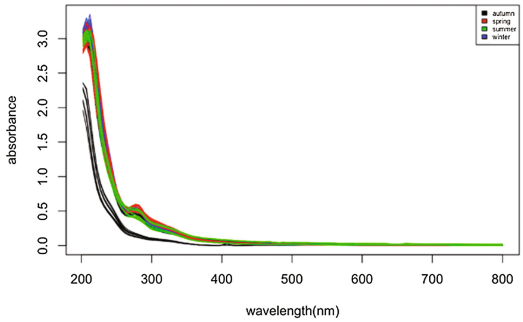


Fig. 3. UV-Vis spectroscopic profiles (λ = 200–800 nm - Au) of 12 representative samples of aqueous extracts of banana peels, collected during the seasons of 2015 (summer, fall, winter and spring) and 2016 (summer) in southern Brazil.

Further, PCA of the spectroscopic profiles, an unsupervised multivariate statistical technique, revealed a clear seasoning effect on the grouping of samples. PC1 (58.1%) and PC2 (37.3%) comprised 95.4% of the total variance of the dataset, making possible to explain the data variability with a few latent orthogonal variables. Overall, the results showed a certain degree of discrimination among the samples over the seasons regarding their chemical composition (Fig. 4).

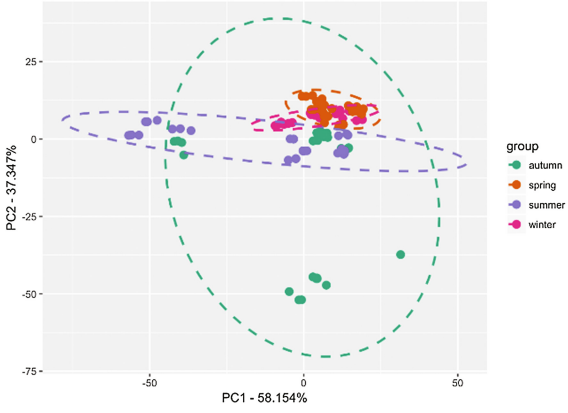


Fig. 4. Score scatter plot of the UV-Vis spectral data (λ = 200–800 nm) on the PC1 and PC2 axes of samples (n = 108) of aqueous extracts of banana peels. Each set of points of the same coloration represents a season.

In a follow-up experiment, hierarchical cluster analysis (HCA) was applied to the spectroscopic dataset and a better sample discrimination seems to be found in comparison to PCA. Again, the findings revealed that unsupervised methods end up merging the data into a pattern that fits well into the natural groups in the sample season (Fig. 5).

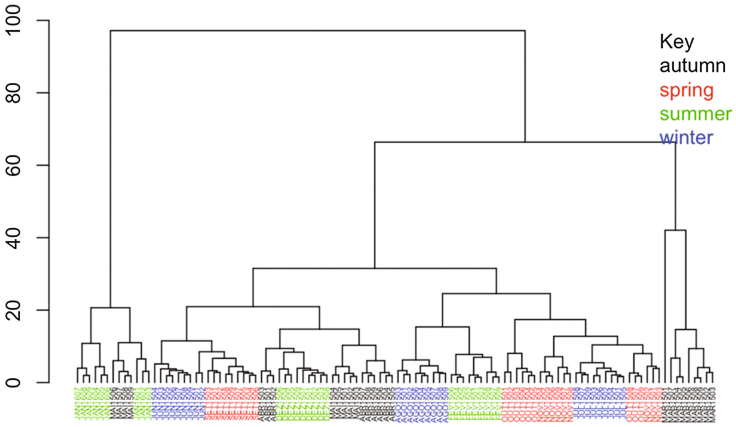


Fig. 5. Hierarchical cluster analysis (HCA) of the UV-Vis absorbances (λ = 200–800 nm) of aqueous extracts of banana peels.

Since phenolic compounds are usually the majoritarian ones in AE's of banana peels, we further performed PCA and HCA aiming at to extract additional information correlating the seasonality with the maximum absorption of peaks of those secondary metabolites (λ = 280–320 nm). In the PCA, PC1 (96.9%) and PC2 (2.8%) explained 99.7% of the total variance of the data set, but an improved sample discrimination regarding the full spectroscopic dataset (λ = 200–800 nm) was not achieved.

Thus, HCA using the same spectral dataset was done affording a better discrimination (Fig. 6). From the root, the first split separates the MAR/2015 samples (fall season) from the remaining, which corroborates the biochemical assays where the lowest amounts of secondary metabolites were found. Navigating the tree downwards, the samples from the second group are clustered by similarity in 3 groups, which have a majority of winter, summer and spring samples in each. Winter samples are grouped integrally, giving a similarity between the JUN/15, JUL/15 and AUG/15

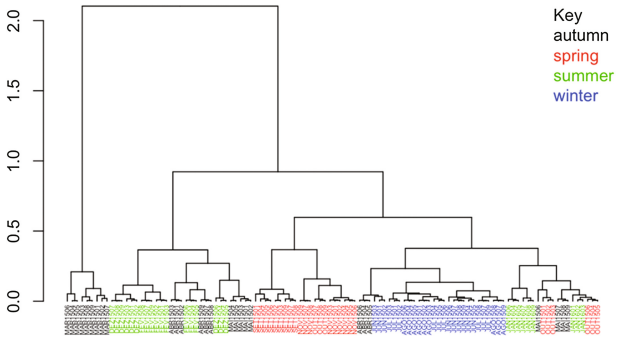


Fig. 6. Hierarchical cluster analysis of the UV-Vis absorbance region of phenolic compounds (λ = 280–320 nm) of crude aqueous extract of banana peels.

samples with respect to the spectral region of interest. In the right, we can see a cluster that mixes summer, autumn, and spring samples.

4 Conclusions

The analytical approach employed in this work, supported by bioinformatics tools, allowed a better understanding of the chemical variability of the banana peels collected during seasons associated to the climatic factors. Low temperatures typically found in the winter were determinant to modulate the banana metabolism for the production of increased amounts of phenolic compounds, also improving the antioxidant activity of AE's. Thus, PCA and HCA demonstrated a discrimination of samples collected in the winter as promising for future biological studies. Besides, HCA allowed identifying autumn-collected samples, i.e., MAR/2015, which showed reduced amounts of bioactive secondary metabolites.

Acknowledgements. To CNPq (National Counsel of Technological and Scientific Development) for financial support (Process nº 407323/2013-9), to CAPES (Coordination for the Improvement of Higher Education Personnel), and EPAGRI (Agricultural Research and Rural Extension Company of Santa Catarina). The research fellowship from CNPq on behalf of M. Maraschin is acknowledged. The work is partially funded by Project PropMine, funded by the agreement between Portuguese FCT and Brazilian CNPq.

References

1. Anyasi, T.A., Jideani, A.I.O., Mchau, G.A.: Morphological, physicochemical, and antioxidant profile of noncommercial banana cultivars. Food Sci. Nutr. **3**, 221–232 (2015)
2. EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária), Brasília, Brazil. https://www.embrapa.br/mandioca-e-fruticultura/cultivos/banana. Accessed 08 Feb 2017
3. Vu, H.T., Scarlet, C.J., Vuong, Q.V.: Optimization of ultrasound-assisted extraction conditions for recovery of phenolic compounds and antioxidant capacity from banana (*Musa cavendish*) peel. J. Food Process. Preserv. **40**, 1–14 (2016)
4. Yuan, Y., Zhao, Y., Yang, J., Jiang, Y., Lu, F., Jia, Y., Bao, Y.: Metabolomic analyses of banana during postharvest senescence by ^{1}H-high resolution-NMR. Food Chem. **218**, 406–412 (2017)
5. Tsamo, C.V.P., Andre, C.M., Ritter, C., Tomekpe, K., Newilah, G.N., Rogez, H., Larondelle, Y.: Characterization of *Musa* sp. fruits and plantain banana ripening stages according to their physicochemical attributes. J. Agric. Food Chem. **62**, 8705–8715 (2014)
6. Pereira, A., Maraschin, M.: Banana (*Musa* sp.) from peel to pulp: Ethnopharmacology, source of bioactive compounds and its relevance for human health. J. Ethnopharmacol. **160**, 149–163 (2015)
7. Tsamo, C.V.P., Herent, M., Tomekpe, K., Emaga, T.H., Quetin-Leclercq, J., Rogez, H., Larondelle, Y., Andre, C.: Phenolic profiling in the pulp and peel of nine plantain cultivars (*Musa* sp.). Food Chem. **167**, 197–204 (2015)

8. Pereira, A.: Determinação do perfil químico e da atividade cicatrizante de extratos de casca de banana cultivar prata anã (Musa sp.) e o desenvolvimento de um curativo para pequenas lesões. 223 f. Tese (Doutorado em Biotecnologia e Biociências) – Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina (2014)
9. Randhir, R., Shetty, P., Shetty, K.: L-DOPA and total phenolic stimulation in dark germinated fava bean in response to peptide and phytochemical elicitors. Process Biochem. **37**, 1247–1256 (2002)
10. Woisky, R.G., Salatino, A.: Analysis of propolis: some parameters and procedures for chemical quality control. J. Apic. Res. **37**, 99–105 (1998)
11. Ribeiro, S.M.R., Barbosa, L.C.A., Queiroz, J.H., Knodler, M., Schieber, A.: Phenolic compounds and antioxidant capacity of Brazilian mango (*Mangifera indica* L.) varieties. Food Chem. **110**, 620–626 (2008)
12. Costa, C., Maraschin, M., Rocha, M.: An R package for the integrated analysis of metabolomics and spectral data. Comput. Methods Programs Biomed. **129**, 117–124 (2015)
13. Hanson, A.B.: ChemoSpec: an R package for chemometric analysis of spectroscopic data and chromatograms (Package Version 1.51-0) (2012)
14. Beleites, C.: Import and export of spectra files. Vignette for the R package hyperSpec (2011)

Influence of Solar Radiation on the Production of Secondary Metabolites in Three Rice *(Oryza sativa)* Cultivars

Eva Regina Oliveira[1(✉)], Ester Wickert[2], Fernanda Ramlov[1], Rodolfo Moresco[1], Larissa Simão[1], Bruno B. Navarro[1], Claudia Bauer[1], Débora Cabral[1], Miguel Rocha[3], and Marcelo Maraschin[1]

[1] Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianopolis, Brazil
ginagro@gmail.com
[2] Santa Catarina State Agricultural Research and Rural Extension Agency (EPAGRI), Experimental Station of Itajaí, Santa Catarina, Brazil
[3] Centre Biological Engineering, School of Engineering, University of Minho, Braga, Portugal

Abstract. Rice (*Oryza sativa* L.) is one of the most produced and consumed cereals worldwide and has its importance highlighted mainly in developing countries, where it plays a strategic economic and social role. Due to the importance of rice in the diet, its composition and nutritional characteristics are directly related to the health of the population. In the rice production systems, some climatic factors are determinants for the good performance of the crop, inducing the biosynthesis of primary and secondary metabolites. The present study determined the metabolic profiles through UV-visible spectrophotometry of leaf samples of three rice cultivars (Marques – white, Ônix – black, and Rubi – red pericarp) throughout the rice's vegetative stages in two experimental times, from September to December 2015 and from January to April 2016. Solar radiation was recorded along the experimental period. To the organosolvent extracts of leaf samples, UV-vis spectrophotometric techniques were applied and the quantitative results of certain metabolites, e.g., chlorophylls, carotenoids, phenolics, flavonoids, and sugars, as well the antioxidant activity, which were analyzed by chemometrics tools. The results showed that biochemical parameters carotenoids, chlorophylls and sugars are more affected by the intensity of the radiation do que as variáveis phenolics, flavonoids and these alterations may be detected through statistical analysis of biochemical concentrations and UV-vis spectra.

