


NEW TRENDS IN GEOMETRY

THEIR ROLE IN THE NATURAL AND LIFE SCIENCES

The background of the cover is an abstract composition of numerous overlapping circles and splatters in various colors, including gold, red, blue, purple, and pink, set against a light, textured background. The circles vary in size and opacity, creating a sense of depth and movement.

Claudio Bartocci
Luciano Boi
Corrado Sinigaglia

Editors

Imperial College Press

NEW TRENDS IN GEOMETRY

THEIR ROLE IN THE NATURAL AND LIFE SCIENCES

NEW TRENDS IN GEOMETRY

THEIR ROLE IN THE NATURAL AND LIFE SCIENCES

Claudio Bartocci

Università di Genova, Italy

Luciano Boi

École des Hautes Études en Sciences Sociales, France

Corrado Sinigaglia

Università degli Studi di Milan, Italy

Editors



Imperial College Press

This page is intentionally left blank

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

NEW TRENDS IN GEOMETRY
Their Role in the Natural and Life Sciences

Copyright © 2011 by Imperial College Press

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-1-84816-642-4
ISBN-10 1-84816-642-7

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore.

CONTENTS

Preface by <i>Claudio Bartocci, Luciano Boi and Corrado Sinigaglia</i>	vii
Part I: Geometry, Theoretical Physics and Cosmology	1
Chapter 1: Claudio Bartocci and Ugo Bruzzo, <i>The Emergence of Algebraic Geometry in Contemporary Physics</i>	3
Chapter 2: Mauro Carfora, <i>Quantum Gravity and Quantum Geometry</i>	17
Chapter 3: Ugo Moschella, <i>The de Sitter and Anti-de Sitter Universes</i>	35
Chapter 4: Jean-Pierre Luminet, <i>Geometry and Topology in Relativistic Cosmology</i>	81
Part II: The Problem of Space in Neurosciences	105
Chapter 5: Leonardo Fogassi, <i>Space Coding in the Cerebral Cortex</i>	107
Chapter 6: Anna Berti and Alessia Folegatti, <i>Action and Space Representation</i>	127
Chapter 7: Claudio Brozzoli and Alessandro Farnè, <i>The Space Representations in the Brain</i>	137
Chapter 8: Corrado Sinigaglia and Chiara Brozzo, <i>The Enactive Constitution of Space</i>	157
Part III: Geometrical Methods in the Biological Sciences	171
Chapter 9: Francis Bailly and Giuseppe Longo, <i>Causes and Symmetries in Natural Sciences: The Continuum and the Discrete in Mathematical Modelling</i>	173
Chapter 10: Riccardo Broglia, <i>Topological Invariants of Geometrical Surfaces and the Protein Folding Problem</i>	211

Chapter 11: Jean-François Sadoc, <i>The Geometry of Dense Packing and Biological Structures</i>	221
Chapter 12: Luciano Boi, <i>When Topology and Biology Meet 'For Life': The Interactions Between Topological Forms and Biological Functions</i>	243
About the Contributors	307
Index	313

C. Bartocci, L. Boi, C. Sinigaglia (Eds.)

I. Scope and Aims of the Book

There are two major motivations for proposing this book. The first is the conviction that the integration of mathematics and physics, molecular and cell biology, and neurosciences will constitute the new frontier and challenge for twenty-first century science. The second relates to the fact that the exciting and appealing science in the twenty-first century is likely to evolve among, not within, traditional disciplines. Therefore, the book and *More geometrico* project focus on the interactions between mathematics, physics, biology and neurosciences, and are aimed at exploring new geometrical and topological modelling in a variety of physical, biological and neuroscience fields.

The authors concentrate on some new, extremely valuable interfaces of mathematical methods and modelling with the physical and life sciences. Its major aim is to further our understanding of the multilevel and scale-change phenomena in these disciplines. One of the main goals of the various contributions is to study the central role of a multilevel and scale-change approach in different fields such as neuroscience, systems biology, and quantum physics and geometry.

We think it is more and more important to inject ideas and methods from modern differential geometry and algebraic topology into neuroscience and molecular and cellular biology and to inspire new directions in these mathematical fields from discussions on the major problems in macroscopic physics and the biological sciences.

The contributions are aimed at giving a precise idea of some far-reaching connections between mathematics, physics, biology and cognition. Its essential scope is to develop rigorous interdisciplinary and integrative research. One important goal is to show that several methods and techniques, especially from algebraic topology and differential geometry, are profoundly involved at different scales and various levels of organisations in the physical and biological processes. Various contributions emphasise the urgent need for developing new mathematical methods, models and techniques suited to work out a mathematical dynamical theory of the emergence of natural and living patterns and behaviours. For example, in

our view, one particularly interesting task would consist of explaining to what extent the mathematical structure and spatial-temporal events that constitute the natural frame of living organisms may influence their bio-chemical, physiological and cognitive organisation. In this perspective, the most important and complex work to be carried out is to construct both qualitative and quantitative mathematical and physical models capable of describing and explaining the effective dynamics, properties and invariants of phenomena at different levels of organisation and action. This is, we think, one of the most important conceptual and philosophical challenges of today's interdisciplinary research.

2. Three Fundamental Themes

2.1. *Mathematical modelling in the neurosciences*

One important question raised in the book is the following: what differentiates the topological and metric structures of the 'physical' space, which is embodied into (or internalised by) the neurophysiological space in the brain, from the properties of the perceptive and cognitive space obtained very likely by deforming the first according to certain general (mathematical and/or physical) laws? The answer to this crucial questions is at the core of the present research in the neurosciences, the major issues of which are: (i) The mechanisms by which our complex sensorial systems such as the visual, the sensorimotor and the vestibular systems drive our perception of the surrounding space and the movements required in order to reach it. (ii) The perception of movement and its neurophysiological basis, and the role of action in the perception of the third dimension and in the cognitive grasping of the properties and qualities of objects localised in our (near and far) space. (iii) The mathematical and physical grounds of human cognition and, reciprocally, the biological and cognitive roots of our mathematical skills.

2.2. *Geometrical and topological methods in the life sciences*

A fundamental goal of the contributions in this field is to show that there are effective models and techniques from mathematical sciences, which can be used to describe many fundamental properties and behaviours observed in biological systems. More precisely, an important part of the present research is aimed at demonstrating that the complex topology and dynamics of DNA-proteins complexes are closely linked to the multilevel epigenetic regulation and to the cell's spatial and functional organisation. It has been emphasised that the geometrical structure and topological form of nuclear components (DNA, nucleosome, chromatin, chromosome etc.) play an important role in the cell differentiation and organism growth. For example, at the molecular and supramolecular level, enzyme topoisomerases, which convert DNA from one topological form to another, appear to have a profound role in the central

genetic events of replication, transcription, recombination and repair. Second, certain topological mechanisms are involved in the fundamental biological process of the compaction of chromatin into the chromosome during the interphase and the metaphase. Third, new mathematical methods and models have been suggested relating to the cells' differentiation and their complex spatial organisation during the different phases of the development of the embryos. It appears an essential task to provide mathematical models and techniques which are more global and dynamic, in order to be able to rethink successfully the causal connection between form and function in the biological sciences.

2.3. *Open theoretical problems and mathematical perspectives in relativistic and quantum physics*

We think it is important to address the issue of the role of algebro-geometric and topological methods in the developments of quantum physics, especially in gauge theory as well as string theory, and tried to show how these methods can provide a deeper understanding of physical phenomena at different scales. We consider that the unification of gravitation and quantum physics requires some fundamental breakthroughs in our understanding of the relationship between space-time and quantum processes. In particular, the superstring theories lead to guessing that the usual structure of space-time at the Planck scale must be dropped from physical thought. A very interesting hypothesis, which is discussed in some chapters of this book (in Part I), is that the global geometrical properties of the manifold model of space-time (either Lorentzian or Riemannian) play a major role in quantum field theories and that, consequently, several physical quantum effects arise from the non-local metrical and topological structures of these manifolds. Modern Kaluza–Klein theories, superstring theory and, in a different way, quantum gravity and non-commutative geometry, showed that space-time symmetries and physical symmetries might be unified through the introduction of new structures (dimensions, invariants, supersymmetries) of space with a different topology. That essentially means that 'hidden' symmetries of fundamental physics can be related to the phenomenon of topological change of certain classes of non-smooth manifolds.

3. Brief Conclusions

We hope this book will contribute to showing the increasingly fundamental role played by geometrical and topological ideas and methods in several relevant fields of research relating to a number of important recent developments in the natural and life sciences. In fact, it is more and more emerging that some geometrical and topological concepts could help to describe and explain a variety of amazing structures and behaviours characterising physical reality, living systems and human

cognition. Indeed, many processes and organisational principles in these fields seem to be closely related to the invariant action of certain fundamental geometrical and topological objects and structures.

Let us conclude by stressing once again the very purpose of this book, which is to explore and understand some far-reaching interfaces where geometry, topology, physics, dynamics, biology and neurosciences seem to interact profoundly, in a great and significant overlap of knowledge among mathematicians, physicists, biologists, neurophysiologists and philosophers of science. Last but not least, one unifying theme characterises this book: Geometry may best be interpreted as a way of thinking rather than a mere formal language or a collection of specific subject areas. There is, perhaps, no branch of mathematics that cannot be considered a part of geometry, and there is, most likely, no field in the natural and life sciences unrelated to the enlightening influence of geometry, when approached in the right, open spirit. Resting upon some geometric-minded theories and methods, the authors propose various insights and prospects for future research. We expect they will contribute to drawing a more comprehensive and meaningful landscape of natural and living phenomena.

The Editors
Paris, Milan and Genoa
February 2010

PART 1

**Geometry, Theoretical Physics and
Cosmology**

This page is intentionally left blank

CHAPTER I

The Emergence of Algebraic Geometry in Contemporary Physics

CLAUDIO BARTOCCI

*Dipartimento di Matematica, Università di Genova,
Via Dodecaneso 35, 16146 Genova, Italy
bartocci@dima.unige.it*

UGO BRUZZO

*Scuola Internazionale Superiore di Studi Avanzati,
Via Beirut 2-4, 34014 Trieste, Italy, and
Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, Italy
bruzzo@sissa.it*

I. Introduction

Starting with Einstein's general relativity, differential geometry has started playing a major role in physics. General relativity describes the gravitational fields as a metric property of the spacetime manifold. More precisely, spacetime (i.e., the manifold the points of which are *events*; we may intuitively say that an event is 'something that happens in a given point in space at a certain time') is supposed to be endowed with a Lorentzian metric. This means that spacetime has pointwise the same structure as the Minkowski space of special relativity but in general is not flat, as on the contrary Minkowski space is. Indeed, out of the metric tensor one can construct another tensor field, the *curvature* field, which measures how far the geometry of spacetime is from that of a flat space. The celebrated Einstein equations prescribe how the matter in our universe determines the curvature of spacetime, and in turn the curvature determines how matter (particles, light rays, extended bodies...) moves.

In this sense, general relativity reaches a complete geometrisation of the gravitational field. A similar goal is achieved by gauge theories in relation to the other

fundamental physical interactions (electromagnetic and nuclear forces). The formulation of gauge theories started in the 1950s, and nowadays they play a central role in the modelisation of fundamental physical interactions — it is not too far-fetched to say that they are the paradigm of contemporary high energy physics. The basic mathematical structure of gauge theories is, again, differential geometry.

However, since the mid 1970s, the development of gauge theories disclosed new perspectives. Researchers started realising that a number of physical features could be captured only by considering global, not just local, properties of the involved geometrical entities. For instance, the charge of an instanton is naturally interpreted as a cohomology class, more precisely a characteristic class of the bundle associated with the instanton; quantum anomalies are cohomology classes of a certain cohomology theory associated with a quantum field theory; in string theory, a deeply geometrised theory, many physical quantities are actually cohomology classes. And not only cohomology is relevant, other global properties of manifolds, such as K-theory, may be brought into play.

A precursor of this trend may be found in the work of Felix Klein. This is very well described in the words of H. Poincaré in his *La valeur de la science*:

Voyez au contraire M. Klein: il étudie une des questions les plus abstraites de la théorie des fonctions; il s'agit de savoir si sur une surface de Riemann donnée, il existe toujours une fonction admettant des singularités données. Que fait le célèbre géomètre allemand? Il remplace sa surface de Riemann par une surface métallique dont la conductibilité varie suivant certaines lois. Il met deux de ses points en communication avec les deux pôles d'une pile. Il faudra bien, dit-il, que le courant passe, et la façon dont ce courant sera distribué sur la surface définira une fonction dont les singularités seront précisément celles qui sont prévues par l'énoncé. [13, p. 28]

On the other hand, look at Professor Klein: he is studying one of the most abstract questions of the theory of functions; to determine whether on a given Riemann surface there always exists a function admitting the given singularities. What does the celebrated German geometer do? He replaces his Riemann surface by a metallic surface whose electric conductivity varies according to certain laws. He connects two of its points with the two poles of a battery. The current, say he, must pass, and the distribution of the current on the surface will define a function whose singularities will be precisely those called for by the enunciation. [4]

Other, more recent, occurrences that come to mind are:

- the twistor programme of the Penrose school in geometry and mathematical physics (from the late 1960s onward);
- the exploitation of techniques from complex geometry in general relativity;
- the interpretation of physical observables in a large class of quantum field theories as geometric invariants.

A distinguished feature of the interplay between geometry and physics in the last, say, 20 years, is that the usual pattern of interaction between the two

disciplines (mathematics provides the language to formulate the models of physical problems — and the techniques needed to solve the resulting equations) has evolved into a much more intricate web of connections. What is especially new in this fact is that physics has been shown to be able to serve as a powerful source of inspiration of new mathematical ideas and techniques. In this paper we shall try to give an introduction to this beautiful circle of ideas.

This paper is articulated in three sections: we first try to give a very rough idea of what algebraic geometry is, then we discuss gauge theory, and in a final section we talk about strings, in particular discussing how a physical theory like string theory is able to produce highly nontrivial mathematical results such as the enumeration of curves in algebraic varieties.

2. Algebraic Geometry in a Nutshell

It is not easy to say in a concise way what algebraic geometry is. Perhaps the sentence that best captures its essence is the following: algebraic geometry is the geometric study of the solutions of systems of algebraic equations. Typical questions in algebraic geometry are the enumerative problems. The simplest one is of course this one: *How many lines go through two given points in the plane?* Of course the answer is one, if the points are distinct; an infinite number if the two points coincide (if we want to be precise, we may say that here ‘plane’ means ‘complex projective plane’).

We may generalise this problem by considering curves of higher degree, where we say that an (algebraic) curve has degree d if it is described by an equation of the type $P(x, y, z) = 0$, where P is a homogeneous polynomial of degree d (and x, y and z are to be thought of as homogeneous coordinates in the plane). Thus, the previous problem was dealing with the case $d = 1$. A curve of degree $d = 2$ is called a *conic*. It is an elementary fact that five points determine a conic; this may be traced to the fact that a conic has equation

$$ax^2 + by^2 + cz^2 + dxy + exz + fyz = 0 \quad (1)$$

This equation contains six coefficients, but an overall factor is irrelevant, so that in order to choose a conic we need to fix five coefficients. However we may run into problems if the five points are collinear. For instance, consider the five points

$$\begin{aligned} P_1 &= (-1, -1, 1), & P_2 &= (0, 0, 1), & P_3 &= (1, 1, 1), \\ P_4 &= (2, 2, 1), & P_5 &= (3, 3, 1) \end{aligned}$$

The conic going through these points has equation $x^2 - y^2 = 0$, i.e., it is the union of two lines intersecting at the point $(0, 0, 1)$, as one can understand by writing the equation as $(x + y)(x - y) = 0$. One can easily check that whenever the five points

are not collinear, the resulting conic is ‘nondegenerate’, i.e., it is not the union of two lines. Thus we should avoid points in ‘degenerate’ position, and consider only points in *generic* position, i.e., not collinear.

Let us go to $d = 3$, i.e., to cubics. Nine points in the plane determine a cubic, *but two cubics intersect at nine points!* This is the so-called Euler–Cramer paradox: how come that nine points determine a cubic, but if we take two cubics, both of them go through the same array of nine points? Things get even worse for $d > 3$: in this range, two curves have more points in common than the number of their coefficients; two curves of degree d meet indeed at d^2 points, but have $\frac{1}{2}(d^2 + 3d - 2)$ independent coefficients.

Again, it is a question of genericity: nine points in generic position determine a unique cubic; nine points at which two cubics intersect are not in generic position.

Now a problem arises: how to detect points in generic position? This can be reduced to a question in linear algebra (Cramer’s theorem about linear systems). This generalises to the classical enumerative problem in algebraic geometry: for a given array of points in generic position in a variety, compute how many curves of a given degree go through them.

We have thus learned some features of algebraic geometry:

- it makes a sharp distinction between generic and nongeneric situations;
- algebraic geometry is able to make numerical predictions;
- algebraic geometry usually deals with finite problems.

It should be emphasised that algebraic geometry turns out to be a powerful tool also in dealing with problems in differential geometry. How can this be possible? The answer lies in two key results, Kodaira’s embedding theorem and Serre’s G.A.G.A. principle. According to Kodaira’s theorem (1954), any compact Riemannian manifold having holonomy group $U(n)$ (equivalently, any compact Kähler manifold), provided that a certain cohomological condition is satisfied, can be given the structure of projective algebraic variety (i.e., an algebraic subvariety of a complex projective space). On the other hand, Serre’s celebrated principle [15] says that any global analytic object on an algebraic variety is algebraic.

3. Gauge Theories

The notion of gauge invariance is already contained, at least *in nuce*, in Maxwell’s formulation of electromagnetic theory. Actually, in his *Treatise on Electricity and Magnetism* (1873), Maxwell noted that the vector potential A (i.e., the vector field whose curl, $\nabla \times A$, is the magnetic field B) can be transformed according to the equation $A = A_0 + \nabla \chi$ and that ‘the quantity χ disappears from the equations [...] and it is not related to any physical phenomenon.’ However, he did not state the associated equation prescribing the transformation of the scalar potential, $\Phi =$

$\Phi_0 - \frac{1}{c} \frac{\partial \chi}{\partial t}$. The interdependence of the scalar and the vector potential had already been remarked upon, in certain particular cases, by Gustav Kirchhoff and Hermann von Helmholtz, and by the Danish physicist Ludvig Valentin Lorenz, who was the first (1867) to impose the condition $\nabla A + \frac{1}{c} \frac{\partial \Phi}{\partial t} = 0$ to ensure that the potentials are solutions of the wave equation.¹ So, what it is usually called the ‘Lorentz condition’ ought to be more appropriately termed the ‘Lorenz condition’. Hendrik Antoon Lorentz has to be credited with the statement (1904) that ‘every admissible pair A and Φ ’ is related via the transformations

$$A = A_0 + \nabla \chi, \quad \Phi = \Phi_0 - \frac{1}{c} \frac{\partial \chi}{\partial t} \quad (2)$$

where the scalar function χ is a solution for the inhomogeneous wave equation

$$\nabla^2 \chi - \frac{1}{c^2} \frac{\partial^2 \chi}{\partial t^2} = \nabla \cdot A_0 + \frac{1}{c} \Phi_0$$

But the conceptual origin of modern gauge theories is to be found in Hermann Weyl’s pioneering work. Weyl tried to unify gravitation and electromagnetism by postulating that the metric tensor $g_{\mu\nu}$ of Einstein’s general relativity was defined up to a change of scale of the kind $g_{\mu\nu} \rightarrow e^\lambda g_{\mu\nu}$ for some function λ of the space-time coordinates. His idea — thoroughly discussed in the first edition (1918) of his book *Raum, Zeit, Materie* [18] — was that ‘the electromagnetic conservation law [was] connected with the new scale-invariance, expressed through a fifth arbitrary function’ (namely, the function λ) [19]. Though this early implementation of the ‘principle of scale invariance (*eichinvarianz*)’ did not work, it proved to be a rather fruitful insight.

In 1926, within the framework of the recently developed quantum mechanics, Vladimir Aleksandrovič Fock discovered that the system of a relativistic particle of charge e interacting with an electromagnetic field is invariant under the transformations (2) together with the transformation $\psi = \psi_0 \exp(\frac{ie\chi}{\hbar c})$ of the wave function ψ . The same result was obtained by Fritz London in 1927, and by Weyl, who fully established the ‘principle of gauge invariance’ in his 1928 book *Gruppentheorie und Quantumechanik* [20] and, a year later, in the two papers *Gravitation and the electron* and *Elektron und Gravitation*. According to Weyl,

This principle of gauge invariance is quite analogous to that previously set up by the author, on speculative ground, in order to arrive at a unified theory of gravitation and electricity. But I now believe that this gauge invariance does not tie together electricity and gravitation, but rather electricity and matter in the manner described above. [20, Engl. transl., p. 100, 101]

¹See [8] for a rather detailed historical account of the early development of the notion of gauge invariance.

Though quantum electrodynamics (QED) is now regarded as a key example of gauge theory, the principle of gauge invariance played almost no role in its development, as observed by David Gross, since at the time ‘it was largely regarded as a complication and a technical difficulty’ [7, p. 956]. On the contrary, gauge symmetry was of crucial importance as a guiding principle in the 1954 historic paper by Chen-Ning Yang and Robert L. Mills, *Conservation of isotopic spin and isospin gauge invariance*. Yang and Mills succeeded in generalising the abelian, $U(1)$ -gauge theory of electrodynamics to a non-abelian, $SU(2)$ -gauge theory²; in 1956 their ideas were generalised by Ryoyu Utiyama [16] to include any arbitrary finite-dimensional Lie group. Within this framework, during the 1960s, thanks to the efforts of a number of researchers (among which Sheldon Lee Glashow, Peter W. Higgs, Sidney Coleman, Jeffrey Goldstone, John C. Ward) emerged a new theory of unification of weak and electromagnetic interactions: this programme was completed in 1967 by Steven Weinberg and Abdus Salam [17, 14], who formulated a rigorous $SU(2) \times U(1)$ -gauge theory with massless gauge bosons, combined with a Higgs mechanism for generating W and Z masses by means of a spontaneous symmetry breaking.³ In the decade 1964–1974 a non-abelian gauge theory of strong interactions was devised, the quantum chromodynamics (QCD), as it was christened by one of its creators, the American physicist Murray Gell-Mann. In this case, the group of gauge invariance is $SU(3)$, which expresses the fact that each quark of a given flavour (u, d, b, t, s, c) has three different colours (red, yellow, blue).

Before discussing how algebraic geometry comes into play, we shall briefly sketch the main mathematical features of classical (i.e., nonquantum) Yang–Mills theories. The basic tool is the theory of fibre bundles, origins of which can be traced back to Weyl’s and Élie Cartan’s work and which was virtually completed in the early 1950s. Let G be a Lie group. A G -principal fibre bundle $P \rightarrow M$ over a manifold M can be locally identified with the Cartesian $U \times G$, where U is a sufficiently small open subset of M ; on the intersection $U \cap V$, the two cylinders $U \times G$ and $V \times G$ are glued by means of a transition function $\varphi_{UV} : U \cap V \rightarrow G$. A connection D is a differential operator (with geometrical meaning) on the tangent space of P defining a notion of horizontality for tangent vectors. The curvature F_D of the connection D , which is a differential 2-form taking values in the Lie algebra of G , describes an obstruction to integrability of the distribution of horizontal subspaces. For example, if M is an oriented Riemannian manifold of dimension n , we can take $G = SO(n)$ (the special orthogonal group) and consider the principal

²One should mention that Oskar Klein was the first, in 1938, to propose a non-abelian, $SU(2) \times U(1)$ gauge theory in the attempt to unify gravity and nuclear forces on a six-dimensional space-time (actually, the topological product of the Minkowski space and the two-dimensional sphere).

³See [11, Chap. 21] for a concise account of the development of quantum field theory in the period 1960–1983; for a far more detailed account see [10].

fibre bundle of orthonormal frames over M ; in this case, there is a distinguished connection, namely the Levi-Civita connection, whose curvature encodes many important geometric features of M .

In a gauge theory, M is thought of as the underlying spacetime, while the Lie group G describes the kinematical symmetries of the theory. The group \mathcal{G} of gauge transformations is the infinite-dimensional group of automorphisms of the G -fibre bundle $P \rightarrow M$; connections correspond to gauge potentials, and there is a natural action of \mathcal{G} on the space \mathcal{A} of all connections on P . The curvature F_D can be interpreted as the field strength, and the Lagrangian density of the theory is expressed by $\text{Tr}(F_D \wedge * F_D)$, where Tr denotes an invariant quadratic form on the Lie algebra of G and $*$ denotes the Hodge duality operator. The essential fact is that the Lagrangian density is invariant under gauge transformations, so that one can introduce the Yang–Mills functional

$$S(D) = \frac{1}{4} \int_M \text{Tr}(F_D \wedge * F_D) \quad (3)$$

which is defined on the space \mathcal{A}/\mathcal{G} of gauge equivalence classes of connections. For instance, in electromagnetism M is the Minkowski spacetime, the group G is $U(1)$, a connection corresponds to a 4-potential (A, ϕ) , the associated curvature represents the electromagnetic field, and the Lagrangian density is the usual Maxwell’s source-free Lagrangian.

The Yang–Mills equations are the Euler–Lagrange equations for the functional (3). In the case of electromagnetism, these are just Maxwell’s equations in absence of charges and currents, i.e. $dF_D = 0 = d * F_D$. When the group is not abelian, we have to replace Cartan’s exterior differential d by the covariant differential d_D defined by the connection, so that Yang–Mills equations read

$$d_D F_D = 0 \quad (4)$$

$$d_D * F_D = 0 \quad (5)$$

The former is automatically satisfied, since it is nothing but the Bianchi identity for the curvature F_D . The latter corresponds to a system of nonlinear partial differential equations (in contrast with Maxwell’s equations, which are linear); hence, it is hard to solve and, in general, only a few exact solutions are known.

On a Riemannian four-manifold,⁴ the Hodge operator $*$ satisfies the relation $*^2 = 1$, so that any differential 2-form ω can be decomposed as the sum $\omega = \omega^+ + \omega^-$, where ω^+ is self-dual and ω^- is anti-self-dual (i.e. $*\omega^\pm = \pm\omega^\pm$). A connection whose curvature is self-dual or anti-self-dual is said to be an *instanton*. Quite

⁴Note that in gauge theory, as intended by mathematicians, the metric on the base manifold is not Lorentzian, but definite-positive. This assumption can be given a physical justification by invoking the so-called ‘Wick rotation’.

clearly, instantons are solutions to the Yang–Mills equations; furthermore, it can be proved that they correspond to the absolute minima of the Yang–Mills functional. In 1977, Michael Atiyah and Roger Penrose’s Ph.D. student Richard Ward established a one-to-one correspondence between instantons over the four-sphere and certain holomorphic vector bundles on the complex projective space $\mathbb{C}P^3$ [3]. Thanks to this result, it became possible to apply the machinery of algebraic geometry to gauge theory. Actually, by using tools introduced by Wolf Barth and Geoffrey Horrook to study holomorphic vector bundles on projective spaces, the problem of classifying instantons over the four-sphere was solved, independently, by Atiyah and Nigel Hitchin in Oxford and by Vladimir G. Drinfel’d and Yuri I. Manin in Moscow: the four mathematicians agreed to publish their construction — by then known as ADHM construction — in a joint paper [1].

More or less in the same period, Michael Atiyah, Nigel Hitchin and Isadore Singer, in their pioneering paper [2], were able to compute the dimension of the moduli space parametrizing the instantons over a compact four-dimensional Riemannian manifold; the key ingredient for the computation was the index theorem proved by Atiyah and Singer in the early 1960s.

A substantial breakthrough was obtained by Simon Donaldson in the early 1980s. Inspired by Atiyah and Bott’s paper *The Yang–Mills equations over Riemann surfaces*, Donaldson started studying the space parametrizing gauge equivalence classes of instantons over a four-dimensional manifold. This space — called the moduli space of instantons — may be regarded as a deep invariant of the manifold M . In particular, using this moduli space one can associate with the manifold M a set of invariants, called *Donaldson invariants* [6]. Using these techniques from gauge theory, Donaldson and others were able to provide a classification of four-manifolds — in some sense the analogue of what for three-manifolds is the Poincaré conjecture.

One has here an important link with algebraic geometry. If the four-manifold X happens to be a two-dimensional complex manifold, i.e., a complex surface, then one has the so-called Hitchin-Kobayashi correspondence. This relates instantons with *holomorphic* bundles, i.e., bundles structure of which is somehow compatible with the complex structure of the base manifold. If in addition X is algebraic, then one can use the powerful techniques of algebraic geometry to compute the Donaldson invariants. A good reference about this aspect is [9].

4. String Theory

Quantum electrodynamics and quantum chromodynamics are the two main constituents of what is called the ‘standard model’ of particle physics. This model unifies the strong and weak nuclear interactions and electromagnetism, and is able to obtain astoundingly good theoretical predictions. (For an introduction to the

standard model we refer the reader to [5]). However, this model is not free of inconsistencies and drawbacks. The main problem is perhaps the fact that it provides no explanation of the way things are, for instance, why there are three families of elementary particles. It gives no theoretical prediction of the many constants present in the theory (for example, why the ratio between the masses of the muon and of the electron is 203.7). And, most worrying of all, the standard model conflicts with general relativity.

String theory, which at present is the best, or perhaps the unique, candidate to serve the role of a unified model of all interactions, originated — rather serendipitously — as an alternative model of strong interactions. In the middle 1960s, the physicist Gabriele Veneziano, working at the *Centre Européen pour la Recherche Nucléaire* (CERN) in Geneva, had the idea of describing the strong nuclear interaction in terms of a vibrating extended one-dimensional string, instead of a particle. Despite its originality, and the simplifications that it carried along, Veneziano's attempt — together with additional contributions by Yoichiro Nambu, Leonard Susskind and others — very soon met seemingly unsurmountable difficulties. However, in the middle 1970s, John Schwarz and Joël Scherk built upon Veneziano's intuition to construct a quantum theory of gravity. The basic property of a string — be it open, or closed as a ring — is to vibrate in infinite different ways; the different modes of vibration of the string give rise, according to the relation between mass and energy of special relativity, to a number of particles of different masses. More complicated mechanisms give also rise to particles that mediate the fundamental (gauge) forces. In particular, if the string is closed, and its length is of the order of Planck's scale (10^{-33} cm), its vibration spectrum includes a particle of zero mass and spin two, which may be interpreted as a graviton — the quantum of gravitational interaction.

The basic mathematics of string theory has been mainly developed in the 1980s by John Schwarz, Michael Green and Edward Witten. Closed strings evolve in a high-dimensional spacetime, describing two-dimensional surfaces⁵ that mathematicians call *Riemann surfaces*.

This fits into a general feature of theoretical physics in the last thirty years. Indeed, starting basically in the 1970s, it was understood that many physical observables may be given a geometric interpretation. As examples, we may cite magnetic charges, instanton charges, quantum anomalies A very interesting idea in this connection is that quantum expectation values of quantum field theories may be regarded as fine invariants of the geometry of the spaces over which the quantum field theory is formulated. Thus, the celebrated Donaldson invariants are expectation values of a supersymmetric Yang–Mills theory, the Gromov–Witten invariants can be analogously understood in terms of quantum strings, etc.

⁵One of the two dimensions is a parameter along the string, the other is time.

The interesting feature here is that algebraic geometry often provides a way for computing such invariants.

Let us briefly discuss this situation in the case of string theory. As we hinted a few lines earlier, classically a string is a loop travelling through space, describing, while time evolves, the *string worldsheet*, a two-dimensional surface in spacetime (Figure 1). This theory may be quantised; the quantum-mechanical string is a very rich theory and aims at providing a unified theory of all fundamental interactions. In particular, strings may undergo quantum processes, e.g., a string can split into two (see Figure 2). According to Feynman's approach to quantum mechanics, in order to compute the probability of transition from a given quantum state to another, it is necessary to average over all possible intermediate states, suitably weighted.

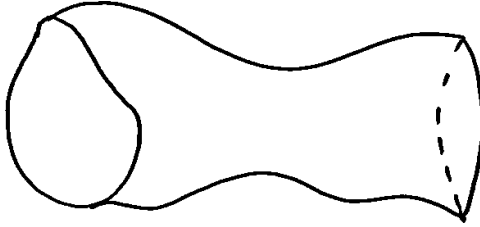


Figure 1 A string worldsheet.

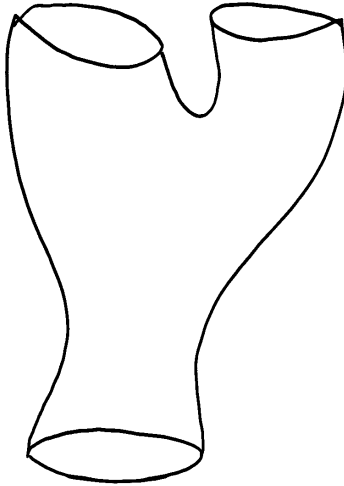


Figure 2 A string splits into two strings.

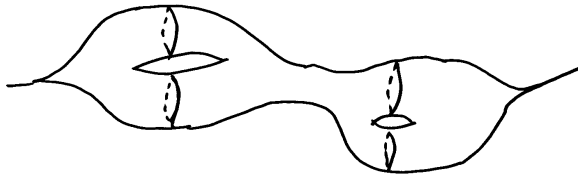


Figure 3 A one-particle state evolves into another one-particle state via string interactions.

Now, the worldsheet of a string may be regarded as a (punctured) Riemann surface, where the punctures correspond to ingoing or outgoing particles. As a consequence, the calculation of the probability amplitude requires an integration over the moduli space of all possible (punctured) Riemann surfaces. Thus, for instance, a Riemann surface of genus two with two punctures (Figure 3) describes a contribution to the physical process where a one-particle state evolves into another one-particle state through a sequence of intermediate string states with different numbers of strings.

It is a very important feature of string theory that the consistency of its quantum theory corresponds to precise geometric features of the spaces involved. The basic question here is that to avoid inconsistencies (that technically correspond to *quantum anomalies*) one needs to assume that spacetime is not four-dimensional, but rather ten-dimensional. To gain contact with ‘real’ physics, one assumes that the ten-dimensional space has a product structure, its factors being a four-dimensional manifold, corresponding to the universe that we observe macroscopically (for instance, we might assume that it is the Minkowski manifold of special relativity), and a six-dimensional manifold whose dimension is so small, to make it unobservable at the usual energy scales (one can indeed prove that the dimension of this space is in inverse proportion to the energy required to see it). This space is called a ‘compactification space’, and it turns out to be equipped with a very rich geometry structure (technically, it is a *Calabi–Yau manifold*).

Now, some string theory models have shown to possess new remarkable properties and these have correspondents in *new geometric structures of the compactification spaces*. The latter helped to shed light on old, basic questions in geometry. A highly nontrivial instance of this situation is *mirror symmetry*. It may happen that two different string models, compactified on different Calabi–Yau manifolds X and Y , are physically equivalent. When this takes place, we say that X and Y are mirror symmetric. (Note that this is a physical definition of an equivalence between two mathematical objects!). From the mathematical viewpoint, the relation between X and Y may be explained in terms of a notion called *quantum cohomology*. The latter is in turn related to problems in enumerative geometry, in particular, enumeration of curves in Calabi–Yau manifolds. This is the reason why quantum string-theoretic computations have counterparts in enumerative geometry.

Thus, mirror symmetry allows one to compute the number of curves of genus g and degree n on a Calabi–Yau manifold (to understand the notion of genus, one should note a complex curve may be regarded as real surface, and the genus is then the number of ‘holes’ or ‘handles’ in the surface). The startling nature of these predictions can be appreciated by browsing the following table.

Numbers of curves of genus g on a quintic hypersurface as predicted by mirror symmetry

Degree	$g = 0$	$g = 1$
$n = 1$	2875	0
$n = 2$	609250	0
$n = 3$	317206375	609250
$n = 4$	242467530000	3721431625
$n = 5$	229305888887625	12129909700200
$n = 6$	248249742118022000	31147299732677250
$n = 7$	295091050570845659250	71578406022880761750
$n = 8$	375632160937476603550000	154990541752957846986500
$n = 9$	503840510416985243645106250	324064464310279585656399500
\vdots	\vdots	\vdots
large n	$a_0 n^{-3} (\log n)^{-2} e^{2\pi n \alpha}$	$a_1 n^{-1} e^{2\pi n \alpha}$

It is interesting to note that these numbers have been first computed by physicists using string theory, and only afterwards the results were confirmed by computations done by mathematicians using entirely different and purely mathematical tools — and with a lot of effort!

References

- [1] Atiyah, M., Hitchin, N., Drinfel’d, V. G. *et al.*, 1978, Construction of instantons, *Phys. Lett.* A65, 185–187.
- [2] Atiyah, M., Hitchin, N., and Singer, I. 1978, Self-duality in four-dimensional Riemannian geometry, *Proc. Roy. Soc. London, Ser. A*, 362, 425–461.
- [3] Atiyah, M., Ward, R. 1977, Instantons and algebraic geometry, *Comm. Math. Phys.* 55, 11–124.
- [4] Bruce Halsted, 1913, The value of science, in *The Foundations of Science: Science and Hypothesis the Value of Science, Science and Method*, The Science Press, New York and Garrison, p. 211.
- [5] Cheng, T. P., Ling, L. F. 1988, Gauge theory of elementary particle physics, Oxford University Press, New York.

- [6] Donaldson, S. K., Kronheimer, P. B. 1990, *The geometry of four-manifolds*, Oxford University Press, New York.
- [7] Gross, D. J. 1992, Gauge theory: past, present, and future? *Chinese Journal of Physics*, 30, 995–972.
- [8] Jackson, J. D., Okun, L. B. 2001, Historical roots of gauge invariance, *Rev. Mod. Phys.* 73, 663–680.
- [9] Le Potier, J. 1995, *Faisceaux semi-stables et systèmes cohérents*, London Math. Soc. Lecture Note Ser. 208, Cambridge University Press, Cambridge.
- [10] O’Raifeartaigh, L. 1997, *The dawning of gauge theory*, Princeton University Press, Princeton.
- [11] Pais, A. 1988, *Inward bound: of matter and forces in the physical world*, Clarendon Press, Oxford/Oxford University Press, New York.
- [12] Penrose, R. 2004, *The road to reality*, Jonathan Cape, London.
- [13] Poincaré, H. 1970 [1905], The value of science, in *The Foundations of Science: Science and Hypothesis, the Value of Science, Science and Method*, translated by B. Halsted, The Science Press, New York and Garrison, 1913.
- [14] Salam, A., Ward, J. C. 1964, Electromagnetic and weak interactions, *Phys. Lett.* 13, 168–171.
- [15] Serre, J.-P. 1956, Géométrie algébrique et géométrie analytique, *Ann. Inst. Fourier* 6, 1–42.
- [16] Utiyama, R. 1956, Invariant theoretical interpretation of interaction, *Phys. Rev.*, II. Ser. 101, 1597–1607.
- [17] Weinberg, S. 1967, A model for leptons, *Phys. Rev. Lett.* 19, 1264–1266.
- [18] Weyl, H. 1918a, *Raum, Zeit, Materie*, Springer, Berlin, Space, time and matter, Dover, New York 1952, translation of the 4th edn.
- [19] Weyl, H. 1918b, Gravitation und Elektrizität, *Sitz. preuss. Akad. Wiss.* 465.
- [20] Weyl, H. 1928, *Gruppentheorie und Quantumechanik*, The theory of groups and quantum mechanics, Dover, New York 1950.
- [21] Ward, R. S., Wells, R. O. 1991, *Twistor Geometry and Field Theory*, Cambridge University Press, Cambridge.

This page is intentionally left blank

CHAPTER 2

Quantum Gravity and Quantum Geometry

MAURO CARFORA

*Dipartimento di Fisica Nucleare e Teorica,
Università degli Studi di Pavia,
via A. Bassi 6, I-27100 Pavia, Italy*

and

*Istituto Nazionale di Fisica Nucleare, Sezione di Pavia,
via A. Bassi 6, I-27100 Pavia, Italy
mauro.carfora@pv.infn.it*

I. Introduction

The deep connections between physics and geometry have a long historical record and have been the cause of important paradigmatic shifts in both disciplines. This cross-fertilisation, no matter how highly specialised and audacious it may appear nowadays, has its origins in ancient times when the desire to make more and more accurate land and astronomical measurement strongly influenced the development of geometry, and when Hellenistic astronomers realised that Euclidean geometry brought order into the vagaries of celestial phenomena. These two situations can be considered as typical along the whole history of physics and mathematics: (i) New mathematics is developed in connection with the quest of understanding important physical questions; (ii) New physics is developed from known mathematics. Examples abound, to wit: Newton mechanics and the theory of gravitation required the development of calculus; Maxwell's electrodynamics and analytical mechanics paralleled the development of the theory of partial differential equations and complex analysis; Einstein's relativity found its natural language in terms of the differential geometry developed by K. F. Gauss, B. Riemann, and T. Levi-Civita; The Standard Model of elementary particle physics, Yang–Mills theory, builds upon the theory of connections over fibre bundles and pays back high dividends to mathematics by bringing techniques of non-linear field theory in geometry and



Figure 1 Classical mechanics, here epitomized by Galileo's Principle of Relativity, required the development of calculus.

differential topology; quantum mechanics cannot be separated from the theory of Hilbert spaces and functional analysis and its methods have provided a continuous source of inspiration for modern analysis and geometry.

In recent years methods from quantum field theory have provided, for reasons which are not yet fully understood, effective means for solving problems in geometry and topology that were previously considered quite intractable. In particular, the use of Feynman's *sum over histories* formalism (i.e., functional integration) in gravitational physics has allowed the development of strategies for successfully addressing basic questions in the theory of moduli space of Riemann surfaces, in algebraic geometry, in knots theory, and in the topology of three-manifolds. This fact is quite surprising since geometrical functional integration does not have a mathematically proven existence. Methods of this sort provide fanciful expressions built up out of a non-existent invariant measure on infinite-dimensional sets of

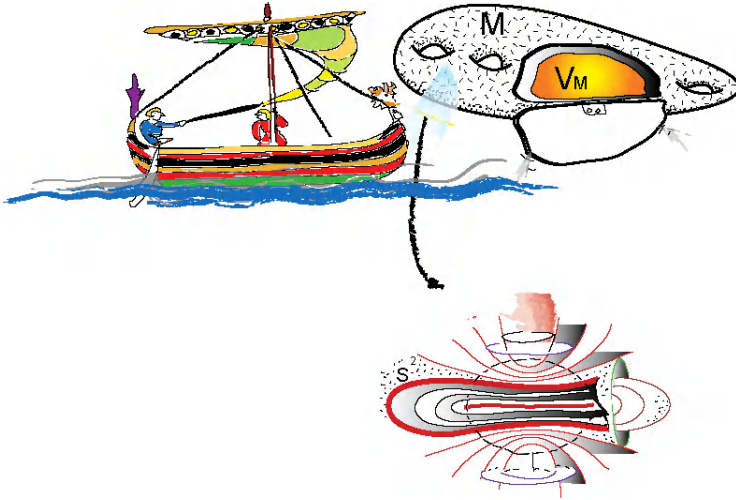


Figure 2 In our journey to a deeper understanding of the nature of gravity we are witnessing indications of the existence of a new territory: quantum geometry.

spacetime geometries and out of an action which is unbounded when evaluated on the geometrical configurations the measure would generically select. From an algorithmic point of view they can be considered, at best, as a book-keeping device (i.e., a generating functional) for the quantum fluctuations of geometrical structure around a given (classical) spacetime geometry. Leaving these technicalities aside, one has to admit that to whatever degree of significance one is willing to accept the status of functional integration, we have to grant it a basic role in solving geometrical problems. This latter remark indicates that *quantum geometry* may exist as a mathematical category and that its development and proper understanding cannot be disentangled from the analysis of one of the basic problems of modern theoretical physics, the quantization of gravity.

2. Glimpses of Quantum Gravity

A basic tenet of quantum theory is that forces between particles are due to exchanges of quanta, whereas the cornerstone of the modern theory of gravitation, general relativity, is that the gravitational field is not a force but rather a manifestation of the non-trivial dynamics of spacetime geometry. In such a deceptively simple observation we have a glimpse of the root of the difficulties in making sense of a quantum theory of gravity.

One can argue that in order to understand the nature of quantum geometry and of its relation to quantum gravity we should try to establish a direct connection

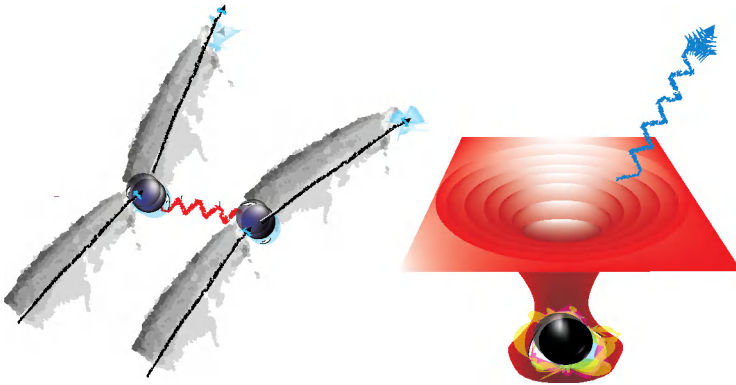


Figure 3 According to quantum mechanics all natural forces are due to an exchange of quanta. However, according to general relativity, gravitation, no matter of how intense, is only the apparent manifestation of the deformation of spacetime geometry.

between geometry and quantum mechanics. To some extent, this can be done by quantizing the motion of test particles on Riemannian manifolds, an analysis which exploits the fact that the geometry of a Riemannian manifold M can be probed by the field of its geodesics. The underlying strategy is to reconstruct the (classical) geometry of M and its (quantum) deformations out of the quantum dynamics of its geodesic flow. Such an approach can be considered as the starting point of A. Connes' non-commutative geometry program.

Actually, non-commutative geometry is just an aspect of the many tight requirements that quantum mechanics and relativity put on the spacetime arena. From a mathematical point of view, most of these requirements naturally come about in dealing with one of the leading candidates for a full-fledged quantum theory of gravity: *string theory*. The rationale of this remark is twofold: (i) String theory requires the quantization of extended objects (open and closed strings, and branes); (ii) The quantization of extended objects put strong constraints on the possible spacetime geometries where these objects can be geometrically realised. These constraints are generated by the quantum dynamics and by the underlying symmetries of the extended objects themselves. Roughly speaking, strings give rise to their own ambient (quantum) spacetime. The theory, in this sense, is by its very nature a generator of natural candidates for quantum geometries. Without any doubts we have here a framework which is rather detached from general relativity, which, notwithstanding its name, is a rather rigid theory. The spacetime geometry of general relativity is necessarily four-dimensional, of Lorentzian signature and it is dominated by the invariance under the action of the diffeomorphisms group: the group of all smooth, invertible point transformations. An invariance, this latter, rooted into Einstein's equivalence principle and which makes spacetime

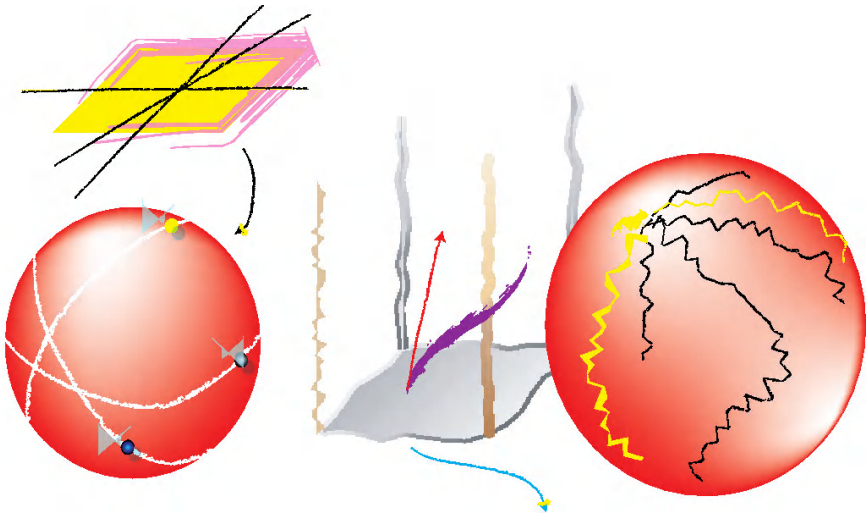


Figure 4 The geometry of a Riemannian manifold can be probed by the geodesic motion of test particles (left). The quantization of this motion (right) can provide a first indication of the nature of quantum geometry: the strategy is to recover the classical geometry of the underlying manifolds and the possible deformations of such a geometry out of the quantization of the geodesic field.

points (*events*) the basic objects of the theory. It is possible to develop a quantum theory of gravity which is well-adapted to these rigid kinematical structures of general relativity: *Loop quantum gravity*. Its quantum dynamics stresses more or less directly the basic role of the diffeomorphisms group (hence of points). However, its impact on the development of quantum geometry is rather limited being strictly confined to geometrical aspects which concern the constraints that an underlying four-dimensional diffeomorphisms invariant theory imposes on quantum fluctuations of the spacetime geometry. Even if, from the point of view of basic physics, this is certainly a positive characteristic of loop quantum gravity, it puts strong and limitative constraints when addressing quantum geometry issues. Thus, in the remaining part of my discussion I will try to focus on some of the more flexible aspects of the correspondence between string theory and quantum geometry.

3. Strings and Geometry

Let us start by recalling some of the basic characteristics of the motion of a point particle in a given spacetime geometry. Such a motion is described by a parameterized curve which maps the oriented real line in a four-dimensional

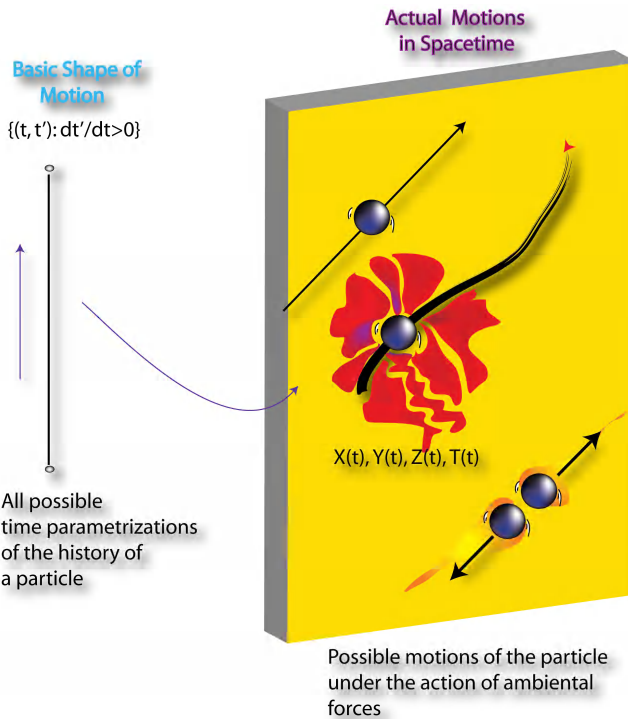


Figure 5 The geometry of motion.

spacetime. In this picture, the real line plays the role of describing all possible parameterisations of the history of a particle, i.e., the *basic shape of the motion* of a point particle. Whereas the curve in spacetime represents the *actual physical motion* of the particle under the action of the external forces.

Note that there is just one possible basic shape of motion for a particle, the real line. Indeed, the real line which parameterizes the motion of a point particle cannot split or join: if this happens then it would imply that we were dealing with two or more point particles and, in order to recover their dynamics, we have to provide further information. In particular, we need to know the nature of the interaction forces acting among such particles when they split or join, (i.e., besides a basic shape of motion, we have to know the nature of the *interaction vertices*).

There is an equivalent description of the motion of point particles: We can consider the spacetime coordinates X , Y , Z , and T describing the actual motion in spacetime as (Poincaré-valued) fields living on the manifold which parameterizes the motion, i.e., on the real line. This is a perfectly sensible way of studying the

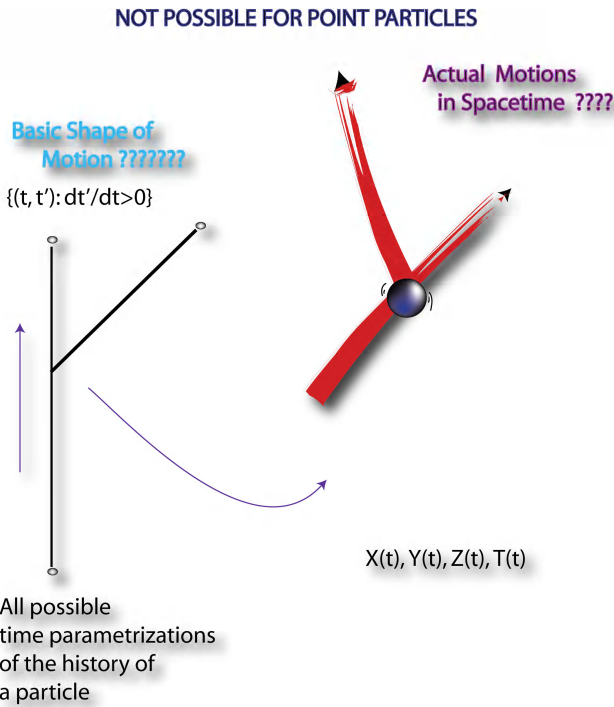


Figure 6 A kinematical set up which is not possible for point particles.

dynamics (in particular in absence of external forces), perhaps not very familiar in elementary mechanics but well-known in relativity. String theory is, in a rather precise sense, a natural extension of such a description.

The basic idea of string theory is to replace point-like objects by string-like objects: one-dimensional (closed or open) strings which (classically!) we can think of as evolving in a given spacetime M . From a mathematical point of view this is strictly related to the theory of minimal (or maximal) surfaces in M , (*harmonic map theory*).

Also for strings it is worth recalling some basic features of the characters of their classical motion in a given spacetime geometry. In the simplest situation, the motion of a string is described by a parameterized surface which maps an *oriented cylinder* in a four-dimensional spacetime. The cylinder plays here the role of describing all possible parameterizations of the history of a free closed string, i.e., the *basic shape of the motion* of a closed string is itself an *abstract surface*. Here, I am emphasizing the adjective abstract since the cylinder is not in relation with any larger dimensional space in which is to be immersed. The *immersion*

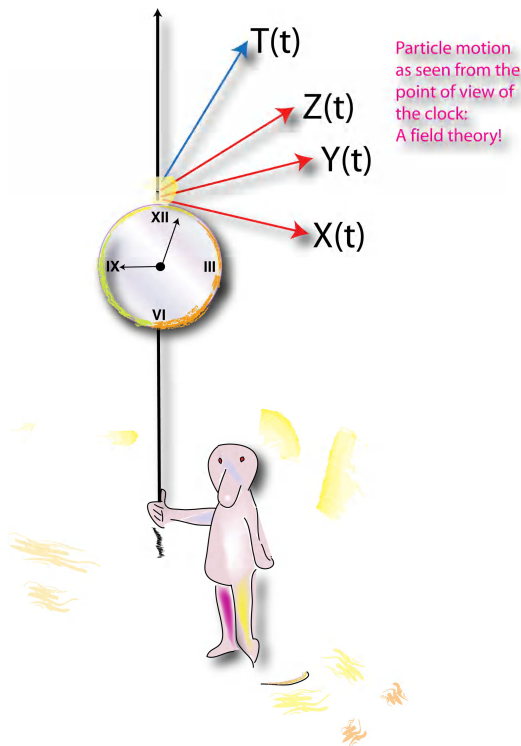


Figure 7 An equivalent field-theoretic description of the motion of a particle.

comes only when describing the surface in spacetime and represents the *actual physical motion* of the string. This is a subtle but important point with far reaching consequences in the quantum theory.

As in the case of point dynamics, we have an equivalent *field theoretic description* of the motion of a string that allows us to interpret the embedding coordinates, describing the string as a two-dimensional surface in spacetime, as (Poincaré) fields on an abstract two-dimensional surface.

Note that now nothing prevents the parameters space of the theory from being topologically non-trivial. Geometrically, this follows from the fact that the parameters space is two-dimensional and what really matters in its characterization is the existence of a local product structure (roughly allowing a time and a space parameterization) and its global shape. In more technical words, what is needed is a *conformal structure* on the abstract surface describing the parameters space. Moreover, from the physical point of view we do not need to provide any longer information on the nature of the interaction vertex between strings which apparently occurs

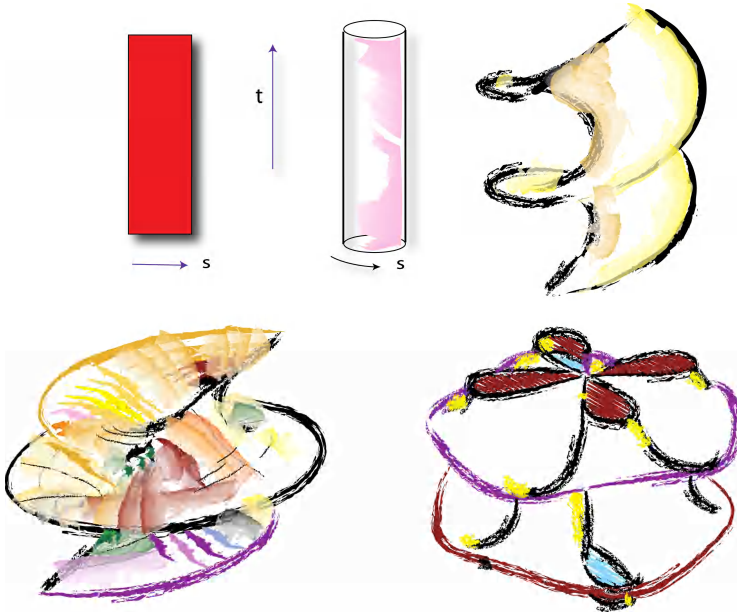


Figure 8 Minimal surfaces.

when the topology of the surface changes. This is so because in the enveloping spacetime there is no actual interaction vertex: Lorentz invariance delocalizes it.

What is the basic strategy in quantizing strings in such a rich kinematical set up? Well, nobody really knows how to write down a full-fledged quantum string theory, if nothing else because we are so used to the notion of point that it is very difficult to figure out how the quantum dynamics of extended objects like strings should deform a space(time) generated by points. However, strings can be studied on the perturbative level. According to the quantum democracy principle (associated with Feynman's sum over histories) we have to sum over all possible immersions of the surface in the spacetime M , with a bias related to an action associated with the immersion. As we have recalled before, this is equivalent to study the quantum theory of matter fields on the abstract surface providing all equivalent parameterizations of the immersion. Such a quantum field theory is an example of what is technically known as a *quantum conformal field theory*. The adjective conformal emphasizes the fact that such quantum theories should depend only on the conformal structure of the surface on which the matter fields live.

Roughly speaking, such models of conformal field theories should describe the classical solutions of string theory and their perturbative deformations, more or less as the quantization of the geodesic flow on a Riemannian manifold may allow to

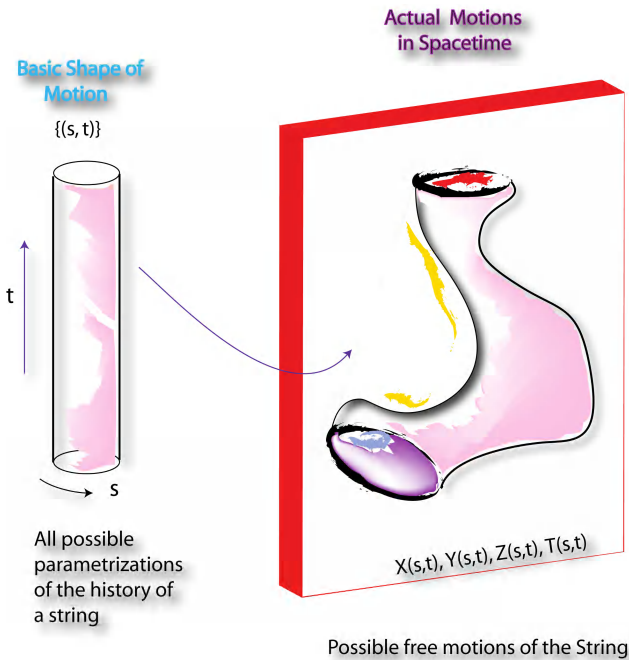


Figure 9 The kinematics of (closed) string motion.

recover the classical geometry of the manifold and some of its relevant deformation (e.g., non-commutative geometry). This aspect of string theory is related to the behaviour of the theory on large scales, where quantum gravity effects basically have a perturbative nature around a classical background spacetime geometry. We get a non-trivial quantum dynamics already at this step. A renormalization of the spacetime geometry is indeed required since the quantum fluctuations of the matter fields introduce infinite counter-terms modifying the original spacetime metric.

It is quite amazing that in such a perturbative framework one gets Einstein equations of classical general relativity as a condition for a sensible physics.

These remarks raise the question of how we probe the effective spacetime geometry in string theory. In this connection, the basic observation is that the fluctuations of the spacetime fields tend to introduce a fundamental length scale for each spacetime dimension. The net effect is that we never see small radii in effective spacetime geometry. This is a basic feature of string theory.

The next step in quantizing (perturbatively) string theory is to sum the matter conformal field theory over all possible shapes of the surface describing the string parameterizations. As we have stressed, in string theory what really matters is the

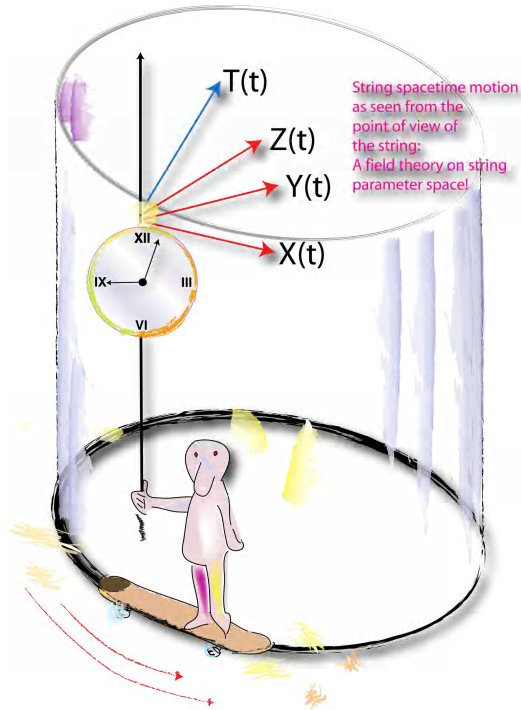


Figure 10 The motion of a string as a field theory on the cylinder surface providing string parameters space.

conformal structure of the abstract surface over which the fields live. There can be a value of distinct conformal structures for each given surface topology, (think, for instance, of the difference between *thin* and *fat* two-dimensional tori, without referring to their actual size). The actual size of the surface can be described by a field which tells us how to locally rescale the shape of the surface until it reaches the size we like. Classically this field is irrelevant, as the choice of a particular time rate in describing the motion of a point particle is irrelevant. Such irrelevance should hold also at the quantum level. However, from the quantum point of view, this field couples with surface curvature and introduces another length scale into the game!

These two length scales induced by quantum fluctuations generate competing effects on the dynamics of strings and in general destroy the basic conformal symmetry of the theory. However they compensate in particular spacetime dimensions! The existence of such *critical dimensions*, some of which are compact, is a rather impressive property but it has been overemphasized: it is not the subtler property of string theory.

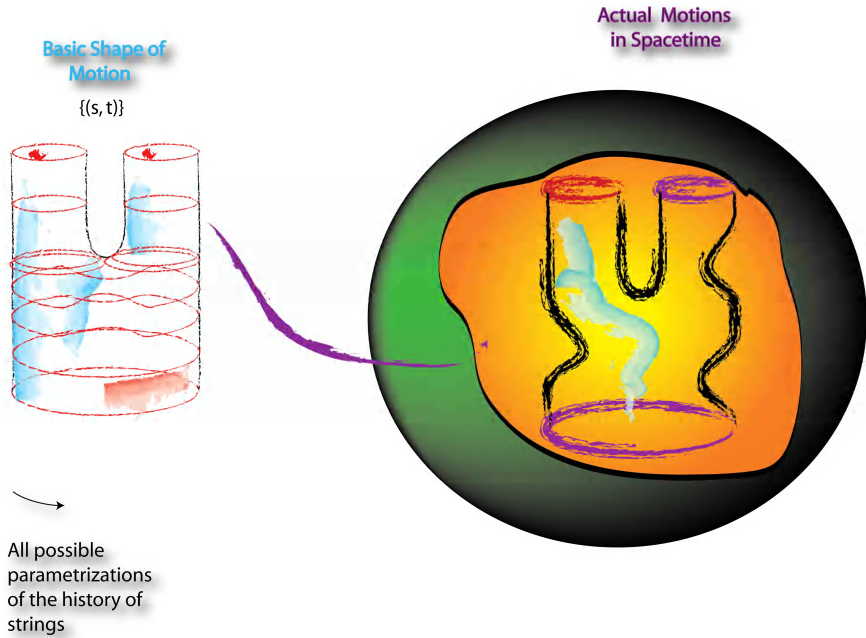


Figure 11 Non-trivial topology in string parameter space.

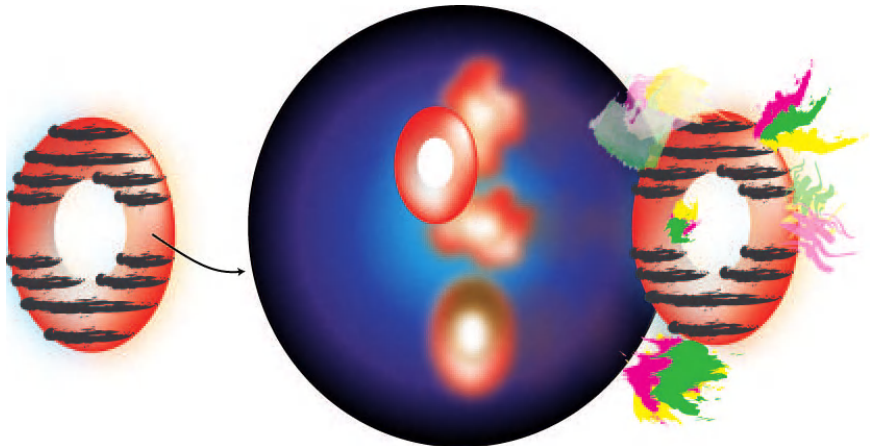


Figure 12 Summing over different embedding fields is equivalent to a quantum conformal field theory on the abstract surface defining the parameters space of the string. The matter fields associated with the embedding variables are here depicted as fluctuating colour patches.

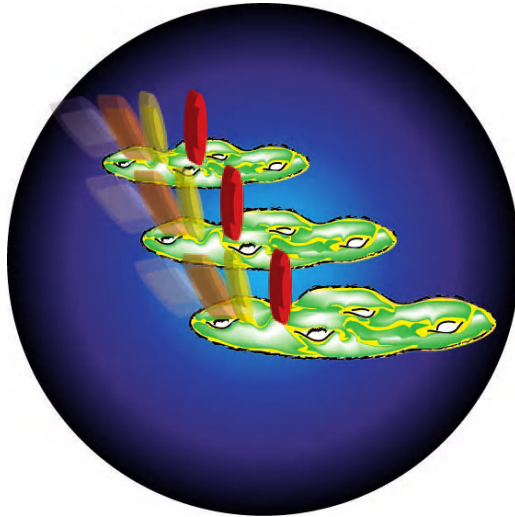


Figure 13 Quantum fluctuations in the matter fields modify spacetime geometry already at the perturbative level.

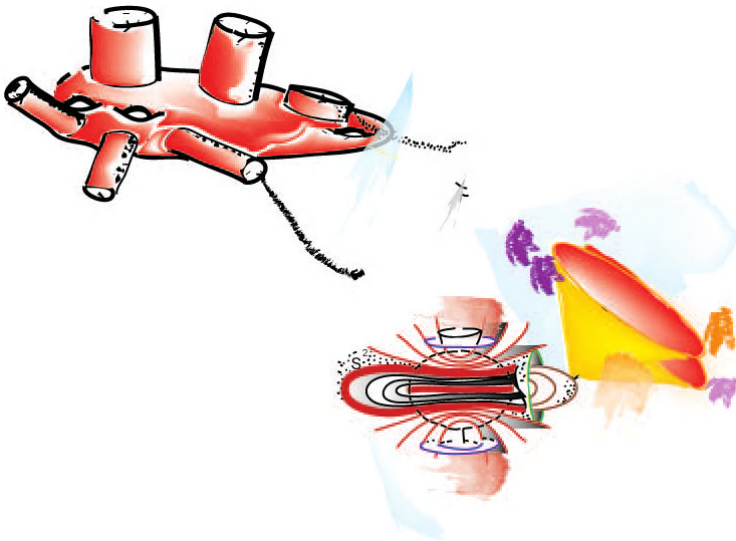


Figure 14 Einstein equations are obtained as a necessary condition for having a sensible (first-order) perturbation theory of the quantum fluctuations of the matter fields living on a string.

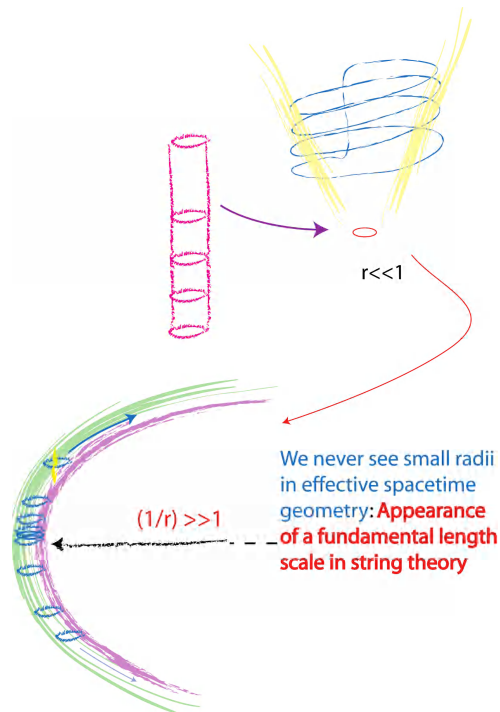


Figure 15 How we probe spacetime geometry in string theory: for large scales strings behave like point particles (left). When probing small scales they tend to wrap around (right). The two pictures turn out to be dual to each other since strings are extended objects and (quantum) momentum states can be exchanged with (quantum) winding states. This duality is not possible for point particles.

Perhaps the most important aspect of string geometry is that distinct spacetimes can be dual descriptions of the same Physics. This comes about when evaluating how the quantum field fluctuations located at different points of the string are correlated to each other. These quantities (*correlation functions*) provide global information (often of topological origin) on the nature of the ambient spacetime (defined by the quantum dynamics of the fields living on the surface) and define the physical system the string theory describes. It often happens that two distinct physical systems have the same correlation functions. In other words, strings moving on a given spacetime may behave as strings moving on a different spacetime. This is one of the most amazing properties of string theory, indicating the existence of distinct manifolds having a common *quantum geometry ancestor*. An example is afforded by mirror symmetry, which relates topologically distinct pairs of

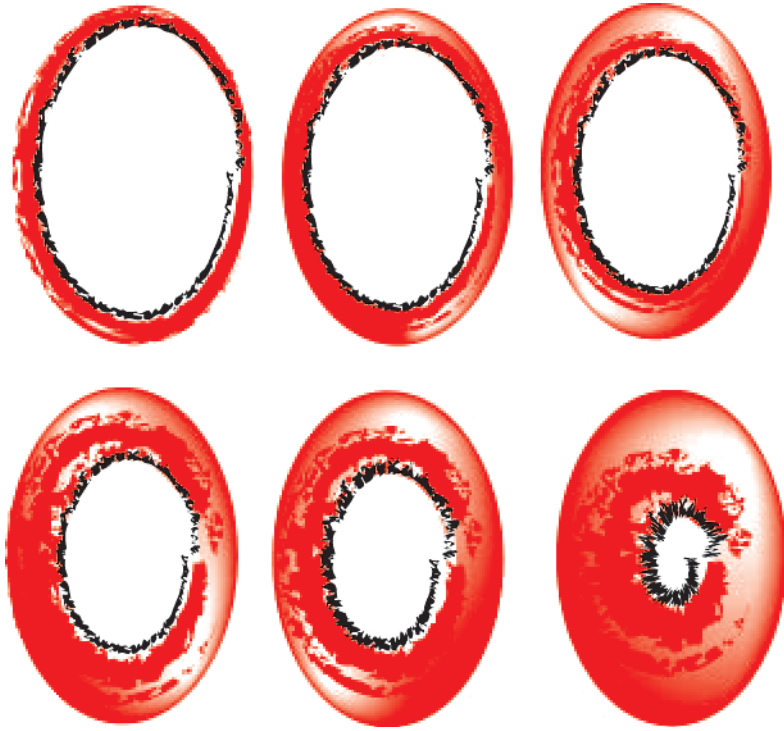


Figure 16 Conformal geometry provides the geometrical information coded into the local shape of the surface, i.e., how to tell when two infinitesimal triangles on the surface are similar to each other. Whereas for a point particle there is just one basic shape (a line associated with proper time evolution), for a surface we have many non-equivalent shapes. The actual size of a surface can be recovered if we provide the further information contained in a field, living on the surface, which tells us how to locally rescale the size of infinitesimal triangles while maintaining invariant their local shape.

Calabi–Yau manifolds. From the geometrical point of view such mirror conjugation uncovers unexpected common structures among distinct manifolds and gives rise to mathematical identities with applications in enumerative geometry. It must be stressed that these dualities in string theory are very difficult to prove, even at the level of physical rigour. In a rather definite sense, string theory often only offers a natural guessing ground for quantum geometry. However, the deep nature of the geometrical and topological relations it allows us to uncover pays back and mitigates the hard work needed for their rigorous proof. One has to admit that to whatever degree of significance one is willing to accept the status of string theory in physics, one has to grant it a basic role at least in experimental mathematics!

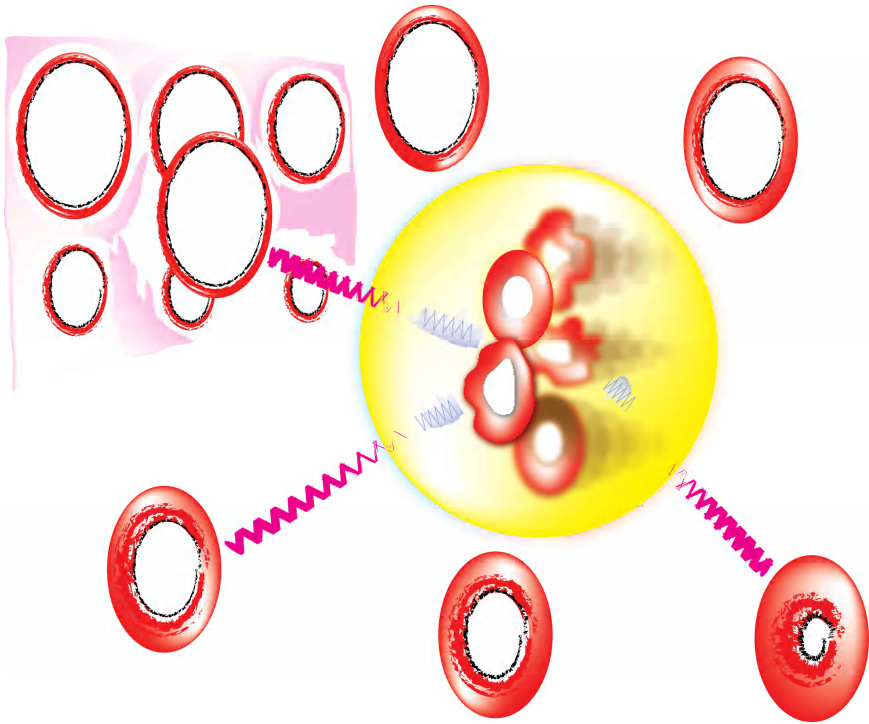


Figure 17 Competition among shape quantization and matter field quantization.

There is no room here even for a sketchy indication of how duality symmetries work, nonetheless, I hope to have stimulated your interest in a new fascinating chapter of the relation between physics and geometry. The unification scheme which emerges from this chapter indicates that string theory has already led to a profound paradigm shift in geometry and that most likely it will provide the natural language for quantum geometry.

Recommended Reading List

A very good and readable elementary introduction to string theory is:

Zwiebach, B. 2004, *A first course in String Theory*, Cambridge University Press, Cambridge.

Loop quantum gravity is nicely discussed in:

Rovelli, C. 2004, *Quantum Gravity*, Cambridge Monographs on Mathematical Physics, Cambridge University Press, Cambridge.

An interesting account of the interaction between mathematics and physics is described in: Vafa, C. 1999, On the future of Mathematics/Physics Interaction, in *Mathematics: Frontiers and Perspectives*, Arnold, V. *et al.* eds., Am. Math. Soc., pp. 321–328.

Vafa, C. Unifying Themes in Topological Field Theories, hep-th/0005180, Talk presented in conference on Geometry and Topology in honour of Atiyah, M., Bott, R., Hirzebruch, F. and Singer, I. Harvard University, May 1999; see also *Geometric Physics*, hep-th/9810149, *Documenta Mathematica*, Extra Volume ICM 1998, 537–556.

A crystal clear presentation of conformal field theory and of its relation to strings can be found in:

Frohlich, J., Gawedzki, K. *Conformal Field Theory and Geometry of Strings*, hep-th/9310187.

This page is intentionally left blank

CHAPTER 3

The de Sitter and Anti-de Sitter Universes

UGO MOSCHELLA

*Dipartimento di Fisica e Matematica
Università dell'Insubria, 22100 Como
INFN, Sezione di Milano
Via Valleggio 11, 22100 Como, Italy
Ugo.Moschella@uninsubria.it*

I. Introduction

The year 1998 has witnessed two major revolutions in physics that have a crucial feature in common: both of them are based on a nonvanishing cosmological constant.

The first revolution comes from some important progress in the astronomical observations [1, 2] which have led to the surprising conclusion that the recent universe is dominated by an almost spatially homogeneous exotic form of energy density to which there corresponds a negative pressure. Such negative pressure acts repulsively at large scales, opposing itself to the gravitational attraction. This effect may explain the accelerated expansion of the universe and may account for an important part of the missing mass. It has become customary to characterise such energy density by the term 'dark'.

The simplest and best known candidate for the dark energy is the cosmological constant. It was Einstein himself who introduced a constant term in the equations for the gravitational field as a mechanism to obtain static cosmological solutions [3]. However, this possibility was immediately set aside because the static solution so obtained was unstable and, even more, because the observations performed shortly after by Edwin Hubble pointed towards a non static expanding universe. If there is no quasi-static world then away with the cosmological constant (postcard from Einstein to Hermann Weyl — 1923) and the cosmological constant was to be downgraded to a mere mathematical curiosity for half a century.