Keywords: Rice · Spectroscopy · Metabolic profiles · Statistical models · UV-vis spectrophotometry

1 Introduction

Rice (*Oryza sativa* L.) is one of the most produced and consumed cereals in the world, being socially and economically important mostly in developing countries [1]. Due to

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_36

the importance of rice in the diet, its composition and nutritional characteristics are directly related to the health of the population. This cereal is capable of supplying 20% of energy and 15% of the daily need of an adult's protein, as well as containing vitamins, lipids, minerals, phosphorus, calcium, and iron [2].

In Brazil, where the annual consumption is on average 25 kg/inhabitant [3], the southern region accounts for most of the national rice production [4]. In the state of Santa Catarina, the guarantee of the economic viability of the pre-germinated rice crop results from relevant technologies developed by public research and rural extension efforts, notably based on the actions of the Agricultural Research and Extension Company of Santa Catarina (EPAGRI). Two new cultivars of rice were introduced by EPAGRI, with peculiar characteristics that, besides the nutritional attributes of the traditional grains (white), are characterized by the accumulation in the pericarp of pigments of great nutraceutical importance [5, 6]. Thus, the cultivars Rubi (red pericarp) and Ônix (black pericarp) are considered special due to the coloring of the grains, attributed to the presence of compounds beneficial to health [7, 8]. In the production, climatic factors, isolated or in association, are determinants for the good performance of the rice culture [9] and production of primary and secondary metabolites. In this sense, two important factors are the temperature and the average insolation over the growth stages of the plants [10].

The present study determined the metabolic profiles of leaf samples of three rice varieties developed by EPAGRI along the vegetative stage in two periods: (*i*) September to December 2015 (spring – summer, southern Brazil) and (*ii*) January to April 2016 (summer – autumn). Insolation, i.e., the amount of solar energy/cm^2/min reaching the leaf surface, has been daily measured over the experimental period and was further correlated with the metabolic profiles through chemometrics tools. The biochemical and climatic datasets were further related aiming to build statistical models to better understand the regulatory effect of the solar radiation on the *O. sativa* secondary metabolism. For that, spectrophotometric techniques were adopted, since the UV-vis spectrophotometry allows the rapid and low cost acquisition of qualitative and quantitative data from the plant metabolism whose contents can be altered in response to external stimuli. To the biochemical dataset, bioinformatics tools developed by our research group were applied, using multivariate statistical techniques as further described.

2 Materials and Methods

2.1 Biological Material

In a greenhouse at EPAGRI, Itajaí Experimental Station (26°57′57″S and 48°48″01′W, southern Brazil), pre-germinated rice seeds of three varieties were sown: Rubi (red pericarp), Onyx (black pericarp), and Marques (white pericarp) in two periods: (*i*) September to December-2015 and (*ii*) January to April-2016. Along the vegetative stage of the plants, samples of adult leaves were collected in regular intervals (3) and taken to the laboratory for the biochemical analyzes.

The radiation data for the studied period were provided by the Information Center for Environmental Resources and Hydrometeorology of Santa Catarina (CIRAM/EPAGRI).

2.2 Biochemical Analyzes - Total Phenolic and Flavonoid Compounds and Antioxidant Activity (DPPH Assay)

Samples of rice leaves (1 g, fresh weight, n = 4) were macerated in crucible with liquid N_2 and added of 5 V methyl alcohol (MeOH). The organosolvent extract was recovered by filtration on cellulose filter under vacuum, followed by the biochemical analyzes. The total content of phenolic compounds was determined by the Folim-Ciocalteau colorimetric method [11], recording the absorbance of the reactions in an UV-visible spectrophotometer (Gold Spectrum lab 53 UV-Vis spectrophotometer, BEL photonics, Brazil) at $\lambda = 750$ ηm.

To determine the total flavonoid contents, the methodology described by Zacarias *et al*. (2007) was adopted, with modifications. An aliquot of 0.5 mL of the MeOH extract was added to 0.5 mL of methanolic aluminum chloride solution (2% w/v) and to 2.5 mL analytical standard ethanol. After one hour of incubation, the absorbance was measured at 420 ηm. The results were expressed as mg of quercetin per g of dry mass.

The reduction potential of the DPPH radical by the MeOH extracts of the leaf samples was determined as described by Kim et al. (2002). To that end, the absorbance of a DPPH methanolic solution (1 mM in 80% methanol) was measured at wavelength 530 nm. The DPPH-methanolic extract mixture was incubated for 30 min in the dark and the antioxidant reaction measured at 530 ηm.

2.3 Extraction and Quantification of Total Chlorophylls and Carotenoids

Rice leaf samples (100 mg, fresh weight, n = 4) were incubated in a water bath at 65° C with 7 mL dimethylsulfoxide (DMSO) for two hours. The extract was recovered by filtration and the final volume adjusted to 10 mL with DMSO (Hiscox & Israelstam, 1979). The absorbance values at $\lambda = 480$, 649, and 665 ηm were obtained through an UV-vis spectrophotometer (Gold Spectrum lab 53 UV-Vis spectrophotometer, BEL photonics, Brazil). For purpose of calculation of the amounts of chlorophylls *a* and *b,* the *Wellburn* formulas (1994) were used, being the data expressed as mg/g dry mass.

2.4 Extraction and Quantification of Total Soluble Sugars

The extraction of soluble sugars was done as proposed by Shannon (1968). The rice leaf samples (100 mg, fresh weight, n = 4) were crushed in liquid nitrogen and macerated in MCW solution (methanol: chloroform: water, 12:5:3, v/v/v). Total soluble sugars were measured according to Umbreit & Burris (1964). The absorbance readings were taken at 630 ηm in an UV-vis spectrophotometer (UV-2000A, Instrutherm). The content of total soluble sugars was calculated from the standard glucose curve (1 to 200 μg mL^{-1}, $y = 0.008x$, $r^2 = 0.99$). The results were expressed as mg glucose per g dry mass.

2.5 UV-Visible Scanning Spectrophotometry

DMSO extracts of leaf samples were UV-vis scanned (Gold Spectrum lab 53 UV-Vis spectrophotometer, BEL photonics, Brazil) in their absorbances over the spectral window ($\lambda = 480 - 665$ ηm). The data set was exported as a.*csv* file format for further chemometrics analysis.

2.6 Statistical and Chemometric Analysis

The biochemical and UV-vis data sets of the leaf extracts investigated were processed considering the respective wavelengths of interest. Further, the data matrix was exported as a.*csv* format file and subjected to univariate and multivariate statistical analysis, using principal component analysis (PCA). PCA can help one to extract relevant features from a given dataset, minimizing the redundant information and characterizing the relationship between the variables studied.

For that, scripts were written in R language using tools defined by our research group, through the *specmine* package, and some functions from the packages Chemospec [11] and HyperSpec [12]. The scripts, raw data, and chemometrics analysis are available in supplementary material, at http://darwin.di.uminho.pt/pacbb2017/rice-cultivars. The report of analysis generated from the scripts provided by the R Markdown is also available at this site, allowing the computational experiments details to be analysed in detail and fully reproducible.

3 Results and Discussion

The results from the spectroscopic and biochemical analyzes of the primary (sugars) and secondary (chlorophylls, carotenoids, phenolic compounds, and flavonoids) metabolites, as well as the antioxidant activity allowed identifying discrepancies of leaf's metabolic profiles of the three varieties investigated regarding the effect of accumulated solar radiation and the average daily radiation over each experimental interval studied, *i.e.*, September to December-2015 and January to April-2016.

The one-way analysis of variance (ANOVA) of the biochemical data revealed discrepancies ($p < 0.05$) among the rice varieties, mostly for the contents of phenolic and flavonoid compounds, followed by the antioxidant activity (DPPH assay), sugars, and chlorophylls. On the other hand, the rice genotypes do not differ significantly in their carotenoids concentrations over the years (see report in supplementary material for the details).

For radiation data, linear regression analysis was performed first starting with the mean daily radiation, and then considering their accumulated values. The most visible effects occurred in the variables carotenoids, chlorophyll, and soluble sugars, followed by DPPH Inhibition. For carotenoids, the R2 values show that over one third of the variance can be explained by the radiation levels (regression analysis results are available in supplementary material). The phenolics and flavonoids showed high p-values, thus do not seem to be affected by the radiation levels.

In a follow-up experiment, PCA was applied to the biochemical data aiming to discriminate the rice genotypes. PCA shows that mostly of the data set variability (65.6%) has been explained by the first two principal components.

In this analysis, the variables contents of sugars, phenolics, and flavonoids, as well as the inhibition (%) of DPPH are in line with PC1, whereas chlorophylls and carotenoids spread over the PC2 axis. The results revealed a clear separation of the genotypes according to their metabolic profiles. Ônix samples grouped in PC1+/PC2+, influenced by the higher concentration of sugar. On the other hand, Rubi genotype grouped in PC1− due to their higher amounts of chlorophylls, carotenoids, phenolic compounds and higher inhibition activity of the DPPH radical. The Marques variety was found between the groups of the other two cultivars at PC1+ and PC2− (Fig. 1).

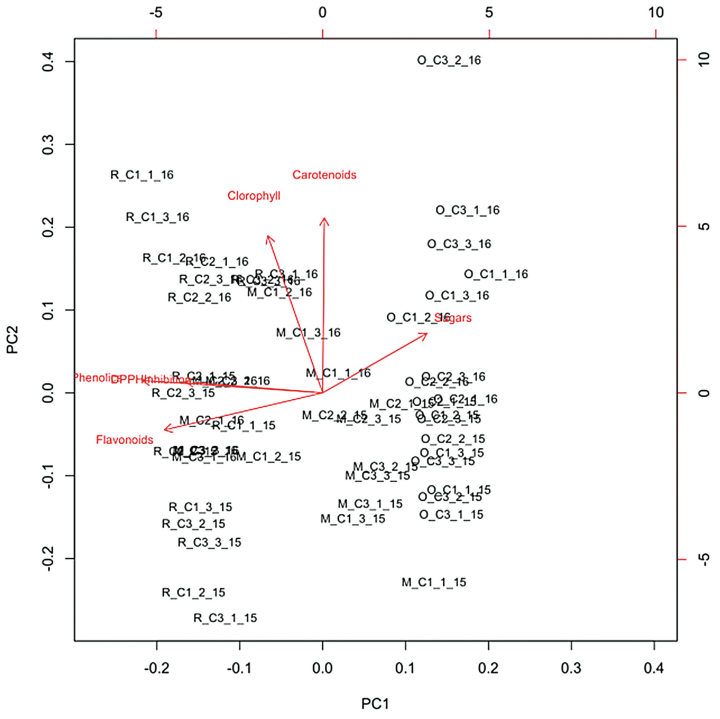


Fig. 1. Resulting bi-plot of the PCA results (PC1-39.3% and PC2-26.3%) showing the quantitative data variables (carotenoids, chlorophylls, phenolics, flavonoids, and inhibition of DPPH radical) in red, and the different scores of the samples (the reference is given in black for each sample).

In order to obtain a better understanding of the data dispersion (rice harvest seasons in southern Brazil), the results of the PCA were further analysed, considering the effects of the years of collection on those variables (Fig. 2A). The most of the 2016-collected samples grouped in PC2−, as the opposite has been detected for the samples collected in 2015. In a second approach, PCA results were interpreted aiming to correlated them

with the rice genotypes. Interestingly, as already seen above, PCA showed marked discrimination between Ônix (black pericarp, PC1+) and Rubi (red pericarp, PC1−) genotypes over the PC1 axis, while Marques (white pericarp) appears to be intermediate (Fig. 2B).

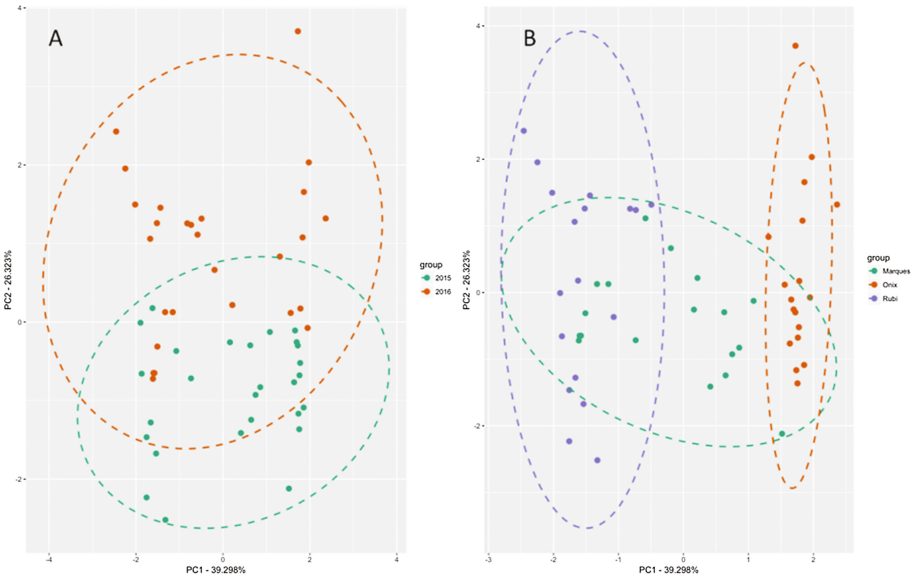


Fig. 2. Principal component analysis scoring scatter plots showing the effects of the year of collection on the biochemical variables of rice leaves (A) and among the rice varieties Marques, Ônix, and Rubi (B).

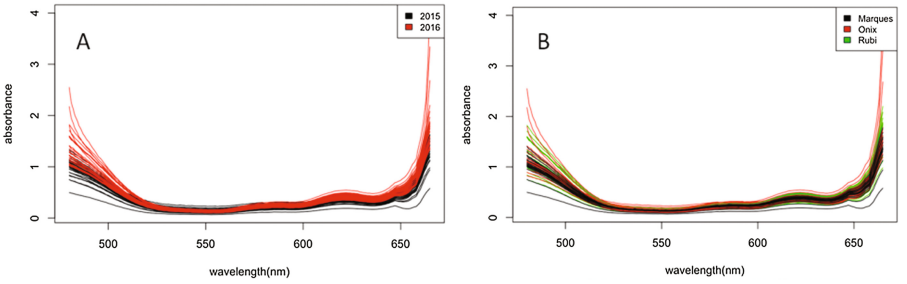


Fig. 3. UV-vis spectrophotometric profiles ($\lambda = 480 - 665$ ηm, DMSO) of leaf samples of rices genotypes. **A** – years 2015 and 2016. **B** – cultivars Marques, Onix, and Rubi.

Regarding the UV-vis spectroscopic profiles ($\lambda = 480 - 665$ ηm), a general vision to the class and contents of secondary metabolites is allowed, also revealing differences resulting from the genotypes and harvest times. All the studied samples showed intensive absorbance signals in the corresponding wavelengths of chlorophylls, carotenoids, and anthocyanins, with higher peaks for the 2016-harvest samples (Fig. 3A). Among the rice

genotypes, higher amounts were found in the Rubi leaf samples, followed by Ônix and Marques (Fig. 3B).

Taking into account the similarity of the UV-vis profiles among the samples and the eventual occurrence of redundant information, PCA was adopted again as a data reduction technique, in order to extract latent information from the spectroscopic data set. Again, the UV-vis spectral profile of 2015-collected samples seems to differ from that of 2016-collected ones (Fig. 4A), as a less clear separation has been found for the rice genotypes through the spectroscopic data set (Fig. 4B).

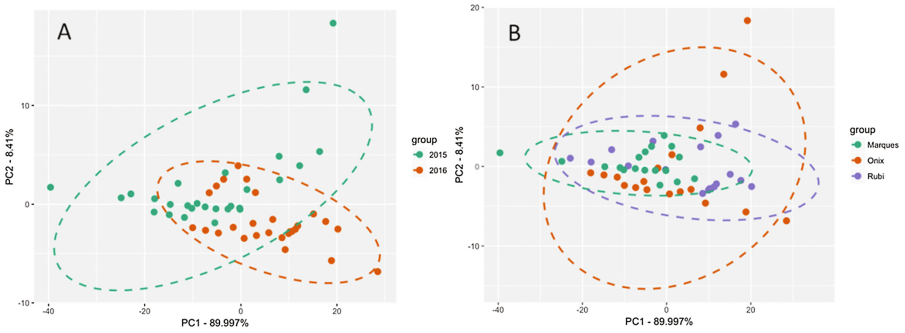


Fig. 4. Principal component analysis scores scatter plots (principal components 1 and 2) of the spectral data set (UV-vis, $\lambda = 480 - 665$ ηm, DMSO extract) colored according to the year (2015- and 2016-collected samples) (A) and to the rice genotypes Marques, Onix and Rubi.

4 Conclusions

In rice cultivation, climatic factors such as temperature and levels of solar radiation are determinant for the yield of the crop. Biochemical parameters may reflect possible physiological changes, *e.g.,* energetic and metabolic gains of plants throughout the crop cycle. The results obtained from the biochemical and UV-vis spectroscopic analyzes revealed the influence of the solar radiation on the metabolic profiles of the rice cultivars investigated.

For example, higher levels of chlorophyll, carotenoids, and sugars, important compounds associated to the photosynthetic apparatus, were shown in the 2016 harvest, when the solar radiation accumulated was larger than that found in 2015. Additionally, the rice genotypes respond differently to the insolation as noted for their discrepant secondary metabolites composition over the years. The chemometrics approach adopted allowed us to better discriminate the genotypes behavior over the years, by applying unsupervised multivariate statistical methods to the biochemical and UV-vis spectroscopic dataset. Taken together, the PCA findings suggest that different compounds may be used for building statistic monitoring models to better understand the rice genotypes answers to the solar radiation over the harvesting times in southern Brazil.

Acknowledgements. To CNPq (National Counsel of Technological and Scientific Development) for financial support (Process no. 407323/2013-9), to CAPES (Coordination for the Improvement of Higher Education Personnel), and EPAGRI (Agricultural Research and Rural Extension Company of Santa Catarina). The research fellowship from CNPq on behalf of M. Maraschin is acknowledged. The work is partially funded by Project PropMine, funded by the agreement between Portuguese FCT and Brazilian CNPq.

References

1. Marchezan, W.M., Avila, E., Antonio, L.: Arroz: composição e características nutricionais. Ciência Rural **38**, 1184–1192 (2008)
2. FAO. Food and Agriculture Organization of the United Nations, Rome, Italy. http://www.fao.org. Access 10 June 2015
3. MAPA: www.agricultura.gov.br. Access 21 Aug 2016
4. Klering, E.V., et al.: Modelagem agrometeorológica do rendimento de arroz irrigado no Rio Grande do Sul. Pesquisa Agropecuária Bras. **43**, 549–558 (2008)
5. Walter, M., et al.: Antioxidant properties of rice grains with light brown, red and black pericarp colors and the effect of processing. Food Res. Int. **50**, 698–703 (2013)
6. Zhang, M.W., et al.: Phenolic profiles and antioxidant activity of black rice bran of different commercially available varieties. J. Agric. Food Chem. **58**, 7580–7587 (2010)
7. Zhou, Z., et al.: The distribution of phenolic acids in rice. Food Chem. **87**, 401–406 (2004)
8. Muntana, N., et al.: Study on total phenolic contents and their antioxidant activities of Thai white, red and black rice bran extracts. Pak. J. Biol. Sci. **13**, 170–176 (2010)
9. Nam, S.H., et al.: Antioxidative, antimutagenic, and anticarcinogenic activities of rice bran extracts in chemical and cell assays. J. Agric. Food Chem. **53**, 816–822 (2005)
10. Morimitsu, Y., et al.: Inhibitory effect of anthocyanins and colored rice on diabetic cataract formation in the rat lenses. In: International Congress Series, p. 503–508. Elsevier (2002)
11. Costa, C., Maraschin, M., Rocha, M.: An R package for the integrated analysis of metabolomics and spectral data. Comput. Methods Programs Biomed. **129**, 117–124 (2015)
12. Randhir, R., Preethi, S., Kalidas, S.: L-DOPA and total phenolic stimulation in dark germinated fava bean in esponse to peptide and phytochemical elicitors. Process Biochem. **37**, 1247–1256 (2002)

Cryfa: A Tool to Compact and Encrypt FASTA Files

Diogo Pratas(✉), Morteza Hosseini, and Armando J. Pinho

IEETA, University of Aveiro, Aveiro, Portugal
{pratas,seyedmorteza,ap}@ua.pt

Abstract. NGS (next-generation sequencing) is bringing the need to efficiently handle large volumes of patient data, maintaining privacy laws, such as those with secure protocols that ensure patients DNA confidentiality. Although there are multiple file representations for genomic data, the FASTA format is perhaps the most used and popular. As far as we know, FASTA encryption is being addressed with general purpose encryption methods, without exploring a compact representation. In this paper, we propose Cryfa, a new fast encryption method to store securely FASTA files in a compact form. The main differences between a general encryption approach and Cryfa are the reduction of storage, up to approximately three times, without compromising security, and the possibility of integration with pipelines. The core of the encryption method uses a symmetric approach, the AES (Advanced Encryption Standard). Cryfa implementation is freely available, under license GPLv3, at https://github.com/pratas/cryfa.

Keywords: FASTA encryption · AES · Cryptography · Compression · DNA sequences

1 Introduction

The emergence and advances in the next-generation sequencing (NGS) provided a way to access genome information, at a nucleotide level, namely to identify and study evolutionary events, as well as alterations for clinical purposes [1]. The sensitivity of the data shows the importance for efficiently preserve confidentiality, namely through privacy protocols and methods, such as cryptography [2].

A DNA sequence is a succession of letters, with four possible outcomes (A,C,G,T), that indicate the order and nature of nucleotides within a DNA chemical chain. The process of unveiling the chain is known as DNA sequencing. This process can be seen as a capture of small pieces from a huge puzzle with lots of repeated, changed and missing pieces. Therefore, to obtain a *complete* species sequence, several stages must first be applied, such as the assembly, causing the insertion of several errors or unknown symbols in the final sequences. Unknown symbols are usually represented by 'N' symbols.

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_37

The representation of DNA sequences for storage and analysis purposes is usually set on a specific file format: FASTA. The FASTA files usually contain DNA sequences in an assembled or pre-assembled state. Explicitly, the FASTA format is a text-based format for representing two information streams: DNA sequences and headers. The headers stand for sequence names and comments that precede the sequences. An example of a hypothetical FASTA file is given in Fig. 1.

```
>gi|528476558|ref|NC_018928.2| Homo sapiens chromosome 17
TCCACTCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTC
AGCACATCAGCTCTGAGGCCTCCACTCACTGAGACAATAGACCCAAG
TCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTCCACTC
AACAGCTCTGAGGGCTCCACTCACTGAGACAACAGACCCAAGAACAA
>gi|528476670|ref|NC_018912.2| Homo sapiens chromosome 1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTA
TAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCAA
```

Fig. 1. Example of a hypothetical FASTA file content.

As it can be seen in Fig. 1, the headers always begin with a '>' character followed by some description and with a '\n' character at the end (new line). The DNA sequences are usually restricted to the alphabet {A,C,G,T,N}, although very rarely, there are other symbols, such as {\n,H,M,P,R,T,X,Y,a,c,g,t,n,m}.

Mainly, the compression of FASTA files has been addressed by Deliminate [3], BIND [4], MFCompress [5], LEON [6] and MetaCRAM [7], where for only DNA sequences by XM [8] and GECO [9]. For a state-of-the-art in FASTA and DNA sequence compressors see [10]. Although a compressor efficiency is dependent on the nature of the data, it seems that for FASTA files Deliminate [3] and MFCompress [5] are consistently more efficient. On the other hand, for encrypting the FASTA files, up to our knowledge, there is not a specific method. Actually, FASTA encryption is being addressed with general purpose encryption methods.