In the late 1970s some serious problems with the standard cosmological model led to the idea that a field with negative pressure, producing effects similar to a

cosmological constant, could be at the origin of an exponential expansion of the universe in the first instants of its life [4, 5]. Such an expansion, known as inflation, is at present an essential characteristic of the majority of the Big Bang cosmological models in use. However, the energy density which is required to feed inflation is enormously larger than the dark energy which is observed in recent cosmic epochs ($z < 1$) and it is unclear whether there is a relation between the two phenomena.

At any rate, even when the inflationary paradigm became an integral part of the standard cosmological model, the idea of the physical irrelevance of the cosmological constant in the interpretation of the recent history of the cosmos persisted.

This consolidated belief was shattered and quickly abandoned in 1998, following the discovery that the expansion of the universe in the present epoch is accelerated [1, 2]. This discovery was at first based on measurements of the distance-red shift relation of distant type Ia supernovae, but it has been subsequently confirmed and strengthened by independent observations.

As of today, the Λ CDM (Lambda Cold Dark Matter) model, which is obtained by adding a cosmological constant to the standard model, is the one which is in better agreement with the cosmological observations, the latter being progressively more precise. For example, the SLS (Supernova Legacy Survey) results show that dark energy behaves as a cosmological constant within a few per cent error.

In the above context the de Sitter geometry, which is the homogeneous and isotropic solution of the vacuum Einstein equations with cosmological term, appears to take the role of reference geometry of the universe. In other words, it is the de Sitter geometry, and not the Minkowski one, which would be the geometry of empty spacetime (namely of spacetime deprived of its matter and radiation content). In addition, if the description provided by the Λ CDM model is correct, the remaining energy components must in the future progressively thin out and eventually vanish, thus letting the cosmological constant term alone survive, as it appears evidently from Friedmann's equation:

$$\frac{H^2}{H_0^2} = \Omega_{M0} \frac{a_0^3}{a(t)^3} + \Omega_{R0} \frac{a_0^4}{a(t)^4} + \Omega_{\Lambda 0} + \Omega_{K0} \frac{a_0^2}{a(t)^2}. \quad (1)$$

Therefore, the de Sitter geometry is the one to which the geometry of the universe approaches asymptotically. These considerations show the actuality and the importance that the de Sitter geometry acquires in present-day cosmology, in addition to the traditional role which it plays in the context of inflationary models; the de Sitter geometry takes a role in contemporary cosmology that is in a way more relevant than the one played by the flat Minkowski geometry.

The second revolution has taken place in string theory: this is the AdS/CFT correspondence (Anti-de Sitter/Conformal Field Theory) [6]. The original conjecture is about a correspondence between string theory on $AdS_5 \times S_5$ and a Yang-Mills theory on the conformal boundary of AdS_5 . However, a general simple idea of

holographic type is the common feature at the basis of the vast literature referring to the original Maldacena's paper: the idea is that the spacetime we live in is a 'brane' or a boundary of a higher dimensional manifold whose curvature is essentially provided by a negative cosmological constant, i.e., an anti-de Sitter universe or a portion of it. This idea has far-reaching consequences and has a broad domain of application ranging from calculations of nontrivial amplitudes in quantum field theory (aiming to QCD) to the birth of new ideas for cosmology and quantum gravity. The renewed situation sets out great challenges but also opportunities for new physical and mathematical ideas.

From the geometrical viewpoint, among the cousins of Minkowski spacetime i.e., the class of Lorentzian manifolds, the de Sitter and anti-de Sitter spacetimes are its closest relatives. Indeed, like the Minkowski spacetime, they are maximally symmetric, i.e., they admit kinematical symmetry groups having a maximal number of generators.¹ Maximal symmetry also implies that the curvature is constant (zero in the Minkowski case). Owing to their symmetry, it is possible to give a description of the de Sitter universes without using the formalism of general relativity at all. However, it is worth saying right away that, even if they share important features with Minkowski spacetime, their physical interpretation is quite different and the technical problems to be solved in order to merge de Sitter spacetimes with quantum physics are considerably harder.

Many of the traditional concepts and ideas of quantum field theory have to be reconsidered. In particular, in the traditional formulation of quantum field theory it is of crucial importance the commutativity of spacetime translations which does not hold any more in presence of a cosmological constant. The mathematical problems arising from this simple fact are of considerable difficulty and a true solution seems not to be accessible to heuristic methods. Due to its topical character, the literature on the de Sitter and anti-de Sitter universes is very broad but the results obtained so far do not reach much beyond the construction of free theories. Difficulties persist as regards both the acquisition of general and structural results as well as in operational and computational possibilities: computations which in the Minkowskian case would be simple and occasionally even trivial become quickly prohibitive or even impossible. This, in spite of the fact that one is dealing with maximally symmetric manifolds. The technical and structural reason for this discrepancy lies exactly in those characteristic aspects of Minkowski spacetime which do not persist in the presence of curvature and which, therefore, render ineffective the similarity existing between the de Sitter and anti-de Sitter models and flat spacetime. Of particular importance is foremost the absence of a commutative translation group

¹In the four-dimensional case a Riemannian manifold has an isometry group with at most ten generators. In the Minkowski case the isometry group is the Poincaré group and the ten independent transformations have a familiar physical interpretation: one time translation, three spatial translations, three rotations and three boosts.

and of the consequent Fourier representation of spacetime, the latter being the linear energy-momentum space which, mathematically speaking, is a copy of the same Minkowski spacetime.

In the following we will give an easy introduction to the geometry of the de Sitter manifolds. We will discuss then some aspects of quantum field theory in the above context.

2. A Visual Description of the de Sitter Manifolds

2.1. Curved spaces of constant curvature

One easy way to replace the usual flat geometry of the Euclidean physical space \mathbb{R}^3 with some curved geometry consists in moving to a fictitious four-dimensional flat world and considering there the geometry of convenient three-dimensional hypersurfaces. The simplest curved model of space is the surface of a hypersphere embedded in a four-dimensional Euclidean flat space \mathbb{R}^4 :

$$\mathbb{S}^3 = \{x \in \mathbb{R}^4, x_1^2 + x_2^2 + x_3^2 + x_4^2 = a^2\}. \quad (2)$$

\mathbb{S}^3 is homogeneous, isotropic and has positive curvature with value $6/a^2$. The six-dimensional invariance group of \mathbb{S}^3 is simply the rotation group $\text{SO}(4)$ of the four-dimensional ambient space; it can be interpreted as the group of motions of the spherical space in the same way as the Euclidean group $\text{E}(3)$ (translations and rotations) is the group of motions of \mathbb{R}^3 . The main difference is that there are no commutative ‘translations’ on \mathbb{S}^3 .

All the non-Euclidean geometrical properties of the hypersphere come by restriction to it of the Euclidean geometry of the fictitious ambient space. In particular all geodesics, that are the analog in the curved geometry of what are straight lines in the flat case, can be obtained by intersecting the hypersphere with two-planes passing through the geometrical centre of the sphere (see Figure 1). One recognises immediately that in this geometry ‘straight lines’ are maximal circles.

The second possibility is more elaborated and produces a space with negative curvature. One moves again to a fictitious four-dimensional world, but now this is a four-dimensional Minkowski spacetime \mathbb{M}^4 (loosely speaking, a timelike direction has been added to the Euclidean \mathbb{R}^3 , while in the previous case a spatial direction was added). Here, a model of space with negative constant curvature is the upper sheet of the two-sheeted hyperboloid \mathbb{H}^3 :

$$\mathbb{H}^3 = \{x \in \mathbb{M}^4, x_0^2 - x_1^2 - x_2^2 - x_3^2 = a^2\}. \quad (3)$$

As shown in Figure 2 the lightcone emerging from any point of \mathbb{H}^3 does not meet the surface anywhere else. This means that, in the ambient spacetime, the surface is *spacelike* and, as such, it is a good model for a space. As before the geometry

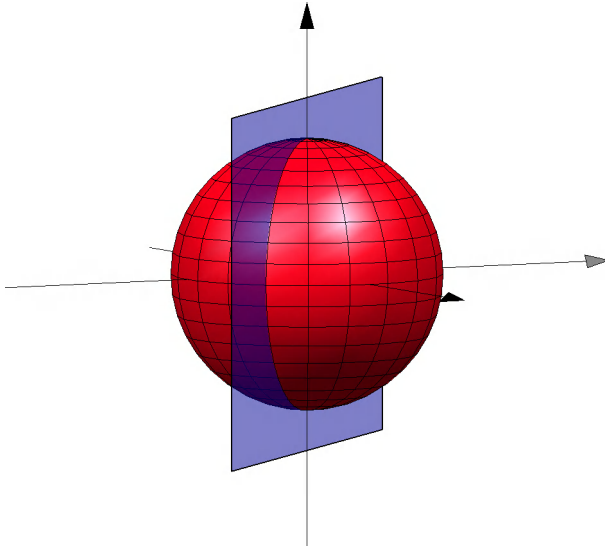


Figure 1 A spherical model of space (positive curvature). Geodesics are great circles and are obtained by intersecting the sphere with two-planes passing through the centre of the sphere in the ambient space.

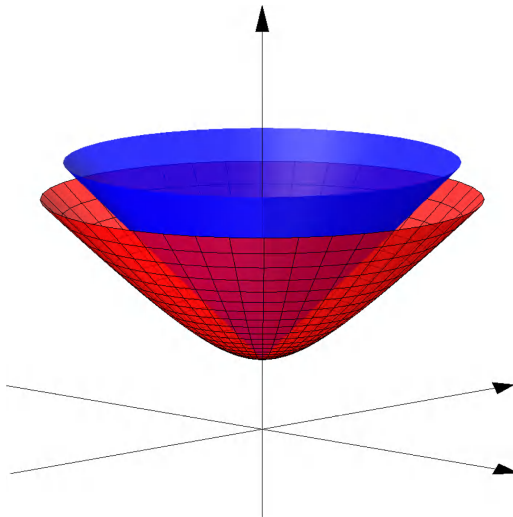


Figure 2 A hyperbolic model of space (negative curvature). \mathbb{H}^3 (the red surface) is spacelike in the ambient Minkowski spacetime. Geodesics are branches of hyperbolae.

of \mathbb{H}^3 is constructed by restriction of the Lorentzian geometry of the ambient Minkowski spacetime \mathbb{M}^4 . In particular, the six-dimensional isometry group of \mathbb{H}^3 is the Lorentz group $\text{SO}(1, 3)$ of the ambient spacetime. Geodesics are branches of hyperbolae, obtained as before by intersecting \mathbb{H}^3 with two-planes containing the centre.

2.2. The de Sitter universe

By analogy we now introduce a five-dimensional Minkowski spacetime \mathbb{M}^5 by adding a spacelike direction to \mathbb{M}^4 (just as we did in the spherical case). In \mathbb{M}^5 we consider the hypersurface with equation

$$dS_4 = \{x \in \mathbb{M}^5, x_0^2 - x_1^2 - x_2^2 - x_3^2 - x_4^2 = -R^2\}. \quad (4)$$

This is the de Sitter spacetime [7] (see Figure 3). It has constant negative curvature $-12/R^2$ (the sign depends on conventions) and reproduces (after a renormalisation) Minkowski spacetime in the limit of zero curvature (i.e., when the radius R tends to infinity).

The causal structure of dS_4 is induced by restriction of the Lorentzian geometry of the ambient Minkowski spacetime \mathbb{M}^5 exactly as the geometry of the sphere was determined by the Euclidean geometry of the ambient \mathbb{R}^4 . In particular, the

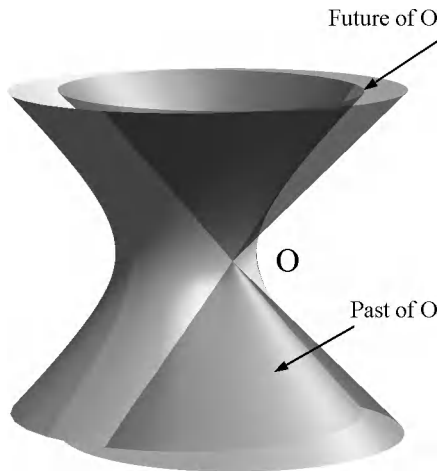


Figure 3 The de Sitter manifold represented as an hyperboloid embedded in a Minkowski spacetime with one dimension more. The future and past cones of the ambient spacetime induce the causal ordering of the de Sitter manifold. The regions shadowed by the five-dimensional cone issued from the event O are indeed the past Γ_O^- and the future Γ_O^+ of O .

de Sitter line element is obtained concretely by restricting the five-dimensional invariant interval to the manifold dS_4 :

$$ds^2 = [(dx_0)^2 - (dx_1)^2 - (dx_2)^2 - (dx_3)^2 - (dx_4)^2]_{dS_4} \quad (5)$$

This line element is the most symmetrical solution of the field equations written down by Einstein in 1917, where he introduced the famous cosmological constant Λ [3]. The radius R corresponding to a given value of Λ is

$$R = \sqrt{\frac{3}{\Lambda}}.$$

A pivotal role is played by the five-dimensional lightcone of the ambient spacetime:

$$C = \{\xi \in \mathbb{M}^5, \xi_0^2 - \xi_1^2 - \xi_2^2 - \xi_3^2 - \xi_4^2 = 0\}. \quad (6)$$

The cone C induces the causal ordering of the events on the de Sitter manifold; it also plays the role of de Sitter momentum space. The de Sitter spacetime has a boundary at timelike infinity (while timelike infinity of the Minkowski manifold is a point). The cone C also provides a description of this boundary, which may be used instead of a Penrose diagram.

The de Sitter kinematical group coincides with the Lorentz group of the ambient spacetime $SO(1, 4)$. As for the sphere, there are no commutative translations on the de Sitter manifold. This fact is a source of considerable technical difficulties in the study of de Sitter quantum field theory. A study of the complex de Sitter manifold with applications to quantum field theory has been described in [8, 9].

The relationship between the de Sitter universe and the geometry of the sphere is deeper than a mere analogy. Indeed, for imaginary times

$$x_0 \rightarrow ix_0$$

the (Euclidean) de Sitter manifold is a sphere and the Euclidean de Sitter group is the rotation group $SO(5)$.

2.3. Anti-de Sitter

Let us now introduce a flat five-dimensional space $\mathbb{E}^{(2,3)}$ by adding a timelike direction to \mathbb{M}^4 (as we did in the hyperbolic case). $\mathbb{E}^{(2,3)}$ has two timelike directions and three spacelike directions and therefore it is not a spacetime in the ordinary sense (a Lorentzian manifold with one temporal and three spatial dimensions). However, the hypersurface with equation

$$AdS_4 = \{x \in \mathbb{E}^{(2,3)}, x_0^2 - x_1^2 - x_2^2 - x_3^2 + x_4^2 = R^2\}, \quad (7)$$

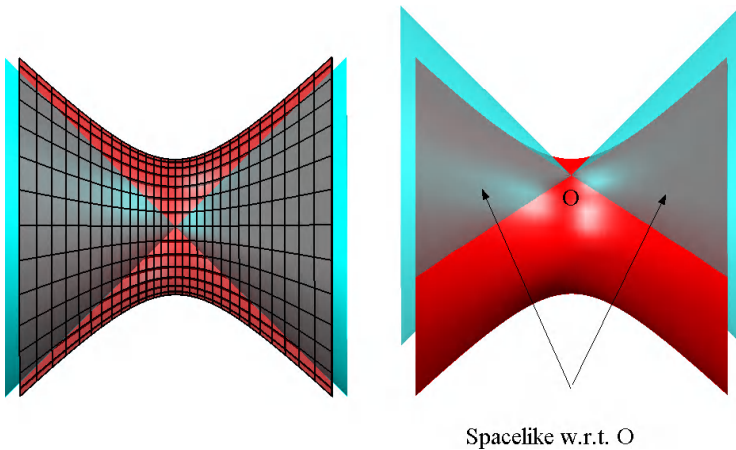


Figure 4 A visualisation of the anti-de Sitter universe. The asymptotic cone plays a crucial role exactly as in the de Sitter case. The regions of AdS_4 that are in the shadow of the five-dimensional cone emerging from an event O are the regions that are not causally connected to the event O . The asymptotic cone in the ambient space can be regarded as a representation of the boundary at spacelike infinity of the AdS manifold and carries a natural action of the conformal group that is the group-theoretical foundation for the AdS-CFT correspondence.

is a spacetime: this is the anti-de Sitter universe (see Figure 4). It has constant positive curvature and reproduces (after a renormalisation) the Minkowski spacetime in the limit when the curvature tends to zero.

The causal structure of AdS_4 is induced by restriction of the geometry of the ambient space $\mathbb{E}^{(2,3)}$ (the analogy is now with the geometry of \mathbb{H}^3 that is determined by the causal structure of the ambient spacetime \mathbb{M}^4). As before the null cone of the ambient space

$$C = \{\xi \in \mathbb{M}^5, \xi_0^2 - \xi_1^2 - \xi_2^2 - \xi_3^2 + \xi_4^2 = 0\} \quad (8)$$

induces the causal ordering on the anti-de Sitter manifold. Anti-de Sitter timelike geodesics are ellipses and are obtained by intersecting the hyperboloid with two-planes passing through the centre of the ambient space. The geodesics passing through a certain event all meet at the antipodal point. Owing to the existence of closed timelike curves the causal ordering is only local. One may construct a globally causal manifold by considering the covering of the anti-de Sitter manifold (recall that the covering of a circle is a line). However even the covering of the anti-de Sitter remembers the ‘periodicity in time’ of the original manifold: the focusing of geodesics remains true also in the covering space and the geodesics issued from an event meet again infinitely many times in the covering.

The anti-de Sitter line element is constructed by restricting the five-dimensional invariant ‘interval’ of the ambient space to the manifold AdS_4 :

$$ds^2 = [(dx_0)^2 - (dx_1)^2 - (dx_2)^2 - (dx_3)^2 + (dx_4)^2]_{AdS_4}. \quad (9)$$

This line element is the maximally symmetrical solution of the cosmological Einstein equations when the cosmological constant Λ is negative. The anti-de Sitter kinematical group coincides with the isometry group $SO(2, 3)$ of the ambient space.

The relationship between the anti-de Sitter universe and the geometry of \mathbb{H}^3 is deeper than a mere analogy. Indeed, for imaginary time

$$x_4 \rightarrow ix_4,$$

the (Euclidean) anti-de Sitter manifold is a copy of \mathbb{H}^4 and the Euclidean de Sitter group is $SO(1, 4)$. A study of the complex anti-de Sitter manifold with applications to quantum field theory has been described in [10].

AdS is not a globally hyperbolic spacetime. In non-globally hyperbolic manifolds knowledge of equations of motion and of initial data is not enough to determine the time evolution of physical quantities. In the anti-de Sitter case, the lack of global hyperbolicity is due to the existence of a boundary at spacelike infinity: information can flow in from infinity. This fact is a source of difficulties in quantising fields on the anti-de Sitter manifolds. However this is also an opportunity since this boundary at infinity offers the very possibility for formulating the famous AdS/CFT correspondence [6].

To present an intuitive idea of this topic let us introduce coordinates on a five-dimensional anti-de Sitter manifold AdS_5 (embedded in a six-dimensional space $\mathbb{E}^{(2,4)}$) obtained by intersecting AdS_5 with hyperplanes $\{X_4 + X_5 = e^v\}$ (see Figure 5). Each slice Π_v of AdS_5 is a copy of Minkowski spacetime \mathbb{M}^4 . Points in each slice Π_v can be thus parametrised by Minkowskian coordinates x_0, x_1, x_2, x_3 (rescaled by e^v on Π_v). This explains why the anti-de Sitter coordinates (v, x_0, x_1, x_2, x_3) are also called Poincaré coordinates.

The coordinate system covers only one-half of the anti-de Sitter manifold; the anti-de Sitter metric takes the following form:

$$\begin{aligned} ds^2 &= [(dX_0)^2 - (dX_1)^2 - (dX_2)^2 - (dX_3)^2 - (dX_4)^2 + (dX_5)^2]_{AdS_5} \\ &= e^{2v}(dx_0^2 - dx_1^2 - dx_2^2 - dx_3^2) - dv^2. \end{aligned} \quad (10)$$

The slices Π_v are often called branes. The Minkowskian geometry of the brane is induced by the ambient anti-de Sitter metric: for instance space-like separation in any slice Π_v can be understood equivalently in the Minkowskian sense of the slice itself or in the sense of the ambient anti-de Sitter universe.

When we consider the limit $v \rightarrow \infty$ we arrive at the anti-de Sitter boundary at spacelike infinity, which therefore may (essentially) be thought of as a four-dimensional Minkowski spacetime. The AdS-CFT correspondence establishes an

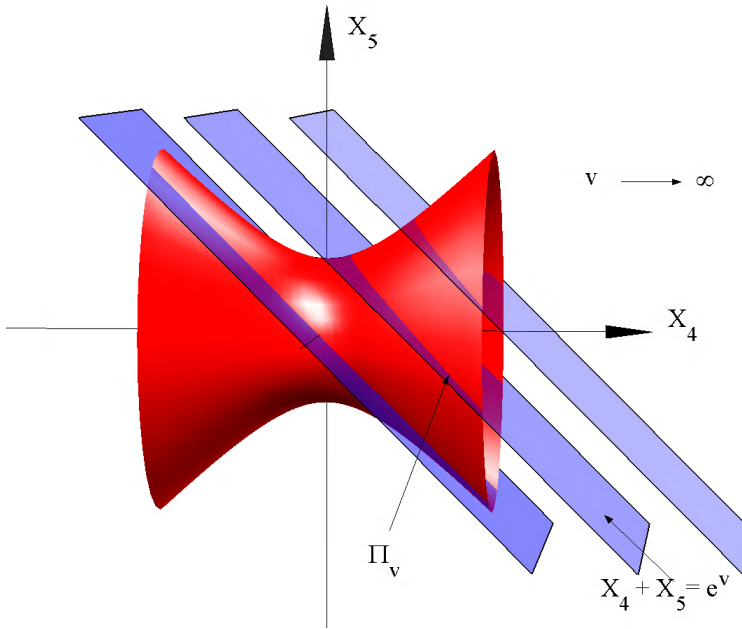


Figure 5 Construction of the AdS-Poincaré coordinates. The limit $v \rightarrow \infty$ describes the boundary of the AdS manifold.

equivalence between a theory on the five-dimensional AdS_5 and a relativistic theory on the boundary \mathbb{M}^4 (this is an instance of another popular idea in contemporary theoretical physics: the *holographic principle*). The theory on the boundary is conjectured to have a larger symmetry group, namely the conformal group [6, 11, 12].

3. De Sitter

The shortest path to understanding the de Sitter geometry in d dimension consists in adding a spacelike dimension to a d -dimensional Minkowski spacetime and considering the hypersurface with equation

$$dS_d = \{x \in \mathbb{M}^{d+1}, x^2 = x \cdot x = x_0^2 - x_1^2 - \dots - x_{d+1}^2 = -R^2\} \quad (11)$$

(we use lower indices for the inertial coordinates x_i of the ambient spacetime just for visual comfort of the formulae). This manifold represents the de Sitter spacetime [7] (see Figure 1). The causal structure of dS_d is induced by restriction of the Lorentzian geometry of the ambient Minkowski spacetime \mathbb{M}^{d+1} exactly as the geometry of the spherical surface is determined by restriction of the Euclidean

geometry of the ambient space \mathbb{R}^{d+1} . In particular, the de Sitter line element is obtained concretely by restricting the invariant interval to the manifold dS_{d+1} :

$$ds^2 = (dx_0^2 - dx_1^2 - \dots - dx_{d+1}^2)|_{dS_d} \quad (12)$$

The radius R corresponding to a given value of Λ is

$$R = \sqrt{\frac{(d-1)(d-2)}{2\Lambda}}.$$

There is a *causal* ordering relation on dS_d induced by that of \mathbb{M}^{d+1} (see Figure 1); let

$$V^+ = \{x \in \mathbb{M}^{d+1} : x_0 > \sqrt{x_1^2 + \dots + x_d^2}\} \quad (13)$$

be the future cone of the origin in the ambient space; then, for

$$x, x' \in dS_d, \quad x > x' \Leftrightarrow x - x' \in V^+. \quad (14)$$

The future and past regions of a given event x in dS_d are given by $\Gamma_x^\pm = \{x' \in dS_d : x' > x (x > x')\}$. A pivotal role in the following construction is played by the lightcone of the ambient spacetime, that in a certain sense plays the role of de Sitter momentum space (see Figure 9):

$$C = \{\xi \in \mathbb{M}^{d+1}, \xi^2 = \xi_0^2 - \xi_1^2 - \dots - \xi_d^2 = 0\}. \quad (15)$$

The de Sitter kinematical group coincides with the Lorentz group of the ambient spacetime $SO_0(1, d)$. There are no commutative translations on the de Sitter manifold and this fact is the source of considerable technical difficulties in the study of de Sitter quantum field theory. The relationship between the de Sitter universe and the geometry of the sphere is deeper than a mere analogy. Indeed, for imaginary times $x_0 \rightarrow ix_0$ the (Euclidean) de Sitter manifold (see Eq. 4) is a sphere and the Euclidean de Sitter group is the corresponding rotation group $SO_0(1, d)$.

3.1. Coordinate systems

The de Sitter geometry finds its most important physical applications in cosmology. In Friedmann's cosmology one usually 'breaks' the general relativistic covariance and singles out a special coordinate system: there is a natural choice of 'cosmic time' that makes the universe appear spatially homogeneous and isotropic at large scales. This property is mathematically encoded in the Friedmann–Robertson–Walker line element:

$$ds^2 = dt^2 - a(t)^2 dl^2. \quad (16)$$

The spatial distance dl^2 describes the geometry of a homogeneous and isotropic space manifold: either a sphere \mathbb{S}^{d-1} , or a hyperplane \mathbb{R}^{d-1} or a Lobatchevski space

\mathbb{H}^{d-1} . In this respect the de Sitter geometry is rather special: due to the maximal symmetry and the topology of the de Sitter manifold, there are suitable choices of cosmic time that make the de Sitter manifold appear like either a spherical, or a flat or a hyperbolic FRW model. Let us choose for instance the following coordinate system:

$$x(t, \omega) = \begin{cases} x_0 = R \sinh\left(\frac{t}{R}\right) \\ x_i = R \cosh\left(\frac{t}{R}\right) \omega_i \quad i = 1, \dots, d \end{cases} \quad (17)$$

with $\sum \omega_i^2 = 1$, so that Equation (4) is easily satisfied. The coordinate x_0 depends only on the cosmic time (see Figure 6); hypersurfaces of constant time are spheres and the coordinate system covers the whole universe. With this choice the de Sitter line element describes a spherical FRW model:

$$ds^2 = (dx_0^2 - dx_1^2 - \dots - dx_{d+1}^2)|_{dS_d} = dt^2 - R^2 \cosh^2\left(\frac{t}{R}\right) d\omega^2. \quad (18)$$

Another possible choice of time is, say, the combination $x_0 + x_d$ (see Figure 7). Homogeneous and isotropic surfaces of constant time Π_t (or ‘horospheres’) are

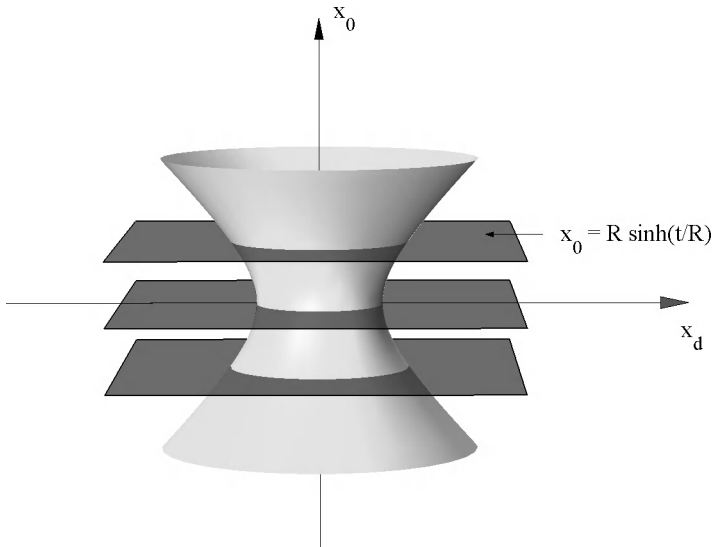


Figure 6 Construction of the coordinate system representing the de Sitter geometry as closed FRW model. Hyperurfaces of equal cosmic time are intersection of the de Sitter manifold with hyperplanes $x_0 = \text{const}$.

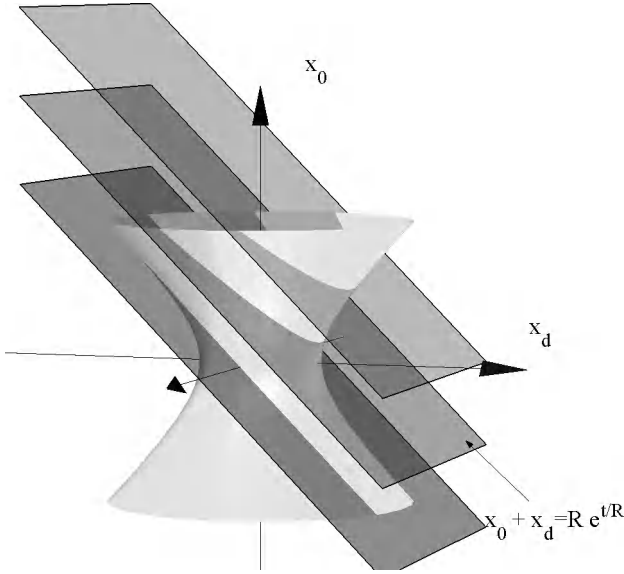


Figure 7 Construction of the coordinate system representing the de Sitter geometry as a flat FRW model. Hypersurfaces of equal cosmic time are intersections of the de Sitter manifold with hyperplanes $x_0 + x_d = \text{const}$. Only one half of the manifold is covered since it has to be $x_0 + x_d > 0$. Of course this restriction does not hold any more if the time t is considered as a complex coordinate.

paraboloids obtained by intersecting dS_d with the hyperplanes $x_0 + x_d = R e^{\frac{t}{R}}$ (the latter relation also introduces the relevant cosmic time):

$$x(t, \mathbf{x}) = \begin{cases} x_0 = R \sinh \frac{t}{R} + \frac{1}{2R} \mathbf{x}^2 \exp \frac{t}{R} \\ x_i = \mathbf{x}_i \exp \frac{t}{R}, \quad i = 1, \dots, d-1, \\ x_d = R \cosh \frac{t}{R} - \frac{1}{2R} \mathbf{x}^2 \exp \frac{t}{R}. \end{cases} \quad (19)$$

This parametrisation is also called the ‘horocyclic parametrisation’. Real values of the coordinates only describe the part Π of the de Sitter manifold which intersects the half-space $\{x_0 + x_d > 0\}$; that region can be thought as the future of an event infinitely far in the past. Any slice Π_t is conformal to a Euclidean plane; indeed the de Sitter universe in these coordinates appears as a flat FRW model with exponentially growing scale factor:

$$ds^2 = dt^2 - \exp \frac{2t}{R} d\mathbf{x}^2. \quad (20)$$

This form of the de Sitter line element was originally introduced by Lemaître in 1925 (see e.g., [13]). It is interesting to note that the first coordinate system used by de Sitter himself was a static coordinate system with closed spatial sections; de Sitter was following Einstein's cosmological idea of a static closed universe, the idea that led to the introduction of the cosmological term in Einstein's equations. A static coordinate system (i.e., a coordinate system where nothing depends explicitly on time) is not the most natural to describe an expanding universe, but it has other interesting properties, mainly in relation to black hole physics: horizons, temperature and entropy.

Static closed coordinates are represented in Figure 8. The Lemaître form of the de Sitter line element is the most useful in cosmological applications. Recent observations point towards the existence of a nonzero cosmological constant and a flat space. For an empty universe (i.e., a universe filled with a pure cosmological constant) this would correspond precisely to the above description of the de Sitter universe.

Finally a third cosmic time coordinate may be introduced the relation $x_d = R \cosh \frac{t}{R}$. With this choice the de Sitter universe appears as an open FRW

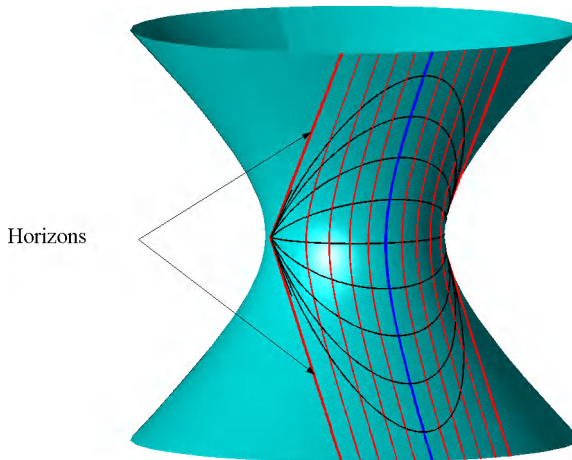


Figure 8 A chart representing static closed coordinates. This is the coordinate system originally used by W. de Sitter in 1917. Vertical timelike curves are obtained by intersecting the hyperboloid with parallel two-planes. Only the central hyperbola is a geodesic because it is the only one lying on a plane that contains the origin of the ambient spacetime. The other timelike curves are accelerated trajectories. They have been coloured in red because there is a redshift for light sources moving along these world lines; this effect was called the de Sitter effect and was thought to have some bearing on the redshift results obtained by Slipher.

cosmological model:

$$ds^2 = dt^2 - R^2 \sinh^2 \frac{t}{R} d\Omega^2, \quad (21)$$

where dl^2 is the line element of a Lobatchevski space of unit radius. This choice had some popularity in the mid-nineties when open inflationary models were introduced to account for the missing mass [14–16]. The subsequent observations pointing towards the existence of a nonzero cosmological constant greatly reduced the interest in these models, but they can still play some role when the future high precision cosmological observations will tell us more about the actual value of the curvature of the space. Indeed, even a very small negative spatial curvature (but not exactly zero) can still make a noticeable difference in our understanding of the cosmos.

3.2. Boundary at infinity. Geodesics

The de Sitter manifold has a boundary at timelike infinity. One standard way to study that boundary is to make use of a Penrose's compactification. This can be obtained by the following change of the time coordinate in Eq. (17):

$$\sinh \frac{t}{R} = \tan u, \quad \cosh \frac{t}{R} = \frac{1}{\cos u}, \quad -\frac{\pi}{2} < u < \frac{\pi}{2}. \quad (22)$$

With the help of these coordinates the de Sitter metric is written

$$ds^2 = \frac{R^2}{\cos^2 u} (du^2 - d\Omega^2) \quad (23)$$

and the de Sitter universe is conformal to a portion of the Einstein static universe. Events on the past boundary \mathcal{I}^- are given the coordinates $(u = -\pi/2, \omega_i)$ while on the future boundary \mathcal{I}^+ are given by the coordinates $(u = \pi/2, \omega_i)$ (see [17] for further details). Another maybe less usual visualisation can be obtained by taking the large t asymptotics in Eq. (17):

$$\begin{cases} x_0 \simeq \pm R e^{|t|}, \\ x_i \simeq R e^{|t|} \omega_i. \end{cases} \quad (24)$$

It follows that the light-cone C of the ambient spacetime \mathbb{M}^{d+1} can also be regarded as a projective representation of the boundary of the de Sitter manifold at timelike infinity. The invariance group of the cone C is also copy of $\text{SO}_0(1, d)$. One proposal is to interpret that group as the Euclidean conformal group. Although it may appear unnatural at this point, we choose to work only with the forward cone $C^+ = \{\xi \in C, \xi_0 > 0\}$ to represent both the future and past infinities (as opposed to the choice of the two-sheeted cone $C = C^+ \cup C^-$). We can then identify events at

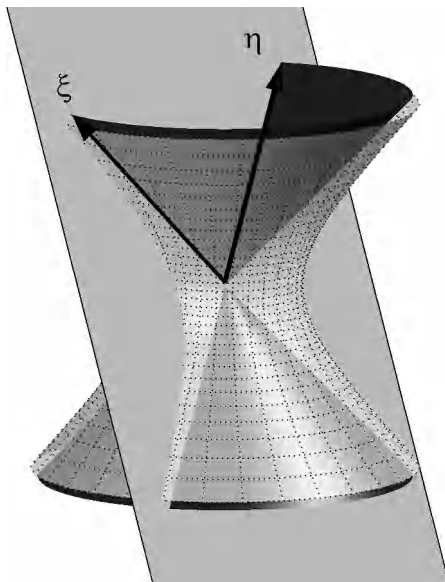


Figure 9 Represented are the de Sitter universe and the lightcone of the $(d+1)$ -dimensional ambient spacetime. The timelike geodesics are the hyperbolae obtained by intersecting the hyperboloid by any Lorentzian two-plane passing through the origin of the ambient spacetime. Any two-plane associated with a timelike geodesic can be identified by specifying two null vectors ξ and η in the future lightcone C^+ ; ξ and η can be used also to parametrise the geodesic itself. In flat spacetime, geodesics are labelled by their four-momentum. By analogy, the lightcone C^+ can be interpreted as the space of directions of momentum vectors in the de Sitter universe. In particular, de Sitter plane waves are constructed using vectors belonging to the future lightcone C^+ .

future infinity in the Penrose diagram with equivalence classes of vectors on C^+ as follows ($\lambda > 0$):

$$(u = \pi/2, \omega_i) \longleftrightarrow \xi = (\lambda, \lambda\omega_i), \quad (25)$$

while events at past infinity are described by the map:

$$(u = -\pi/2, \omega_i) \longleftrightarrow \eta = (\lambda, -\lambda\omega_i). \quad (26)$$

Particular representatives of the equivalence classes are $\hat{\xi} = (1, \boldsymbol{\omega})$ and $\hat{\eta} = (1, -\boldsymbol{\omega})$.

The above description of infinity allows for a particularly simple description of the de Sitter timelike geodesics, which is useful for the interpretation of the role of the lightcone of the ambient spacetime. Indeed, a generic timelike geodesic can be obtained as the intersection of the de Sitter hyperboloid with a certain

Lorentzian two-plane containing the origin of the ambient space. Let (ξ, η) be an ordered pair of forward null vectors generating such a two-plane. We associate to (ξ, η) the geodesic obtained by the intersection of such two plane with the de Sitter hyperboloid and contained in the wedge generated by ξ and $-\eta$. Such a geodesic can be parametrised as follows:

$$x(\tau) = \frac{R}{\sqrt{2\xi \cdot \eta}} \left(e^{\frac{\tau}{R}\xi} - e^{-\frac{\tau}{R}\eta} \right), \quad (27)$$

where τ is the proper time. If we consider an event on \mathcal{I}^- with coordinates $(-\pi/2, \omega)$ and an event on \mathcal{I}^+ with coordinates $(\pi/2, \omega')$ the unique geodesics joining them can be parametrised as follows:

$$x(\tau) = \frac{R}{\sqrt{2 + 2\omega \cdot \omega'}} \left(e^{\frac{\tau}{R}\hat{\xi}} - e^{-\frac{\tau}{R}\hat{\eta}} \right). \quad (28)$$

Thus a timelike geodesic is uniquely determined by specifying one event on each boundary. If we specify only one vector ξ we obtain the family of geodesics focusing at the corresponding event (either in the future or in the past).