In this paper, we consider the challenge of compacting and encrypting FASTA files using a fixed-block transformation followed by AES (Advanced Encryption Standard) a symmetric system with 128 bits. The AES, a variant of the Rijndael cipher [11], is based on a design principle known as a substitution-permutation network, a combination of both substitution and permutation. The AES operates on a 4×4 order matrix of bytes. The key size used for an AES cipher specifies the number of repetitions of transformation rounds that convert the plaintext into the ciphertext. The number of cycles of repetition is 10 cycles for 128-bit keys, 12 cycles for 192-bit keys and 14 cycles for 256-bit keys. For longer bit keys, the computation time needed is larger. As such, since the size of the files to encrypt is already large, we opted to use a 128-bit key. Nevertheless, we provide an implementation with an easy way to change to 256-bits.

The AES requires a separate 128-bit round key block for each round plus one more. An initial round is required, where each byte of the matrix is combined with a block of the round key using bitwise xor. Then it starts the phase of rounds, having the operations: sub bytes, a non-linear substitution where each byte is replaced by another giving a lookup table; shift rows, a circular shift given a certain number of steps of the last three rows of the matrix; mix columns, a mixing operation, combining the four bytes in each column; add round key, the subkey is combined with the matrix. After the corresponding number of cycles being applied, it uses a final round, applying the sub bytes, shift rows and add round key operations.

Some attacks are known, such as those based on brute-force, simple key schedule [12], side-channel attacks [13] - encryption after compression [14], among others. Nevertheless, with a proper methodology set, the best known attacks on AES-128 bits need billions of years using current hardware.

Specifically, our purpose is to compress, as efficiently as possible, and encrypt, as secure as possible, the FASTA files using efficient computational times. Since we are dealing with large files, should we compress the data after encryption? No, because the purpose of an AES encryption is to distribute the data uniformly, therefore it is useless. However, should we compress the data before encryption? This was believed to be true, until the appearance of a new type of attacks. This kind of attacks, such as CRIME and BREACH, explore the variable size of the encrypted text [15], given by its redundancy, to deduce the key. They had, for example, a huge impact on SSL/TLS applications [16]. On the AES, they can be used to reduce the time in a brute-force attack, as well as to estimate sequence redundancy. Therefore, in this paper we use a fixed size compaction, independently from its redundancy, to reduce the size of the representability of the DNA sequences and, then, encrypt them. We call compaction and not compression because we are reducing the storage without exploring redundancies of the data or else we would, as explained before, put the security of the files in risk.

The rest of the paper is organized as follows. In the next Sect. 2 we describe the method. In Sect. 3, we show the computational time needed and the improvement regarding space, comparing with a general purpose encryption and specific FASTA compressors. Finally, in Sect. 4 we make some conclusions.

2 Method

We have used the human GRC reference genome (build 37) in FASTA format to study the patterns of distribution of the existing characters. Figure 2 depicts these patterns, where we can see the symbols A,C,G,N,T and \n with higher large proportions relatively to the others. Therefore, we are able to explore these characteristics in the method.

Accordingly, in the following subsections, we describe the method used and how we have implemented it in a computational tool.

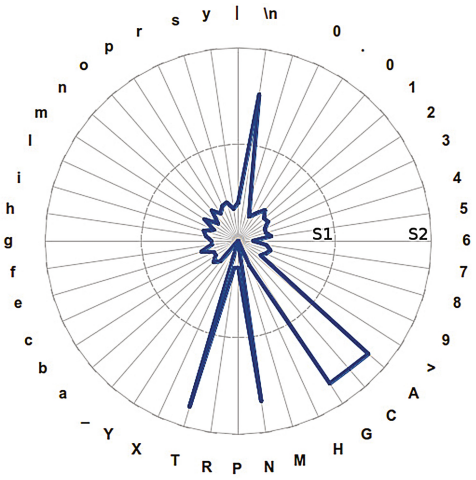


Fig. 2. Patterns of distribution of the complete assembled human genome (GRCv37) stored as a FASTA file. The scale gives the number of occurrences for each symbol in a logarithmic format. The S1 stands for 10^6, while S2 for 10^{10}. The blank place stands for a space (ASCII number 10).

2.1 Description

The core of the method involves several transformations of the data before the encryption, in order to reduce the file size. According to Fig. 3, the FASTA file is split into two streams, the headers and the DNA sequences. Then, the DNA sequences are transformed into packed sequences, that we further will explain. Finally, the streams are joined and the encryption is applied.

The packing transformation is applied to each triplet of DNA bases, where each one is converted into a number contained in the interval $[0; 6^3 - 1]$, giving the alphabet {A,C,G,T,N,X}. Any symbol outside the alphabet is mapped into an 'X'. After the numerical attribution, for each symbol that was mapped into an 'X', it is used an extra byte to describe which symbol was (given its ASCII representation). Since these are not frequent symbols, the output will have a small penalty according to its length.

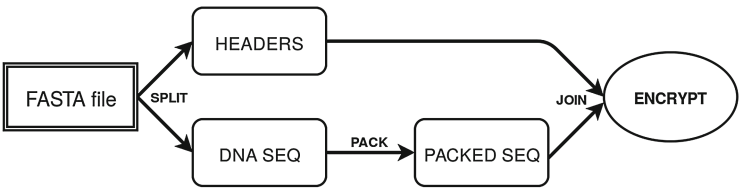


Fig. 3. Diagram showing the FASTA transformation phases (split, pack and join) before encryption.

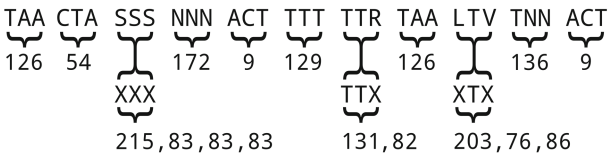


Fig. 4. Packing transformation for an exemplifying set of triplets of DNA bases. The characters outside {A,C,G,T,N,X} are transformed in X and, for each, an extra byte is spend to represent the character is ASCII mode.

Figure 4 shows and example of a DNA sequence being packed. It shows that the sequence, instead of using 264 bits for representation (33 symbols $\times$ 8 bits), needed only 136 bits (11 triplets $\times$ 8 bits + 6 extra $\times$ 8 bits). Notice that in this example the compaction factor was approximately two. The reason is that we have concentrated the infrequent symbols only for the example. Usually, these symbols are rare as it can be seen in Fig. 2. Therefore, asymptotically, the method has approximately a compaction factor of three.

Finally, the encryption is held, using AES-128 bits. For the purpose, a password is set (by the user) and hashed into a numerical value between $[0; 2^{64} - 1]$. The initialization of the matrix is also set pseudo-randomly according to a different hash function, given by a seed that provides from the password. Then, the encryption rounds starts, according to the phases described in the introduction. The final output is the encrypted text.

The decryption process is done using the reverse process, given its symmetric property. Briefly, the text is decrypted and, then, the unpacking is done using the reverse mapping of Fig. 4. As such, each number representing the triplet is converted to the three bases. If there are at least one X in each triplet, the corresponding characters will be read in order to disambiguate the original character.

2.2 Implementation

We have implemented the method in a fully automatic command line tool (Cryfa), written in C++ language, for multiple operating systems. The tool uses Crypto++, a free C++ library class for the cryptography AES scheme (https://www.cryptopp.com/). The tool can be applied to any FASTA file.

Cryfa writes the output data to standard output for a direct integration with bioinformatic analysis systems. Currently, the password is set by reading it from a file. Mainly, because interactive loading can be problematic in a pipeline and set the password as a command argument can be dangerous given the history of commands loading. Nevertheless, optimization of a more secure way will be a subject of further studies.

Cryfa can be downloaded, under GPLv3 license (free for research purposes), at https://github.com/pratas/cryfa.

3 Results

We have used two state-of-the-art FASTA compressors, Deliminate [3] and MFCompress [5], to compare the number of the bits and time need to represent several FASTA files. The FASTA files used are all the human GRC reference chromosomes (build 37) from NCBI [17], totaling approximately 3 GB of data. We have included a script, available at https://raw.githubusercontent.com/pratas/cryfa/master/scripts/run.sh, that allows replicating the results under a Linux OS.

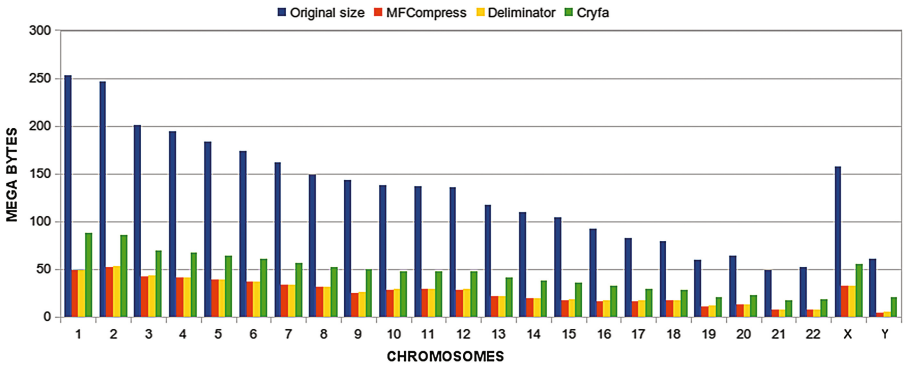


Fig. 5. Number of bytes needed for MFCompress, Deliminate and Cryfa methods to store each human chromosome. The original file size is included as a reference and it can be seen as an approximation of the number of bytes needed by a general purpose encryption method.

The Fig. 5 shows the number of bytes needed to represent each chromosome (FASTA file) according to different methods. We are able to see that MFCompress and Deliminate use fewer bits for storage than Cryfa. However, when compared with the original file size, Cryfa is relatively near state-of-the-art compressors. Notice that MFCompress and Deliminate do not encrypt the data. General purpose encryption methods represent the files with approximately the same size as the original size, given its uniform distribution randomization, therefore these values represent general purpose encryption without compaction.

The Fig. 6 shows the time needed for the mentioned methods to run on all chromosomes. As it can be seen, Cryfa is the second fastest method. Moreover, both MFCompress and Deliminate used parallelization, while Cryfa ran in a single core CPU. Furthermore, Cryfa, besides compaction, also encrypts the FASTA files, showing very fast running times.

Consider now that we would have information that a collection of bacteria, such as *Escherichia coli*, was sequenced and encrypted in different files according to an AES cipher. If we would apply a variable-size compression, for example using MFCompress to explore the redundancy of the data, and after an encryption, we would have a clue of what would be the most redundant/complex

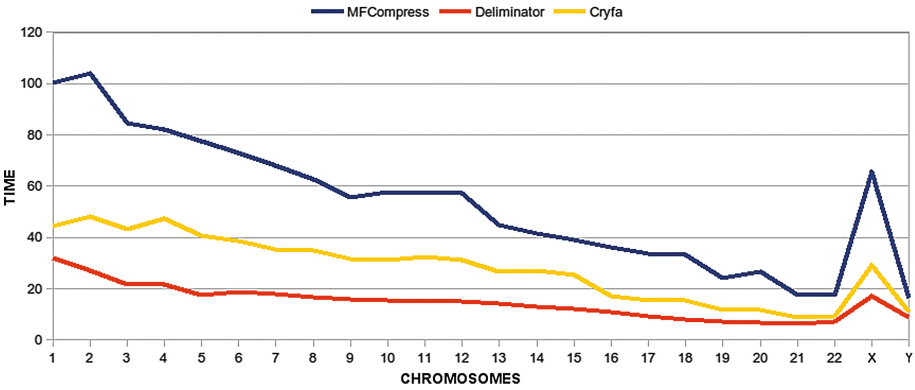


Fig. 6. Time needed for MFCompress, Deliminate and Cryfa methods to store each human chromosome. Time values are in seconds.

sequences. As such, we would have access to the content properties (an indication) without decryption. With the method that we present here, this does not happen, given its fixed block size compaction and, hence, we are not able to know the complexity of the files without decryption.

4 Conclusions

In order to preserve the confidentiality of DNA sequences, specifically FASTA files, we have proposed a method to encrypt efficiently the data. The public available implementation, Cryfa, compacts each triplet of DNA bases into one character, using a fixed block size packing, then it uses AES symmetric encryption. When compared with general encryption tools, it allows reducing the storage approximately three times, without creating security problems, such as those derived from compression before encryption. On the other hand, we have shown that, regarding compression, relatively to its original storage size it is not far from FASTA state-of-the-art compression methods. Moreover, it uses very fast processing times.

Low-data complexity has been explored in AES attacks [18]. Therefore, in future works, we will uniformly permute the plaintext, before encryption, to transform the plaintext into high-complexity data. Moreover, we will extend Cryfa to FASTQ files.

Acknowledgments. This work was partially funded by FEDER (Programa Operacional Factores de Competitividade - COMPETE) and by National Funds through the FCT - Foundation for Science and Technology, in the context of the projects `UID/CEC/00127/2013`, `PTCD/EEI-SII/6608/2014`.

References

1. Mardis, E.R.: The impact of next-generation sequencing technology on genetics. Trends Genet. **24**(3), 133–141 (2008)
2. Schlosberg, A.: Data security in genomics: a review of australian privacy requirements and their relation to cryptography in data storage. J. Pathol. Inform. **7**, 6 (2016)
3. Mohammed, M.H., Dutta, A., Bose, T., Chadaram, S., Mande, S.S.: DELIMINATE - a fast and efficient method for loss-less compression of genomic sequences. Bioinformatics **28**(19), 2527–2529 (2012)
4. Bose, T., Mohammed, M.H., Dutta, A., Mande, S.S.: BIND-an algorithm for loss-less compression of nucleotide sequence data. J. Biosci. **37**(4), 785–789 (2012)
5. Pinho, A.J., Pratas, D.: MFCompress: a compression tool for fasta and multi-fasta data. Bioinformatics **30**, 117–118 (2013)
6. Benoit, G., Lemaitre, C., Lavenier, D., Drezen, E., Dayris, T., Uricaru, R., Rizk, G.: Reference-free compression of high throughput sequencing data with a probabilistic de bruijn graph. BMC Bioinformatics **16**(1), 288 (2015)
7. Kim, M., Zhang, X., Ligo, J.G., Farnoud, F., Veeravalli, V.V., Milenkovic, O.: MetaCRAM: an integrated pipeline for metagenomic taxonomy identification and compression. BMC Bioinformatics **17**(1), 94 (2016)
8. Cao, M.D., Dix, T.I., Allison, L., Mears, C.: A simple statistical algorithm for biological sequence compression. In: Proceedings of the Data Compression Conference, DCC-2007, Snowbird, Utah, pp. 43–52, March 2007
9. Pratas, D., Pinho, A.J., Ferreira, P.: Efficient compression of genomic sequences. In: Proceedings of the Data Compression Conference, DCC-2016, Snowbird, Utah, pp. 231–240, March 2016
10. Hosseini, M., Pratas, D., Pinho, A.J.: A survey on data compression methods for biological sequences. Information **7**(4), 56 (2016)
11. Daemen, J., Rijmen, V.: AES proposal: Rijndael. EC (1999)
12. Biryukov, A., Khovratovich, D., Nikolić, I.: Distinguisher and related-key attack on the full AES-256. In: Halevi, S. (ed.) CRYPTO 2009. LNCS, vol. 5677, pp. 231–249. Springer, Heidelberg (2009). doi:10.1007/978-3-642-03356-8_14
13. Ashokkumar, C., Giri, R.P., Menezes, B.: Highly efficient algorithms for AES key retrieval in cache access attacks. In: IEEE European Symposium on Security and Privacy (EuroS&P), 261–275. IEEE (2016)
14. Gullasch, D., Bangerter, E., Krenn, S.: Cache games-bringing access-based cache attacks on AES to practice. In: IEEE Symposium on Security and Privacy, pp. 490–505. IEEE (2011)
15. Keerthi, V.K., et al.: Taxonomy of SSL/TLS attacks. Int. J. Comput. Netw. Inf. Secur. **8**(2), 15 (2016)
16. Meyer, C., Schwenk, J.: Lessons learned from previous SSL/TLS attacks-a brief chronology of attacks and weaknesses. IACR Cryptology ePrint Archive **2013**, 49 (2013)
17. Church, D., Deanna, M., Schneider, V., et al.: Modernizing reference genome assemblies. PLoS Biol. **9**(7), e1001091 (2011)
18. Bouillaguet, C., Derbez, P., Dunkelman, O., Fouque, P.A., Keller, N., Rijmen, V.: Low-data complexity attacks on aes. IEEE Trans. Inf. Theory **58**(11), 7002–7017 (2012)

An Automated Colourimetric Test by Computational Chromaticity Analysis: A Case Study of Tuberculosis Test

Marzia Hoque Tania[1(✉)], K.T. Lwin[1], Kamal AbuHassan[1], Noremylia Mohd Bakhori[2], Umi Zulaikha Mohd Azmi[2], Nor Azah Yusof[2,3], and M.A. Hossain[1]

[1] Anglia Ruskin IT Research Institute (ARITI), Anglia Ruskin University, Chelmsford, UK
marzia.hoque@pgr.anglia.ac.uk,
{khin.lwin,kamal.abu-hassan,alamgir.hossain}@anglia.ac.uk
[2] Institute of Advance Technology, Universiti Putra Malaysia, Serdang, Malaysia
noremyliamb@gmail.com, umizulaikha.ika@gmail.com,
azahy@upm.edu.my
[3] Department of Chemistry, Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia

Abstract. This paper presents an investigation into a novel approach for an automated universal colourimetric test by chromaticity analysis. This work particularly focuses on how a well-adjusted harmony between computational complexity and biochemical analysis can reduce the associated cost and unlock the limit on conventional chemical practice. The proposed research goal encompasses the potential to the criteria- anytime anywhere access, low cost, rapid detection, better sensitivity, specificity and accuracy. Our method includes obtaining the amount of colour change for each instance by delta E calculation. The system can provide the result in any ambient condition from the trajectory of colour change using Euclidean distance in LAB colour space. The strategy is verified on plasmonic ELISA based diagnosis of tuberculosis (TB). TB detection by plasmonic ELISA is a challenging, demanding and a time-consuming diagnosis. Completing the computation in real time, we circumvent the obstacle liberating the TB diagnosis in less than 15 min.

Keywords: Colourimetric test · Plasmonic ELISA · TB test · Chromatic analysis · Delta E

1 Introduction

The colourimetric test provides a decisive analysis for the present elements or concentration of chemical compound facilitated by a colour agent. The procedure can be inclusive or exclusive of the enzyme. When it comes to colourimetric assays for medical diagnosis, a wide range of rapid, visual readout, quantitative detection, low cost and robust system have been utilised in the literature [1, 2].

For quantification of colourimetric test, Yetisen et al. (2014) developed a cross platform smartphone application featuring interphone repeatability to quantify the concentration of glucose, protein, pH, replacing the requirement of spectrometer and microplate

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_38

reader [3]. The capturing process starts with calibration at given lighting condition followed by user input of sensor type, target analyte, unit of concentration and number of data points. Sample image of corresponding test zone is processed utilising electromagnetic radiation from coloured zone, concentration of the analyte and corresponding value on screen is returned, the image is transformed and the result is produced from measured value vs. calibration curve and at last the information is required to synchronise

The scope of colourimetric analysis is certainly not limited to health domain. The recent approaches for different applications include Rapid Diagnostic Test (RDT) to promote point-of-care (POC), opto-mechanical reader, paper based and Digital holographic microscopy (DHM) on the mobile platform with geo-tagging facility, hardware based colourimetric sensor array for medical diagnosis, molecular biology, detection of elements and monitoring environmental factors (such as air and water quality) [1, 4–6]. Shen et al. (2012) presented the potential of smartphone based colourimetric tests, not to eradicate the conventional method, but to provide portable, transferable, immediate, low-cost diagnosis to huge population with limited access [7]. They compensated ambient lighting environment and formed the calibration curve of concentration from the chromaticity value to measure pH. They envisioned their colour conversion analysis techniques to be useful to any POC diagnosis with colourimetric response, even for florescence data.

The colourimetric analysis is not a new concept. The wide-range of its application is clearly evident from the above discussion. With the recent advancement of mobile phone camera and demand for POC [8], there is a need for an one-size-fits-all approach to perform an automated universal colourimetric test with economic and technical feasibility, at anywhere and anytime. This work includes such an approach with a case study of TB test. However, this work does not focus on fluorescence test and ultraviolet (UV) method.

2 Literature Review

Globally, TB is one of the leading causes of death. In 2015, 35% of the HIV positive people died due to TB [9]. In the same year, 10.4 million reported to suffer from TB and TB was fatal to 17% of them. The Sustainable Development Goals (SDGs) for 2030 includes ‘End TB strategy’ to diminish the global epidemic of TB [10]. TB can be cured; due to in time adequate treatment 49 million deaths were prevented within 2000–2015. Several approaches exist for TB test either being expensive, time consuming or ineffectual. The conventional methods include sputum smear microscopy, which can take up to three days, rapid molecular tests (WHO recommended Xpert® MTB/RIF assay takes up to two hours [11]); culture methods taking up to 12 weeks. UK Visas and Immigration uses radiometric method for TB screening [12]. UK National Health Service (NHS) conducts chest X-ray, blood test and tuberculin skin test for different diagnosis of TB [13]. There are also lateral flow tests (LFTs) to diagnose TB.

Tsai et al. (2013) strategised a colourimetric sensing using unmodified gold nanoparticles (AuNps) and single-stranded detection oligonucleotides for TB test [14]. A

smartphone was utilised to collect the multiple detection results of colour variation from the concentration on cellulose paper and transmit to the cloud. The result showed, 2.6 nM tuberculosis mycobacterium could be detected by this method. The turnaround time was 1 h, after DNA extraction from the patient. The smartphone was used in here as a data sharing media, the RGB value was analysed by Java based open source software, Image J. Osman et al. (2010) proposed a tuberculosis bacilli detection technique from the tissue sample by Ziehl-Neelsen staining method [15]. The prepared sample image from optical microscope was segmented by moving k-mean clustering for tuberculosis bacilli extraction. Both RGB and C-Y colour were utilised to acquire a robust and improved segmentation under various staining condition. The hybrid multilayered perceptron network (HMLP) selected the feature among the geometrical features of Zernike moments to detect tuberculosis bacilli. The result showed 98.0%, 100% and 96.19% of accuracy, sensitivity and specificity respectively to find the class of definite and possible TB.

The challenging part of TB test is to acquire the required of sensitivity and specificity. Moreover, it has to be cost-effective for long-term post-treatment monitoring and infected population in developing countries. The World Health Organization (WHO) also prefers diagnostic tools which are inexpensive, disposable and easy-to-use [16, 17]. However, the smart devices are yet to earn its reputability for colourimetric analysis for its constraint on quantitative measurements [7]. The degree of freedom is limited by system provided small colour change and insufficiency of RGB intensity value. Like any other colour image processing, mitigating the impact of lighting condition in colourimetric analysis of diagnostic assays is to blame for cell phones not reaching its full potential. Contrariwise, the quest of providing properly distinguishable colour, complexity of chemical method might be minimised by a powerful algorithm for colour detection. Incorporation of mobile phone can not only facilitate easy and automatic colour detection but can also enable disease decision using machine learning techniques. Thus, this paper focuses on the algorithm to eliminate the dependency on lighting environment and superior camera.