The conserved quantities associated with the geodesical motion are the components of the two-form²

$$K = K_{(\xi, \eta)} = mc \frac{\xi \wedge \eta}{\xi \cdot \eta}. \quad (29)$$

A theory of classical scattering on the de Sitter manifold can be built by using the above conserved quantities [18].

4. De Sitter Quantum Field Theory

4.1. Plane waves

Let us consider the de Sitter Klein–Gordon equation

$$\square\phi + m^2\phi = 0, \quad (30)$$

where \square is the Laplace–Beltrami operator relative to the de Sitter metric. It is possible to solve Eq. (30) by separating the variables in any of the coordinate systems that we have presented in the previous section. Let us choose to work for instance in the flat space coordinate system (19). By posing $\phi = \chi(t) \exp i\mathbf{k} \cdot \mathbf{x}$ the

² ξ and η denote here the covariant one-forms associated to the null vectors; we use the same symbol for a vector and its dual.

Klein–Gordon equation is solved if χ satisfies the following equation:

$$\partial_t^2 \chi + \frac{d-1}{R} \partial_t \chi + \left(e^{-\frac{2t}{R}} \mathbf{k}^2 + m^2 \right) \chi = 0. \quad (31)$$

By introducing the conformal time $s = -Re^{-\frac{t}{R}}$ and the rescaled function $f(s) = s^{\frac{1-d}{2}} \chi$, the previous equation is transformed into the Bessel equation:

$$s^2 \partial_s^2 f + s \partial_s f + (s^2 k^2 + \nu^2) f = 0, \quad (32)$$

where

$$\nu^2 = m^2 R^2 - \left(\frac{d-1}{2} \right)^2. \quad (33)$$

There are therefore two regimes, corresponding either to real values of ν or to purely imaginary values such that

$$|\nu| < \left(\frac{d-1}{2} \right). \quad (34)$$

Thus, by separating variables in horocyclic coordinates one obtains the general solution of the de Sitter Klein–Gordon equation in the following way:

$$\phi(t, \mathbf{x}) = s^{\frac{1-d}{2}} B_\nu(ks) \exp i\mathbf{k} \cdot \mathbf{x}, \quad (35)$$

where B_ν is a solution of the Bessel equation. From here one can proceed and try to quantise the Klein–Gordon according to the recipes of canonical quantisation [19]; what particular choice of Bessel function is physically meaningful is another story. Things are not so obvious because of the lack of a global energy operator; one has to advocate some other principle like the adiabatic prescription (see e.g., [19]) or something else. Of course the de Sitter manifold is maximally symmetric and many other possible choices are in principle equally good.

There exists also an alternative approach: it is possible to introduce global waves in a manifestly coordinate-independent way by insisting on the embedding of the de Sitter hyperboloid in the Minkowski ambient space [8, 9, 22]. Let us introduce the waves

$$\psi(x, \xi) = |x \cdot \xi|^\lambda, \quad (36)$$

where λ is a complex number and $\xi = (\xi_0, \dots, \xi_d) \in C^+$ a future directed null vector of the ambient space (in clear: $\xi \cdot \xi = 0$ and $\xi_0 > 0$). It is not difficult to see that

$$\square |x \cdot \xi|^\lambda = \frac{1}{R^2} \lambda(\lambda + d - 1) |x \cdot \xi|^\lambda \quad (37)$$

provided $x \cdot \xi \neq 0$. The parameter ν that we have introduced previously is related to the exponent λ as follows:

$$i\nu = \lambda + \frac{d-1}{2}, \quad (38)$$

physical values of the parameter λ thus correspond to unrestricted real or purely imaginary ν with $|\nu| \leq \frac{d-1}{2R}$ so that

$$m^2 R^2 = \lambda(1-d-\lambda) = \left(\frac{d-1}{2}\right)^2 + \nu^2 > 0. \quad (39)$$

Here comes in the physical interpretation of ‘momentum space’ that we have given to the future lightcone C^+ establishing the waves (36) as the strict de Sitter analogues of the exponential plane waves of the flat Minkowski spacetime. Indeed, as the exponentials $\exp \frac{i}{\hbar} p \cdot x$ that are labelled by the momentum vector, the waves are labelled by the ‘momentum direction’ ξ and by the ‘modulus’ ν .

There is however an unexpected difficulty: the plane waves are singular on $(d-1)$ -dimensional light-like submanifolds of dS_d that are the intersection of dS_d with the hyperplane tangent to the cone along ξ . To deal with that singularity let us introduce the complexification of the de Sitter spacetime, that can be represented as the complex hyperboloid

$$dS_d^{(c)} = \{z = x + iy \in \mathbb{C}^{d+1} : z_0^2 - z_1^2 - \dots - z_d^2 = -R^2\}. \quad (40)$$

This implies of course that an event of the complex de Sitter manifold is such that $x^2 - y^2 = -R^2$ and $x \cdot y = 0$. The physically relevant global waves can be defined as analytic functions for z in the tubular domains \mathcal{T}^+ or \mathcal{T}^- of $dS_d^{(c)}$:

$$\begin{aligned} \mathcal{T}^+ &= (\mathbb{R}^{d+1} + iV^+) \cap dS_d^{(c)} = \{z = x + iy \in \mathbb{C}^{d+1} : y^2 > 0, y_0 > 0\}, \\ \mathcal{T}^- &= (\mathbb{R}^{d+1} + iV^+) \cap dS_d^{(c)} = \{z = x + iy \in \mathbb{C}^{d+1} : y^2 > 0, y_0 > 0\}, \end{aligned} \quad (41)$$

where $\mathbb{R}^{d+1} \pm iV^+$ are the forward and backward tubes in the ambient complex Minkowski space \mathbb{C}^{d+1} (with $x \in \mathbb{R}^{d+1}$ and respectively $\pm y \in V^+$). In Minkowski space, the physical meaning of such domains is linked to the positivity of the spectrum of the energy-momentum vector operator in Minkowski QFT’s in \mathbb{R}^{d+1} . Here is the definition of de Sitter plane waves. For $z \in \mathcal{T}^+$ or $z \in \mathcal{T}^-$ we define

$$\psi_{i\nu}(z, \xi) = (z \cdot \xi)^{-\frac{d-1}{2} + i\nu}, \quad (42)$$

where ν and ξ satisfy the above conditions. The phase is chosen to be zero when the argument is real and positive.

4.2. Two-point functions of the Klein–Gordon quantum field

We now briefly outline the main features of dS Klein–Gordon QFT which we will use later in the study of particle decay (see [8] for further details). As usual for free fields, the theory is completely encoded in the two-point function $W(x, x')$. Generally speaking $W(x, x')$ should be a distribution on $dS_d \times dS_d$ satisfying the following conditions:

1. *Locality*: $W(x, x') = W(x', x)$ for every space-like separated pair (x, x') .
2. *de Sitter invariance*: $W(gx, gx') = W(x, x')$, $\forall g \in \text{SO}_0(1, d)$.
3. *Positive-definiteness*: $\forall f \in \mathcal{D}(dS_d)$

$$\int_{dS_d \times dS_d} W(x, x') \bar{f}(x) f(x') d\sigma(x) d\sigma(x') \geq 0. \quad (43)$$

where $d\sigma$ is the invariant measure on the de Sitter manifold. For the Klein–Gordon field it should also be that

$$(\square_x + m^2)W_m(x, x') = 0, \quad (\square_{x'} + m^2)W_m(x, x') = 0. \quad (44)$$

Although there are infinitely many inequivalent theories satisfying all these requirements, there is one preferred theory (for each value of the mass m) which is often referred to as the ‘Euclidean’ or Bunch–Davies vacuum [20, 21]. What is perhaps not yet so well known is that the corresponding preferred theory can be directly constructed in a manifestly de Sitter invariant way [8, 22] by exploiting the previously introduced basis of analytical de Sitter plane waves; it is possible to give a spectral analysis of the corresponding two-point functions very similar to the usual Fourier representation of the two-point function of the Poincaré invariant two-point function of a Klein–Gordon field satisfying the Wightman axioms [23].

The construction goes as follows: by analogy with the flat case [23] we introduce a two-point function defined in the complex domain $z \in \mathcal{T}^-$, $z' \in \mathcal{T}^+$ [8] which is a superposition of plane waves:

$$W_\nu(z, z') = c_{d,\nu} \int_\gamma \psi_{i\nu}(z, \xi) \psi_{-i\nu}(z', \xi) d\mu_\gamma(\xi). \quad (45)$$

We used the parameter ν as a label according with Eq. (33); the constant $c_{d,\nu}$ has to be determined by imposing the local Hadamard condition or, equivalently, the canonical commutation relations (CCR’s).

W_ν manifestly solves the (complex) de Sitter Klein–Gordon equation in both variables, and is analytic in the domain $\mathcal{T}^- \times \mathcal{T}^+$ (*normal analyticity property*). The integration at the RHS can be performed along any basis submanifold γ of the cone C^+ w.r.t. a corresponding measure $d\mu_\gamma$ that is induced by the invariant measure on the cone. Particular instances are the spherical basis $\gamma_0 = \{\xi \in C^+ : \xi_0 = 1\}$,

where the measure $d\mu_\gamma$ is the rotation invariant measure on the sphere, or the hyperbolic basis $\gamma_d = \{\xi \in C^+ : |\xi_d| = 1\}$, where the measure $d\mu_\gamma$ is the Lorentz invariant measure on a two-sheeted unit mass shell.

With these premises, the fundamental result is that, by Stoke's theorem, the integral at the RHS of Eq. (45) is independent on the chosen integration submanifold. This immediately implies that $W_\nu(z, z')$ is a de Sitter invariant bi-solution of the Klein–Gordon equation and therefore is actually a function of the de Sitter invariant variable

$$\zeta = \frac{z \cdot z'}{R^2}. \quad (46)$$

Now we can explicitly compute the integral defining $W_\nu(z, z')$:

$$W_\nu(z, z') = \frac{\Gamma\left(\frac{d-1}{2} + i\nu\right) \Gamma\left(\frac{d-1}{2} - i\nu\right)}{2(2\pi)^{\frac{d}{2}} R^{d-2}} (\zeta^2 - 1)^{-\frac{d-2}{4}} P_{-\frac{1}{2}+i\nu}^{-\frac{d-2}{2}}(\zeta), \quad (47)$$

where P is an associated Legendre function of the first kind [24]. The choice of normalisation assures that the canonical commutation relations hold. Eq. (47) shows that $W_\nu(z, z')$ can be analytically continued in the “cut-domain”

$$dS_d^{(c)} \times dS_d^{(c)} \setminus \{(z, z') \in X_d^{(c)} \times X_d^{(c)} : (z - z')^2 \geq 0\} \quad (48)$$

where it satisfies the complex covariance condition: $W_\nu(gz, gz') = W_\nu(z, z')$ for all g belonging to the complex de Sitter group.

The correlation function $W_\nu(x, x') = \langle \Omega, \phi(x)\phi(x')\Omega \rangle$ between two *real* events x and x' is then the boundary value of the analytic function $W_\nu(z, z')$ from the domain $\mathcal{T}^- \times \mathcal{T}^+$. The knowledge of $W_\nu(x, x')$ permits to use the standard methods of the Wightman's reconstruction theorem [23] (or either the GNS construction) which give the Fock space of the theory, the quantum field operators and the associated representation of the de Sitter group. In our case it turns out that the restriction of such representation to the one-particle subspace of the Fock space are unitary and irreducible. For ν real these are the representations of the principal series; for imaginary ν such that $0 < |\nu| < \frac{d-1}{2}$ these are the representation of the complementary series.

4.3. Generalised free fields

Actually, these analyticity properties are not restricted to Klein–Gordon fields, but they can be shown to hold for any de Sitter invariant two-point Wightman function W satisfying the normal analyticity spectral condition [8]. The correlation function $W(x, x') = \langle \Omega, \phi(x)\phi(x')\Omega \rangle$ between two *real* events x and x' is then the boundary value of the analytic function $W(z, z')$ from the domain $\mathcal{T}^- \times \mathcal{T}^+$.

The ‘permuted Wightman function’ $W(x', x) = \langle \Omega, \phi(x')\phi(x)\Omega \rangle$ is the boundary value of $W(z, z')$ from the domain $\mathcal{T}^+ \times \mathcal{T}^-$. From here one can construct the commutator

$$C(x, x') = W(x, x') - W(x', x) \quad (49)$$

and the Green functions. In particular the retarded propagator $R(x, x')$ plays an important role:

$$R(x, x') = i\theta(x, x')C(x, x'), \quad (50)$$

where θ is the characteristic function of the ‘future cone’ $\Gamma^+(x')$ of the event x' :

$$\theta(x, x') = \begin{cases} 1 & \text{if } x \geq x', \\ 0 & \text{otherwise.} \end{cases} \quad (51)$$

It is worthwhile to stress that the thermal properties of de Sitter free fields (in the sense of the Gibbons–Hawking temperature [21]) can be proven easily in this analytic framework and the maximal analyticity property of the two-point function is indeed equivalent to those thermal properties [8]. The same thermal properties have been shown to hold [9] also in the interacting case if the n -point functions enjoy suitable analyticity properties.

5. Lifetime of a de Sitter Particle

In this section we outline an application [27–29] of the previous formalism to the study of instable particles on the de Sitter manifold. First of all we point out two important properties.

5.1. Two properties that are crucial

In our application to particle decays:

1. The projector identity

This is a statement concerning the convolution on the de Sitter manifold of a pair of two-point functions belonging to the principal series

$$\int_{dS_d} W_\nu(x, u) W_{\nu'}(u, y) d\sigma(y) = 2\pi |\coth(\pi\nu)| \delta(\nu^2 - \nu'^2) W_\nu(x, y). \quad (52)$$

2. The Källén–Lehmann type representation

Consider the product of n different two-point functions of the principal series. There exist an integral representation for the product as superposition of kernels of the principal series as follows:

$$\prod_{j=1}^n W_{\nu_j}(x, y) = \int da^2 \rho(a^2; \nu_1, \dots, \nu_n) W_a(x, y). \quad (53)$$

Similar properties hold also in the Minkowski case. The proof of these statements is however elementary in the Minkowski case while nontrivial for de Sitter theories. An explicit calculation of the weights ρ is easy in the Minkowski case for $n = 2$, while difficult or impossible in other cases.

5.2. The model

Let $\phi_0, \phi_1, \dots, \phi_N$ be $N + 1$ independent free neutral scalar fields, with mass parameters $\nu_0, \nu_1, \dots, \nu_N$ respectively, operating in a Fock space \mathcal{H} , and

$$(\Omega, \phi_j(x)\phi_k(y)\Omega) = \delta_{jk} W_{\nu_j}(x, y). \quad (54)$$

Vectors of the form

$$\int f(x_{01}, \dots, x_{0q_0}, \dots, x_{N1}, \dots, x_{Nq_N}) : \prod_{j=0}^N \prod_{k=1}^{q_j} \phi_j(x_{jk}) d\sigma(x_{jk}) : \Omega, \quad (55)$$

where f is a smooth and fast decreasing test function, generate the closed subspace $\mathcal{H}_{q_0, \dots, q_N}$ of q_0 particles of type 0, \dots q_N particles of type N . We now switch on an interaction term

$$\gamma \int g(x) \mathcal{L}(x) d\sigma(x), \quad \mathcal{L}(x) =: \phi_0(x) \phi_1(x)^{q_1} \dots \phi_N(x)^{q_N} :, \quad (56)$$

where g is a smooth real rapidly decreasing function which, in the end, must be made to tend to the constant 1. Self-interactions $\mathcal{L}(x) =: \phi(x)^n :$ are a special case of this coupling. At first order the transition amplitude between two orthogonal normalised states ψ_0 and ψ_1 in \mathcal{H} is given by

$$(\psi_0, iT_1(\gamma g)\psi_1), \quad T_1(\gamma g) = \int \gamma g(x) \mathcal{L}(x) d\sigma(x). \quad (57)$$

Let ψ_0 be a one-particle state of the form $\int f(x)\phi_0(x)\Omega dx$; the smooth test function f contains the physical details about the quantum state of the unstable particle whose disintegration we aim to study. Let $\mathcal{H}_{0,q}$ be the space of all states containing q_1 particles of type 1, \dots , q_N particles of type N , and $P_{0,q}$ be the projector onto this space, with $q = (q_1, \dots, q_n)$. If ψ_0 has norm 1, Wick's theorem gives the probability of its transition to any possible q -particle state of $\mathcal{H}_{0,q}$:

$$\begin{aligned} \Gamma(1_0; q_1, \dots, q_N) &= (\psi_0, T_1(\gamma g) P_{0,q} T_1(\gamma g)^* \psi_0) \\ &= \gamma^2 \int \overline{f(x)} f(y) g(u) g(v) W_{\nu_0}(x, u) \left(\prod_{j=1}^N q_j! W_{\nu_j}(u, v)^{q_j} \right) \\ &\quad \times W_{\nu_0}(v, y) d\sigma(x) d\sigma(y) d\sigma(u) d\sigma(v). \end{aligned} \quad (58)$$

We now replace one of the switching-on factors, say $g(v)$, by 1 in the above expression. By using Eqs. (53) and (52) we find the following general formula for the transition probability:

$$\begin{aligned} \Gamma(1_0; q_1, \dots, q_N) &= \frac{\gamma^2 2\pi |\coth(\pi v)| \delta(v^2 - v'^2) \int g(x) |F(x)|^2 dx}{\int \overline{f(x)} W_{v_0}(x, y) f(y) dx dy} \\ &\times \left(\prod_{j=1}^N q_j! \right) \rho(v_0^2; v_1, \dots, v_1, \dots, v_N, \dots, v_N). \end{aligned} \quad (59)$$

Here

$$F(x) = \int W_{v_0}(x, y) f(y) dy; \quad (60)$$

is the smooth classical solution of the KG equation corresponding to the wavepacket f ; the denominator is the squared norm of ψ_0 which is no longer assumed to be one. This formula has an interesting simple structure: the first factor does not depend on the number or nature of the decay particles but only on the wavefunction of the incoming unstable particle. The infrared problem is contained in this factor and has to be overcome when letting the remaining $g(x)$ tend to 1 (adiabatic limit). The second factor is the relevant Källén–Lehmann weight times the right combinatorial factor.

5.3. Decay $1_\kappa \rightarrow 2_\nu$

Let us now focus on the decay of a particle of mass $v_0 = \kappa$ into two identical particles of mass $v_1 = \nu$. The first task is to compute the Källén–Lehmann weight $\rho(\kappa^2; \nu, \nu) \equiv \rho_\nu(\kappa)$. This is a hard question to be solved. To do that we use the following (suitably normalised) generalised Mehler–Fock transform of the squared two-point function:

$$\begin{aligned} \rho_\nu(\kappa) &= \frac{\left(\Gamma\left(\frac{d-1}{2} + i\nu\right) \Gamma\left(\frac{d-1}{2} - i\nu\right) \right)^2 \sinh \pi \kappa}{2(2\pi)^{1+\frac{d}{2}} R^{d-2}} \\ &\times \int_1^\infty P_{-\frac{1}{2}+i\kappa}^{-\frac{d-2}{2}}(x) \left[P_{-\frac{1}{2}+i\nu}^{-\frac{d-2}{2}}(x) \right]^2 (x^2 - 1)^{-\frac{d-2}{4}} dx. \end{aligned} \quad (61)$$

This integral is well defined for masses such that $|\text{Im}\nu| < \frac{d-1}{4}$; this includes the principal series and a portion of the complementary series. Inversion [25]

gives precisely

$$W_v^2(x, y) = \int_0^\infty d\kappa^2 \rho_v(\kappa) W_\kappa(x, y) = \int_{-\infty}^\infty \kappa d\kappa \rho_v(\kappa) W_\kappa(x, y). \quad (62)$$

The integral (61) can be directly computed for odd d . For even spacetime dimensions computing (61) is far from obvious. We have devised a method based on Mellin transform techniques [26] that allows the computation for any dimension d (real or complex). Here is the result:

$$\begin{aligned} \rho_v(\kappa) &= \frac{R^{2-d} \sinh \pi \kappa}{(4\pi)^{\frac{d+2}{2}} \sqrt{\pi} \Gamma\left(\frac{d-1}{2}\right)} \frac{\Gamma\left(\frac{d-1}{4} + \frac{i\kappa}{2}\right) \Gamma\left(\frac{d-1}{4} - \frac{i\kappa}{2}\right)}{\Gamma\left(\frac{d+1}{4} + \frac{i\kappa}{2}\right) \Gamma\left(\frac{d+1}{4} - \frac{i\kappa}{2}\right)} \\ &\times \prod_{\epsilon, \epsilon' = \pm} \Gamma\left(\frac{d-1}{4} + i\epsilon v + \frac{i\epsilon' \kappa}{2}\right). \end{aligned} \quad (63)$$

The striking result is that, contrary to what happens in the Minkowski spacetime, the weight ρ never vanishes. This means that for masses of the principal series decay processes into heavier particles are always possible. In particular, in that range of masses one is not allowed to draw conclusions about the stability of a certain particle just from its being the lightest in a hierarchy. This result has nothing to do with the standard thermal interpretation of the de Sitter ‘vacuum’. A similar computation in flat thermal field theory does not exhibit this phenomenon in two-particle decays. The standard Minkowskian result is recovered in the limit of zero curvature that is achieved by setting $\kappa = m_0 R$ and $v = m_1 R$:

$$\lim_{R \rightarrow \infty} \rho(\kappa^2; v, v) d\kappa^2 = \rho(m_0^2; m_1, m_1) dm_0^2, \quad (64)$$

where

$$\rho(m_0^2; m_1, m_1) = \frac{1}{2^d \pi m_0 \Gamma\left(\frac{d-1}{2}\right)} \left(\frac{m_0^2 - 4m_1^2}{4\pi}\right)^{\frac{d-3}{2}} \theta(m_0^2 - 4m_1^2). \quad (65)$$

Note the appearing in the limit of the Heaviside function θ that forbids the decay of a particle if the decay products are globally heavier.

All these effects are of course extremely small with the current value of the cosmological constant. What about particle physics at inflation? At that epoch $mR \sim m \times 10^{-15} \text{GeV}^{-1} \ll \frac{3}{4}$ for every particle of reasonable mass. Our results should therefore be extended to the remaining portion of the complementary series $|\text{Im } v| > \frac{d-1}{4}$ where all scalar particles lie at the inflation era (but there is no complementary series in the Fermionic case). By analytic continuation of

Eq. (63) in ν ,

$$W_\nu^2(x, y) = \int_{-\infty}^{\infty} \kappa d\kappa \rho_\nu(\kappa) W_\kappa(x, y) + \sum_{n=0}^{N-1} A_n(\nu) W_{i(\mu+2iv+2n)}$$

where $\mu = (d-1)/2$ and

$$A_n(\nu) = \frac{8\pi(-1)^n}{n!2^d\pi^{\frac{1+d}{2}}R^{d-2}\Gamma(\mu)} \frac{\Gamma(\mu+2iv+n)\Gamma(-2iv-n)}{\Gamma(\mu+2iv+2n)\Gamma(-\mu-2iv-2n)} \\ \times \frac{\Gamma(\mu+n)\Gamma(-iv-n)\Gamma(\mu+iv+n)}{\Gamma(-iv-n+\frac{1}{2})\Gamma(\mu+iv+n+\frac{1}{2})}.$$

The number of discrete terms is the largest N satisfying $N < 1 + |\text{Im}\nu| - \mu/2$, or 0 if this is negative. A particle of the complementary series with parameter $\kappa = i\beta$ can only decay into two particles with parameter $\nu = \frac{1}{2}(|\beta| + \mu + 2n)$, where n is any integer such that $0 \leq 2n < \mu - |\beta|$, and the decay is instantaneous. A particle with mass $m \ll m_c$ can only decay into two particles of mass $m_1 \sim m/\sqrt{2}$. Even if the geometry of the universe at inflation was not exactly de Sitterian, this example indicates that quantum field theoretical arguments concerning particle physics at inflation might need revision.

We now turn to the adiabatic limit and its meaning in the de Sitter context, in the case when all particles are in the principal series. A first complication is the existence of several choices of cosmic time, having different physical implications and the result might depend on one's preferred choice. We have seen that in the closed model the cosmic time t is related to the ambient space coordinates by the relation $x_0 = R \sinh(t/R)$; $g(x)$ can be chosen as the indicator function of some cosmic time interval T , say $g(x) = g_T(x) = \theta(T/2 - |t|)$.

In the flat model the situation is a bit more tricky. We saw that cosmic time is defined by the relation $x^0 + x^d = R \exp(t/R)$ and flat coordinates cover only half of the de Sitter manifold, all the events such that $x^0 + x^d > 0$. If we introduce the characteristic function $h_T(x) = \theta(Re^{T/2R} - x^0 - x^d)\theta(x^0 + x^d - Re^{-T/2R})$ then we have to add the contribution coming from the other half, i.e., $g(x) = g_T(x) = h_T(x) + h_T(-x)$. With these premises we have found [27, 28] that in both models the first factor in (59) diverges like T ; thus it has to be divided by T to extract a finite result which is the same in both models:

$$\lim_{T \rightarrow \infty} \frac{\gamma^2 C(\kappa) \int g(x) |F(x)|^2 dx}{T \int \bar{f}(x) W_\kappa(x, y) f(y) dx dy} = \frac{\gamma^2 \pi \coth(\pi\kappa)^2}{|\kappa|}. \quad (66)$$

Here the second (unforeseen) result comes in: in contrast to the Minkowskian case the limiting probability per unit of time does not depend on the wavepacket! This result seems to contradict what we see everyday in laboratory experiments, the well known effect of special relativity of time dilation. Furthermore, in contrast with

the violation of particle stability that is exponentially small in the de Sitter radius, this phenomenon does not depend on how small the cosmological constant is. How can we solve this paradox and reconcile the result with everyday experience? The point is that the idea of probability per unit time (Fermi's golden rule) has no scale-invariant meaning in de Sitter: if we use the limiting probability to evaluate amplitudes of processes that take place in a short time we get a grossly wrong result. This is in strong disagreement with what happens in the Minkowski case where the limiting probability is attained almost immediately (i.e., already for finite T). Therefore to describe what we are really doing in a laboratory we should not take the limit $T \rightarrow \infty$ and rather use the probability per unit of time relative to a laboratory-consistent scale of time. In that case we will recover all the standard wisdom even in the presence of a cosmological constant. But, if an unstable particle lives a very long time ($\gg R$) and we can accumulate observations, then a nonvanishing cosmological constant would radically modify the Minkowski result and de Sitter invariant result will emerge. This result should not be shocking: after all erasing any inhomogeneity is precisely what the quasi de Sitter phase is supposed to do at the epoch of inflation; in the same way, from the viewpoint of an accelerating universe, all the long-lived particles look as if they were at rest and so their lifetime would not depend on their peculiar motion.

6. Anti-de Sitter

The study of quantum fields on the AdS spacetime has begun with the pioneering approach of [30] whose main concern was to specify boundary conditions such that the difficulties arising by the lack of global hyperbolicity of the underlying AdS manifold could be circumvented and the resulting QFT be well defined. Another, earlier, approach was also given on the basis of group-theoretical methods [31] following ideas that can be traced back to Dirac [32].

Both of these approaches have influenced very much the recent research on the AdS/CFT subject. However, their applicability is more or less limited to free AdS QFTs (even if they can produce useful ingredients for perturbative calculations) and one may feel necessary setting the AdS/CFT debate on a more general basis [11, 12] in which both AdS quantum fields and boundary CFTs would be treated from the viewpoint of the structural properties of their n -point correlation functions.

6.1. Notations and geometry

We consider the vector space \mathbb{R}^{d+2} equipped with the following pseudo-scalar product:

$$X \cdot X' = X^0 X'^0 - X^1 X'^1 - \dots - X^d X'^d + X^{d+1} X'^{d+1}. \quad (67)$$

The $(d + 1)$ -dimensional AdS universe can then be identified with the quadric

$$AdS_{d+1} = \{X \in \mathbb{R}^{d+2}, X^2 = R^2\}, \quad (68)$$

where $X^2 = X \cdot X$, endowed with the induced metric

$$ds_{AdS}^2 = (dX^0{}^2 - dX^1{}^2 - \dots + dX^{d+1}{}^2)|_{AdS_{d+1}}. \quad (69)$$

The AdS relativity group is $G = SO_0(2, d)$, that is the component connected to the identity of the pseudo-orthogonal group $O(2, d)$. Two events X, X' of AdS_{d+1} are space-like separated if $(X - X')^2 < 0$, i.e., if $X \cdot X' > R^2$.

We will also consider the complexification of AdS_{d+1} :

$$AdS_{d+1}^{(c)} = \{Z = X + iY \in \mathbb{C}^{d+2}, Z^2 = R^2\}. \quad (70)$$

In other terms, $Z = X + iY$ belongs to $AdS_{d+1}^{(c)}$ if and only if $X^2 - Y^2 = R^2$ and $X \cdot Y = 0$. In the following we will put for notational simplicity $R = 1$.

Two parametrisations for the AdS manifold have a special status:

The ‘covering parametrization’ $X = X[r, \tau, e]$: it is obtained by intersecting AdS_{d+1} with the cylinders with equation $\{X^{0^2} + X^{d+1^2} = r^2 + 1\}$, and is given by

$$\begin{cases} X^0 = \sqrt{r^2 + 1} \sin \tau \\ X^i = r e^i \quad i = 1, \dots, d \\ X^{d+1} = \sqrt{r^2 + 1} \cos \tau \end{cases} \quad (71)$$

with $e^2 \equiv e^{1^2} + \dots + e^{d^2} = 1$ and $r \geq 0$. For each fixed value of r , the corresponding ‘slice’

$$C_r = AdS_{d+1} \cap \{X^{0^2} + X^{d+1^2} = r^2 + 1\} \quad (72)$$

of AdS_{d+1} is a manifold $\mathbb{S}_1 \times \mathbb{S}_{d-1}$. The complexified space $AdS_{d+1}^{(c)}$ is obtained by giving arbitrary complex values to r, τ and to the coordinates $e = (e^i)$ on the unit $(d - 1)$ -sphere.

The parametrisation (71) allows one to introduce relevant coverings of AdS_{d+1} and $AdS_{d+1}^{(c)}$ by unfolding the 2π -periodic coordinate τ (resp. $\text{Re}\tau$), interpreted as a time-parameter: these coverings are denoted respectively by \widehat{AdS}_{d+1} and $\widehat{AdS}_{d+1}^{(c)}$. A privileged ‘fundamental sheet’ is defined on these coverings by imposing the condition $-\pi < \tau < \pi$ (resp. $-\pi < \text{Re}\tau < \pi$). This procedure also associates with each manifold C_r its covering \widehat{C}_r which is a cylinder $\mathbb{R}_\tau \times \mathbb{S}_{d-1}$. We will use the symbols X, Z, \dots , also to denote points of the coverings.

Similarly one introduces a covering \hat{G} of the group G by taking in G the universal covering of the rotation subgroup in the $(0, d + 1)$ -plane. By transitivity, AdS_{d+1} and \widehat{AdS}_{d+1} are respectively generated by the action of G and \hat{G} on the base point $B = (0, \dots, 0, 1)$.

The physical reason which motivates the introduction of the covering \widehat{AdS}_{d+1} , that is the requirement of nonexistence of closed time-loops, also leads us to specify the notion of space-like separation in \widehat{AdS}_{d+1} as follows: let $X, X' \in \widehat{AdS}_{d+1}$ and let g an element of \hat{G} such that $X' = gB$; define $X_g = g^{-1}X$.

X and X' are space-like separated if X_g is in the fundamental sheet of \widehat{AdS}_{d+1} and $(X - X')^2 \equiv (g^{-1}X - g^{-1}X')^2 < 0$. This implies that $X_g = X_g[r, \tau, e]$ with $-\pi < \tau < \pi$ and $\sqrt{r^2 + 1} \cos \tau > 1$.

It is also interesting to note that on each manifold C_r the condition of space-like separation between two points $X = X[r, \tau, e]$ and $X' = X'[r, \tau', e']$ reads (in view of (71):

$$(X - X')^2 = 2(r^2 + 1)(1 - \cos(\tau - \tau')) - r^2(e - e')^2 < 0, \quad (73)$$

and that the corresponding covering manifold \hat{C}_r therefore admits a global causal ordering which is specified as follows:

$$(\tau, e) > (\tau', e') \quad \text{iff} \quad \tau - \tau' > 2\text{Arcsin} \left(\frac{(e - e')^2}{4} \frac{r^2}{r^2 + 1} \right)^{\frac{1}{2}}. \quad (74)$$

The ‘Poincaré parametrization’ $X = X(v, x)$: it only covers the part Π of the AdS manifold which belongs to the half-space $\{X^d + X^{d+1} > 0\}$ of the ambient space and is obtained by intersecting AdS_{d+1} with the hyperplanes $\{X^d + X^{d+1} = e^v = \frac{1}{u}\}$ ³, each slice Π_v (or ‘horosphere’) being an hyperbolic paraboloid:

$$\begin{cases} X^\mu = e^v x^\mu = \frac{1}{u} x^\mu & \mu = 0, 1, \dots, d - 1 \\ X^d = \sinh v + \frac{1}{2} e^v x^2 = \frac{1 - u^2}{2u} + \frac{1}{2u} x^2 \\ X^{d+1} = \cosh v - \frac{1}{2} e^v x^2 = \frac{1 + u^2}{2u} - \frac{1}{2u} x^2 \end{cases} \quad (75)$$

with $x^2 = x^0^2 - x^1^2 - \dots - x^{d-1}^2$. In each slice $\Pi_v, x^0, \dots, x^{d-1}$ can be seen as coordinates of an event of a d -dimensional Minkowski spacetime \mathbb{M}^d with metric

³The coordinate $u = e^{-v}$ is frequently called z in the recent literature. We are forced to change this notation because we reserve the letter z to complex quantities. By also allowing negative values for u the coordinate system (75) covers almost all the real manifold AdS_{d+1} .

$ds_M^2 = dx^{0^2} - dx^{1^2} - \dots - dx^{d-1^2}$ (here and in the following where it appears, an index M stands for Minkowski). This explains why the horocyclic coordinates (v, x) of the parametrisation (75) are also called Poincaré coordinates. The scalar product (1) and the AdS metric can then be rewritten as follows:

$$X \cdot X' = \cosh(v - v') - \frac{1}{2}e^{v+v'}(x - x')^2, \quad (76)$$

$$ds_{AdS}^2 = e^{2v}ds_M^2 - dv^2 = \frac{1}{u^2}(ds_M^2 - du^2). \quad (77)$$

Equation (76) implies that

$$(X(v, x) - X(v, x'))^2 = e^{2v}(x - x')^2. \quad (78)$$

This in turn implies that space-like separation in any slice Π_v can be understood equivalently in the Minkowskian sense of the slice itself or in the sense of the ambient AdS universe.

Equation (77) exhibits the region Π of AdS_{d+1} as a warped product with warping function $\omega(v) = e^v$ and fibers conformal to \mathbb{M}^d .

Finally, the representation of Π by the parametrisation (71) is specified by considering Π as embedded in the fundamental sheet of $\widehat{AdS}_{d+1}^{(c)}$; it is therefore described by the following conditions on the coordinates r, τ, \mathbf{e} :

$$-\pi < \tau < \pi; \quad re^d + \sqrt{r^2 + 1} \cos \tau > 0. \quad (79)$$

The ‘Euclidean’ submanifold E_{d+1} of $\widehat{AdS}_{d+1}^{(c)}$ is the set of all points $Z = X + iY$ in $\widehat{AdS}_{d+1}^{(c)}$ such that $X = (0, X^1, \dots, X^{d+1})$, $Y = (Y^0, 0, \dots, 0)$ and $X^{d+1} > 0$. It is therefore represented by the upper sheet (characterised by the condition $X^{d+1} > 0$) of the two-sheeted hyperboloid with equation $X^{d+1^2} - Y^{0^2} - X^{1^2} - \dots - X^{d^2} = 1$. E_{d+1} is equally well represented in both parametrisations (71) and (75) as follows:

$$Z = Z[r, \tau = i\sigma, \mathbf{e}]; \quad (r, \sigma, \mathbf{e}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{S}_{d-1} \quad (80)$$

or

$$Z = Z(v, (iy^0, x^1, \dots, x^{d-1})); \quad v \in \mathbb{R}, \quad (y^0, x^1, \dots, x^{d-1}) \in \mathbb{R}^d. \quad (81)$$

In view of (80), E_{d+1} is contained in the fundamental sheet of $\widehat{AdS}_{d+1}^{(c)}$.

For each v , the complexification $\Pi_v^{(c)}$ of the horosphere Π_v is parametrised by formulae (6) in which x is replaced by the complex Minkowskian vector $z = x + iy = (z^0, \dots, z^d)$; the Euclidean submanifold of this complex Minkowskian manifold is obtained as the intersection $\Pi_v^{(c)} \cap E_{d+1}$.

6.2. Quantum field theory

Let us consider now a general QFT on \widehat{AdS}_{d+1} ; for simplicity we limit the present discussion to one scalar field $\Phi(X)$; we present a simplified version of the theory exposed in [10]. A theory is completely determined by the set of all n -point vacuum expectation values (or ‘Wightman functions’) of the field Φ , given as distributions on the corresponding product manifolds $(\widehat{AdS}_{d+1})^n$:

$$\mathcal{W}_n(X_1, \dots, X_n) = \langle \Omega, \Phi(X_1) \cdots \Phi(X_n) \Omega \rangle. \quad (82)$$

These distributions are supposed to be tempered when represented in the variables of the covering parametrisation $X_j = X_j[r_j, \tau_j, e_j]$ and to satisfy a set of general requirements, namely *AdS invariance*, *positive-definiteness*, *hermiticity*, *local commutativity*, *analyticity corresponding to an appropriate spectral condition* and *‘dimensional boundary conditions’ at infinity*.

The requirement of AdS invariance (corresponding to the scalar character of the field) can be written as follows:

$$\mathcal{W}_n(gX_1, \dots, gX_n) = \mathcal{W}_n(X_1, \dots, X_n) \quad \text{for any } g \in \hat{G}. \quad (83)$$

The usual positivity and hermiticity properties [23] are valid for scalar QFTs on any spacetime and we do not spell them out.

(a) *Local commutativity*. $\Phi(X)$ commutes (as an operator-valued distribution) with $\Phi(X')$ for X, X' space-like separated in the sense of the covering space \widehat{AdS}_{d+1} , as defined above. As in the Minkowskian case, this postulate is equivalent to the coincidence of permuted Wightman functions at space-like separation of consecutive arguments X_j, X_{j+1} [23].

(b) *Analyticity corresponding to energy spectrum condition*. Since the parameter of the covering group of the rotations in the $(0, d+1)$ -plane is interpreted as a genuine time-translation for the observers in all the corresponding Killing trajectories, and since the complexifications of these trajectories do not exhibit any geometrical periodicity in $\widehat{AdS}_{d+1}^{(c)}$, we consider QFTs for which the corresponding infinitesimal generator $J_{0,d+1}$ is represented by a self-adjoint operator whose spectrum is bounded from below. By using a standard Laplace transform argument in the corresponding time-variables τ_1, \dots, τ_n , one is led to formulate this spectral condition by the following analyticity property of the Wightman functions:

Each tempered distribution $\mathcal{W}_n(X_1[r_1, \tau_1, e_1], \dots, X_n[r_n, \tau_n, e_n])$ is the boundary value of a holomorphic function $W_n(Z_1, \dots, Z_n)$ which is defined in a complex neighbourhood of the set $\{Z = (Z_1, \dots, Z_n); Z_j = X_j + iY_j \in \widehat{AdS}_{d+1}^{(c)}; Z_j = Z_j[r_j, \tau_j, e_j]; \text{Im}\tau_1 < \text{Im}\tau_2 < \dots < \text{Im}\tau_n\}$.

As a by-product, the Schwinger function S_n , that is the restriction of each W_n to the Euclidean submanifold $\{(Z_1, \dots, Z_n) \in (E_{d+1})^n; \sigma_1 < \sigma_2 < \dots < \sigma_n\}$, is well-defined. See [10] for a more general setting.

(c) *Dimensional boundary conditions at infinity.* By making use of the coordinates (71) the following limits should exist in the sense of distributions:

$$\begin{aligned} & \lim_{(r_1, \dots, r_n) \rightarrow +\infty} (r_1 \cdots r_n)^\Delta \mathcal{W}_n(X_1[r_1, \tau_1, \mathbf{e}_1], \dots, X_n[r_n, \tau_n, \mathbf{e}_n]) \\ & = \mathcal{W}_n^\infty([\tau_1, \mathbf{e}_1], \dots, [\tau_n, \mathbf{e}_n]). \end{aligned} \quad (84)$$

It is not true in general that a distribution $\mathcal{W}_n(X_1, \dots, X_n)$ can be restricted to the submanifold $\prod_{j=1}^n \hat{C}_{r_j}$ of $(\widehat{AdS}_{d+1})^n$ (C_r was defined in Eq. (72)). The above spectral condition b) implies that this can be done in the present framework.

Moreover, it is then natural to assume that *the limit in Eq. (84) can be extrapolated to the holomorphic functions W_n in their tube domains T_n* so that the corresponding limits W_n^∞ are themselves holomorphic in T_n and admit the corresponding distributions \mathcal{W}_n^∞ as their boundary values on the reals. By restricting all these holomorphic functions to the Euclidean manifolds $\tau_j = i\sigma_j$, $j = 1, \dots, n$, one then obtains a similar condition for the Schwinger functions S_n and the corresponding limits S_n^∞ .

As a special application of the previous framework, it is meaningful to consider the restrictions of the distributions \mathcal{W}_n to the submanifolds $\hat{C}_r^{\times n}$ of $\widehat{AdS}_{d+1}^{\times n}$ (i.e., to the case when all variables r_j are equal to r). One then notices that the positivity conditions satisfied by assumption by the distributions \mathcal{W}_n on \widehat{AdS}_{d+1} can be extended to test-functions of the variables τ_j and \mathbf{e}_j localized in these submanifolds $r_1 = \dots = r_n = r$. The standard reconstruction procedure allows to say that in each slice \hat{C}_r the given field on \widehat{AdS}_{d+1} yields by restriction a well-defined quantum field $\Phi_r(\tau, \mathbf{e})$. This field is obviously invariant under the product of the translation group with time-parameter τ by the orthogonal group $SO(d)$ of space transformations acting on the sphere \mathbb{S}_{d-1} of the variables \mathbf{e} . Moreover, it follows from the locality postulate a) together with Eqs. (73) and (74) that the field Φ_r also satisfies local commutativity in the sense of the spacetime manifold \hat{C}_r . Finally, in view of b), the n -point functions of Φ_r are (for each r) boundary values of holomorphic functions of the complex variables τ_1, \dots, τ_n in the tube T_n , which shows that these theories satisfy a spectral condition with respect to the generator of time-translations.

7. Correspondence with Conformal Field Theories on $\widehat{\mathcal{C}}_{2,d}$ à la Lüscher–Mack

We shall now introduce the asymptotic cone $\mathcal{C}_{2,d}$ (resp. $\mathcal{C}_{2,d}^{(c)}$) of AdS_{d+1} (resp. $AdS_{d+1}^{(c)}$) and wish to identify the limit (in the sense of Eq. (84)) of a QFT on \widehat{AdS}_{d+1} satisfying the previous properties with a QFT on the corresponding covering $\widehat{\mathcal{C}}_{2,d}$ of $\mathcal{C}_{2,d}$. To do this, we first notice that by adapting the covering parametrisation (71)

of \widehat{AdS}_{d+1} to the case of its asymptotic cone, $\mathcal{C}_{2,d} = \{\eta = (\eta^0, \dots, \eta^{(d+1)}); \eta^{0^2} - \eta^{1^2} - \dots - \eta^{d^2} + \eta^{(d+1)^2} = 0\}$, one readily obtains the following parametrisation:

$$\begin{cases} \eta^0 = r \sin \tau \\ \eta^i = r e^i \quad i = 1, \dots, d \\ \eta^{d+1} = r \cos \tau \end{cases} \quad (85)$$

with $e^{1^2} + \dots + e^{d^2} = 1$ and $r \geq 0$, or in brief: $\eta = \eta[r, \tau, e]$.

The parametrisation (85) allows one to introduce the coverings $\widehat{\mathcal{C}}_{2,d}$ and $\widehat{\mathcal{C}}_{2,d}^{(c)}$ of $\mathcal{C}_{2,d}$ and $\mathcal{C}_{2,d}^{(c)}$ by again unfolding the 2π -periodic coordinate τ (resp. $\text{Re } \tau$). A privileged ‘fundamental sheet’ is defined on these coverings by imposing the condition $-\pi < \tau < \pi$ (resp. $-\pi < \text{Re } \tau < \pi$).

We also note that the standard condition of space-like separation on $\mathcal{C}_{2,d}$ is similar to the condition chosen on the *AdS* spacetime, namely

$$\begin{aligned} (\eta - \eta')^2 &= r^2 \left[4 \left(\sin \left(\frac{\tau - \tau'}{2} \right) \right)^2 - (e - e')^2 \right] \\ &= -2r^2 (\cos(\tau - \tau') - e \cdot e') < 0, \end{aligned} \quad (86)$$

and yields the corresponding global causal ordering on $\widehat{\mathcal{C}}_{2,d}$

$$(\tau, e) > (\tau', e') \quad \text{iff} \quad \tau - \tau' > 2 \text{Arcsin} \left(\frac{(e - e')^2}{4} \right)^{\frac{1}{2}}. \quad (87)$$

Note that in the space of variables (τ, τ', e, e') , the region described by Eq. (87) is exactly the limit of the region given by Eq. (74) when r tends to infinity.

By taking the intersection of $\mathcal{C}_{2,d}$ with the family of hyperplanes with equation $\eta^d + \eta^{d+1} = e^v$, one obtains the analogue of the horocyclic parametrisation (75), namely:

$$\begin{cases} \eta^\mu = e^v x^\mu \quad \mu = 0, 1, \dots, d-1 \\ \eta^d = \frac{1}{2} e^v (1 + x^2) \quad x^2 = x^{0^2} - x^{1^2} - \dots - x^{d-1^2} \\ \eta^{d+1} = \frac{1}{2} e^v (1 - x^2), \end{cases} \quad (88)$$

which implies the following identity (similar to (76)) between quadratic forms:

$$(\eta - \eta')^2 = e^{v+v'} (x - x')^2. \quad (89)$$

By taking Eq. (85) into account, one then sees that these formulae correspond (in dimension d) to the embedding of Minkowski space into the covering of the cone

$\mathcal{C}_{2,d}$, namely one has (in view of the identification $\eta^d + \eta^{d+1} = e^v = r(e^d + \cos \tau)$):

$$x^0 = \frac{\sin \tau}{\cos \tau + e^d}, \quad x^i = \frac{e^i}{\cos \tau + e^d}, \quad (90)$$

with

$$\cos \tau + e^d > 0, \quad -\pi < \tau < \pi. \quad (91)$$

Let us now consider a general QFT on \widehat{AdS}_{d+1} whose Wightman functions \mathcal{W}_n satisfy AdS invariance together with the properties (a), (b) and (c) described in the previous section. In view of (c), we can associate with the latter the following set of n -point distributions $\widetilde{\mathcal{W}}_n(\eta_1, \dots, \eta_n)$ on $\widehat{\mathcal{C}}_{2,d}$:

$$\widetilde{\mathcal{W}}_n(\eta_1, \dots, \eta_n) = (r_1 \cdots r_n)^{-\Delta} \mathcal{W}_n^\infty([\tau_1, \mathbf{e}_1], \dots, [\tau_n, \mathbf{e}_n]). \quad (92)$$

At first, one can check that the set of distributions $\widetilde{\mathcal{W}}_n$ satisfy the required positivity conditions for defining a QFT on $\widehat{\mathcal{C}}_{2,d}$. This is because, in view of postulate (c) (applied with all r_j equal to the same r), the distributions \mathcal{W}_n^∞ appear as the limits of the n -point functions of the QFT's on the spacetimes $\widehat{\mathcal{C}}_r$ when r tends to infinity. The positivity conditions satisfied by the latter are then preserved in the limit, in terms of test-functions of the variables τ_j and \mathbf{e}_j , and then extended in a trivial way into the radial variables r_j as positivity conditions for the distributions on the cone $\widehat{\mathcal{C}}_{2,d}$ (by using the appropriate test-functions homogeneous in the variables r_j [33]).

It follows from the reconstruction procedure [23] that the set of distributions $\widetilde{\mathcal{W}}_n$ define a quantum field $\widetilde{\mathcal{O}}(\eta)$ on $\widehat{\mathcal{C}}_{2,d}$. $\widetilde{\mathcal{O}}(\eta)$ enjoys the following properties:

Local commutativity: Since the region (87) is the limit of (74) for r tending to infinity, it results from the boundary condition (c) and from the local commutativity of all fields Φ_r in the corresponding spacetimes $\widehat{\mathcal{C}}_r$ that the field $\widetilde{\mathcal{O}}(\eta)$ satisfies local commutativity on $\widehat{\mathcal{C}}_{2,d}$.

Spectral condition: In view of our postulate (c) extended to the complex domain T_n in the variables τ , we see that the n -point distributions $\widetilde{\mathcal{W}}_n(\eta_1, \dots, \eta_n)$ are boundary values of holomorphic functions in the same analyticity domains of $(\widehat{\mathcal{C}}_{2,d}^{(c)})^n$ as those of the Lüscher–Mack field theories [33].

It is possible to show [11] that the \widehat{G} -invariance (83) of the AdS n -point functions, together with the properties (a), (b), (c), imply the conformal invariance of the field $\widetilde{\mathcal{O}}(\eta)$; more precisely, the Wightman functions $\widetilde{\mathcal{W}}_n$ of this field are invariant under the action on $\widehat{\mathcal{C}}_{2,d}$ of the group \widehat{G} , now interpreted as in [33] as the ‘quantum mechanical conformal group’, namely that one has:

$$\widetilde{\mathcal{W}}_n(g\eta_1, \dots, g\eta_n) = \widetilde{\mathcal{W}}_n(\eta_1, \dots, \eta_n) \quad (93)$$

for all g in \widehat{G} .

We can then summarise the results of [11] this section by the following statement: *The procedure we have described (expressed by Eqs. (84) and (92)) displays a general AdS/CFT correspondence for QFTs:*

$$\Phi(X) \rightarrow \tilde{\mathcal{O}}(\eta) \quad (94)$$

between a scalar (AdS invariant) quantum field $\Phi(X)$ on the covering $\widehat{\text{AdS}}_{d+1}$ of AdS_{d+1} whose Wightman functions satisfy the properties (a), (b), (c), and a conformally invariant local field $\tilde{\mathcal{O}}(\eta)$ on the covering $\widehat{\mathcal{C}}_{2,d}$ of the cone $\mathcal{C}_{2,d}$, enjoying the Lüscher–Mack spectral condition; the degree of homogeneity (dimension) Δ of $\tilde{\mathcal{O}}(\eta)$ is equal to the asymptotic dimension of the AdS field $\Phi(X)$.

Of course, from this general point of view, the correspondence may *a priori* be many-to-one. Finally, according to the formalism described in [33], the correspondence (94) can be completed by saying that there exists a unique conformal (Minkowskian) local field $\mathcal{O}(x)$ of dimension Δ whose n -point functions \mathcal{W}_n^M are expressed in terms of those of $\tilde{\mathcal{O}}(\eta)$ by the following formulae:

$$\begin{aligned} \mathcal{W}_n^M(x_1, \dots, x_n) &= e^{(v_1 + \dots + v_n)\Delta} \tilde{\mathcal{W}}_n(\eta_1, \dots, \eta_n) \\ &= \prod_{1 \leq j \leq n} (\eta_j^d + \eta_j^{d+1})^\Delta \tilde{\mathcal{W}}_n(\eta_1, \dots, \eta_n). \end{aligned} \quad (95)$$

In the latter, the Minkowskian variables x_j are expressed in terms of the cone variables η_j by inverting (88), which yields:

$$x_j^\mu = \frac{\eta_j^\mu}{\eta_j^d + \eta_j^{d+1}}. \quad (96)$$

8. Two-Point Functions

8.1. The analytic structure of two-point functions on the AdS spacetime

It turns out that in all field theories on $\widehat{\text{AdS}}_{d+1}$ satisfying the general requirements described in subsection 2.2, the two-point function enjoys *maximal analyticity properties* in all the coordinates, as it is the case for the Minkowski [23] and de Sitter cases [8]. A full proof of these results will be found in [10]. We shall only give here a descriptive account of them, needed for further applications. Since, in particular, AdS covariance and the ‘energy spectrum condition’ are responsible for this maximal analytic structure; we shall consider this general class of two-point functions as ‘preferred’.

There are two distinguished complex domains of $\text{AdS}_{d+1}^{(c)}$, invariant under real AdS transformations, which are of crucial importance for a full understanding of

the structures associated with two-point functions. They are given by:

$$\begin{aligned} T^+ &= \{Z = X + iY \in AdS_{d+1}^{(c)}; Y^2 > 0, \epsilon(Z) = +1\}, \\ T^- &= \{Z = X + iY \in AdS_{d+1}^{(c)}; Y^2 > 0, \epsilon(Z) = -1\}, \end{aligned} \quad (97)$$

where

$$\epsilon(Z) = \text{sign}(Y^0 X^{d+1} - X^0 Y^{d+1}). \quad (98)$$

T^+ and T^- are the AdS version of the usual forward and backward tubes T_M^+ and T_M^- of complex Minkowski spacetime, obtained in correspondence with the energy-momentum spectrum condition [23]; let us recall their definition (in arbitrary spacetime dimension p):

$$\begin{aligned} T_M^+ &= \{z = x + iy \in \mathbb{M}^{p(c)}; y^2 > 0, y^0 > 0\}, \\ T_M^- &= \{z = x + iy \in \mathbb{M}^{p(c)}; y^2 > 0, y^0 < 0\}. \end{aligned} \quad (99)$$

In the same way as these Minkowskian tubes are generated by the action of real Lorentz transformations on the ‘flat’ (one complex time-variable) domains $\{z = x + iy; y = (y^0, \vec{0}); y^0 > 0(\text{resp. } y^0 < 0)\}$, the domains (97) of $AdS_{d+1}^{(c)}$ are generated by the action of the group G on the flat domains obtained by letting τ vary in the half-planes $\text{Im}\tau > 0$ or $\text{Im}\tau < 0$ and keeping r and e real in the covering parametrisation (71) of the AdS quadric. In fact, by using the complex extension of this parametrisation and putting $r = \sinh(\psi + i\phi)$, $\tau = \text{Re}\tau + i\sigma$ one can represent the domains (97) by the following semi-tubes (invariant under translations in the variable $\text{Re}\tau$):

$$\pm \sinh \sigma > \left[\frac{(\sin \phi)^2 + ((\cosh \psi)^2 - (\cos \phi)^2)(\text{Im } e)^2}{(\cosh \psi)^2 - (\sin \phi)^2} \right]^{\frac{1}{2}} \quad (100)$$

This representation (which clearly contains the previously mentioned flat domains) can be thought of, either as representing the domains (97) of $AdS^{(c)}$ if τ is identified to $\tau + 2\pi$, or coverings of the latter embedded in $\widehat{AdS}_{d+1}^{(c)}$, which we denote by \hat{T}^+ and \hat{T}^- , if one does not make this identification.

One typical property of Wightman’s QFT [23] is that any two-point distribution $\mathcal{W}_M(x, x')$ satisfying the spectral condition is the boundary value of a function $W_M(z, z')$ holomorphic for $z \in T_M^-$ and $z' \in T_M^+$. An analogous property also holds for n -point functions.

It is a consequence of AdS invariance together with the spectrum assumption (b) [10] that, also in the AdS spacetime, general two-point functions can be

characterised by the following global analyticity property which plays the role of a *G-invariant spectral condition*:

(b^(inv)) *Normal analyticity condition for two-point functions: the two-point function $\mathcal{W}(X, X')$ is the boundary value of a function $W(Z, Z')$ which is holomorphic in the domain $\hat{T}^- \times \hat{T}^+$ of $\widehat{AdS}_{d+1}^{(c)} \times \widehat{AdS}_{d+1}^{(c)}$.*

A further use of AdS invariance implies that $W(Z, Z')$ is actually a function $w(\zeta)$ of a single complex variable ζ ; this variable ζ can be identified with $Z \cdot Z'$ when Z and Z' are both in the fundamental sheet of $\widehat{AdS}_{d+1}^{(c)}$; AdS invariance and the normal analyticity condition together imply the following:

Maximal analyticity property: $w(\zeta)$ is analytic in the covering $\widehat{\Theta}$ of the cut-plane $\Theta = \{\mathbb{C} \setminus [-1, 1]\}$.

For special theories which are periodic in the time coordinate τ , $w(\zeta)$ is in fact analytic in Θ itself. One can now introduce all the usual Green functions. The ‘permuted Wightman function’ $\mathcal{W}(X', X) = \langle \Omega, \Phi(X')\Phi(X)\Omega \rangle$ is the boundary value of $W(Z, Z')$ from the domain $\{(Z, Z') : Z \in \hat{T}^+, Z' \in \hat{T}^-\}$. The commutator function is then $\mathcal{C}(X, X') = \mathcal{W}(X, X') - \mathcal{W}(X', X)$. The retarded propagator $\mathcal{R}(X, X')$ is introduced by splitting the support of the commutator $\mathcal{C}(X, X')$ as follows

$$\mathcal{R}(X, X') = i\theta(\tau - \tau')\mathcal{C}(X, X'). \tag{101}$$

The other Green functions are then defined in terms of \mathcal{R} by the usual formulae: the advanced propagator is given by $\mathcal{A} = \mathcal{R} - i\mathcal{C}$ while the chronological propagator is given by $\mathcal{F} = -i\mathcal{A} + \mathcal{W}$.

Note finally that, as a function of the single variable $\zeta = X \cdot X'$, the jump $i\delta w(\zeta)$ of $w(\zeta)$ across its cut $(-\infty, +1]$ coincides with the retarded propagator $\mathcal{R}(X, X')$ (or the advanced one); in the periodic (i.e., ‘true AdS’) case, the support of δw reduces to the compact interval $[-1, +1]$.

8.2. The simplest example revisited: Klein–Gordon fields in the AdS/CFT correspondence

The Wightman functions of fields satisfying the Klein–Gordon equation AdS_{d+1}

$$\square_{AdS}\Phi + m^2\Phi = 0. \tag{102}$$

display the simplest example of the previous analytic structure:

$$W_\nu(Z, Z') = w_\nu(\zeta) = \frac{e^{-i\pi\frac{d-1}{2}}}{(2\pi)^{\frac{d+1}{2}}} (\zeta^2 - 1)^{-\frac{d-1}{4}} Q_{\nu-\frac{1}{2}}^{\frac{d-1}{2}}(\zeta). \tag{103}$$

Here Q is a second-kind Legendre's function; the parameter ν is linked to the field's mass by the relation

$$\nu^2 = \frac{d^2}{4} + m^2. \quad (104)$$

and the normalisation of W_ν is chosen by imposing the short-distance Hadamard behaviour.

Since $W_\nu(Z, Z')$ and $W_{-\nu}(Z, Z')$ are solutions of the same Klein–Gordon equation (and share the same analyticity properties), the question arises if these Wightman function both define acceptable QFTs on AdS_{d+1} . The answer [34] is that only theories with $\nu \geq -1$ are acceptable and there are therefore two regimes: for $\nu > 1$ there is only one field theory corresponding to a given mass while for $|\nu| < 1$ there are two theories. The case $\nu = 1$ is a limit case. Eq. (103) shows clearly that the only difference between the theories parametrised by opposite values of ν is in their large distance behaviour:

$$w_{-\nu}(\zeta) = w_\nu(\zeta) + \frac{\sin \pi \nu}{(2\pi)^{\frac{d+1}{2}}} \Gamma\left(\frac{d}{2} - \nu\right) \Gamma\left(\frac{d}{2} + \nu\right) (\zeta^2 - 1)^{-\frac{d-1}{4}} P_{-\frac{1}{2}-\nu}^{-\frac{d-1}{2}}(\zeta). \quad (105)$$

Now we notice that in this relation (where all terms are solutions of the same Klein–Gordon equation) the last term is *regular on the cut* $\zeta \in [-1, 1]$. This entails (reintroducing the AdS radius R) that, in the two theories, the c -number commutator $[\Phi(X), \Phi(X')]$ takes the same value for all (time-like separated) vectors (X, X') such that $|X \cdot X'| < R^2$. Therefore we can say that *the two theories represent the same algebra of local observables at short distances (with respect to the radius R)*. But since the last term in the latter relation grows the faster the larger is $|\nu|$ we see that the two theories drastically differ by their long range behaviours.

The existence of the two regimes above has given rise to two distinct treatments of the AdS/CFT correspondence in the two cases and symmetry breaking had been advocated to explain the difference.

In the present context, by applying the correspondence as given in Eq. (94), the two regimes can be treated in one stroke. Indeed the large ζ behaviour of the Legendre's function Q (valid for any complex ν):

$$Q_{\nu-\frac{1}{2}}^{\frac{d-1}{2}}(\zeta) \simeq e^{i\pi\frac{d-1}{2}} 2^{-\nu-\frac{1}{2}} \frac{\Gamma\left(\nu + \frac{d}{2}\right)}{\Gamma(\nu + 1)} \pi^{\frac{1}{2}} \zeta^{-\frac{1}{2}-\nu}. \quad (106)$$

It follows that the two-point function (103) and thereby all the n -point functions of the corresponding Klein–Gordon field satisfy the dimensional boundary conditions

at infinity with dimension $\Delta = \frac{d}{2} + \nu$. Indeed, let τ and τ' be complex and such that $\text{Im}\tau < \text{Im}\tau'$. It follows that

$$\begin{aligned} W_\nu^\infty([\tau, e], [\tau', e']) &= \lim_{r, r' \rightarrow \infty} (rr')^{\frac{d}{2} + \nu} W_\nu(Z[\tau, r, e], Z'[\tau', r', e']) \\ &= \frac{2^{-\nu-1} \Gamma\left(\nu + \frac{d}{2}\right)}{(2\pi)^{\frac{d}{2}}} \frac{1}{\Gamma(\nu + 1)} \frac{1}{[\cos(\tau - \tau') - e \cdot e']^{\frac{d}{2} + \nu}}. \end{aligned} \quad (107)$$

This equation expresses nothing more than the behaviour of the previous Legendre's function at infinity. Not only all the ν 's are treated this way in one stroke but, also, one can study the boundary limit for theories corresponding to $\nu < -1$, even if the corresponding QFT may have no direct physical interpretation.

The two-point function of the conformal field $\widehat{\mathcal{O}}(\eta)$ on the cone $\widehat{\mathcal{C}}_{2,d}$ corresponding to (107) is then constructed by following the prescription of Eq. (92), which yields

$$\begin{aligned} \widetilde{W}_\nu(\eta, \eta') &= (rr')^{-\frac{d}{2} - \nu} W_\nu^\infty([\tau, e], [\tau', e']) \\ &= \frac{1}{2\pi^{\frac{d}{2}}} \frac{\Gamma\left(\nu + \frac{d}{2}\right)}{\Gamma(\nu + 1)} \frac{1}{[-(\eta - \eta')^2]^{\frac{d}{2} + \nu}}. \end{aligned} \quad (108)$$

Correspondingly, we can deduce from (108) the expression of the two-point function of the associated Minkowskian field on \mathbb{M}^d , given by formula (95); by taking Eq. (89) into account, we obtain:

$$\begin{aligned} W_\nu^M(z, z') &= e^{(\nu + \nu')\left(\frac{d}{2} + \nu\right)} \widetilde{W}_\nu(\eta(v, z), \eta'(v', z')) \\ &= \frac{1}{2\pi^{\frac{d}{2}}} \frac{\Gamma\left(\nu + \frac{d}{2}\right)}{\Gamma(\nu + 1)} \frac{1}{[-(z - z')^2]^{\frac{d}{2} + \nu}}. \end{aligned} \quad (109)$$

In the latter, the Poincaré coordinates z and z' must be taken with the usual $i\epsilon$ -prescription ($\text{Im} z^0 < \text{Im} z'^0$), which can be checked to be implied by the spectral condition (b) of section 2 through the previous limiting procedure.

Let us now describe how the previous limiting procedure looks in the Poincaré coordinates (75). These coordinates offer the possibility of studying directly the boundary behaviour of the AdS Wightman functions in a larger domain of the complex AdS spacetime. This fact is based on the following simple observation: consider the parametrisation (75) for two points with complex parameters specified by

$$\begin{aligned} Z &= Z(v, z), & v &\in \mathbb{R}, & z &\in T_M^- \\ Z' &= Z'(v', z'), & v' &\in \mathbb{R}, & z' &\in T_M^+. \end{aligned} \quad (110)$$

It is easy to check that this choice of parameters implies that $Z \in T^-$ and $Z' \in T^+$. It follows that, given an AdS invariant two-point function satisfying locality and the normal analyticity condition $b^{(inv)}$, the following restriction automatically generates a local and (Poincaré) covariant two-point function on the slice Π_v , which satisfies the spectral condition [23] (in short: the two-point function of a general Wightman QFT):

$$W_{\{v\}}^M(z, z') = W(Z(v, z), Z'(v, z')). \quad (111)$$

On the basis of the dimensional boundary condition (84), and of the fact (obtained by comparing (71) and (75)) that $\frac{e^v}{r} = \sqrt{1 + \frac{1}{r^2}} \cos \tau + e^d$ tends to the finite limit $\cos \tau + e^d$ when r tends to infinity, one sees that the following limit exists and that it yields (in view of (92) and (95)):

$$\lim_{v \rightarrow +\infty} e^{2v\Delta} W_{\{v\}}^M(z, z') = W^M(z, z'). \quad (112)$$

The limiting two-point function $W^M(z, z')$ then automatically exhibits locality, Poincaré invariance and the spectral condition. The invariance under special conformal transformations and scaling property would necessitate a special check, but they result from the general statement of conformal invariance of the limiting field $\hat{\mathcal{O}}(\eta)$.

When applied to the Wightman functions of Klein–Gordon fields (i.e., with $\Delta = \frac{d}{2} + v$), the latter presentation of the limiting procedure gives immediately the result obtained in Eq. (109) but in a larger complex domain:

$$\lim_{v \rightarrow \infty} e^{2v(\frac{d}{2}+v)} W_v(Z(v, z), Z'(v, z')) = \frac{1}{2\pi^{\frac{d}{2}}} \frac{\Gamma\left(v + \frac{d}{2}\right)}{\Gamma(v+1)} \frac{1}{[-(z-z')^2]^{\frac{d}{2}+v}}. \quad (113)$$

In a completely similar way one can compute the bulk-to-boundary correlation function by considering a two-slice restriction $W_v(Z(v, z), Z'(v', z'))$ of W_v . The bulk-to-boundary correlation function is obtained by sending $v' \rightarrow \infty$ while keeping v fixed, by the following limit:

$$\begin{aligned} & \lim_{v' \rightarrow \infty} e^{v'(\frac{d}{2}+v)} W_v(Z(v, z), Z'(v', z')) \\ &= \frac{1}{2\pi^{\frac{d}{2}}} \frac{\Gamma\left(v + \frac{d}{2}\right)}{\Gamma(v+1)} \frac{1}{(e^{-v} - e^v(z-z')^2)^{\frac{d}{2}+v}} \\ &= \frac{1}{2\pi^{\frac{d}{2}}} \frac{\Gamma\left(v + \frac{d}{2}\right)}{\Gamma(v+1)} \left(\frac{u}{u^2 - (z-z')^2} \right)^{\frac{d}{2}+v}. \end{aligned} \quad (114)$$

The above results suggest the following alternative approach to the AdS/CFT correspondence. Starting from a given set of AdS invariant n -point functions satisfying general requirements it is (at least formally) possible to obtain a set of Poincaré invariant n -point functions in one-dimension less by taking the following restrictions [10]:

$$\mathcal{W}_{n\{v\}}^M(x_1, \dots, x_n) = \mathcal{W}_n(X_1(v, x_1), \dots, X_n(v, x_n)). \quad (115)$$

On the basis of the requirement of asymptotic dimensionality (c) the boundary will be obtained by taking the following limits:

$$\mathcal{W}_n^M(x_1, \dots, x_n) = \lim_{v \rightarrow \infty} e^{nv\Delta} \mathcal{W}_{n\{v\}}^M(x_1, \dots, x_n). \quad (116)$$

One can also consider a many-leaf restriction as follows:

$$\begin{aligned} & \mathcal{W}_{n\{v_{m+1}, \dots, v_n\}}(X_1, \dots, X_m, x_{m+1}, \dots, x_n) \\ &= \mathcal{W}_n(X_1, \dots, X_m, X_{m+1}(v_{m+1}, x_{m+1}), \dots, X_n(v_n, x_n)), \end{aligned} \quad (117)$$

and get various bulk-to-boundary correlation functions by taking the limit as before:

$$\begin{aligned} & \mathcal{W}_n(X_1, \dots, X_m, x_{m+1}, \dots, x_n) \\ &= \lim_{v_{m+1}, \dots, v_n \rightarrow \infty} e^{(v_{m+1} + \dots + v_n)\Delta} \mathcal{W}_{n\{v_{m+1}, \dots, v_n\}}(X_1, \dots, X_m, x_{m+1}, \dots, x_n). \end{aligned} \quad (118)$$

Restricting ourselves here to the limiting procedure described by Eq. (116), we then see that the general AdS/CFT correspondence for QFTs described in Section 3 can alternatively be presented purely in terms of a limit of Minkowskian fields, denoted as follows:

$$\Phi(X) \rightarrow \{\varphi_v(x)\} \rightarrow \mathcal{O}(x), \quad (119)$$

where each field $\varphi_v(x)$ is the scalar Minkowskian field whose n -point correlation functions are those given by (115).

There is a substantial difference between two-point and n -point functions. In fact, in view of their maximal analyticity property the two-point functions admit restrictions to the slices Π_v which are themselves boundary values of holomorphic functions in relevant Minkowskian complex domains of the corresponding complexified slices $\Pi_v^{(c)}$: in this case there is therefore no problem of restriction of the distribution \mathcal{W}_2 to $\Pi_v \times \Pi_v$.

As regards the n -point correlation functions, the existence of the restrictions (115) as distributions on $(\Pi_v)^n$ is not an obvious consequence of the requirements (a), (b), (c) of Section 2. Only the existence of the corresponding restrictions at Euclidean points of $(\Pi_v^{(c)})^n$ (namely the Schwinger functions of

these Minkowskian theories) are direct consequences of the spectral condition (b) we have assumed: this is because changing τ into $i\sigma$ in (71) or changing x^0 into iy^0 in (75), all other parameters being kept real, yield two equivalent representations of the Euclidean points of $\widehat{\text{AdS}}_{d+1}^{(c)}$.

As a matter of fact, in order to be able to define the restrictions (115) as distributions enjoying the full structure of Minkowskian n -point functions, namely as distribution boundary values of holomorphic functions in relevant domains of $(\Pi_v^{(c)})^n$, one is led to use instead of (b) an alternative spectral condition in which the positivity of the spectrum refers to the representation of a d -dimensional Abelian subgroup of G playing the role of the Minkowskian translation group with respect to the slices Π_v .

Let us briefly sketch the construction. Using the horocyclic parametrisation of Eq. (75), we can lift the action of the Poincaré group as follows. Consider the standard action of the Poincaré group on the Minkowski spacetime coordinates: $x'^\mu = \Lambda_v^\mu x^\nu + a^\mu$, $\mu = 0, 1, \dots, d-1$. By plugging this relation into Eq. (75) we promptly obtain the following relation:

$$\begin{cases} X'^\mu = \Lambda_v^\mu X^\nu + (X^d + X^{d+1})a^\mu \\ X'^d = \left(1 + \frac{a^2}{2}\right) X^d + a_\mu \Lambda_v^\mu X^\nu + X^{d+1} \frac{a^2}{2} \\ X'^{d+1} = \left(1 - \frac{a^2}{2}\right) X^{d+1} - a_\mu \Lambda_v^\mu X^\nu - X^d \frac{a^2}{2} \end{cases}, \quad (120)$$

where Greek indices are raised and lowered with the standard Minkowski metric. In matrix form we get

$$g(\Lambda, a) = \begin{pmatrix} \Lambda & a & a \\ \Lambda a^T & \left(1 + \frac{a^2}{2}\right) & \frac{a^2}{2} \\ -\Lambda a^T & -\frac{a^2}{2} & \left(1 - \frac{a^2}{2}\right) \end{pmatrix}. \quad (121)$$

Among such transformations there is the Abelian subgroup of Poincaré translations $g(\mathbb{I}, a)$. The corresponding generators

$$P_\mu \equiv (X^d + X^{d+1}) \frac{\partial}{\partial X^\mu} + X_\mu \left(\frac{\partial}{\partial X^d} - \frac{\partial}{\partial X^{d+1}} \right) \quad (122)$$

of these transformations form an Abelian algebra. The AdS spectral condition (b) of Section 2 should then be supplemented by the following one:

(b') *Spectral condition: the infinitesimal generators P^μ are represented by (commuting) self-adjoint operators whose joint spectrum is contained in the*

forward light-cone $V^+ = \{p^\mu p_\mu \geq 0, p^0 \geq 0\}$ of a d -dimensional Minkowski momentum space.

By using a Laplace transform argument [23] in the corresponding vector variables x_1, \dots, x_n one can see that this spectral condition implies the following analyticity property of the Wightman functions:

Analyticity corresponding to the spectrum of Poincaré translations: each AdS distribution $\mathcal{W}_n(X_1(v_1, x_1), \dots, X_n(v_n, x_n))$ is the boundary value of a holomorphic function $W_n(Z_1(v_1, z_1), \dots, Z_n(v_n, z_n))$ which is defined in the tube

$$\mathcal{T}_n = \{Z = (Z_1, \dots, Z_n) \in AdS_{d+1}^{(c)}; Z_j = Z_j(v_j, z_j); v_1, \dots, v_n \in \mathbb{R}, \text{Im}(z_{j+1} - z_j) \in V^+, j = 1, \dots, n-1\}. \quad (123)$$

Property (b') implies in particular that it is meaningful to consider the restricted distributions $\mathcal{W}_{n\{v\}}^M$ given in Eq. (115). The Poincaré invariance of $\mathcal{W}_{n\{v\}}^M$ follows immediately by Eq. (120). Furthermore, the positive-definiteness of this family of distributions is induced as before by the analogous property satisfied by the distributions \mathcal{W}_n on \widehat{AdS}_{d+1} .

Under these conditions the reconstruction procedure is now justified and the given field on \widehat{AdS}_{d+1} yields by restriction a well-defined quantum field $\varphi_v(x)$.

Moreover, it follows from the locality postulate (a) together with Eq. (78) that the field φ_v also satisfies standard local commutativity in Π_v . Finally, in view of (b'), the n -point functions of φ_v are (for each v) boundary values of holomorphic functions in the tube domains T_n^M of Wightman's QFT. This shows that these theories satisfy a standard energy-momentum spectrum condition (with respect to the generators of spacetime translations). The conformal covariance of the boundary field $\mathcal{O}(x)$ results from the general analysis of the previous section.

References

- [1] Riess, A. G., Filippenko, A. V., Challis, P. *et al.*, 1998, Observational evidence from supernovae for an accelerating universe and a cosmological constant, *Astronomical Journal* 116, 1009–1038.
- [2] Perlmutter, S., Aldering, G., Goldhaber, G. *et al.*, 1999, Measurements of omega and lambda from 42 high redshift supernovae, *Astrophys. J.* 517, 565–586.
- [3] Einstein, A. 1917, *Kosmologische betrachtungen zur allgemeinen relativitätstheorie*. *Sitzungsber. Preuß. Akad. Wiss. Berlin*, 142–152.
- [4] Guth, A. H. 1981, Inflationary universe: a possible solution to the horizon and flatness problems. *Phys. Rev. D* 23, 347–356.
- [5] Linde, A. 1990, *Particle physics and inflationary cosmology*. Harwood, Chur.
- [6] Maldacena, J. 1998, The large n limit of superconformal field theories and supergravity. *Adv. Theor. Math. Phys.* 2, 231–252.

- [7] De Sitter, W. 1917, On the curvature of space. Proc. Kon. Ned. Akad. Wet. 20, 229–243.
- [8] Bros, J., Moschella, U. 1996, Two-point functions and quantum fields in de sitter universe. Rev. Math. Phys. 8, 327–392.
- [9] Bros, J., Epstein, H. and Moschella, U. 1998, Analyticity properties and thermal effects for general quantum field theory on de Sitter space-time. Commun. Math. Phys. 196, 535–570.
- [10] Bros, J., Epstein, H. and Moschella, U. 2002, Towards a general theory of quantized fields on the anti-de Sitter space-time. Commun. Math. Phys. 231, 481–528.
- [11] Bertola, M., Bros, J., Moschella, U. *et al.*, 2000, A general construction of conformal field theories from scalar anti-de Sitter quantum field theories. Nucl. Phys. B 587, 619–644.
- [12] Rehren, K. H. 2000, Algebraic holography. Annales Henri Poincaré 1, 607–623.
- [13] Moller, C. 1951, The theory of relativity. Oxford University Press, Oxford.
- [14] Ratra, B., Peebles, P. J. E. 1995, Inflation in an open universe. Phys. Rev. D 52, 1837–1894.
- [15] Sasaki, M., Tanaka, T. and Yamamoto, K. 1995, Euclidian vacuum mode functions for a scalar field on open de Sitter spacetime. Phys. Rev. D 51, 2979–2995.
- [16] Moschella, U., Schaeffer, R. 1997, Quantum fluctuations in the open universe. Phys. Rev. D 57, 2147–2151.
- [17] Hawking, S. W., Ellis, G. F. R. 1973, The large scale structure of space-time. Cambridge University Press, Cambridge, UK.
- [18] Cacciatori, S., Gorini, V., Kamenshchik, A. *et al.*, 2008, Conservation laws and scattering for de Sitter classical particles. Class. Quant. Grav. 25, 075008.
- [19] Birrell, N. D., Davies, P. C. W. 1982, Quantum fields in curved space. Cambridge University Press, Cambridge, UK.
- [20] Bunch, T. S., Davies, P. C. W. 1978, Quantum field theory in de-Sitter space: renormalization by point-splitting. Proc. Roy. Soc. Lond. A 360, 117–134.
- [21] Gibbons, G. W., Hawking, S. W. 1977, Cosmological event horizons, thermodynamics, and particle creation. Phys. Rev. D 15, 2738–2751.
- [22] Bros, J., Moschella, U. and Gazeau, J. P. 1994, Quantum field theory in the de Sitter universe. Phys. Rev. Lett. 73, 1746–1749.
- [23] Streater, R. F., Wightman, A. S. 1964, PCT, spin and statistics, and all that. Addison-Wesley, Redwood City, p. 207 (Advanced book classics).
- [24] Erdelyi, A. 1953, The Bateman manuscript project: Higher transcendental functions. Vol. 1. McGraw-Hill Book Company, New York.
- [25] Magnus, W., Oberhettinger, F. and Soni, R. 1966, Formulas and theorems for the special functions of mathematical physics. Springer-Verlag, Berlin.
- [26] Marichev, O. I. 1982, Handbook of integral transforms of higher transcendental functions. Horwood, Chichester.
- [27] Bros, J., Epstein, H. and Moschella, U. 2008, Lifetime of a massive particle in a de Sitter universe. JCAP 0802; 003.
- [28] Bros, J., Epstein, M. and Moschella, U. 2009, Particle decays and stability on the de Sitter universe, Ann. Inst. Henri Poincaré 11, 611–658.
- [29] Bros, J., Epstein, H., Gaudin, M. *et al.*, 2010, Triangular invariants, three-point functions and particle stability on the de Sitter universe, Commun. Math. Phys. 295, 261–288.