3 Methodology

3.1 Plasmonic ELISA Test and Experimental Setup

Colourimetric analysis using plasmonic nanoparticles (NPs) are well utilised in the field of medicine and environmental monitoring by various accessories [18]. A batch of 96-well plates went through series of processes as illustrated in Fig. 1 for automatic diagnosis of TB based on plasmonic ELISA [19]. Three separate experiments were conducted on three sample plates (P_1, P_2, P_3) and all of the experiments were recorded. There were 31 samples in total- 24 in P_1, 1 in P_2 and 6 in P_3. Each column of P_1 varies in concentration. The videos were recorded using iPhone 7 plus (12 MP, wide-angle: *ƒ*/1.8 aperture, telephoto: *ƒ*/2.8 aperture) and iPhone 4 (5 MP).

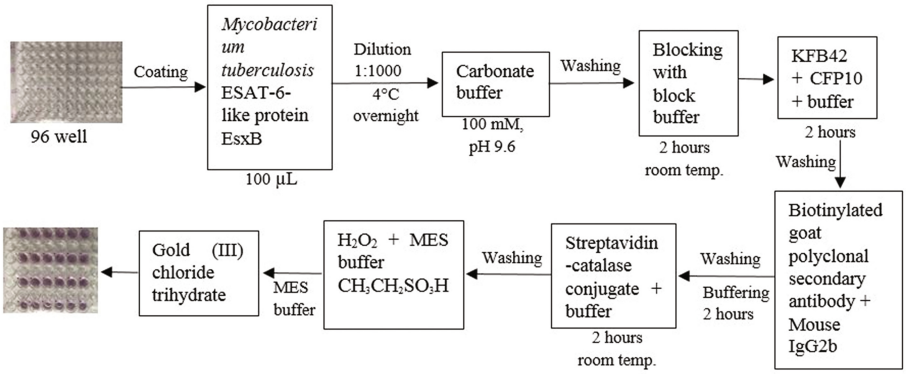


Fig. 1. Stepwise plasmonic ELISA based TB test

3.2 Image Processing and Chromatic Analysis

For chromatic analysis, the colour difference is measured by delta E (ΔE), to be specific CIE76. With known colour space coordinates, the International Commission on Illumination (CIE) 1976 formula delivers the colour difference. It was the first formula to provide ΔE in LAB.

In LAB colour space, if $\left(L_1^*, a_1^*, b_1^*\right)$ and $\left(L_2^*, a_2^*, b_2^*\right)$ are two colour coordinates at t_1 and t_2 s respectively, then colour difference is given by the following formulae [20],

$$\Delta E_{ab}^* = \sqrt{\{(L_2^* - L_1^*)^2 + (a_2^* - a_1^*)^2 + (b_2^* - b_1^*)^2\}} \tag{1}$$

During the complete reaction taking t seconds, the instances (S) of the reaction is divided in t_i time interval. If the ambient condition is consistent for the total test time (for each S), the impact of lighting environment will have a negligible significance on the system. Thus, the TB testing method can be implemented anywhere anytime. Later, a high pass filter is applied on the calculated colour difference for each instance.

$$\mathrm{S} = \{1, 2, \ldots, \mathrm{N}\}, \mathrm{S} \in \mathrm{Z} \tag{2}$$

$$\Delta E(S)_{ab}^* > JND, \text{ where } \Delta E(S)_{ab}^* = \sqrt{(\Delta L^{*2} + \Delta a^{*2} + \Delta b^{*2})} \tag{3}$$

$$\text{maximum amount of colour change at any instant, } \alpha = \max(\Delta E(S)_{ab}^*) \tag{4}$$

An imperative parameter for ΔE calculation is the just-noticeable difference (*JND*) or differential threshold. According to experimental psychology, "it is the amount something must be changed for a difference to be noticeable, detectable at least half the time''. As an intuitive value, some suggested ΔE to be 1.0, but a widely acceptable value is 2.3 [21, 22].

The objective of this experiment is to explore the time response on colour change. Primarily, the instantaneous time to attain maximum ΔE was calculated. The colour

difference can be calculated for other colour spaces as well. However, a validation is required for the colour space coordinates (L*, a*, b*) with classified labels. For concentration based quantitative and semi-quantitative colourimetric analysis, the transition should be linear, which can be further interpreted with the statistical model. In case of colour transformation based studies e.g. TB test, the transition phase is non-linear, where the number of transition phase,

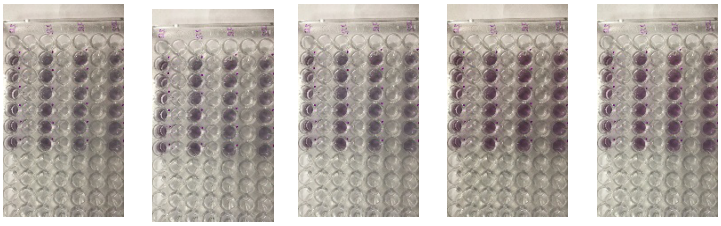
$$\phi = \{0, 1, 2, \ldots, N\}, \text{ where } \phi \in Z^*, Z^* = \{0\} \cup Z^+. Z^+ \text{denotes positive integer.} \quad (5)$$

4 Experimental Results and Discussion

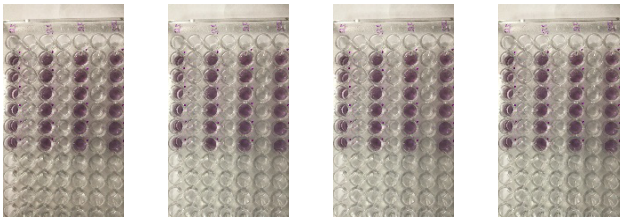
The total TB test was conducted in 845.4 s. The video was converted from MOV to JPEG based images taking time interval (t_i) as 0.0332 (every frame), 1, 5, 10, 20 and 50 s. The analysis can be performed by direct video acquisition as well. As the procedure relies on the difference, not the (x, y, z) values itself, both of the phone camera played adequate role. Thus, the dependency on high configuration camera can be avoided. All of the sample plate contained only one class. For sample plate containing mixture of positive and negative specimen, one would require to perform another simple step splitting the wells in separate images.

For P_1 with $t_i = 10$ s, the total number of images, $f(S) = 84$. Eliminating the number of images where the wells were being filled, number of images become 58. From Fig. 2, the change of colour is visible; however, data interpretation and quantification with precision is difficult. The amount of colour change (ΔE) is evident for each instance in Fig. 3(a). As a result, ΔE eased the shortcoming of naked eye. In this experiment, the sample began to turn into blue colour at 7.67 min (Fig. 2(b)). The colour change strongly continued for next ten seconds; the transition is evident at 7.83 min as well (Fig. 3(c)). However, α is achieved at 11.33 min, when the sample was turning into pink- which implies, rest of the experiment can be excluded. It can be stated with certainty that the time response analysis is capable of salvaging minimum 2.67 min (19% of the total time) of the full experiment time. It is almost impossible to visualise the transition from $S = 68$ to $S = 73$ (Fig. 2(d–i)) with naked eye, let alone quantification. Thus, Eq. (1) is utilised (Fig. 3(a)). From Fig. 3(b), it can be observed that $\phi = 2$. A hypothetical line is drawn for TB positive test for better illustration of the observation. $\Delta\phi$ is the time taken for the specimen to turn from blue to pink. Equation (5) helps to realise the time response of Eq. (1) (Fig. 3(b)). Equation (3) was implemented for better visualization of Eq. (5). Equation (5) can be written for plasmonic ELISA based TB test as-

$$\phi_{TB} = \begin{cases} 1, & TB - ve \\ 2, & TB + ve \end{cases}, \text{ where } \phi_{TB} \neq f(\alpha) \quad (6)$$



(a) S= 46; 7.50min; ΔE=0.4003 (b) S= 47; 7.67min; ΔE= 8.9025 (c) S= 48; 7.83min; ΔE= 7.2441 (d) S= 68; 11.167min; ΔE= 0.2793 (e) S= 69; 11.33min; ΔE=11.6616



(f) S=70; 11.50min; ΔE=10.5856 (g) S= 71; 11.67min; ΔE= 5.8642 (h) S=72; 11.83min; ΔE= 7.4725 (i) S= 73; 12.00min; ΔE= 0.1381

Fig. 2. Progress of TB test at various instances (S) in LAB colour space (P = 1 and t_i = 10 s). ΔE = Colour difference. The images were taken with iPhone 7 plus

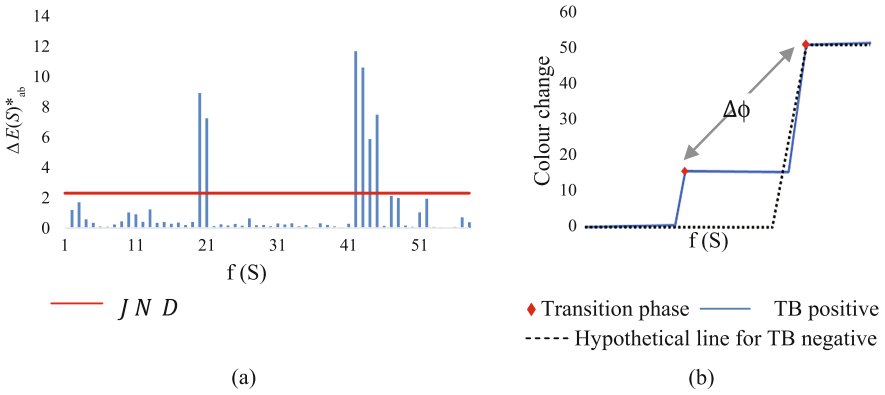


Fig. 3. Time response analysis for P = 1 and t_i = 10 s. (a) $\Delta E(S)^*_{ab}$ and (b) Progression of colour change with respect to time, where $\Delta E(S)^*_{ab} > JND$

The result followed the pattern of Fig. 3(b) at various lighting condition. The impact of Eq. (4) was found insignificant on the pattern. Varying t_i, it was observed that transition is smoother for lower t_i. However, taking $t_i < 1$ s is irrational as the

ELISA test becomes more time consuming without providing any useful new information. The result showed consistency when the plasmonic ELISA test was repeated. It was also observed with naked eye that higher concentration provides faster result.

5 Conclusion

In this work, we have presented a novel approach to perform colourimetric tests by analysing chromaticity. We calculated ΔE during the course of the colourimetric test to attain the colour change at each intense and tracked how ΔE progressed over the time. Intense investigation of time response revealed that the test result can be specified with more feasibility by tracking the colour transition, instead of colour detection or clustering.

We have tested and verified our method on plasmonic ELISA based TB test. For AuNP TB negative, the final colour output is pink, and for TB positive there is another intermediate colour transition to blue. The conventional method is either time consuming, expensive or not efficient. We have demonstrated that the TB test time can be decreased - after defining α, taking even less than the original diagnosis time. It took less than 15 min from filling up the well to get the TB result, whereas conventional method takes few hours to few days. The computation was completed in ~34 s. As the system works on the difference of colours in LAB space with time variation, the dependency on camera quality was reduced. The work being independent of lighting environment is another paramount advantage. This method is implacable to a wide range of colourimetric examinations. The chemical kinetics can be also considered as the potential field of application.

This computerised automatic analysis increase the flexibility and freedom of choice concerning biochemical components, lowering the complexity to it. The presented method enhanced the accessibility, accuracy and precision of colourimetric test, compensating the limitation of naked eye, with reduced cost and time. This technique is going to be employed on mobile enable platforms, followed by clinical testing and validation, to make the system POC prone.

Acknowledgement. This research is supported by the Erasmus Mundus FUSION project, British Council Newton Institutional Links and Newton-Ungku Omar Fund. This is a collaborative research project between Anglia Ruskin University (UK) and Universiti Putra Malaysia (Malaysia).