- [30] Avis, S. J., Isham, C. J. and Storey, D. 1978, Quantum field theory in anti-de Sitter space-time. *Phys. Rev. D* 18, 3565–3576.
- [31] Fronsdal, C. 1974, Elementary particles in a curved space. II. *Phys. Rev. D* 10, 589–598.
- [32] Dirac, P. A. M. 1935, The electron wave equation in de Sitter space. *Ann. Math.* 36, 657–669.
- [33] Lüscher, M., Mack, G. 1975, Global conformal invariance in quantum field theory. *Commun. Math. Phys.* 41, 203–234.
- [34] Breitenlohner, P., Freedman, D. Z. 1982, Stability in gauged extended supergravity. *Ann. Phys.* 144, 249–281.

This page is intentionally left blank

CHAPTER 4

Geometry and Topology in Relativistic Cosmology

JEAN-PIERRE LUMINET

*Laboratoire Univers et Théories, CNRS-UMR 8102,
Observatoire de Paris, F-92195 Meudon cedex, France
jean-pierre.luminet@dospm.fr*

Overview

General relativity does not allow one to specify the topology of space, leaving the possibility that space is multiply rather than simply connected. We review the main mathematical properties of multiply connected spaces, and the different tools to classify them and to analyse their properties. Following their mathematical classification, we describe the different possible multiconnected spaces which may be used to construct Friedmann–Lemaître universe models. Observational tests concern the distribution of images of discrete cosmic objects or more global effects, mainly those concerning the cosmic microwave background. According to the 2003–2006 WMAP data releases, various deviations from the flat infinite universe model predictions hint at a possible non-trivial topology for the shape of space. In particular, a finite universe with the topology of the Poincaré dodecahedral spherical space fits remarkably well with the data and is a good candidate for explaining both the local curvature of space and the large angle anomalies in the temperature power spectrum. Such a model of a small universe, whose volume would represent only about 80% the volume of the observable universe, offers an observational signature in the form of a predictable topological lens effect on the one hand, and raises new issues about the physics of the early universe on the other hand.

I. The Four Scales of Geometry

The forms which nature takes are limited by certain constraints. The first constraint is imposed by the three-dimensional character of space (I am referring here to the

usual three dimensions of length, width and depth, while being aware that recent theories invoke the existence of extra spatial dimensions which are only detectable on very small distance scales). Space is not a passive background, rather it has a structure which influences the shape of all existing objects. Every material form pays tribute to the rules dictated by the architecture of space.

The true architecture of space, and the constraints which it imposes, are still unknown. We can however reach a better understanding of the Universe by delving into the large range of abstract spaces arising in geometry, and by studying their local as well as their global structure. It is true that a mental image of non-Euclidean space eludes most of laymen, but geometry provides us with a consistent mathematical description.

Which mathematical space is capable of representing real physical space? The problem is much more complicated than it would appear. The microscopic and macroscopic worlds are profoundly different from the space of our immediate surroundings. The question of a geometric representation of space arises on four different levels, or, as the physicists say, four scales. These are microscopic, local, macroscopic and global.

On a local scale, that is to say for distances of between 10^{-18} metre (the distance now accessible to experimentation in particle accelerators) and 10^{11} metres (approximately the Earth–sun distance), the geometry of space is very well described by that of ordinary, three-dimensional Euclidean space E^3 . ‘Very well’ means that this mathematical structure serves as a natural framework for those physical theories, like classical mechanics and special relativity, which account satisfactorily for the quasi totality of natural phenomena.

On a macroscopic scale, that is to say for distances between 10^{11} and 10^{25} metres, the geometry of space is better described as non-Euclidean, or, more accurately, as a continuous Riemannian manifold (a three-dimensional generalisation of a surface with variable curvature). Such a space is curved to a greater or lesser extent by massive bodies (in the vicinity of exceptionally massive or dense bodies, like black holes, the effects of curvature can be felt over distances of a few metres only.) The physical framework is Einstein’s General Theory of Relativity, in which the spacetime structure is more satisfactorily explained in terms of a supple, elastic fabric, gravitational phenomena being the manifestation of the non-zero curvature of the manifold.

On infinitesimally small distance scales, that is for distances less than 10^{-18} metre, we are into the realm of unexplored microscopic space. Neither powerful electron microscopes nor high energy particle accelerators can probe its most detailed structure. Here, geometric models only exist in the form of speculative theories. This microscopic space could reveal special geometric properties. What is it really made of? Do ‘grains’ of space, analogous to the grains of energy in quantum physics, actually exist? Imaginative theorists, like Paul Dirac and John Wheeler, took this idea further by treating space like a collection of grains or soap

bubbles. In their view, space is not simply a passive coordinate system. As a magical substance, whose curvature, granularity and excitations determine the masses, charges and fields of particles, it plays an active role in creating the material world. For example, space may be perturbed by fluctuations which permanently modify its shape, and make it extremely complicated — unstable, discontinuous and chaotic. It might even possess extra hidden dimensions.

These highly speculative topics, whose study is now well underway, will be among the stakes in tomorrow's physics. I shall not be looking into them here, as we are mainly concerned with the widest perspective on the universal fabric of space. There, no lesser surprises await us. It is not yet known whether space is infinite, with zero or negative curvature; or whether it is finite, with a positive curvature, like a multidimensional sphere. Strangest of all would be a 'wraparound' space, that is one folded back on itself. Such a space could be finite while being flat or negatively curved. Treating these global aspects of space requires a new discipline, a mixture of advanced mathematics and subtle cosmological observations: Cosmic Topology.

2. Curvature vs. Topology

The origins of topology go back to a riddle posed by the idle rich Prussians of the city of Königsberg, constructed around the branches of the Pregel river. The riddle consisted of deciding if, from any point in the city, it was possible to take a stroll in a closed loop while crossing once, and once only, each of the seven bridges which span the branches of the Pregel. The riddle was solved by the famous mathematician Leonhard Euler, who, in 1736, gave the necessary conditions which would allow such a route and, since the configuration of the bridges did not satisfy these rules, he proved that it was impossible to cross all seven bridges in a single trip.

Most important, Euler pointed out that, for the first time in the history of mathematics, one was dealing with a geometrical problem which had nothing to do with the metrics. The only important factors were the relative positions of the bridges. Indeed, if we trace the map of the city on a rubber sheet, and if we stretch or squeeze it in any direction without puncturing, cutting or tearing it, the nature of the problem is absolutely unchanged.

The solution given by Euler perfectly illustrates the two complementary aspects of geometry as the science of space: the 'metric' part deals with the properties of distance, while the 'topological' part studies the global properties, without introducing any measurements. The topological properties are those which remain insensitive to deformations, provided that these are continuous: with the condition of not cutting, piercing or gluing space, one can stretch it, crush it, or knead it in any way, and one will not change its topology, for example the fact that it is finite or infinite, the fact that it has holes or not, the number of holes if it has them, etc. It is easy to

see that although continuous deformations may move the holes in a surface, they can neither create nor destroy them. Thus for a topologist, there is no difference between a rugby ball and a soccer ball. Worse, a ring and a coffee cup are one and the same object, characterised by a hole through which one can pass one's finger. On the other hand, a mug and a bowl are radically different on the level of topology, since a bowl does not have a handle.

Topology holds quite a few surprises. Let us take the Euclidean plane: it is an infinite two-dimensional page, that one visualises most often within a three-dimensional space, although it has no need for this embedding to be perfectly well defined in an intrinsic way. The local geometry of the plane is determined by its metric, that is to say by the way in which lengths are measured. Here, it is sufficient to apply the Pythagorean theorem for a system of two rectilinear coordinates covering the plane: $ds^2 = dx^2 + dy^2$. This is a local measurement which says nothing about the finite or infinite character of space. Now let us change the topology. To do so, we take the plane and cut a strip of infinite length in one direction and finite width in the other. We then glue the two sides of the strip: we obtain a cylinder, a tube of infinite length. In this operation, the metric has not changed: the Pythagorean theorem still holds for the surface of the cylinder. The 'intrinsic' curvature of the cylinder is therefore zero. This may appear surprising, since one has the impression that there is a non-zero curvature 'somewhere', whose radius would be the radius of the cylinder. However, this 'somewhere' calls into play a space exterior to the cylinder: the one in which we visualise it. In this sense, the cylinder has a so-called 'extrinsic' curvature. Nevertheless, a flat creature, some sort of geometric paramecium living on the surface, would have access neither to this exterior space of higher dimension, nor to the extrinsic curvature of the cylinder. Tied to its two-dimensional space, it could make all of the necessary verifications (for instance measuring the sum of the angles in a triangle or the ratio of the circumference of a circle to its radius), and it would detect no difference with respect to the Euclidean metric of the infinite plane. The cylinder is said to be *locally Euclidean*.

Nevertheless, the cylinder differs from the plane in many respects. Certainly, its area is infinite, just like the plane, but it possesses a finite circumference in the direction perpendicular to its symmetry axis. In other words, the cylinder is anisotropic: not all directions are equivalent; following the length of a straight line parallel to the axis, one moves off toward infinity, while if one moves in the perpendicular direction, one returns to the departure point. In the operation of constructing a cylinder from a section of the plane, some of the global properties have changed; the cylinder thus has a different topology than that of the plane, while having the same metric. Its most remarkable characteristic is the existence of an infinite number of 'straight lines' which join two arbitrary distinct points on the cylinder: those which make 0, 1, 2 . . . turns around the cylinder. Viewed in three dimensions, these straight lines are helices with constant spacing.

Let us continue with our cutting and gluing game. Take a tube of stretchable rubber, of finite length, and glue its two ends edge-to-edge. This is strictly equivalent to starting with a rectangle and gluing its opposite edges two by two. We obtain a torus, a surface having the shape of a ring or an inner tube. Here, a new difficulty arises. A real inner tube, just like the cylindrical tube, can be materialised in normal three-dimensional space; it therefore has an extrinsic curvature. However, in contrast to the tube, the inner tube also has a non-zero intrinsic curvature, which varies in different regions: sometimes positive, sometimes negative. However the toric surface obtained by identifying the opposite sides of a rectangle has an intrinsic curvature which is everywhere zero. This *flat torus*, a surface whose global properties are identical to those of a ring but whose curvature is everywhere zero, cannot be viewed within our usual three-dimensional space (it can only be embedded into E^4). Yet one can describe all of its properties without exception; its area is finite in the sense that it is impossible to move infinitely far away from one's departure point, and it is not isotropic, since two of its directions, named the principal directions, are privileged. Let us imagine a creature living on a flat torus, moving straight ahead along a principal direction; she communicates via light rays with her departure point, in such a way that she can calculate the distance travelled; at a certain moment, this distance attains a maximum, and then begins to decrease; after having made a complete circuit, the creature has returned to her point of departure. She would conclude from this that she lives in a space of finite extent. Nevertheless, by having measured the sum of the angles in a triangle in these surroundings, she has still found 180 degrees, because of which she would also deduce that she lives in a Euclidean plane. The metric (local geometry) of the flat torus is still given by the Pythagorean theorem, just like that of the plane and the cylinder.

Through simple cutting and re-gluing of parts of the plane, we have thus defined two surfaces with different topologies than the plane: the cylinder and the flat torus, which however belong to the same family, the locally Euclidean surfaces. The gluing method becomes extremely fruitful when the surfaces are more complicated. Let us take two tori and glue them to form a 'double torus'. As far as its topological properties are concerned, this new surface with two holes can be represented as an eight-sided polygon (an octagon), which can be understood intuitively by the fact that each torus was represented by a quadrilateral. But this surface is not capable of paving the Euclidean plane, for an obvious reason: if one tries to add a flat octagon to each of its edges, the eight octagons will overlap each other. One must therefore curve in the sides and narrow the angles, in other words pass to a hyperbolic space: only there does one succeed in fitting eight octagons around the central octagon, and starting from each of the new octagons one can construct eight others, *ad infinitum*. By this process one paves an infinite space: the Lobachevsky hyperbolic plane.

More generally, a two-dimensional n -torus T_n is a torus with n holes. T_n can be constructed as the connected sum of n simple tori. The n -torus is therefore

topologically equivalent to a connected sum of n squares whose opposite edges have been identified. This sum is itself topologically equivalent to a $4n$ -gon where all the vertices are identical with each other and the sides are suitably identified by pairs. Such an operation is not straightforward when $n \geq 2$. All the vertices of the polygon correspond to the same point of the surface. Since the polygon has at least eight edges, it is necessary to make the internal angles thinner in order to fit them suitably around a single vertex. This can only be achieved if the polygon is represented in the hyperbolic plane H^2 instead of the Euclidean plane E^2 : this increases the area and decreases the angles. The more angles to fit together, the thinner they have to be and the greater the surface. The n -torus ($n \geq 2$) is therefore a compact surface of negative curvature. This type of surface is most commonly seen at bakeries, in the form of pretzels. We call them ‘hyperbolic pretzels’. They all have the same local geometry, of hyperbolic type; however, they do not have the same topology, which depends on the number of holes.

When one deals with more than two dimensions, the gluing method remains the simplest way to visualise spaces. By analogy with the two-dimensional case, the three-dimensional simple torus T^3 (also referred to as the *hypertorus*) is obtained by identifying the opposite faces of a parallelepiped. The resulting volume is finite. Let us imagine a light source at our position, immersed in such a structure. Light emitted backwards crosses the face of the parallelepiped behind us and reappears on the opposite face in front of us; therefore, looking forward we can see our back. Similarly, we see in our right our left profile, or upwards the bottom of our feet. In fact, for light emitted isotropically, and for an arbitrarily large time to wait, we could observe ghost images of any object viewed arbitrarily close to any angle. The resulting visual effect would be comparable (although not identical) to what could be seen from inside a parallelepiped of which the internal faces are covered with mirrors. Thus one would have the visual impression of infinite space, although the real space is closed.

3. Basics of Topology

3.1. Simple vs. multiple connectedness

Let us now formalise a little bit more the topological notions introduced above. The strategy for characterizing the shape of a space M is to produce invariants which capture the key features of the topology and uniquely specify each equivalence class. The topological invariants can take many forms. They can be just numbers, such as the dimension of the manifold, the degree of connectedness or the Poincaré–Euler characteristic. They can also be whole mathematical structures, such as the homotopy groups. The latter are defined in an elegant way from the tightening of laces. A lace is a closed curve traced on a surface. On the infinite plane, we

can draw an arbitrary lace, however large, from an arbitrary point; this lace can always be retightened and reduced to a point without encountering any obstacles. The topologists call such a space *simply connected*. Formally, a lace at x in \mathbf{M} is any path which starts at x and ends at x . Two laces g and g' are homotopic if g can be continuously deformed into g' . The manifold \mathbf{M} is simply-connected if every lace is homotopic to a point. Obviously, the Euclidean spaces E^1, E^2, \dots, E^n , and the spheres S^2, S^3, \dots, S^n are simply-connected.

On the other hand, the circle S^1 , the cylinder $S^1 \times E^1$ and the torus $S^1 \times S^1$ do not have this property. Of course, there are laces which can be completely retightened, as in the plane; but some of them cannot: a circle which wraps around the cylinder or which is traced around the torus, for example, cannot be continuously shrunk to a point. For such spaces, the topology is said to be *multiply connected*.

The study of homotopic laces in a manifold \mathbf{M} is a way of detecting holes or handles. Moreover the equivalence classes of homotopic laces can be endowed with a group structure, essentially because laces can be added by joining them end to end. The group of laces is called the first homotopy group at x or, in the terminology originally introduced by Poincaré, the fundamental group $\pi_1(\mathbf{M}, x)$. The fundamental group is independent of the base point: it is a topological invariant of the manifold.

For surfaces, multi-connectedness means that the fundamental group is non trivial: there is at least one lace that cannot be shrunk to a point. But in higher dimensions the problem is more complex because laces, being only one-dimensional structures, are not sufficient to capture all the topological features of the manifolds. The purpose of algebraic topology, extensively developed during the twentieth century, is to generalise the concept of homotopic laces and to define higher homotopy groups. However the fundamental group (the first homotopy group) remains essential.

3.2. Fundamental domain and holonomy group

In the nineteenth century, mathematicians discovered that it is possible to represent any surface whatsoever with a polygon whose sides one identifies, two by two. The torus is topologically equivalent to a rectangle with opposite edges identified. The rectangle is called a *fundamental domain* (hereafter FD) of the torus. From a topological point of view (namely without reference to size), the FD can be chosen in different ways: a square, a rectangle, a parallelogram, even a hexagon (since the plane can be tiled by hexagons, the flat torus can be also represented by a hexagon with suitable identification of edges).

The FD distinctly characterises a certain aspect of the topology. But this is not enough; we must also specify the geometric transformations which identify the points. Indeed, starting from a square, one could identify the points diametrically opposite with respect to the centre of symmetry of the square, and the

surface obtained will no longer be a flat torus; it will no longer even be Euclidean, but spherical, a surface called the projective plane. The mathematical transformations used to identify points form a group of symmetries, called the *holonomy group*.

This group is discrete, i.e., there is a non zero shortest distance between any two homologous points, and the generators of the group (except the identity) have no fixed point. This last property is very restrictive (it excludes for instance the rotations) and allows the classification of all possible holonomy groups. Due to the fact that the holonomy group is discrete, the FD is always convex and has a finite number of faces. In two dimensions, it is a surface whose boundary is constituted by lines, thus a polygon. In three dimensions, it is a volume bounded by faces, thus a polyhedron.

3.3. Universal covering

Starting from the fundamental domain and acting with the transformations of the holonomy group on each point, one creates a number of replicas of the FD; we produce a sort of tiling of a larger space, called the *universal covering space* (hereafter UC) M^* . By construction, M^* is locally indistinguishable from M . But its topological properties can be quite different. The UC is necessarily simply connected: any lace can be shrunk to a point. Thus, when M is simply-connected, it is identical to its universal covering space M^* . But when M is multiply connected, each point of M generates replicas of points in M^* . The universal covering space can be thought of as an unwrapping of the original manifold. For instance, the UC of the flat torus is the Euclidean plane E^2 , which indeed reflects the fact that the flat torus is a locally Euclidean surface.

3.4. Spaceforms

To summarise: the shape of a homogeneous space is entirely specified if one is given a fundamental domain; a particular group of symmetries, the holonomies, which identify the edges of the domain two by two; and a universal covering space that is paved by fundamental domains. Classifying the possible shapes thus reduces, in part, to classifying symmetries.

Let us apply this recipe in order to list all homogeneous surfaces: two-dimensional spaces with no boundaries and no sharp points. As far as the curvature is concerned, homogenous surfaces are of three types: spherical surfaces, with positive curvature (like the surface of a rugby ball); Euclidean surfaces, with zero curvature (whose planar geometry is taught in high school); and hyperbolic surfaces, of negative curvature (like certain parts of a saddle or of a trumpet's horn). Within each of these basic types, mathematicians have classified all possible topologies — also referred to as *spaceforms*.

There are only two forms for spherical surfaces, both of them finite: the sphere, which can be given a wide variety of different metrical aspects depending on what sort of continuous stretching it is subjected to, and the projective plane. The sphere is simply connected, the projective plane is not. The latter surface is not easily visualised; the simplest way to do so is to pass through the intermediary of its fundamental domain, a disk, whose diametrically opposite points are identified.

Euclidean surfaces can come in five possible shapes: the plane, of course, which is the simply connected prototype, but also the cylinder, the Möbius band (which is an infinitely wide Möbius strip), the flat torus, and the Klein bottle, all of which are multiply connected. The first three are infinite, the other two finite. These surfaces, although conceptually simple, are not all easy to visualise; thus, although the Klein bottle has no curvature, it is closed in on itself and has neither inside nor outside; it is said to be ‘non-orientable’.

Finally, the hyperbolic surfaces, with negative curvature, have an infinite number of topologies. Only one of them, equivalent to the Lobachevsky plane, is simply connected. All others are multiply connected, characterised by the number of holes. We have seen, for example, that the surface of a generalised pretzel is hyperbolic.

One conclusion that we can quickly draw from this classification is that, in the infinite set of homogeneous surfaces, they are all hyperbolic, up to only seven exceptions.

4. Three-Dimensional Manifolds of Constant Curvature

The passage from two dimensions to three dimensions in no way reduces to a simple generalisation, but leads to the appearance of radically new properties. Every regular surface can be homogenised so as to be described by a metric of constant curvature; this means that there are only three prototypical simply connected surfaces (which serve as universal coverings), to which all other surfaces are necessarily related. Things are not the same for three-dimensional spaces: there are eight possible universal covering spaces (see Thurston, 1997, for a synthesis). Only three of these are homogeneous and isotropic, the remaining five are homogeneous but not isotropic, meaning that at a given point the measurement of the curvature depends on direction.

Three-dimensional cylinders are some relatively simple examples of these. In the same way that the usual cylinder can be considered as the ‘product’ $S^1 \times E^1$ of a circle S^1 and a straight line E^1 (in the sense that if one slides a circle along a straight line perpendicular to its center one creates a cylinder), the ‘three-dimensional spherical cylinder’ can be pictured as the product $S^2 \times E^1$ of a sphere S^2 and a straight line E^1 . However, while the cylindrical surface could be described with the metric of the Euclidean plane E^2 , the cylindrical-spherical

space is fundamentally distinct from the Euclidean space E^3 . The curvatures measured are different depending on the orientations of the referential planes used to cut it. Similarly, the ‘cylindrical-hyperbolic’ space $H^2 \times E^1$, obtained by stacking Lobachevsky planes, is fundamentally distinct from E^3 .

Cosmology, however, focuses mainly on locally homogeneous and isotropic spaces, namely those admitting one of the three geometries of constant curvature. Any compact 3-manifold M with constant curvature k can thus be expressed as the quotient $M = M^*/\Gamma$, where the universal covering space M^* is either:

- the Euclidean space E^3 if $k = 0$
- the 3-sphere S^3 if $k > 0$
- the hyperbolic 3-space H^3 if $k < 0$

and Γ is a subgroup of isometries of M^* acting freely and discontinuously.

4.1. Euclidean space forms

Simply-connected Euclidean space, E^3 , with uniformly zero curvature, is infinite in every direction. Its full isometry group is $G = ISO(3) = E^3 \times SO(3)$, and the generators of the possible holonomy groups Γ (i.e., discrete subgroups without fixed point) include the identity, the translations, the glide reflections and the screw motions (combinations of a rotation and a translation parallel to the axis of rotation) occurring in various combinations. The multiply connected Euclidean spaces are characterised by their fundamental polyhedra and their holonomy groups. The fundamental polyhedra are either a finite or infinite parallelepiped, or a prism with a hexagonal base, corresponding to the two ways of tiling Euclidean space. The various different combinations generate 17 distinct multiply connected Euclidean spaces (for an exhaustive study, see Riazuelo *et al.*, 2004a).

Seven of these spaces are open (of infinite volume). Two of these, called *slab spaces*, are made of a slab that extends infinitely in two directions, but has finite thickness. The two ends are identified by a translation or after a rotation of 180° . The five others, called *chimney spaces*, are made of a rectangular chimney of infinite height, whose front and back (and left and right) surfaces are identified by a translation and appropriate rotations.

Ten other Euclidean spaces are closed (of finite volume). The first six spaces are orientable hypertori. The simplest hypertorus T^3 is constructed by identifying the opposite faces of a parallelepiped by translations. The other hypertori are obtained after gluing with a quarter turn, a half-turn, a one-sixth turn and a one-third turn, while the Hantzsche–Wendt space has a more complicated structure. It is these six compact, orientable Euclidean spaces that present a particular interest for cosmology, since they could perfectly model the spatial part of the so-called ‘flat’ universe models.

Eventually, four closed Euclidean spaces are non-orientable generalisations of the Klein bottle: Klein space, Klein space with a horizontal flip, Klein space with a vertical flip and Klein space with a half-turn.

4.2. Spherical space forms

The simply-connected spherical space S^3 , with positive curvature, is the hypersphere. Einstein attempted to give an intuitive image of such a finite yet limitless three-dimensional space, that a little bit of exercise suffices to render familiar to our thinking. A way to visualise the hypersphere consists in imagining the points of the hypersphere as those of a family of two-dimensional spheres which grow in radius from 0 to a maximal value R , then shrink from R back to 0 (in the same way that a sphere can be cut into planar slices which are circles of varying radius). Another possibility is to view the hypersphere as composed of two spherical balls embedded in Euclidean space, glued along their boundaries in such a way that each point on the boundary of one ball is the same as the corresponding point on the other ball.

The full isometry group of S^3 is $SO(4)$. The holonomies that preserve the metric of the hypersphere, i.e., the admissible subgroups G of $SO(4)$ without fixed point, acting freely and discontinuously on S^3 , belong to three categories:

1. the cyclic groups of order p , Z_p ($p \geq 2$), made up of rotations by an angle $2\pi/p$ around a given axis, where p is an arbitrary integer;
2. the dihedral groups of order $2m$, D_m ($m > 2$), which are the symmetry groups of a regular plane polygons of m sides;
3. the binary polyhedral groups, which preserve the shapes of the regular polyhedra.

The group T^* preserves the tetrahedron (4 vertices, 6 edges, 4 faces), of order 24; the group O^* preserves the octahedron (6 vertices, 12 edges, 8 faces), of order 48; the group I^* preserves the icosahedron (12 vertices, 30 edges, 20 faces), of order 120. There are only three distinct polyhedral groups for the five polyhedra, because the cube and the octahedron on the one hand, the icosahedron and the dodecahedron on the other hand are duals, so that their symmetry groups are the same.

If one identifies the points of the hypersphere by holonomies belonging to one of these groups, the resulting space is spherical and multiply connected. For an exhaustive classification, see Gausmann *et al.* (2001). There is a countable infinity of these, because of the integers p and m which parametrise the cyclic and dihedral groups.

The spaces with cyclic group are called *lens spaces*, denoted thus because their fundamental polyhedra have the shapes of lenses. For instance, the projective (also

called elliptic) space $\mathbf{P}^3 = \mathbf{S}^3/\mathbf{Z}_2$ is obtained by identifying diametrically opposite points on \mathbf{S}^3 . It was used by de Sitter (1917) and Lemaître (1931) as the space structure of their cosmological models, while Einstein (1917) selected the simply connected hypersphere.

The spaces with dihedral group are called *prism spaces*, because of the shape of their fundamental polyhedra. Finally, the spaces with polyhedral groups are called *polyhedral spaces*. Among them, the *Poincaré Dodecahedral Space* \mathbf{S}^3/Γ^* is obtained by identifying the opposite pentagonal faces of a regular spherical dodecahedron after rotating by $1/10^{\text{th}}$ turn in the clockwise direction around the axis orthogonal to the face. This configuration involves 120 successive operations and gives some idea of the extreme complication of such multiply connected topologies. Its volume is 120 times smaller than that of the hypersphere with the same radius of curvature, and it is of particular interest for cosmology, giving rise to fascinating topological mirages (see below).

Since the universal covering \mathbf{S}^3 is compact, all the multiply connected spherical spaces are also compact. As the volume of \mathbf{S}^3 is $2\pi^2\mathbf{R}^3$, the volume of $\mathbf{M} = \mathbf{S}^3/\Gamma$ is simply $\text{vol}(\mathbf{M}) = 2\pi^2\mathbf{R}^3/|\Gamma|$ where $|\Gamma|$ is the order of the group Γ . For topologically complicated spherical 3-manifolds, $|\Gamma|$ becomes large and $\text{vol}(\mathbf{M})$ is small. There is no lower bound since Γ can have an arbitrarily large number of elements (for lens and prism spaces, the larger p and m are, the smaller the volume of the corresponding spaces). Hence $0 < \text{vol}(\mathbf{M}) \leq 2\pi^2\mathbf{R}^3$. In contrast, the diameter, i.e., the maximum distance between two points in the space, is bounded below by $\sim 0.326\mathbf{R}$, corresponding to the dodecahedral space.

4.3. Hyperbolic space forms

Locally hyperbolic manifolds are less well understood than the other homogeneous spaces. However, according to the pioneering work of Thurston, almost all 3-manifolds can be endowed with a hyperbolic structure. The universal covering space, \mathbf{H}^3 , is the three-dimensional analog of the Lobachevsky plane \mathbf{H}^2 , and extends to infinity in every direction. Its group of isometries is isomorphic to $\text{PSL}(2, \mathbf{C})$, namely the group of fractional linear transformations acting on the complex plane. Finite subgroups are discussed in Beardon (1983). The mathematicians have not succeeded in classifying all of them, but they know an infinite number of examples. Some of these spaces are closed (with finite volume), and others are open (with infinite volume).

In hyperbolic geometry there is an essential difference between the two-dimensional case and higher dimensions. A surface of genus $g \geq 2$ supports uncountably many non equivalent hyperbolic metrics. But the so-called *rigidity theorem* proves that a connected oriented n -dimensional manifold supports *at most one* hyperbolic metric as soon as $n \geq 3$. In simple terms, this means that if one fixes a hyperbolic topology, there is only a single metric compatible with this topology.

From this, it follows that the volume of space (in units of the curvature radius R) is fixed by its topology. It is thus possible to classify the closed hyperbolic spaces by increasing volumes, which could have seemed, at a first glance, contradictory with the very purpose of topology.

However the volumes cannot be made arbitrarily small by gluing operations. The absolute lower bound is $V_{min} = 0.16668$, but no space has been constructed having precisely this volume. Until now, the smallest known hyperbolic space (that is to say one whose fundamental polyhedron and holonomy group were able to be completely calculated) is *Weeks space*, with a volume equal to 0.94272. Its FD is a polyhedron with 26 vertices and 18 faces, of which 12 are pentagons and 6 are quadrilaterals. Its outer structure, the Klein coordinates of the vertices and the 18 matrix representations of the generators of the holonomy group are given in Lehoucq *et al.* (1999).

In cosmology, the Weeks manifold leaves room for many topological lens effects, since the volume of the observable universe is about 200 times larger than the volume of Weeks space for $\Omega_0 = 0.3$. Indeed, any compact hyperbolic space have geodesics shorter than the curvature radius, leaving room to fit a great many copies of a fundamental polyhedron within the horizon radius, even for manifolds of volume ~ 10 . The publicly available program SnapPea, available on the internet (Weeks) classifies all known spaces by increasing volumes, and gives their properties: the structure of the fundamental polyhedron, the nature of the transformations in the holonomy group, the characteristic topological lengths, etc. Several millions of compact hyperbolic spaces with volume less than ten could be calculated.

5. Topology and Cosmology

General relativity has successfully passed a number of experimental tests, but, like any physical theory, it is incomplete. One of the limits on its validity is well known: it does not take into account the microscopic properties of matter, described by quantum physics. Einstein was well aware of this, since, after putting the finishing touches on his gravitational theory in 1916, he passed the rest of his days attempting to unify gravity with the other physical interactions, in vain. Present day attempts at unification, whether ‘superstrings’, ‘M-theory’ or ‘quantum loop gravity’, tend to run into the same difficulties (see e.g., Smolin, 2002). What is less known is that general relativity is also incomplete on the large scale: is space finite or infinite, oriented or not? What is its global shape? Gravitation does not by itself decide the overall form taken by space. The preceding examples have indeed shown that the curvature of space does not necessarily allow one to come to any conclusions about its finite or infinite character.

These basic cosmological questions come from the global topology of the Universe, about which general relativity is silent. Einstein’s theory in fact only

allows one to deal with the local geometric properties of the Universe. Its partial differential equations have as a solution a metric tensor g_{ab} , or, equivalently, the infinitesimal element of distance ds^2 separating two events in space-time. This leads the study of the Universe, of its content, and of its physical properties to problems of differential geometry on a pseudo-Riemannian manifold.

It is presently believed that our Universe is correctly described at large scale by a Friedmann–Lemaître (hereafter FL) model. The FL models are homogeneous and isotropic solutions of Einstein’s equations, of which the spatial sections have constant curvature; they include the de Sitter solution, as well as those incorporating a cosmological constant, or a non standard equation of state. The FL models fall into three general classes, according to the sign of their spatial curvature $k = -1, 0,$ or $+1$. The spacetime manifold is described by the metric $ds^2 = c^2 dt^2 - R^2(t) d\sigma^2$, where $d\sigma^2 = d\chi^2 + S_k^2(\chi)(d\theta^2 + \sin^2\theta d\phi^2)$ is the metric of a three-dimensional homogeneous manifold, flat [$k = 0$] or with curvature [$k \pm 1$]. The function $S_k(\chi)$ is defined as $\sinh(\chi)$ if $k = -1$, χ if $k = 0$, $\sin(\chi)$ if $k = 1$; $R(t)$ is the scale factor, chosen equal to the spatial curvature radius for non flat models.

The spatial topology is usually assumed to be the same as that of the corresponding simply connected, universal covering space: the hypersphere, Euclidean space or the three-dimensional hyperboloid, the first being finite and the other two infinite. However, there is no particular reason for space to have a simply connected topology. In any case, general relativity says nothing on this subject; it is only the strict application of the cosmological principle, added to the theory, which encourages a generalisation of locally observed properties to the totality of the Universe. Likewise, an ant in the middle of the desert would be convinced that the entire world is made of grains of sand. However, to the metric element given above there are several, if not an infinite number, of possible topologies, and thus of possible models for the physical Universe. For example, the hypertorus and familiar Euclidean space are locally identical, and relativistic cosmological models describe them with the same FL equations, even though the former is finite and the latter infinite; likewise, the equations for a Universe of negative curvature make no distinction between a finite or an infinite space. In fact, only the boundary conditions on the spatial coordinates are changed. Thus the multi-connected cosmological models share exactly the same kinematics and dynamics as the corresponding simply connected ones (for instance, the time evolution of the scale factor $R(t)$ is identical).

At this stage, it is wise to recall that cosmological models do not reduce to three dimensions, but are four-dimensional space-times. Thus, to the problem of the topology of space is added that of the topology of time. What can be said about the space-time topology? An infinite spectrum of possibilities offer themselves as models. Nevertheless, some brief consideration of the physical properties of the Universe allows us to rapidly isolate a good number of inadmissible topologies. Here is why. Models of the big bang are homogeneous, meaning that their

spatial part has a curvature which is everywhere uniform, and expanding. These two properties allow one to unambiguously distinguish slices of simultaneous space and the axis of cosmic time. We can therefore describe space-time as the mathematical product of a three-dimensional Riemannian space and the time axis. This foliation considerably simplifies things. Time is represented by a one-dimensional space whose points represent instants: a single number suffices to determine a particular time. Time possesses an ordered structure: on a line, one point is necessarily situated either before or after another point. The topology of time is in the end rather poor; in contrast to that of multidimensional space, it only offers two cases: the line E^1 and the circle S^1 . These two forms in fact correspond to two great philosophical conceptions, linear time and cyclic time. The latter has long prevailed in myths, such as that of the eternal return, but today it has been abandoned by physics because it violates the principle of causality, according to which cause must precede effect. As a consequence, any identification of points along the time axis is forbidden. In the framework of cosmological models of expansion followed by contraction, one could, certainly, think to identify the big bang and the big crunch, that is to say the beginning of time with its end; but this operation is unlawful, for these points are singularities which are not even part of the Universe.

The question of cosmic topology therefore reduces principally to that of the spatial component of the Universe. For each type of possible curvature, as we have seen, there are various FL models with multiply connected topologies. In relativistic cosmology, the curvature of physical space depends on the way the total energy density of the Universe may counterbalance the kinetic energy of the expanding space. The normalised density parameter Ω_0 , defined as the ratio of the actual density to the critical value that an Euclidean space would require, characterises the present-day contents (matter, radiation and all forms of energy) of the Universe. If Ω_0 is greater than 1, then space curvature is positive and geometry is spherical; if Ω_0 is smaller than 1 the curvature is negative and geometry is hyperbolic; eventually Ω_0 is strictly equal to 1 and space is Euclidean.

The next question about the shape of the Universe is to know whether space is finite or infinite — equivalent to know whether space contains a finite or an infinite amount of matter–energy, since the usual assumption of homogeneity implies a uniform distribution of matter and energy through space. From a purely geometrical point of view, all positively curved spaces are finite whatever their topology, but the converse is not true: flat or negatively curved spaces can have finite or infinite volumes, depending on their degree of connectedness (Ellis, 1971; Lachièze-Rey & Luminet, 1995).

From an astronomical point of view, it is necessary to distinguish between the ‘observable universe’, which is the interior of a sphere centered on the observer and whose radius is that of the cosmological horizon (roughly the radius of the last scattering surface), and the physical space. There are only three logical possibilities. First, the physical space is infinite — like for instance the simply-connected

Euclidean space. In this case, the observable universe is an infinitesimal patch of the full universe and, although it has long been the preferred model of many cosmologists, this is not a testable hypothesis. Second, physical space is finite (e.g. an hypersphere or a closed multiconnected space), but greater than the observable space. In that case, one easily figures out that if physical space is much greater than the observable one, no signature of its finitude will show in the observable data. But if space is not too large, or if space is not globally homogeneous (as is permitted in many space models with multiconnected topology) and if the observer occupies a special position, some imprints of the space finitude could be observable. Third, physical space is smaller than the observable universe. Such an apparently odd possibility is due to the fact that space can be multiconnected and have a small volume. There is a lot of possibilities, whatever the curvature of space. Small universe models may generate multiple images of light sources, in such a way that the hypothesis can be tested by astronomical observations. The smaller the fundamental domain, the easier it is to observe the multiple topological imaging. Lehoucq *et al.* (1998) have calculated, for a given catalog of observable cosmic sources (discrete or diffuse) with a given depth in redshift, the approximate number of topological images in locally hyperbolic and locally spherical spaces as a function of the cosmological parameters Ω_m and Ω_Λ . How do the present observational data constrain the possible multi-connectedness of the universe and, more generally, what kinds of tests are conceivable? The following sections deal with these matters (see Luminet, 2001, for a non-technical book about all the aspects of topology and its applications to cosmology).

6. The Drumhead Universe

The topology and the curvature of space can be studied by using specific astronomical observations. For instance, from Einstein's field equations, the space curvature can be deduced from the experimental values of the total energy density and of the expansion rate. If the Universe was finite and small enough, we should be able to see 'all around' it, because the photons might have crossed it once or more times. In such a case, any observer might identify multiple images of a same light source, although distributed in different directions of the sky and at various redshifts, or to detect specific statistical properties in the apparent distribution of faraway sources such as galaxy clusters. To do this, methods of 'cosmic crystallography' have been devised (Lehoucq *et al.*, 1996; Uzan *et al.*, 1999), and extensively studied by the Brazilian school of cosmic topology (see e.g. Gomero *et al.*, 2002). The main limitation of cosmic crystallography is that the presently available catalogs of observed sources at high redshift are not complete enough to perform convincing tests.

Fortunately, the topology of a small Universe may also be detected through its effects on such a Rosetta stone of cosmology as is the cosmic microwave background (hereafter CMB) fossil radiation (Levin, 2002). If you sprinkle fine sand

uniformly over a drumhead and then make it vibrate, the grains of sand will collect in characteristic spots and figures, called Chladni patterns. These patterns reveal much information about the size and the shape of the drum and the elasticity of its membrane. In particular, the distribution of spots depends not only on the way the drum vibrated initially but also on the global shape of the drum, because the waves will be reflected differently according to whether the edge of the drumhead is a circle, an ellipse, a square, or some other shape. In cosmology, the early Universe was crossed by real acoustic waves generated soon after the big bang. Such vibrations left their imprints 380 000 years later as tiny density fluctuations in the primordial plasma. Hot and cold spots in the present-day 2.7 K CMB radiation reveal those density fluctuations. Thus the CMB temperature fluctuations look like Chladni patterns resulting from a complicated three-dimensional drumhead that vibrated for 380 000 years. They yield a wealth of information about the physical conditions that prevailed in the early Universe, as well as present geometrical properties like space curvature and topology. More precisely, density fluctuations may be expressed as combinations of the vibrational modes of space, just as the vibration of a drumhead may be expressed as a combination of the drumhead's harmonics. The shape of space can be heard in a unique way. Lehoucq *et al.* (2002) calculated the harmonics (the so-called 'eigenmodes of the Laplace operator') for most of the spherical topologies, and Riazuelo *et al.* (2004a) did the same for all 18 Euclidean spaces. Then, starting from a set of initial conditions fixing how the universe originally vibrated (the so-called Harrison–Zeldovich spectrum), it is possible to evolve the harmonics forward in time to simulate realistic CMB maps for a number of flat and spherical topologies (Uzan *et al.*, 2004).

The 'concordance model' of cosmology describes the Universe as a flat infinite space in eternal expansion, accelerated under the effect of a repulsive dark energy. The data collected by the NASA satellite WMAP (Spergel *et al.*, 2003) have produced a high resolution map of the CMB which showed the seeds of galaxies and galaxy clusters and allowed to check the validity of the dynamic part of the expansion model. However, combined with other astronomical data (Tonry *et al.*, 2003), they suggest a value of the density parameter $\Omega_0 = 1.02 \pm 0.02$ at the 1σ level. The result is marginally compatible with strictly flat space sections. Improved measurements could indeed lower the value of Ω_0 closer to the critical value 1, or even below to the hyperbolic case. Presently however, taken at their face value, WMAP data favour a positively curved space, necessarily of finite volume since all spherical spaceforms possess this property.

Now what about space topology? There is an intriguing feature in WMAP data, already present in previous COBE measurements (Hinshaw *et al.*, 1996), although at a level of precision that was not significant enough to draw firm conclusions. The power spectrum depicts the minute temperature differences on the last scattering surface, depending on the angle of view. It exhibits a set of peaks when anisotropy is measured on small and mean scales (i.e., concerning regions of the sky of relatively modest size). These peaks are remarkably consistent with the infinite flat space

hypothesis. At large angular scales, the concordance model predicts that the power spectrum should follow the so-called ‘Sachs–Wolfe plateau’. However, WMAP measurements fall well below the plateau for the quadrupole and the octopole moments (i.e., for CMB spots typically separated by more than 60°). Since the flat infinite space model cannot explain this feature, it is necessary to look for an alternative.

CMB temperature anisotropies essentially result from density fluctuations of the primordial Universe: a photon coming from a denser region will lose a fraction of its energy to compete against gravity, and will reach us cooler. On the contrary, photons emitted from less dense regions will be received hotter. The density fluctuations result from the superposition of acoustic waves which propagated in the primordial plasma. Riazuelo *et al.* (2004a) have developed complex theoretical models to reproduce the amplitude of such fluctuations, which can be considered as vibrations of the Universe itself. In particular, they simulated high resolution CMB maps for various space topologies (Riazuelo *et al.*, 2004b) and were able to compare their results with real WMAP data. Depending on the underlying topology, the distribution of the fluctuations differs. For instance, in an infinite flat space, all wavelengths are allowed, and fluctuations must be present at all scales.

The CMB temperature fluctuations can be decomposed into a sum of *spherical harmonics*, much like the sound produced by a music instrument may be decomposed into ordinary harmonics. The ‘fundamental’ fixes the height of the note (as for instance a 440 hertz acoustic frequency fixes the *A* of the pitch), whereas the relative amplitudes of each harmonics determine the tone quality (such as the *A* played by a piano differs from the *A* played by a harpsichord). Concerning the relic radiation, the relative amplitudes of each spherical harmonics determine the power spectrum, which is a signature of the space geometry and of the physical conditions which prevailed at the time of CMB emission.

The first observable harmonics is the quadrupole (whose wavenumber is $l = 2$). WMAP has observed a value of the quadrupole seven times weaker than expected in a flat infinite Universe. The probability that such a discrepancy occurs by chance has been estimated to 0.2% only. The octopole (whose wavenumber is $l = 3$) is also weaker (72% of the expected value). For larger wavenumbers up to $l = 900$ (which correspond to temperature fluctuations at small angular scales), observations are remarkably consistent with the standard cosmological model.

The unusually low quadrupole value means that long wavelengths are missing. Some cosmologists have proposed to explain the anomaly by still unknown physical laws of the early universe (Tsujikawa *et al.*, 2003). A more natural explanation may be because space is not big enough to sustain long wavelengths. Such a situation may be compared to a vibrating string fixed at its two extremities, for which the maximum wavelength of an oscillation is twice the string length. On the contrary, in an infinite flat space, all the wavelengths are allowed, and fluctuations must be

present at all scales. Thus this geometrical explanation relies on a model of finite space whose size *smaller* than the observable universe constrains the observable wavelengths below a maximum value.

Such a property has been known for a long time, and was used to constrain the topology from COBE observations (Sokolov, 1993). Preliminary oversimplified analyses (de Oliveira-Costa & Smoot, 1995) suggested that any multi-connected topology in which space was finite in at least one space direction had the effect of lowering the power spectrum at large wavelengths. Weeks *et al.* (2004) reexamined the question and showed that indeed, some finite multiconnected topologies do lower the large-scale fluctuations whereas others may elevate them. In fact, the long wavelengths modes tend to be relatively lowered only in a special family of closed multiconnected spaces called ‘well-proportioned’. Generally, among spaces whose characteristic lengths are comparable with the radius of the last scattering surface R_{lss} (a necessary condition for the topology to have an observable influence on the power spectrum), spaces with all dimensions of similar magnitude lower the quadrupole more heavily than the rest of the power spectrum. As soon as one of the characteristic lengths becomes significantly smaller or greater than the other two, the quadrupole is boosted in a way not compatible with WMAP data. In the case of flat tori, a cubic torus lowers the quadrupole whereas an oblate or a prolate torus increase the quadrupole; for spherical spaces, polyhedral spaces suppress the quadrupole whereas high order lens spaces (strongly anisotropic) boost the quadrupole. Thus, well-proportioned spaces match the WMAP data much better than the infinite flat space model.

7. The Dodecahedral Universe

Among the family of well-proportioned spaces, the best fit to the observed power spectrum is the *Poincaré Dodecahedral Space*, hereafter PDS (Luminet *et al.*, 2003). Recall that this space is positively curved, and is a multiconnected variant of the simply-connected hypersphere S^3 , with a volume 120 times smaller for the same curvature radius. The associated power spectrum, namely the repartition of fluctuations as a function of their wavelengths corresponding to PDS, strongly depends on the value of the mass-energy density parameter. Luminet *et al.* (2003) computed the CMB multipoles for $l = 2, 3, 4$ and fitted the overall normalisation factor to match the WMAP data at $l = 4$, and then examined their prediction for the quadrupole and the octopole as a function of Ω_0 . There is a small interval of values within which the spectral fit is excellent, and in agreement with the value of the total density parameter deduced from WMAP data (1.02 ± 0.02). The best fit is obtained for $\Omega_0 = 1.016$. Since then, the properties of PDS have been investigated in more details by various authors. Lachièze-Rey (2004) found an analytical expression of the eigenmodes of PDS, whereas Aurich *et al.* (2005) and Gundermann (2005)

computed numerically the power spectrum up to the $l = 15$ mode and improved the fit with WMAP data. The result is quite remarkable because the Poincaré space has no degree of freedom. By contrast, a three-dimensional torus, constructed by gluing together the opposite faces of a cube and which constitutes a possible topology for a finite Euclidean space, may be deformed into any parallelepiped: therefore its geometrical construction depends on 6 degrees of freedom.

The values of the matter density Ω_m , of the dark energy density Ω_Λ and of the expansion rate H_0 fix the radius of the last scattering surface R_{lss} as well as the curvature radius of space R_c , thus dictate the possibility to detect the topology or not. For $\Omega_m = 0.28$, $\Omega_0 = 1.016$ and $H_0 = 62 \text{ km/s/Mpc}$, $R_{lss} = 53 \text{ Gpc}$ and $R_c = 2.63 R_{lss}$. It is to be noticed that the curvature radius R_c is the same for the simply-connected universal covering space S^3 and for the multiconnected PDS. Incidentally, the numbers above show that, contrary to a current opinion, a cosmological model with $\Omega_0 \sim 1.02$ is far from being ‘flat’ (i.e., with $R_c = \infty$)! For the same curvature radius than the simply-connected hypersphere S^3 , the smallest dimension of the fundamental dodecahedron is only 43 Gpc, and its volume about 80% the volume of the observable universe (namely the volume of the last scattering surface). This implies that some points of the last scattering surface will have several copies. Such a lens effect is purely attributable to topology and can be precisely calculated in the framework of the PDS model. It provides a definite signature of PDS topology, whereas the shape of the power spectrum gives only a hint for a small, well-proportioned universe model.

To be confirmed, the PDS model (sometimes popularised as the ‘soccerball universe model’) must satisfy two experimental tests:

- (1) New data from the future European satellite ‘Planck Surveyor’ (scheduled 2007) could be able to determine the value of the energy density parameter with a precision of 1%. A value lower than 1.009 would discard the Poincaré space as a model for cosmic space, in the sense that the size of the corresponding dodecahedron would become greater than the observable universe and would not leave any observable imprint on the CMB, whereas a value greater than 1.01 would strengthen its cosmological pertinence.
- (2) If space has a non trivial topology, there must be particular correlations in the CMB, namely pairs of ‘matched circles’ along which temperature fluctuations should be the same (Cornish *et al.*, 1998). The PDS model predicts 6 pairs of antipodal circles with an angular radius comprised between 5° and 55° (sensitively depending on the cosmological parameters).

Such circles have been searched in WMAP data by several teams, using various statistical indicators and massive computer calculations. First, Cornish *et al.* (2004) claimed to have found no matched circles on angular sizes greater than 25° , and thus rejected the PDS hypothesis. Next, Roukema *et al.* (2004) performed the same

analysis for smaller circles, and found six pairs of matched circles distributed in a dodecahedral pattern, each circle on an angular size about 11° . This implies $\Omega_0 = 1.010 \pm 0.001$ for $\Omega_m = 0.28 \pm 0.02$, values which are perfectly consistent with the PDS model. Finally, Aurich *et al.* (2006a) performed a very careful search for matched circles and found that the putative topological signal in the WMAP data was considerably degraded by various effects, so that the dodecahedral space model could be neither confirmed nor rejected. This shows in passing how delicate the statistical analysis of observational data is, since different analyses of the same data can lead to radically opposed conclusions!

The controversy still went up a tone when Key *et al.* (2006) claimed that their negative analysis was not disputable, and that accordingly, not only the dodecahedral hypothesis was excluded, but also any multiply-connected topology on a scale smaller than the horizon radius. Since such an argument of authority, a fair portion of the academic community believes the WMAP data has ruled out multiply-connected models. However, at least the second part of the claim is wrong. The reason is that they searched only for antipodal or nearly-antipodal matched circles. But Riazuelo *et al.* (2004b) have shown that for generic multiply-connected topologies (including the well-proportioned ones, which are good candidates for explaining the WMAP power spectrum), the matched circles are generally not antipodal; moreover, the positions of the matched circles in the sky depend on the observer's position in the fundamental polyhedron. The corresponding larger number of degrees of freedom for the circles search in the WMAP data generates a dramatic increase of the computer time, up to values which are out-of-reach of the present facilities. It follows that the debate about the pertinence of PDS as the best fit to reproduce CMB observations is fully open.

The new release of WMAP data (Spergel *et al.*, 2006), integrating two additional years of observation with reduced uncertainty, strengthened the evidence for an abnormally low quadrupole and other features which do not match with the infinite flat space model (this explains the unexpected delay in the delivery of this second release, originally announced for February 2004). Besides the quadrupole suppression, an anomalous alignment between the quadrupole and the octopole was put in evidence along a so-called 'axis of evil' (Land and Magueijo, 2005). Thus the question arose to know whether, since non-trivial spatial topology can explain the weakness of the low- l modes, might it also explain the quadrupole-octopole alignment? Until then no multiply-connected space model, either flat (Cresswell *et al.*, 2006) or spherical (Aurich *et al.*, 2006b; Weeks and Gundermann, 2006) was proved to exhibit the alignment observed in the CMB sky. This is not a strong argument against such models, since the 'axis of evil' is generally interpreted as due to local effects and foreground contaminations (Prunet *et al.*, 2005).

As a provisional conclusion, since some power spectrum anomalies are one of the possible signatures of a finite and multiply-connected universe, there is still a continued interest in the Poincaré dodecahedral space and related finite universe

models. And even if the particular dodecahedral space is eventually ruled out by future experiments, all of the other models of well-proportioned spaces will not be eliminated as such. In addition, numerical simulations show that, even if the size of a multiply-connected space is larger than that of the observable universe, we could all the same discover an imprint in the fossil radiation, even while no pair of circles, much less ghost galaxy images, would remain. The topology of the universe could therefore provide information on what happens outside of the cosmological horizon! But this is a search for the next decade...

Maybe the most fundamental issue remains to link the present-day topology of space to a quantum origin, since classical general relativity does not allow for topological changes during the course of cosmic evolution. Theories of quantum gravity could allow to address the problem of a quantum origin of space topology. For instance, in the approach of quantum cosmology, some simplified solutions of Wheeler–de Witt equations show that the sum over all topologies involved in the calculation of the wavefunction of the universe is dominated by spaces with small volumes and multiconnected topologies (Carlip, 1993; e Costa and Fagundes, 2001). In the approach of brane worlds (see Brax 2003 for a review), the extra-dimensions are often assumed to form a compact Calabi–Yau manifold; in such a case, it would be strange that only the ordinary dimensions of our 3-brane would not be compact like the extra ones. These are only heuristic indications on the way unified theories of gravity and quantum mechanics could ‘favour’ multiconnected spaces. Whatsoever the fact that some particular multiconnected space models, such as PDS, may be refuted by future astronomical data, the question of cosmic topology will stay as a major question about the ultimate structure of our universe.

References

- [1] Aurich, R., Lustig, S. and Steiner, F. 2005, *Class. Quant. Grav.* 22, 2061.
- [2] Aurich, R., Lustig, S. and Steiner, F. 2006a, *Mon. Not. Roy. Astron. Soc.* 369, 240.
- [3] Aurich, R., Lustig, S., Steiner, F. *et al.*, 2006b, *Class. Quant. Grav.* 24, 1879.
- [4] Beardon, A. 1983, *The Geometry of Discrete Groups*. New York: Springer.
- [5] Brax, P., van de Bruck, C. 2003, *Class. Quant. Grav.* 20, R201.
- [6] Carlip, S. 1993, *Class. Quant. Grav.* 10, 207.
- [7] Cornish, N., Spergel, D. and Starkman, G. 1998, *Class. Quant. Grav.* 15, 2657.
- [8] Cornish, N., Spergel, D., Starkman, G. *et al.*, 2004, *Phys. Rev. Lett.* 92, 201302.
- [9] Cresswell, J., Liddle, A., Mukherjee, P. *et al.*, 2006, *Phys. Rev. D* 73 041302.
- [10] De Oliveira-Costa, A., Smoot, G. 1995, *Astrophys. J.* 448, 447.
- [11] De Sitter, W. 1917, *Mon. Not. Roy. Astron. Soc.* 78, 3.
- [12] Costa, S., Fagundes, H. 2001, *Gen. Rel. Grav.* 33, 1489.
- [13] Einstein, A. 1917, *Preuss. Akad. Wiss. Berlin Sitzber.* 142.
- [14] Ellis, G. F. R. 1971, *Gen. Rel. Grav.* 2, 7.
- [15] Gausmann, E., Lehoucq, R., Luminet, J.-P. *et al.*, 2001, *Class. Quant. Grav.* 18, 5155.

- [16] Gomero, G., Teixeira, A., Reboucas, M. *et al.*, 2002, *Int. J. Mod. Phys. D* 11, 869.
- [17] Gundermann, J. 2005, [arXiv:astro-ph/0503014].
- [18] Hinshaw, G., Banday, A. J., Bennett, C. L. *et al.*, 1996, *Astrophys. J. Lett.* 464, L17.
- [19] Key, J., Cornish, N., Spergel, N. *et al.*, 2006, [arXiv:astro-ph/0604616v1].
- [20] Lachièze-Rey, M. 2004, *Class. Quant. Grav.* 21, 2455.
- [21] Lachièze-Rey, M., Luminet, J. P. 1995, *Phys. Rep.* 254, 135.
- [22] Land, K., Magueijo, J. 2005, *Phys. Rev. Lett.* 95, 071301.
- [23] Lehoucq, R., Lachièze-Rey, M. and Luminet, J. P. 1996, *Astron. Astrophys.* 313, 339.
- [24] Lehoucq, R., Luminet, J.-P. and Uzan, J.-P. 1999, *Astron. Astrophys.* 344, 735.
- [25] Lehoucq, R., Weeks, J., Uzan, J.-P. *et al.*, 2002, *Class. Quant. Grav.* 19, 4683.
- [26] Lemaitre, G. 1931, *Mon. Not. Roy. Astron. Soc.* 91, 490.
- [27] Levin, J. 2002, *Phys. Rep.* 365, 251.
- [28] Luminet, J.-P. 2001, *L'Univers chiffonné*, Fayard, Paris. English translation by E. Novak (2008). *The Wraparound Universe*, AK Peters, Wellesley, MA.
- [29] Luminet, J.-P., Weeks, J., Riazuelo, A. *et al.*, 2003, *Nature* 425, 593.
- [30] Prunet, S., Uzan, J.-P., Bernardeau, F. *et al.*, 2005, *Phys. Rev. D* 71 083508.
- [31] Riazuelo, A., Uzan, J.-P., Lehoucq, R. *et al.*, 2004a, *Phys. Rev. D* 69, 103514.
- [32] Riazuelo, A., Weeks, J., Uzan, J.-P. *et al.*, 2004b, *Phys. Rev. D* 69, 103518.
- [33] Roukema, B., Lew, B., Cechowska, M. *et al.*, 2004, *Astron. Astrophys.* 423, 821.
- [34] Smolin, L. 2002, *Three Roads to Quantum Gravity*. Perseus Books Group, New York.
- [35] Sokolov, I. 1993, *JETP Lett.* 57, 617.
- [36] Spergel, D. N., Verde, L., Peiris, H. V., *et al.*, 2003, *Astrophys. J. Suppl. Ser.* 148, 175.
- [37] Spergel, D. N., Bean, R., Doré, O. *et al.*, 2006, [arXiv:astro-ph/0603449v2].
- [38] Thurston, W. P. 1997, *Three-dimensional Geometry and Topology*. Volume 1. In S. Levy ed., Princeton Mathematical Series, 35. Princeton University Press, Princeton, NJ.
- [39] Tonry, J. L., Schmidt, B. P., Barris, B. *et al.*, 2003, *Astrophys. J.* 594, 1.
- [40] Tsujikawa, S., Maartens, R. and Brandenberger, R. 2003, *Phys. Lett.* B574, 141.
- [41] Uzan, J.-P., Lehoucq, R. and Luminet, J.-P. 1999, *Astron. Astrophys.* 351, 766.
- [42] Uzan, J.-P., Riazuelo, A., Lehoucq, R. *et al.*, 2004, *Phys. Rev. D* 69, 043003.
- [43] Weeks, J., SnapPea. [Online]. Available at: <http://geometrygames.org/SnapPea/> [Accessed on 04 December 2008].
- [44] Weeks, J., Luminet, J.-P., Riazuelo, A. *et al.*, 2004, *Mon. Not. Roy. Astron. Soc.* 352, 258.
- [45] Weeks J., Gundermann, J. 2006, [arXiv:astro-ph/0611640v1].

This page is intentionally left blank

PART 2

The Problem of Space in Neurosciences

This page is intentionally left blank

CHAPTER 5

Space Coding in the Cerebral Cortex

LEONARDO FOGASSI

Dipartimento di Neuroscienze, Via Volturno 39, 43110 Parma

and

Dipartimento di Psicologia, B.go Carissimi 10, 43100 Parma

Dipartimento di Neuroscienze, V. Volturno 39,

43100 Parma, Italy

fogassi@unipr.it

I. Introduction

The concept of space is largely present in our daily life, both in real and metaphoric terms. There is general agreement that, introspectively, space is perceived as unitary. In fact, although we often use words referring to different spatial coordinates, such as up-down, left-right, near-far, we don't easily conceive space, in mental terms, as subdivided in different sectors. At the same time, we move in space, by using our body parts or man-made vehicles. However, this does not give us immediately an insight on the crucial importance of our movements for the formation of our internal representation of space. The idea underlying this chapter is that space is not coded in the brain in a unitary way and that the cortical motor system very likely constitutes the basis for building our cerebral representation of space.

The issue of which coordinate system space is coded in is a crucial one. When speaking about space in physical terms, we need to define the reference axes in respect to which we identify a spatial location. Geometrically, each spatial location is defined by means of three coordinates (x , y , z) in respect to the origin. If instead we define space in biological terms, it can be characterised by the various sensory modalities involved by the application to the body of a spatially organised stimulus. If for example a stimulus is touching our body, we can define the body sector where it is applied as a 'personal' space, linked to the somatosensory modality. If, instead, the stimulus is presented outside the body, we can define an

'extrapersonal' space, that is linked to the visual or acoustic sensation. However, space can be defined also in motor terms, on the basis of the direction and the amplitude of the movements that are needed in order to reach a specific location.

The difference between the various definitions of space corresponds also to a difference in the coordinate system in which the external stimuli are centred. The locations of stimuli in the visual space, certainly the most deeply studied in humans and monkeys, are referred, at the beginning of cortical visual processing, to a coordinate system centred on the retina (retinocentric or oculo-centric frame of reference). We must recognise, however, that if a visual stimulus does not move, when we shift the eyes, the location of that stimulus changes in retinal coordinates in respect to when the eyes had not moved yet. In other words, if a visual stimulus is presented at 20° right and 10° up from the centre of the retina (the fovea centralis), when the eyes move 10° to the left the stimulus will fall in a retinal position that is now 30° right and 10° up in respect to the fovea. The stimulus, of course, is the same, but its retinal location has changed. Note, however, that its position in respect to the head remains the same, independent of the eye movement, provided that the head did not move together with the eyes. This observation enables us to define a new coordinate system, in which visual stimuli are coded in respect to the median axis of the head. If, however, the head also moved with the eyes, the visual stimulus will now have different coordinates with respect to the head axis but its position remains the same when referred to another axis, that is, the median axis of the body. Following this logic, if the body also moved, now the visual stimulus will assume different coordinates with respect to the retina, the head axis and the body axis. However, its position referring to the other visual stimuli in the world remains the same. Summing up, a same visual stimulus can be defined in different frames of reference, namely retinocentric, head-centred, body-centred and allocentric. The existence of these different frames of reference raises many questions:

- (a) How is it possible to pass from one coordinate system to another?
- (b) In how many reference systems sensory stimuli are coded in the brain?
- (c) Can visual stimuli be coded in the brain in a reference frame independent of eye or head movements?

The question (a) is particularly important for planning movements in space. If, for example, I want to reach for an object located at a certain distance from my body, its spatial position is initially registered in retinal coordinates, but the arm movement must be performed quite independently of the eye position and also of the arm position with respect to the object. An object located in the upper right visual field can be reached with the right hand both when the hand is on the right of the body or when it is on the left, crossing, in this latter case, the body axis. The possibility of correctly performing a reaching movement implies a

coordinate transformation process and also a computation of the relative position of the different effectors (body, arm).

Several models have been proposed in order to explain how all these computations can occur. The description of these models, however, is beyond the scope of this article. I will try to provide some answers to questions (b) and (c), that are directly concerned with the way in which the brain codes space.

The general issue of space coding has been directly addressed, from a neurophysiological point of view, more than twenty years ago. Basing on the results of these experiments, I think it is possible to claim that now we have a good knowledge of how space is coded in the cerebral cortex. The three most important achievements are the following:

- (a) space is coded in the brain not as a unique representation, but as subdivided in different sectors, each corresponding to a dedicated anatomo-functional circuit, involved in the sensorimotor transformation for actions performed with a specific effector;
- (b) space is coded, at the single neuron level, in a frame of reference suitable for the function of the brain cortical area to which the neuron belongs;
- (c) space perception is strictly linked to a motor concept of space.

In the next sections I will describe the empirical data supporting these concepts.

2. The Traditional Concept. Space is Coded in Oculocentric Coordinates

A visual stimulus is, in the first stages of its cortical processing, coded in a retinocentric coordinate frame, that is its position in space is strictly linked to eye position. Let us briefly examine the organisation of the visual system. The visual information coming from the retina is conveyed to the cortex by two main pathways: the magnocellular pathway, mainly involved in the analysis of motion and brightness contrast and the parvocellular pathway, involved in the analysis of shape and colour. Both pathways are involved in the analysis of depth and three-dimensional features. In both pathways visual information is elaborated in subsequent hierarchical steps. The highest level of elaboration occurs, in the magnocellular pathway, in the inferior parietal cortex, while in the parvocellular pathway occurs in the inferotemporal cortex. Based on anatomo-functional and clinical data, in 1982 Ungerleider and Mishkin proposed an influential functional subdivision of visual processing. According to their view, after the primary and secondary visual cortex (V1 and V2) the visual system subdivides in a 'ventral stream' ending in the inferotemporal cortex, dedicated to object perception, and a 'dorsal stream', dedicated to space perception. The ventral stream has been called the 'what' system and the dorsal stream, the 'where' system. In support of this theory, patients with damage to the

ventral stream are unable to recognise and discriminate objects, but are still able to indicate their spatial location, while patients with damage to the dorsal stream present the opposite dissociation. The same deficit was shown, although at a lower degree, also in monkeys.

Keeping with this view, the question arises on which properties of the parietal lobe could explain space perception and in particular in which frame of reference space was coded by parietal neurons. As said before, visual stimuli in the cortex are initially coded in a retinocentric frame of reference. Is stimulus spatial position coded in a different frame of reference in the parietal lobe? This latter possibility could be reasonable, because the parietal lobe was classically thought of as an association cortex, where polymodal integration would allow the formation of an abstract representation of space.

The first neurophysiological data on this topic were those recorded by Andersen and its group (Andersen, Essick and Siegel, 1985; Andersen *et al.*, 1990; Barash *et al.*, 1991) who investigated the properties of areas 7a and LIP (lateral intraparietal area) located in the convexity and the lateral bank of the posterior half of the intraparietal sulcus (IPS), respectively. Neurons of both areas show visual responses to spots of stimuli presented in several positions of the visual space and also discharge when a monkey makes a saccadic eye movement toward the visual stimulus or its remembered location (motor response). Very often the visual and motor responses are related to the same space sector. If, for example, a LIP neuron responds to a stimulus presented in the right upper part of the visual field, the neuron activates also when the monkey makes a saccade towards right-up. First of all, these properties make these neurons candidates for a role in sensorimotor transformations. Second, a remarkable feature of these neurons is that their visual and motor discharge are modulated by the orbital eye position. That is, if the neuron of the previous example has a visual and motor response toward right-up, when the monkey fixates in different spatial locations the intensity of the response is modulated by the eye position, provided that the visual stimulus is always presented in the neuron's receptive field (RF).¹ The authors concluded that 7a and LIP neurons combine the information, in retinal coordinates, of the stimulus position with that of the eye position in the orbit. Thus, the encoding of stimulus spatial location results from the interaction between these two factors. In other words, the frame of reference used by these neurons is still retinocentric but, in addition, they can also have information on where the eye is positioned with respect to the head axis. Thus, a space coding independent of eye position is not realised in these areas at the level of single neurons, but, the authors propose, could occur at a population level. That is, the combination of the responses of many neurons in areas LIP and 7a could

¹The term 'receptive field' is used to indicate the sector of visual space that triggers the neuron response, when a stimulus (for instance a light spot) is introduced in it.

give information on the absolute position of the stimulus in space, at least with respect to the head axis. On the other hand, every neuron of these areas has sufficient information to operate a transformation of the stimulus position (in retinal coordinates) in a motor vector starting from the fixation point and directed exactly to that visual location.

In subsequent experiments, in which the monkey could move the head in the horizontal plane, Brotchie *et al.* (1995) also demonstrated that head position in space could modulate the spatial visual response of LIP neurons.

A different result was obtained by the group of Galletti (Galletti *et al.*, 1993), who recorded from area V6A, a high order visual area located in the anterior lip of the parieto-occipital sulcus. They found that V6A neurons responded strongly to visual stimuli, and demonstrated that many of these neurons presented the same gaze modulation effect previously shown in areas 7a and LIP. In addition, however, they found a limited number of neurons that responded to a visual stimulus introduced in a fixed spatial position, independent of monkey gaze. Although this finding is exactly the demonstration of the existence of a frame of reference independent of the retina, the small amount of neurons with this property does not allow robust theories to build on the transformation of coordinates frame at this cortical level.

More recent studies (Andersen *et al.*, 1998) performed in a different parietal area (parietal reaching region, PRR) containing neurons related to both arm and eye movements, demonstrated that also arm movements can be coded in oculocentric coordinates.

The characteristics of all these studied parietal regions is that their neurons are always strongly linked to eye movements. That is, probably the visual responses found in these areas are used to guide several types of eye movements. Very likely also in PRR the coding of arm movements is strictly dependent on eye movement coding. Thus the question becomes what is the code of visual responses in other areas, the neurons of which are related to effectors different from the eyes, such as the arm, the head or the body? The investigation of the properties of motor areas provides an answer to these questions.

3. Coding of Peripersonal Space in the Parieto-Frontal Circuits for Reaching

Before describing the properties of areas encoding spatial position in a non-oculocentric frame of reference, I must point out three concepts:

- (a) the motor cortex is not made of just three subdivisions, as classically thought, but is composed of at least seven distinct cytoarchitectonic areas (see Rizzolatti, 1998; Rizzolatti and Luppino, 2001);

- (b) each of these areas is involved in at least one main parieto-frontal circuit;
- (c) cortical motor neurons appears to code, as their principal role, the goal of motor acts;
- (d) neuroanatomical and neurophysiological data suggest that even the parietal cortex can be considered part of the motor system (Mountcastle *et al.*, 1975; Hyvarinen 1982; Rizzolatti, Fogassi and Gallese, 1997; Rizzolatti, Luppino and Matelli, 1998; Fogassi and Luppino, 2005).

The receptive field of a neuron is that portion of space that, when stimulated, elicits the neuron discharge. A small change in the stimulus location can determine a drastic decrease of the neuron discharge.

Area F4 is an area located in the caudal part of ventral premotor cortex. It can be distinguished, by means of histological and histochemical methods, from the primary motor cortex (F1) and the rostrally located area F5, involved in coding goal-directed hand and mouth motor acts.

One of the electrophysiological methods used for determining which are the movements controlled by a specific area is electrical microstimulation. This technique allows, using very low current intensities, to elicit the activity of a small neuronal population localised within a certain diameter around the stimulating microelectrode. This method gives very good information on the somatotopic representation of a particular cortical sector.

Electrical microstimulation of area F4 elicits trunk, neck, arm and facial movements. In accord with these findings, recordings from this area during movement execution show that its motor neurons are active during reaching, orienting and facial movements (Gentilucci *et al.*, 1988; Fogassi *et al.*, 1996a).

The most interesting properties of this area probably consist in its neuronal responses to sensory stimuli. There are two main categories of sensory neurons. The first is constituted by somatosensory neurons, e.g., neurons activated by the tactile stimulation of the face, the arm and the trunk. The second is formed by bimodal, somatosensory and visual neurons, that discharge not only during the application of somatosensory stimuli, but also to the introduction of three-dimensional visual stimuli in a space sector close to the neuron somatosensory RF. Very often the best response of bimodal neurons is obtained approaching an object to their tactile RF. Their visual RF is peculiar, because it is limited not only in width, but also in depth. Generally, the visual response begins when the stimulus becomes closer to the monkey (no more than 40 cm) and ends when it is near to the tactile RF. This limitation in depth of the visual RF (three-dimensional visual RFs) is very different from what happens in the visual areas, in which the neuron response does not depend from the distance at which the stimulus is presented. Because of this delimitation of the visual space in which the stimulus is effective in evoking the neuronal response of bimodal neurons, these visual RFs have been called 'peripersonal'.

In most bimodal neurons the visual RF is in register with the tactile RF. For example, a neuron the tactile RF of which is on the right hemi-face will respond also to an object approaching the right hemi-face, while a neuron with a tactile RF on the shoulder will be activated by a similar object, but approached to the shoulder.

Coming back to the issue of the frames of reference, it has been demonstrated with specific experiments that the visual responses of F4 bimodal neurons are not retinocentric. This important point has been demonstrated by training monkeys to fixate in different spatial positions, while an object was approached to the tactile RF of each neuron. The results showed that the visual response was always present, independently of eye position (see also Gentilucci *et al.*, 1983; Fogassi *et al.*, 1992; Graziano, Yap and Gross, 1994; Fogassi *et al.*, 1996b; Graziano, Hu and Gross, 1997a).

Thus in the premotor cortex there occurs a transformation of coordinates, from retinocentric to some type of eye-independent frame of reference. Where is the centre of coordinates of this reference system? It is not very easy to respond to this question, because in single neuron recording experiments the monkey's head is generally fixed. However, Graziano and coworkers (1994) showed that, by moving the head or the arm of the monkey, the visual RF of F4 bimodal neurons followed the tactile RF. In this study the monkey was trained to fixate in different spatial position, while a stimulus was moved toward the body part where was the tactile RF of the neuron under investigation. First of all, it was demonstrated that the shift in eye position did not influence the body part-related visual response. Furthermore, when the hand or the head were moved, the visual response could be evoked only when the stimulus was moved towards the head or the arm (depending on the tactile RF of the neuron), and not when it was moved toward the sector of space previously occupied by the arm or the head. That is, the visual stimulus was encoded in a body-part (somatocentred) frame of reference, and not with respect to the body or the head midline.

This concept was further corroborated by two subsequent studies of the Graziano's group. In one of them (Graziano, Reiss and Gross, 1999), it was demonstrated that a category of F4 neurons responded not only to a peripersonal visual stimulus but also to the introduction of an acoustic stimulus in the space near the tactile RF. The acoustic response did not depend on the stimulus intensity, but its presence in the peripersonal space. In most of these trimodal neurons the tactile, visual and acoustic responses were spatially congruent.

In the other study (Graziano, Hu and Gross, 1997b) it was shown that bimodal neurons responding to objects introduced near the tactile RF continued to discharge in the dark, when the monkey could not see the object, but was aware of its presence in the peripersonal space. If the object was then removed, as far as the monkey could see that the stimulus was not there any more, the neuron ceased its

discharge. This latter finding strongly indicates that coding of space in F4 occurs at the representation level (see below).

Area F4 is reciprocally connected with area VIP (Luppino *et al.*, 1999), located in the fundus of the intraparietal sulcus. This area, thanks to its connections with visual motion areas, such as the middle temporal area (MT) and the middle superior temporal area (MST), belonging to the dorsal visual stream (Maunsell and Van Essen, 1983), and with parietal areas endowed with somatosensory properties, contains neurons responding either to visual or to bimodal, visual and somatosensory, stimuli. The most effective visual stimuli are moving stimuli, some of which are also approaching to or going away from the monkey (Colby, Duhamel and Goldberg, 1993; Colby, 1998; Duhamel, Colby and Goldberg, 1998). Bimodal neurons have properties very similar to those of F4 bimodal neurons but, differently from the latter, only a small percentage of them respond to an object introduced in the peripersonal space. The most striking difference between VIP and F4 bimodal neurons consists in the fact that most a visual RFs of VIP neurons are coded in retiocentric coordinates. Interestingly enough, however, the smaller percentage responding to peripersonal visual stimuli coded this stimuli in somatocentred coordinates. Thus, with a visual object approaching the monkey, VIP neurons receive visual information from dorsal stream visual areas and perhaps start a visuomotor transformation for head and arm movements directed to, or going far from, these visual stimuli. The visuomotor transformation for reaching and head orientation is, however, mostly accomplished in area F4. The presence of neurons with peripersonal responses in area VIP could be due to the motor information coming from area F4. Whatever the explanation, the presence of somatocentred neurons in VIP may explain its possible involvement in movement control. In fact, electrical microstimulation of VIP (Thier and Andersen, 1998) elicits head, face and arm movement, although at high intensity current thresholds.

Summing up, the properties of the F4-VIP circuit indicate its involvement in two main, intrinsically linked, functions: (a) visuomotor transformation for axial and proximal actions in space; (b) coding space representation. While the first function follows quite naturally from the cortical location of the two areas and their input-output organisation, the second is not immediately intuitive, if one interprets space coding in a classical sense, that is in strictly perceptual terms. An insight on how these areas code space comes from the peculiar organisation of the three-dimensional visual RFs of their bimodal neurons. In order to build such RFs, it is necessary to combine the tactile RFs of these neurons and the motor properties present in area F4 and, possibly, in area VIP. In principle, the formation of body part-anchored visual RFs could be explained by associative learning, that is by the repetitive association between a visual object approaching a body part and the tactile sensation that it evokes when it arrives in contact with the skin. However, this would not explain why the extent of the three-dimensional visual RFs never exceeds 40 cm and also why, inside this limit, the preferred depth varies so much among

different neurons. These two properties, instead, better account for an interpretation of these RFs in relation to the motor properties, in particular with those of area F4. That is, these RFs are related to the various types of motor acts normally performed inside this space, such as mouth grasping, eye blinking (percutaneous RFs), arm reaching, bringing to the body or trunk orienting (distant peripersonal RFs). Thus, the visual and auditory inputs are instrumental for providing spatial sensory information for the different types of motor acts controlled by area F4. However, one could think that the appropriate sensory information is analysed and that the product of this analysis is then passed to motor neurons that 'decide' the most suitable movement. If this is so, the discharge of bimodal F4 neurons is purely sensory. However, the present view does not restrict the role of premotor areas to a sensorimotor integration, but maintains that their major role is that of coding the goal of motor acts. For example, motor neurons of area F5 (the ventral premotor area located rostrally to F4) code hand and mouth motor acts, such as grasping, biting, tearing, manipulating, etc. This encoding is not simply used for the execution of these acts, but represents a 'motor knowledge' that can be addressed by external sensory input, creating new neuronal categories. For instance, object observation activates F5 'canonical' neurons, a specific category of visuomotor neurons discharging when the monkey grasps an object and when it observes an object congruent with the type of grip that the neuron motorically code (Murata *et al.*, 1997; Raos *et al.*, 2006). Most interestingly, canonical neurons respond also to pure object observation, even in the absence of a grasping movement toward it. This 'visual' response has been interpreted as 'the idea of movement' or, in other words, the object 'motor representation' (Rizzolatti and Fadiga, 1998).