References

1. Yetisen, A.K.: Holographic Sensors. Springer Theses (2014). doi:10.1007/978-3-319-13584-7
2. Cate, D.M., Adkins, J.A., Mettakoonpitak, J., Henry, C.S.: Recent developments in paper-based microfluidic devices. Anal. Chem. **87**, 19–41 (2015). doi:10.1021/ac503968p
3. Yetisen, A.K., Martinez-Hurtado, J.L., Garcia-Melendrez, A., et al.: A smartphone algorithm with inter-phone repeatability for the analysis of colorimetric tests. Sens. Actuators B Chem. **196**, 156–160 (2014). doi:10.1016/j.snb.2014.01.077

4. Suslick, K.S., Rakow, N.A., Sen, A.: Colorimetric sensor arrays for molecular recognition. Tetrahedron **60**, 11133–11138 (2004). doi:10.1016/j.tet.2004.09.007
5. Qin, X., Wang, R., Tsow, F., et al.: A colorimetric chemical sensing platform for real-time monitoring of indoor formaldehyde. IEEE Sens. J. **15**, 1545–1551 (2015). doi:10.1109/JSEN.2014.2364142
6. Luo, W., Greenbaum, A., Zhang, Y., Ozcan, A.: Synthetic aperture-based on-chip microscopy. Light Sci. Appl. (2015). doi:10.1038/lsa.2015.34
7. Shen, L., Hagen, J.A., Papautsky, I.: Point-of-care colorimetric detection with a smartphone. Lab Chip **12**, 4240 (2012). doi:10.1039/c2lc40741h
8. Tania, M.H., Lwin, K.T., Hossain, M.A.: Computational complexity of image processing algorithms for an intelligent mobile enabled tongue diagnosis scheme. In: 10th International Conference on Software, Knowledge, Information Management & Applications, Chengdu, China, p. 15 (2016). doi:10.1109/SKIMA.2016.7916193
9. WHO: Tuberculosis (Fact sheet). In: Media Centre. WHO (2016)
10. UN Sustainable Development Goals: 17 Goals to Transform Our World
11. Blakemore, R., Story, E., Helb, D., et al.: Evaluation of the analytical performance of the Xpert MTB/RIF assay. J. Clin. Microbiol. **48**, 2495–2501 (2010). doi:10.1128/JCM.00128-10
12. UKVI: Tuberculosis tests for visa applicants - GOV. UK (2016)
13. NHS Tuberculosis (TB) - Diagnosis - NHS Choices. http://www.nhs.uk/Conditions/Tuberculosis/Pages/Diagnosis.aspx. Accessed 12 Jan 2017
14. Tsai, T.-T., Shen, S.-W., Cheng, C.-M., Chen, C.-F.: Paper-based tuberculosis diagnostic devices with colorimetric gold nanoparticles. Sci. Technol. Adv. Mater. **14**, 44404 (2013). doi:10.1088/1468-6996/14/4/044404
15. Osman, M.K., Mashor, M.Y., Jaafar, H.: Detection of mycobacterium tuberculosis in Ziehl-Neelsen stained tissue images using Zernike moments and hybrid multilayered perceptron network. In: 2010 IEEE International Conference on Systems, Man, and Cybernetics, pp. 4049–4055. IEEE (2010). doi:10.1109/ICSMC.2010.5642191
16. Khademhosseini, A.: Nano/microfluidics for diagnosis of infectious diseases in developing countries. Adv. Drug Deliv. Rev. **62**, 449–457 (2011). doi:10.1016/j.addr.2009.11.016.Nano/microfluidics
17. Wang, S., Xu, F., Demirci, U.: Advances in developing HIV-1 viral load assays for resource-limited settings. Biotechnol. Adv. **28**, 770–781 (2010). doi:10.1016/j.biotechadv.2010.06.004
18. Shir, D., Ballard, Z.S., Ozcan, A.: Flexible plasmonic sensors. IEEE J. Sel. Top. Quantum Electron. (2016). doi:10.1109/JSTQE.2015.2507363
19. de la Rica, R., Stevens, M.M.: Plasmonic ELISA for the ultrasensitive detection of disease biomarkers with the naked eye. Nat. Nanotechnol. **7**, 821–824 (2012). doi:10.1038/nnano.2012.186
20. Colorimetry - Part 4: CIE 1976 L*a*b* Colour Space. CIE (2007)
21. Mahy, M., Van Eycken, L., Oosterlinck, A.: Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. Color Res. Appl. **19**, 105–121 (1994). doi:10.1111/j.1520-6378.1994.tb00070.x
22. Sharma, G.: Digital Color Imaging Handbook, 1.7.2. CRC Press, Boca Raton (2003)

Characterization of the Chemical Composition of Banana Peels from Southern Brazil Across the Seasons Using Nuclear Magnetic Resonance and Chemometrics

Sara Cardoso[1(✉)], Marcelo Maraschin[2], Luiz Augusto Martins Peruch[3], Miguel Rocha[1], and Aline Pereira[2]

[1] CEB Centre Biological Engineering, University of Minho, Campus of Gualtar, Braga, Portugal
saracardoso501@gmail.com

[2] Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianpolis, SC, Brazil
m.maraschin@ufsc.br

[3] Agricultural Research and Rural Extension Company of Santa Catarina, Criciúma, Brazil

Abstract. Banana peels are a source of important bioactive compounds, such as phenolics, carotenoids, biogenic amines, among others. For industrial usage of that by-product, a certain homogeneity of its chemical composition is claimed, a trait affected by the effect of (a)bioatic ecological factors. In this sense, this study aimed to investigate the banana peels chemical composition, to get insights on eventual metabolic changes caused by the seasons, in southern Brazil. For this purpose, a Nuclear Magnetic Resonance (NMR)-based metabolic profiling strategy was adopted, followed by chemometrics analysis, using the *specmine* package for the R environment. The obtained results show that the different seasons can, in fact, influence the metabolic composition, namely the levels of metabolites extracted from the bananas peels. The analytical approach herein adopted, i.e., NMR-based metabolomics coupled to chemometrics analysis, seems to enable identifying the chemical heterogeneity of banana peels over the harvest seasons, allowing obtaining standardized extracts for further technological purposes of usage.

Keywords: Nuclear Magnetic Resonance · Chemometrics · Banana

1 Introduction

In a worldwide scenario, Brazil is traditionally known as an important banana producer. Banana peel represents about 30% of the fruit and is the main residual biomass (by-product) of the processing industry. Such a by-product has an environmental significance, since it is a rich source of nutrients (e.g. nitrogen and

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_39

phosphorus) which could lead to imbalances in soil and aquatic environments [1]. On the other hand, the use of banana peel for industrial purposes depends on its chemical composition, a trait strongly affected by, e.g., climatic factors, orchard manage practices, genotype, and harvest time.

Banana is well recognized as source of important bioactive compounds, such as phenolics (gallic acid and derivatives [2]), carotenoids (β-carotene and xanthophylls [3]), anthocyanins (delphinidin and cyanidin [4]), biogenic amines (DOPA and L-DOPA [5]), catechins (gallocatechin and epigallocatechnin [6]), and sterols and triterpenes (β-sitosterol, stigmasterol, campesterol, and 24-methylene cycloartanol [7]). Besides, for industrial purposes, large amounts of banana peels must be provided, with homogeneous chemical composition, guaranteeing a continuous furnishment of raw material of high quality.

Thus, this study investigated the banana peel's chemical composition over the seasons, aiming to gain insights regarding eventual metabolic changes occurring along the harvest times of that fruit in southern Brazil. For that, a typical NMR-based metabolic profiling strategy coupled to chemometrics tools has been adopted, where the data analysis workflow includes both univariate (analysis of variance) and multivariate (principal component analysis and hierarchical clustering) statistical analysis.

2 Materials and Methods

Chemicals: Ultra-pure water was obtained through a reverse-osmosis system (Permution E-10, Curitiba, Brazil). The deuterated solvent D_2O was purchased from TediaBrazil (Rio de Janeiro, Brazil) and 3-trimethylsilyl propionic-2, 2, 3, 3-d_4 acid sodium salt (98 atom % D - TSP) and deuterium chloride solution (35 wt. % in D_2O, 99 atom % D) were obtained from Sigma-Aldrich (Saint Louis, MO, USA).

Samples: Thirteen banana peels samples were collected from an agro-ecologically managed orchard, in Biguaçú County (27° 29′ 39″ S; 48° 39′ 20″ W, altitude 2 m), Santa Catarina State, southern Brazil). Three in the autumn (March, April, and May-2011), four in winter (June-2011, July-2010/2011, and August 2011WI), five in spring (September 2010/2011, October 2010/2011, and November-2010), and one in summer (February-2011). The producing region is characterized for well-marked seasons. The sampled biomass was collected from ripe fruits, showing a yellow color throughout the peel, dried at 45 °C until constant weight and crushed in a mortar and pestle, using liquid N_2. Further aqueous extracts (AEs) of the banana peels were obtained as described by Pereira, (2014) [8] and lyophilized.

1D-NMR spectroscopy - spectrum acquisition parameters: Lyophilized AEs were added of 700 uL D_2O, containing 0.024 g % of 3-trimethylsilyl propionic-2, 2, 3, 3-d_4 acid sodium salt (98 atom % D - TSP) as internal standard, vortexed (3x), and centrifuged (4000 rpm/10 min), followed by recovering the supernatant (650 uL)

and transferring it to 5 mm-NMR tubes. The pH of the samples was adjusted to 3.45 with a deuterium chloride solution (35 wt. % in D_2O, 99 atom % D). The unidimensional NMR spectra (^{1}H-NMR) were recorded in a Varian Inova 500 MHz NMR spectrometer and the chemical shifts (δ, ppm) were referenced to the TSP peak at $\delta(^1\text{H})$ 0.00 ppm. Data acquisition used a Dell workstation and the VNMRJ software, running on Windows 7 platform. Briefly, ^{1}H-NMR spectra acquisition parameters were as follows: 300 K, no spinning, spectral window 5995.7 Hz, acquisition time 4 s, complex points 32983, scans 32, steady state 4, receiver gain 10, relaxation delay 6 s, observe pulse 8.18 us at a power compression 59/0.98, mixing time 100 ms for saturation of water ($\delta = 4.87$ ppm, Watergate pulse), and digital resolution ± 0.08657 Hz.

NMR Data Processing: The ^{1}H-NMR spectra were processed using the ACD/NMR processor software (Advanced Chemistry Development, release 12.0) consisting of zero filling, Fourier transforming the 32 K data points, and automatically phased (Ph0 and Ph1). The baseline was manually corrected and all spectra referenced to the internal standard (TSP, $d1_H$ 0.00 ppm). The spectroscopy information of interest was exported as a .csv file containing a matrix with the chemical shifts (^{1}H pmm) and a peak intensity list. Typical resonance regions of the water and internal standard (TSP) signals removed from the dataset for further analysis. Further, each ^{1}H-NMR spectrum was processed using a routine implemented in the R language through the package *specmine* [9]. Peak alignment grouped proximal peaks together according to their position using a moving window of 0.03 ppm. Peaks of the same group were aligned to their median positions across all samples. Also, missing value imputation was done filling with a constant value of 0.0005, and data pre-processing contemplated log transformation and auto-scaling.

Chemometrics: The metadata taken into account was, as previously stated, the seasons. However, as it was only possible to obtain one sample for the summer, only 3 seasons were considered for the purpose of data analysis. The seasons were assigned as follows: the samples from September 2010/2011, October 2010/2011, and November-2010 were considered spring; the February, March, April, and May-2011 samples were considered summer/autumn and, finally, the June-2011, July-2010/2011, and August 2011 samples were considered winter.

The analysis of the obtained data was performed using the *specmine* package, as above, for the R environment [10]. The pipeline used for the data analysis started with one-way analysis of variance (ANOVA), to test the difference in means between the metadata groups for each one of the variables.

Then, multivariate statistical analysis was performed, starting with hierarchical clustering, using an euclidean distance between samples, followed by Principal Components Analysis (PCA).

The data used in the analysis, together with the reports generated using R Markdown are all given in supplementary material available in the URL: http://darwin.di.uminho.pt/pacbb2017/banana-nmr. This allows for the results to be understood in detail and fully reproducible.

3 Results

The spectral profiles obtained for each sample showed that the samples have, approximately, the same peaks along the different samples.

In Fig. 1, we show the mean 1H-NMR spectra of the samples from each of the different seasons considered. The report in supplementary material has all the samples represented.

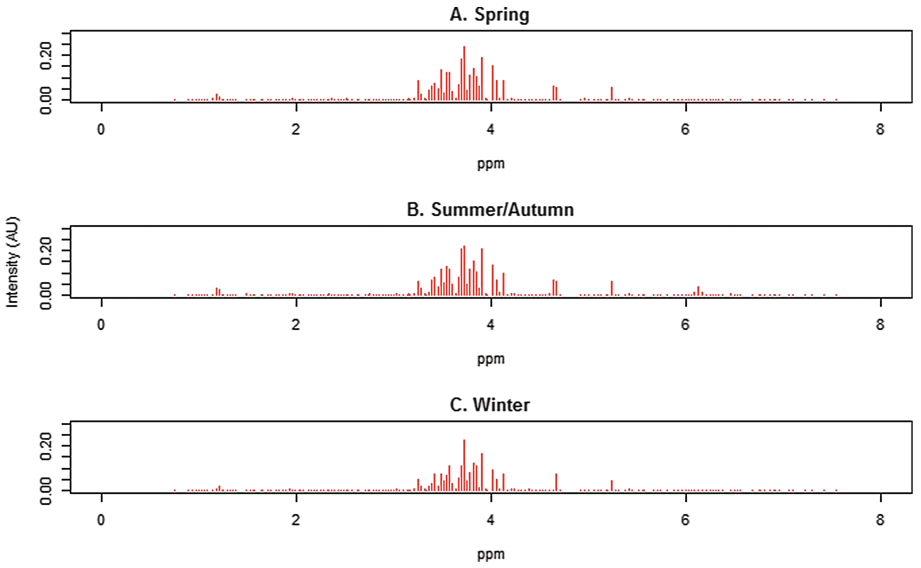


Fig. 1. 1H-NMR mean spectra plots for each season, obtained from the mean of the different samples plots for each season: A - Spring season. B - Summer/autumn season. C - Winter season.

Although the peaks seem not to vary across the different samples and, therefore, the seasons, the intensity of the peaks seems to slightly vary from season to season. This could mean that, despite having the same metabolites across samples, the concentrations of such metabolites vary from season to season.

The result for the hierarchical clustering can be observed in Fig. 2. It was possible to group the samples according to the metadata quite well. The samples from the autumn group were very close, except the sample from July 2011, that was more close to the samples from May 2011 (summer/autumn) and September 2011 (spring).

On the other hand, the other three winter samples seem to be closer to the samples from April 2011 (summer/autumn) and October 2011 (spring). Furthermore, the spring samples were also fairly well grouped, with the exception of the already mentioned samples from April 2011 and October 2011. It is noteworthy

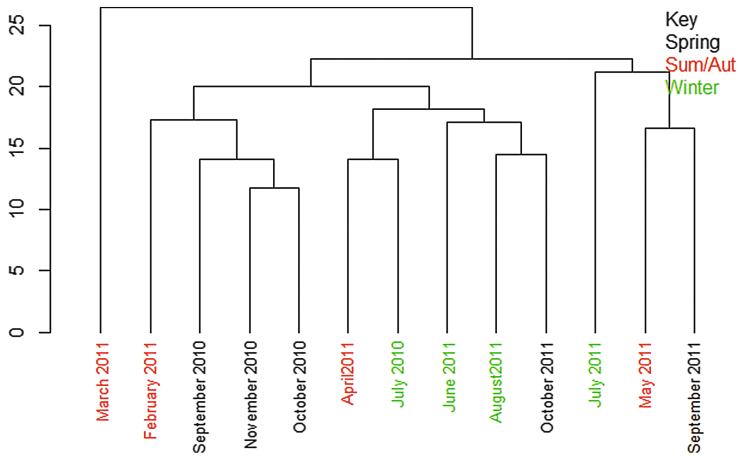


Fig. 2. Dendrogram plot of the result of the hierarchical clustering, with euclidean distance between samples. Spring samples are in black, Summer/Autumn samples in red and Winter samples in green.

that these differences inside each season may be due to months that were hotter or colder than what is usual. Finally, the summer/autumn samples were, in general, very close to the spring samples.

The best results regarding ANOVA, i.e., the peaks whose corrected p-values were below 0.1, are present in the Table 1. The p-values were corrected by using the False Discovery Rate (FDR) method.

Table 1. ANOVA results for the peaks with the best corrected p-values (FDR method), also showing the pair of samples groups that were significantly different in terms of means for each peak.

Peaks	FDR	Tukey result
1.89	0.08090303	Spring-Winter; Sum/Aut-Winter; Sum/Aut-Spring
4.01	0.08090303	Spring-Winter; Sum/Aut-Winter
4.05	0.08090303	Spring-Winter; Sum/Aut-Winter

It is possible to realize that there were few peaks with low p-values. As it could be somewhat expected by observing the dendrogram plot in Fig. 2, all the three peaks with low p-values had means significantly different between the groups spring and winter, and summer/autumn and winter, as they were the group of samples that were further grouped in the dendrogram. In only one peak, significant differences in the means were observed regarding the summer/autumn and spring groups, as it was quite expected, due to the fact that they seemed very close in the dendrogram plot. The identified peaks occurs in the aliphatic

(1.89 ppm) and anomeric (4.01 and 4.05 ppm) regions of the 1H-NMR spectrum, but, as such, do not allow further metabolite identification.

Finally, the results regarding the PCA analysis, present in Figs. 3 and 4, showed that the first principal component is able to explain more than 20% of the data variability, thus allowing to distinguish the groups winter and spring from summer/autumn. The next two components are able to explain, each one, more than 10% of the data variability, leading to an accumulative explanation of more than 50% of the data variability. The second component seems to be able to distinguish the groups spring and summer/autumn from winter, and the third component the groups winter and summer/autumn from spring.

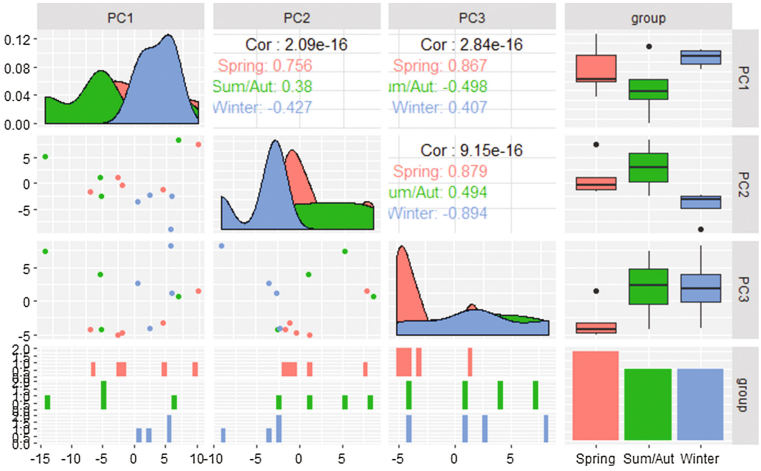


Fig. 3. PCA results.

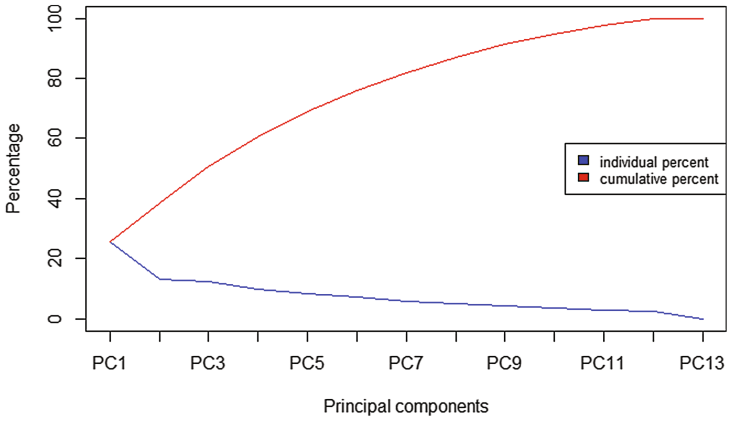


Fig. 4. Screeplot of the PCA results.

4 Conclusions

All the results showed that it is possible to distinguish the banana's peel metabolic composition according to the seasons, mostly due to the peak intensity. This distinction is more noticeable in the winter and summer/autumn groups, as were the ones that were better grouped in the cluster analysis and showed more significant differences in means regarding the ANOVA analysis. Furthermore, the PCA analysis revealed that it is only necessary 3 principal components to explain more than 50% of the data variability. This shows that the different conditions of the seasons can influence the composition of the banana's peel, addressing the need of further studies as ones aims at to explore the potential of banana peel as source of bioactive compounds of interest of health and cosmetics industries, for instance. The NMR-based metabolomic analytical strategy herein shown seems to be capable of identifying the chemical heterogeneity of banana peels over the harvest seasons, allowing obtaining standardized extracts for further industrial applications.

Acknowledgments. To CNPq (National Counsel of Technological and Scientific Development) for financial support (Process n° 407323/2013-9), to CAPES (Coordination for the Improvement of Higher Education Personnel), and EPAGRI (Agricultural Research and Rural Extension Company of Santa Catarina). The research fellowship from CNPq on behalf of M. Maraschin is acknowledged. The work is partially funded by Project PropMine, funded by the agreement between Portuguese FCT and Brazilian CNPq. This study was also partially supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte.

References

1. González-Montelongo, R., Gloria Lobo, M., Gonzfffdfffdlez, M.: Antioxidant activity in banana peel extracts: testing extraction conditions and related bioactive compounds. Food Chem. **119**(3), 1030–1039 (2010)
2. Bich, T., Nguyen, T., Ketsa, S., van Doorn, W.G.: Relationship between browning and the activities of polyphenoloxidase and phenylalanine ammonia lyase in banana peel during low temperature storage. Postharvest Biol. Technol. **30**(2), 187–193 (2003), ISSN 0925-5214, doi:http://dx.doi.org/10.1016/S0925-5214(03)00103-0, http://www.sciencedirect.com/science/article/pii/S0925521403001030
3. Subagio, A., Morita, N., Sawada, S.: Carotenoids and their fatty-acid esters in banana peel. J. Nutr. Sci. Vitaminol. **42**, 553–566 (1996)
4. Seymour, G.B.: Banana. In: Seymour, G.B., Taylor, J., Tucker, G. (eds.) Biochemistry of Fruit Ripening, pp. 95–98. Chapman and Hall, London (1993)
5. Kanazawa, K., Sakakibara, H.: High content of dopamine, a strong antioxidant, in cavendish banana. J. Agric. Food Chem. **48**(3), 844–848 (2000)
6. Someya, S., Yoshiki, Y., Okubo, K.: Antioxidant compounds from bananas (Musa Cavendish). Food Chem. **79**(3), 351–354 (2002)

7. Knapp, F.F., Nicholas, H.J.: The sterols and triterpenes of banana peel. Phytochemistry **8**(1), 207–214 (1969)
8. Pereira, A.: Determinação do perfil químico e da atividade cicatrizante de extratos de casca de banana cultivar prata anã (Musa sp.) e o desenvolvimento de um curativo para pequenas les oes. Ph.D. thesis, Universidade Federal de Santa Catarina (2014)
9. Costa, C., Maraschin, M., Rocha, M.: An R package for the integrated analysis of metabolomics and spectral data. Comput. Methods Programs Biomed. **129**, 117–124 (2016)
10. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008). ISBN 3-900051-07-0, http://www.R-project.org

Author Index

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7