Analogously, an object introduced in the peripersonal space evokes a motor representation of the space in which the object is located, be it seen or heard. As summarised above, area F4 is endowed with neurons that discharge when the monkey executes orienting and reaching motor acts (Gentilucci *et al.*, 1988; Fogassi *et al.*, 1996a; Graziano, Yap and Gross, 1994; Fogassi *et al.*, 1996b; Graziano, Hu and Gross, 1997a). Thus, although the peripersonal 'visual' responses of F4 bimodal neurons are present independently of any impending movement of the monkey, they can be interpreted as a pragmatic representation of the space in which the object evoking this visual response is introduced. For example, a visual response to an object approaching the monkey's face will retrieve the representation of an avoidance or approaching motor act of the trunk, depending on the nature of the stimulus. Similarly, a visual response to an object approaching the monkey's arm could retrieve the representation of a reaching motor act. These representations, depending on the context, can be implemented in an overt motor act or remain in the status of *potential motor acts*, enabling space perception (see next section). Thus, the motor representations of F4 neurons can accomplish two tasks. First, they can play a major role in the sensorimotor transformation for facial, axial and proximal actions. Second, they can code space directly, in motor

terms, using the same coordinate system of the effector acting in that portion of peripersonal space.

The ‘motor representation’ interpretation of the visual response of bimodal neurons is corroborated by a further experiment carried out on F4 bimodal neurons (Fogassi *et al.*, 1996b). In this experiment, after the tactile and visual response of each neuron had been characterised, the visual stimulus (a three-dimensional object) was moved at four different velocities toward the neuron tactile RF. The results showed that the three-dimensional visual RF increased in depth as far as the stimulus velocity increased and shrank as far as the stimulus velocity decreased. In other words, at higher stimulus velocities, the neuron began to respond farther in space and earlier in time. These results are perfectly in line with the notion of the encoding of a potential motor act, because an individual must begin earlier a spatially organised action when the velocity of an approaching target increases, otherwise he could miss the target.

A similar finding has been shown also in humans by Chieffi *et al.* (1992). They asked participants to reach for and grasp a sphere approaching them at different velocities. When the object approached at higher velocities, participants started the forelimb movement earlier in time and farther than at lower velocities. These results, obtained during an overt action execution, appear to be the direct correlate of the potential motor act coded by F4 bimodal neurons.

The strict link between action and space coding is also demonstrated by a clever experiment carried out in a monkey parietal area by Iriki and coworkers (Iriki, Tanaka and Iwamura, 1996). They studied sensory neurons of medial parietal area PEip, endowed with properties similar to those of F4 bimodal neurons. They have tactile RFs on the face or the forelimb and peripersonal visual RFs. In this area there are apparently no motor neurons.

After an initial characterisation of the bimodal features of these neurons, the researchers trained the monkey, from which these neurons had been recorded, to take food out of arm reach by means of a rake. Once the monkey learned the task, they continued to record from PEip neurons and found that the extent of their peripersonal visual RFs were larger than before training, encompassing also the space occupied by the rake, as if this tool had become a prolongation of the arm. After a period in which the monkey did not perform the task anymore, the RFs came back to its original extent. Thus, it is clear that action can plastically model space. Interestingly enough, this mechanism is present also in humans (see below).

4. Further Cortical Areas Involved in Space Coding

Neurons with peripersonal visual responses are not limited to the VIP-F4 circuit, but are present in other cortical and even subcortical regions. Concentrating only on cortical areas, two of them are dorsal premotor area F2 and superior parietal area

MIP (Matelli *et al.*, 1998; Marconi *et al.*, 2001), that are reciprocally connected to form another parieto-frontal circuit coding reaching movements. Area F2, located in the caudal two-thirds of dorsal premotor cortex, contains a sensorimotor representation of the whole body, except the face (Fogassi *et al.*, 1999; Raos *et al.*, 2003). Its sensory input is mostly somatosensory, however in its ventro-rostral part (F2vr) there also visually-driven neurons (Fogassi *et al.*, 1999). Among them there are also bimodal, tactile and visual, neurons, that have large tactile RFs on the forelimb and upper trunk and, as in F4, three-dimensional visual RFs, limited to the peripersonal space and in register with the somatosensory RFs. In contrast to F4, the RFs of F2 neurons are more related to the forelimb, while face-related neurons, that are prominent in F4, are virtually absent in F2. Area MIP is located in the dorsal bank of the intraparietal sulcus. In MIP there are motor neurons related to reaching movements, purely somatosensory neurons, purely visual neurons and neurons with bimodal properties responding to passive touch on the contralateral forelimb and to the presentation of a visual stimulus (see Colby, 1998). Their activity is very high when the monkey executes a reaching movement toward a visual target. It is interesting to note that deep in the sulcus, below these bimodal neurons, there are purely visual neurons whose response increases when the target is moved within reaching distance. Finally, there are motor neurons discharging during reaching movements (Colby and Duhamel, 1991; see also Johnson *et al.*, 1996).

What is the relation between forelimb movements and visual stimulation in the MIP-F2vr circuit? Since most of visual RFs in MIP are large and encompass the periphery of the visual field (Colby and Duhamel, 1991; Galletti *et al.*, 1999), it is likely that coding of space in this circuit could subserve a monitoring function of the reaching movement toward a static or a moving target.

Another cortical sector that probably plays a role in space coding is the rostral part of the inferior parietal lobule (rIPL), namely areas PF and PFG, strongly connected with ventral premotor cortex, including area F4 (Matelli *et al.*, 1986; Cavada and Goldman-Rakic, 1989; Rizzolatti and Luppino, 2001; Rozzi *et al.*, 2006). In this sector there are neurons with somatosensory, visual and motor properties (Leinonen *et al.*, 1979; Hyvarinen, 1981, 1982). Several authors described also many bimodal neurons with tactile RFs on the face, the arm and the upper trunk (Hyvarinen, 1981; Leinonen and Nyman, 1979; Leinonen *et al.*, 1979; Graziano and Gross, 1995; Ferrari *et al.*, 2003) and large visual RFs often very close to the tactile RFs. These bimodal neurons are located in a sector of IPL where mouth and hand motor acts are represented. Thus, it seems quite logical to hypothesise that the activity of bimodal neurons, as already proposed for F4, represents potential motor acts to be executed in the peripersonal space near a specific body part.

Summing up, there are many cortical circuits that can contribute to space coding, in particular of peripersonal space. Although in all these circuits space coding

appears to be strictly related to motor programming and execution, in some circuits the prevailing aspect appears to be the motor representation of the goal of specific motor acts, independently of whether these will be actually executed or not, while in others the on-line control of movement in space seems the most important feature.

5. Lesions Data Confirm the Presence of Different Types of Space Coding

As described in a previous section, a classical theory (Ungerleider and Mishkin, 1982) suggests that the dorsal visual stream, whose higher level of elaboration is located in the parietal cortex, is involved in space perception (the ‘what’ system). Although this theory has been challenged by another one (Goodale *et al.*, 1992), maintaining that the visual information conveyed by the dorsal stream is mainly exploited for action, one cannot deny that one of the most described deficit subsequent to a lesion of the parietal lobe consists in spatial neglect (see Bisiach and Vallar, 2000). More precisely, patients with this syndrome are not visually blind but do not report visual, acoustic or tactile stimuli presented in the left hemispace. The lesion responsible for this deficit, although it can involve either subcortical or cortical structures, in most cases is located in the right inferior parietal lobule. Note that a lesion of a similar sector in the left hemisphere produces, generally, another syndrome, called apraxia. This difference appears to be due to a division of labor, a specialisation, of the two hemispheres. The neglect syndrome is very likely due to an impairment in the elaboration of sensory stimuli that normally allow the individual to become aware of cutaneous and external space. Several theories have been elaborated in order to explain the nature of the neglect syndrome. Some maintain that it consists in a perceptual impairment, others that it is a deficit of mental representation, others — very influential — that it consists in a selective loss of attention for the left hemispace.

Independently of which theory is correct (probably they are not necessarily incompatible one with the other, but each of them just focuses on one aspect of the deficit), the site of the lesion is in a cortical region that, in monkeys, contains neurons with spatial properties that is richly connected with ventral premotor cortex.

Is there evidence of a neglect syndrome in monkeys? Unilateral lesion of ventral premotor cortex produce spatially-related motor and sensory deficits (Rizzolatti, Matelli and Pavesi, 1983; Schieber, 2000; Fogassi *et al.*, 2001). Motor deficits consist in a reluctance to use the arm contralateral to the lesion (in the monkey the deficit is independent of the hemisphere), for example in response to sensory stimuli, in a slowness and inaccuracy in reaching movements and in an impairment in biting food presented in the contralateral hemifield, near the monkey’s

mouth. Sensory deficits consist in a neglect of visual and tactile stimuli introduced in the hemispace contralateral to the lesion, near the monkey face or arm. This impairment is very clear, because when the stimuli are introduced in the hemispace ipsilateral to the lesion, the monkey reacts immediately. Very interestingly, threatening stimuli which, when introduced in the peripersonal space, normally evoke a blinking reaction, elicit this reaction only when introduced in the ipsilateral hemifield. However, when the same stimuli are presented far from the animal (outside the peripersonal space) the blinking reaction is evoked from both the contralateral and the ipsilateral space, showing a clear dissociation between a near and a far space. In addition, eye movements are immediately elicited by stimuli presented far from the monkey, but not when the same stimuli are presented near the monkey in the contralateral visual field.

The opposite dissociation is obtained with a lesion of frontal eye fields (FEF), a cortical area located in front of ventral premotor cortex (Rizzolatti, Matelli and Pavesi, 1983). Monkeys with lesion of FEF do not move the eyes or orient toward stimuli presented in the far space, contralaterally to the lesion side, while they perform normal eye movements toward stimuli presented in the contralateral peripersonal space. Thus, the differential effects provoked by lesion of ventral premotor cortex or FEF confirm the presence of different areas for space processing, namely for processing of peripersonal and extrapersonal space.

Ventral premotor cortex and FEF are anatomically connected with the rostral half of the inferior parietal cortex and with area LIP, respectively. While there are studies reporting some spatial or saccade-related impairment after LIP inactivation (Li, Mazzoni and Andersen, 1999; Wardak, Olivier and Duhamel, 2004), there are no data on VIP inactivation.

I previously described the involvement of IPL in coding peripersonal visual stimuli. In agreement with these data, the lesion of IPL may induce motor deficits such as misreaching and hand clumsiness (Ettlinger and Kalsbeck, 1962; Faugier-Grimaud, Freno and Stein, 1978; Rizzolatti, Gentilucci and Matelli, 1986; Gallese *et al.*, 1994; see also Hyvarinen, 1982) and a kind of neglect or extinction (Denny-Brown and Chambers, 1958; Deuel, 1987). However, the spatial deficits, when present, seem less strong than those observed after ventral premotor lesion.

Although monkey lesion experiments do not allow us to reach a conclusion on the differential contribution of frontal and parietal areas to space processing, it is clear that different sectors of space are processed through different anatomical circuits. This differential processing has also been recently demonstrated in humans. In neglect patients, in most cases the tests have been performed with the patient in his bed. Thus, the typical signs of the syndrome mainly concerned the peripersonal space. However, more recent studies demonstrated a double dissociation effect. For example, there are patients that typically bisect a horizontal line shifting its middle point more to the left when the line is in their peripersonal space, but indicate precisely the middle point when the same test is performed

on a far line, reachable only with a projector light-pen (Halligan and Marshall, 1991; see also Berti and Frassinetti, 2000). Other patients show the opposite dissociation (Cowey, Small and Ellis, 1994; Shelton, Bowers and Heilman, 1990; Cowey, Small and Ellis, 1999). These data speak in favour of a non-unitary space representation, that is of a space coding depending on the type of body part acting in space in a specific behavioural situation. Thus, coding of extrapersonal space is directly related to coding of eye movements, while coding of peripersonal space is linked to the arm, face, and leg movements. Normally, the working space of the various body parts does not change. However, in situations in which stimulus velocity changes, as in the experiment described above (Fogassi *et al.*, 1996b), the working space changes, and probably also its cortical representation.

In human patients it has been demonstrated that spatial representation can directly depend on actions performed in space. Berti and Frassinetti (2000) used the bisection test in a patient showing neglect only for the near space. The patient bisected correctly with a projector light-pen a line presented in the far space, while he shifted the bisection to the right when asked to do it in the peripersonal space with a pencil. When, however, the patient was asked to bisect a line in the far space with a stick held in his hand, he pointed to the right of the middle point, thus showing that the neglected space expanded in depth. As in the above reported study of Iriki *et al.* (1996) in monkeys, tool use extends peripersonal space, probably because it is incorporated in the body schema.

Further data confirming the importance of the motor system in shaping peripersonal space are provided by patients showing extinction. This deficit shares some feature with neglect. Differently from neglect patients, those with extinction normally report the presence of single stimuli presented in the left hemispace, contralateral to the lesion. However, when two stimuli are presented, one to the left and the other to the right of the patient, he/she reports always the stimulus presented ipsilaterally to the lesion. It has been shown that in subjects with tactile extinction, this deficit can be reduced by a visual stimulus presented in the peripersonal, but not in the far, contralateral hemispace. When, however, these patients are trained to use a rake in order to retrieve a distant object, subsequently a visual stimulus presented in the contralateral hemifield, far from the hand but near the tip of the rake, is able to improve the tactile extinction. This effect is not present if the rake is passively held in the patient's hand (see Maravita and Iriki 2004; Farnè, Iriki and Ladavas, 2005).

The neurological data presented above do not easily allow to trace the homology between areas of both species. Recently however, Bremmer *et al.* (2001) in an fMRI experiment, found an activation of ventral premotor cortex and two parietal regions, one in the intraparietal sulcus and one in the inferior parietal gyrus, applying tactile stimuli to the upper face and visual and acoustic stimuli near the same face sector. The activated parietal and frontal regions could correspond to areas VIP and PFG of

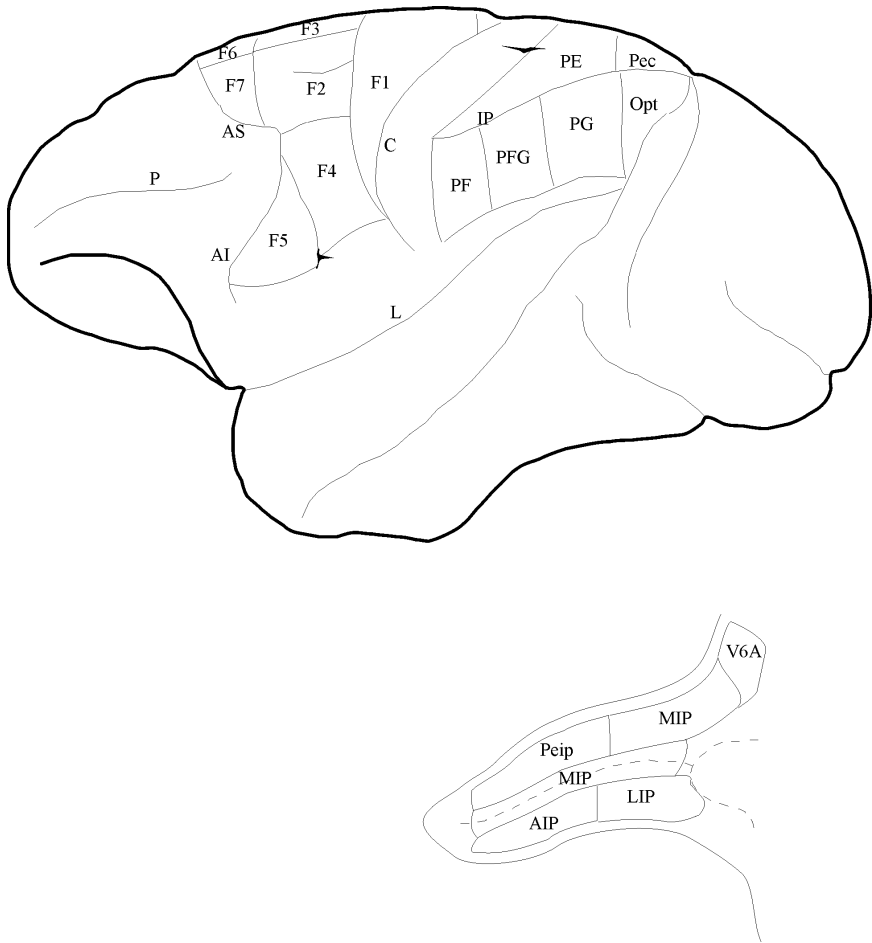


Figure 1 Lateral view of a left hemisphere of a monkey cerebral cortex showing the parcellation of the motor and the posterior parietal cortices. The areas located within the intraparietal sulcus are shown in an unfolded view of the sulcus in the right lower part of the figure. For the nomenclature and definition of posterior parietal and motor areas see Rizzolatti, Luppino and Matelli (1998) and, for a recent redefinition of inferior parietal areas, Gregoriou *et al.* (2006). Abbreviations: AI, inferior arcuate sulcus; AS, superior arcuate sulcus; C, central sulcus; L, lateral fissure; P, principal sulcus.

the monkey, respectively. These findings are consistent with monkey data indicating that peripersonal space is processed by a dedicated neuroanatomical circuit, and support the neuropsychological observation of a double dissociation for near and far space in patients with neglect.

6. Conclusions

In this article I provided evidence that space representation at the brain level is not unitary, but is subdivided among different parieto-premotor circuits, whose activation is strictly related to actions performed in different space sectors. The concept emerging from the organisation of this circuit is that space coding is primarily motor. Although in the adults actions in space are very often triggered by sensory stimuli, so that the first active regions in the cerebral cortex are sensory areas, in infants things can work in a different way. Babies of a few days of age orient their head and trunk in space, move their mouth toward the mother's nipple, perform reaching movements. All these movements allow them to create a space motor representation in their brain, before the visual system, in particular detailed vision, is still completely developed. This 'motor vocabulary' for space will only later be validated by a better elaborated visual input, thanks to the link between premotor and parietal cortex. This view is supported by the above described data demonstrating a crucial role of the adult motor system in plastically changed space representation.

Although neglect patients appear to have a space deficit in different sensory modalities and space sectors, it is still possible to demonstrate a dissociation of the symptoms between far and near space. It would be interesting, in the future, to see which is the neural mechanism allowing the construction in our brain of a representation of space as a whole.

References

- [1] Andersen, R. A., Bracewell, R. M., Barash, S. *et al.*, 1990, Eye position effects on visual, memory and saccade-related activity in area LIP and 7a of macaque. *J. Neurosci.*, 10, 1176–1196.
- [2] Andersen, R. A., Essick, G. K. and Siegel, R. M. 1985, Encoding spatial location by posterior parietal neurons. *Science*, 230, 456–458.
- [3] Andersen, R. A., Snyder, L. H., Batista, A. P. *et al.*, 1998, Posterior parietal areas specialized for eye movements (LIP) and reach (PRR) using a common coordinate frame. *Novartis Found Symp.* 218, 109–122.
- [4] Barash, S., Bracewell, R. M., Fogassi, L. *et al.*, 1991, Saccade-related activity in the lateral intraparietal area. II. Spatial properties. *J. Neurophysiol.*, 66, 1109–1124.
- [5] Berti, A., Frassinetti, F. 2000, When far becomes near: re-mapping of space by tool use. *J. Cog. Neurosci.*, 12, 415–420.
- [6] Bisiach, E., Vallar, G. 2000, Unilateral neglect in humans. In F. Boller and J. Grafman (eds.), *Handbook of Neuropsychology*, 2nd edn., Vol. 1. Amsterdam, Elsevier, pp. 459–502.

- [7] Bremmer, F., Schlack, A., Jon Shah, N. *et al.*, 2001, Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalences between humans and monkeys. *Neuron*, 29, 287–296.
- [8] Brotchie, P. R., Andersen, R. A., Snyder, L. H. *et al.*, 1995, Head position signals used by parietal neurons to encode locations of visual stimuli. *Nature*, 375, 232–235.
- [9] Cavada, C., Goldman-Rakic, P. S. 1989, Posterior parietal cortex in rhesus monkey: II. Evidence for segregated corticocortical networks linking sensory and limbic areas with the frontal lobe. *J. Comp. Neurol.*, 87, 422–445.
- [10] Chieffi, S., Fogassi, L., Gallese, V. *et al.*, 1992, Prehension movements directed to approaching objects: influence of stimulus velocity on the transport and the grasp components. *Neuropsychologia*, 30, 877–897.
- [11] Colby, C. L. 1998, Action-oriented spatial reference frames in cortex. *Neuron*, 20, 15–24.
- [12] Colby, C. L., Duhamel, J. R. 1991, Heterogeneity of extrastriate visual areas and multiple parietal areas in the macaque monkey. *Neuropsychologia*, 29, 517–537.
- [13] Colby, C. L., Duhamel, J. R. and Goldberg, M. E. 1993, Ventral intraparietal area of the macaque: anatomic location and visual response properties. *J. Neurophysiol.*, 69, 902–914.
- [14] Cowey, A., Small, M. and Ellis, S. 1994, Left visuo-spatial neglect can be worse in far than near space. *Neuropsychologia*, 32, 1059–1066.
- [15] Cowey, A., Small, M. and Ellis, S. 1999, No abrupt change in visual hemineglect from near to far space. *Neuropsychologia*, 37, 1–6.
- [16] Denny Brown, D., Chambers, R. A. 1958, The parietal lobe and behavior. *Proc. Ass. Res. Nerv. Ment. Dis.*, 36, 35–117.
- [17] Deuel, R. K. 1987, Neural dysfunction during hemineglect after cortical damage in two monkey models. In M. Jeannerod (ed.), *Neurophysiological and Neuropsychological Aspects of Spatial Neglect*. Amsterdam, North-Holland Co. Elsevier Science Publishers, pp. 315–334.
- [18] Duhamel, J. R., Colby, C. L. and Goldberg, M. E. 1998, Ventral intraparietal area of the macaque: congruent visual and somatic response properties. *J. Neurophysiol.*, 79, 126–136.
- [19] Ettlinger, G., Kalsbeck, J. E. 1962, Changes in tactile discrimination and in visual reaching after successive and simultaneous bilateral posterior parietal ablations in the monkey. *J. Neurol. Neurosurg. Psychiatr.* 25, 256–268.
- [20] Farnè A., Iriki, A. and Ladavas, E. 2005, Shaping multi-sensory action space with tools: evidence from patients with cross-modal extinction. *Neuropsychologia*, 43, 238–248.
- [21] Faugier-Grimaud, S., Frenois, C. and Stein, D. 1978, Effects of posterior parietal lesions on visually guided behavior in monkeys. *Neuropsychologia*, 16, 151–168.
- [22] Ferrari, P. F., Gregoriou, G., Rozzi, S. *et al.*, 2003, Functional organization of the inferior parietal lobule of the macaque monkey. *Soc. Neurosci. Abstr.* 919.7.
- [23] Fogassi, L., Gallese, V., Buccino, G. *et al.*, 2001, Cortical mechanism for the visual guidance of hand grasping movements in the monkey: A reversible inactivation study. *Brain*, 124, 571–586.

- [24] Fogassi, L., Gallese, V., di Pellegrino, G. *et al.*, 1992, Space coding by premotor cortex. *Exp. Brain Res.*, 89, 686–690.
- [25] Fogassi, L., Gallese, V., Fadiga, L. *et al.*, 1996a, Space coding in inferior premotor cortex (area F4): facts and speculations. In F. Laquaniti and P. Viviani (eds.), *Neural Basis of Motor Behavior*. NATO ASI Series. Dordrecht, Kluwer Academic Publishers, pp. 99–120.
- [26] Fogassi, L., Gallese, V., Fadiga, L. *et al.*, 1996b, Coding of peripersonal space in inferior premotor cortex (area F4). *J. Neurophysiol.*, 76, 141–157.
- [27] Fogassi, L., Luppino, G. 2005, Motor functions of the parietal lobe. *Curr. Op. Neurobiol.* 15, 626–631.
- [28] Fogassi, L., Raos, V., Franchi, G. *et al.*, 1999, Visual responses in the dorsal premotor area F2 of the macaque monkey. *Exp. Brain Res.*, 128, 194–199.
- [29] Gallese, V., Murata, A., Kaseda, M. *et al.*, 1994, Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport*, 5, 1525–1529.
- [30] Galletti, C., Battaglini, P. P. and Fattori, P. 1993, Parietal neurons encoding spatial locations in craniotopic coordinates. *Exp. Brain Res.*, 96, 221–229.
- [31] Galletti, C., Fattori, P., Kutz, D. F. *et al.*, 1999, Brain location and visual topography of cortical area V6A in the macaque monkey. *Eur. J. Neurosci.*, 11, 575–582.
- [32] Gentilucci, M., Fogassi, L., Luppino, G. *et al.*, 1988, Functional organization of inferior area 6 in the macaque monkey: I. Somatotopy and the control of proximal movements. *Exp. Brain Res.*, 71, 475–490.
- [33] Gentilucci, M., Scandolara, C., Pigarev, I. N. *et al.*, 1983, Visual responses in the postarcuate cortex (area 6) of the monkey that are independent of eye position. *Exp. Brain Res.*, 50, 464–468.
- [34] Goodale, M. A., Milner, A. D. 1992, Separate visual pathways for perception and action. *Trends in Neurosci.*, 15, 20–25.
- [35] Graziano, M. S. A., Gross, C. G. 1995, The representation of extrapersonal space: a possible role for bimodal visual-tactile neurons. In M. S. Gazzaniga (ed.), *The Cognitive Neurosciences*. Cambridge, MA, MIT Press, pp. 1021–1034.
- [36] Graziano, M. S. A., Hu, X. and Gross, C. G. 1997a, Visuo-spatial properties of ventral premotor cortex. *J. Neurophysiol.*, 77, 2268–2292.
- [37] Graziano, M. S. A., Hu, X. and Gross, C. G. 1997b, Coding the locations of objects in the dark. *Science*, 277, 239–241.
- [38] Graziano, M. S. A., Reiss, L. A. J. and Gross, C. G. 1999, A neuronal representation of the location of nearby sounds. *Nature*, 397, 428–430.
- [39] Graziano, M. S. A., Yap, G. S. and Gross, C. G. 1994, Coding of visual space by premotor neurons. *Science*, 266, 1054–1057.
- [40] Gregoriou, G. G., Borra, E., Matelli, M. *et al.*, 2006, Architectonic organization of the inferior parietal convexity of the macaque monkey. *J. Comp. Neurol.* 496, 422–451.
- [41] Halligan, P. W., Marshall, J. C. 1991, Left neglect for near but not far space in man. *Nature*, 350, 498–500.
- [42] Hyvärinen, J. 1982, Posterior parietal lobe of the primate brain. *Physiol. Rev.*, 62, 1060–1129.
- [43] Hyvärinen, J. 1981, Regional distribution of functions in parietal association area 7 of the monkey. *Brain Res.*, 206, 287–303

- [44] Iriki, A., Tanaka, M. and Iwamura, Y. 1996, Coding of modified body schema during tool use by macaque postcentral neurones. *NeuroRep.*, 7, 2325–2330.
- [45] Johnson, P. B., Ferraina, S., Bianchi, L. *et al.*, 1996, Cortical networks for visual reaching: Physiological and anatomical organization of frontal and parietal lobe arm regions. *Cereb. Cortex*, 6, 102–119.
- [46] Leinonen, L., Hyvarinen, J., Nyman, G. *et al.*, 1979, Functional properties of neurons in lateral part of associative area 7 in awake monkeys. *Exp. Brain Res.*, 34, 299–320.
- [47] Leinonen, L., Nyman, G. 1979, Functional properties of cells in anterolateral part of area 7 associative face area of awake monkeys. *Exp. Brain Res.*, 34, 321–333.
- [48] Li, C. S., Mazzoni, P. and Andersen, R. A. 1999, Effect of reversible inactivation of macaque lateral intraparietal area on visual and memory saccades. *J. Neurophysiol.* 81, 1827–1838.
- [49] Luppino, G., Murata, A., Govoni, P. *et al.*, 1999, Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas AIP and VIP) and the ventral premotor cortex (areas F5 and F4). *Exp. Brain Res.*, 128, 181–187.
- [50] Maravita, A., Iriki, A. 2004, Tools for the body (schema). *Trends in Cog. Sci.*, 8, 79–86.
- [51] Marconi, B., Genovesio, A., Battaglia-Mayer, A. *et al.*, 2001, Eye-hand coordination during reaching. I. Anatomical relationships between parietal and frontal cortex. *Cereb. Cortex*, 11, 513–527.
- [52] Matelli, M., Camarda, R., Glickstein, M. *et al.*, 1986, Afferent and efferent projections of the inferior area 6 in the macaque monkey. *J. Comp. Neurol.*, 251, 281–298.
- [53] Matelli, M., Govoni, P., Galletti, C. *et al.*, 1998, Superior area 6 afferents from the superior parietal lobule in the macaque monkey, *J. Comp. Neurol.*, 402, 327–352.
- [54] Maunsell, J. H. R., Van Essen, D. C. 1983, The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J. Neurosci.*, 3, 2563–2586.
- [55] Mountcastle, V. B., Lynch, J. C., Georgopoulos, A. *et al.*, 1975, Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space. *J. Neurophysiol.*, 38, 871–908.
- [56] Murata, A., Fadiga, L., Fogassi, L. *et al.*, 1997, Object representation in the ventral premotor cortex (area F5) of the monkey. *J. Neurophysiol.*, 78, 2226–2230.
- [57] Raos, V., Franchi, G., Gallese, V. *et al.*, 2003, Somatotopic organization of the lateral part of area F2 (dorsal premotor cortex) of the macaque monkey. *J. Neurophysiol.*, 89, 1503–1518.
- [58] Raos, V., Umiltà, M. A., Murata, A. *et al.*, 2006, Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey. *J. Neurophysiol.*, 95, 709–729.
- [59] Rizzolatti, G., Fadiga, L. 1998, Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5). *Novartis Found Symp.*, 218, 81–95.
- [60] Rizzolatti, G., Fogassi, L. and Gallese, V. 1997, Parietal cortex: from sight to action. *Current Opinion in Neurobiology*, 7, 562–567.

- [61] Rizzolatti, G., Gentilucci, M. and Matelli, M. 1986, Selective spatial attention: one center, one circuit, or many circuits? In M. I. Posner and O. M. Marin (eds.), *Attention and Performance*, Vo. XI. Oxford, Oxford University Press, pp. 251–265.
- [62] Rizzolatti, G., Luppino, G. 2001, The cortical motor system. *Neuron*, 31, 889–901.
- [63] Rizzolatti, G., Luppino, G. and Matelli, M. 1998, The organization of the cortical motor system: new concepts. *Electroencephal. Clin. Neurophysiol.*, 106, 283–296.
- [64] Rizzolatti, G., Matelli, M. and Pavesi, G. 1983, Deficits in attention and movement following the removal of postarcuate (area 6) and prearcuate (area 8) cortex in macaque monkeys. *Brain*, 106, 655–673.
- [65] Rozzi, S., Calzavara, R., Belmalih, A. *et al.*, 2006, Cortical connections of the inferior parietal cortical convexity of the macaque monkey. *Cereb. Cortex*, 16, 1389–1417.
- [66] Schieber, M. 2000, Inactivation of the ventral premotor cortex biases the laterality of motoric choices. *Exp. Brain Res.*, 130, 497–507.
- [67] Shelton, P. A., Bowers, D. and Heilman, K. M. 1990, Peripersonal and vertical neglect. *Brain*, 113, 191–205.
- [68] Thier, P., Andersen, R. A. 1998, Electrical microstimulation distinguishes distinct saccade-related areas in the posterior parietal cortex. *J. Neurophysiol.*, 80, 1713–1735.
- [69] Ungerleider, L. G., Mishkin, M. 1982, Two cortical visual systems. In D. Ingle, M. A. Goodale and R. J. W. Mansfield (eds.), *Analysis of visual behavior*. Cambridge MA, MIT Press, pp. 549–586.
- [70] Wardak, C., Olivier, E. and Duhamel, J. R. 2004, A deficit in covert attention after parietal cortex inactivation in the monkey. *Neuron*, 42, 501–508.

CHAPTER 6

Action and Space Representation

ANNA BERTI* and ALESSIA FOLEGATTI

*Department of Neuropsychology, University of Turin
Via Po 14, 10123 Torino, Italy*

**berti@psych.unito.it*

I. Introduction

One of the more central issues in cognitive neuroscience is how the brain constructs a map of the external world and how this map interacts with the representation of our body, in order to be able to deal with objects placed in the surrounding space. In common language space is defined as ‘the boundless, continuous expanse extending in all directions, within which all material things are contained’ (Webster, 1974). This definition well describes the introspective idea that individuals have of space: something real, fixed and unitary, a kind of ‘container’ in which objects are located.

Is this conventional view correct? Is there indeed a space centre or a space circuit specifically devoted to space perception in the brain? There is now clear evidence that this is not the case.

As already widely reported by Brozzoli and Farnè and by Fogassi in their chapters there are several pieces of evidence both in animal and human studies that space is not homogeneously represented in the brain. Already in 1975 Mountcastle and his co-workers showed that in the monkey inferior parietal lobule there is a ‘neural mechanism generating commands for selective attention to the immediate behavioural surround’ for visual grasping of objects. The idea of the existence of dedicated neural systems for near and far space coding was expanded by Rizzolatti and colleagues (1981a and b, 1983) who proposed, on the basis of electrophysiological findings, a subdivision of the external space into two large sectors: the *peripersonal* or *near* space (coded by neurons in area F4 and VIP) and *extrapersonal* or *far* space (coded in area 8 or FEF and LIP). The peripersonal space is the space for arm and hand action, located around the body and continuing the

personal space. The same neurons that code peripersonal space also code the body surface adjacent to it. On the contrary far space is the region in which oculomotor exploration occurs. It is also the space that can be reached by walking or running or by using tools.

Interestingly, damages of front-parietal circuits, related to near space representation, cause a selective behavioural deficit in near space. In this case monkeys do not respond to stimuli presented around the body, but react in an absolutely normal way to stimuli presented far from the body. On the contrary a monkey with a lesion that selectively affects the circuits related to far space representation can be unaware of the stimuli presented in the contralesional far space, while responding normally to stimuli in near space.

2. Evidence for Discrete Representations of Space in Humans

Indirect evidence of the existence of discrete representations for far and near space in humans has been provided by several neuropsychological studies on brain-damaged patients affected by unilateral neglect (for a similar conclusion in extinction patients see Brozzoli and Farnè, this volume).

Neglect patients do not respond to the presence of contralesional stimuli and do not even explore the contralesional side of space (predominantly the left side, being the lesion located in the right hemisphere), therefore presenting with a behaviour very similar to those described in monkeys with damage to the brain network involved in space representation. In some occasions it is possible to show that the ignored stimuli are, nonetheless, processed up to a high level of analysis (Berti and Rizzolatti, 1992). Neglect has been, therefore, interpreted as an impairment to a complex neural network subserving conscious awareness of space (Berti and Rizzolatti, 2002). Although Bisiach and coworkers (1986) found that right-brain-damaged patients may show neglect for near space (in that case patients did not cancel out the left stimuli drawn on an A4 sheet of paper) but not for personal space and vice versa.

Until rather recently a tacit assumption in neglect studies was that unawareness affects the entire contralesional space. The possibility that unawareness of stimuli could be limited either to the space surrounding the body or to the non-reachable space was not taken into account in clinical evaluation and neglect patients, even now, are usually examined with materials presented only in the space surrounding the body. However, following the seminal studies in animals showing neglect limited to restricted space sectors (Rizzolatti *et al.*, 1983), the possibility of a dissociation between near and far space neglect was addressed also in humans. The first evidence coming from clinical studies were contradictory, showing, as already mentioned, dissociations between near space neglect

and personal neglect (in which what is neglected is the left side of the body) (Bisiach *et al.*, 1986) but not between near and far space neglect (Pizzamiglio *et al.*, 1989).

A clear-cut dissociation between coding of different sectors of extrapersonal space was reported by Halligan and Marshall (1991). They described a patient, who, besides showing neglect in conventional tasks, had also a marked neglect when asked to bisect lines in near space using either an ordinary pen or a projection light pen. However, when the patient was asked to bisect lines presented in the far space by means of a projection light pen, neglect ameliorated and even disappeared. This finding demonstrated that space neglect can be restricted to a specific sector of the external world. A similar dissociation has been described in several following studies (for example Berti and Frassinetti, 2000).

Other studies provided the dissociation opposite to that described by Halligan and Marshall: Cowey, Small and Ellis (1994) investigated bisection of horizontal lines at various viewing distances; in five patients the error was greater for lines further away than for lines of identical angular size within reaching distance. Vuilleumier *et al.* (1998) described a patient with a right temporal haematoma showing marked left visual neglect for far but not near space in a variety of tasks, and even in a reading task.

The evidence that unilateral neglect of far and near visual space may exist independently supports the hypothesis of a segregation of space representations in humans.

If it is possible to show behavioural dissociations between near and far space neglect then it is reasonable to predict different anatomical localisations for space representations. An attempt to localise in normal human volunteers cortical areas active in tasks performed in near and far space was carried out by Weiss *et al.* (2000). Using positron emission tomography (PET), they instructed normal subjects to bisect lines or point to dots in near and far space. The results showed that actions performed in near and far space respectively activated different brain areas. Near space actions determined activation of left dorsal occipital cortex, left intraparietal cortex, and left thalamus, whereas far space actions determined activation of the ventral occipital cortex bilaterally and the right medial temporal cortex. Bjoertomt, Cowey and Walsh (2002) used a TMS (Transcranial Magnetic Stimulation) experiment on normal subjects, inhibiting different part of the posterior cortex. They showed that the inhibitory stimulation of the *right posterior parietal* lobe induces a pseudoneglect (that is a slight leftward bias) in near space, while the stimulation of the *ventral part of right occipital* lobe induces a neglect for far space, therefore suggesting a dorsal (parietal)/ventral (occipital) dichotomy for the anatomical localisation of space coding.

Of greater relevance for the problem of the localisation of the peripersonal space in humans is another fMRI study in which the authors attempted to localise multimodal — tactile, auditory and visual — cortical areas in humans

(Bremmer *et al.*, 2001). Tactile stimuli, moving visual stimuli located near the subject, and auditory stimuli producing the illusion of sound movement were presented. Several cortical areas, related to different stimulus modalities, were found to be activated by the stimuli. However, a multimodal convergence was found only in the depth of the intraparietal sulcus (IPS), in the ventral premotor cortex and in the SII complex. Considering the type of stimuli used (which included the tactile ones), the activated multimodal areas should include the areas coding peripersonal space. The obtained data fit this prediction. On the basis of their location (and properties) the area in the depth of IPS appears to be the homologue of area VIP of the monkeys, while that located in the ventral premotor cortex should be the homologue of area F4. The last area that was activated in this study was SII complex. On the basis of the available evidence in monkeys, it is difficult to draw any firm conclusion on which role (if any) this multimodal area has in coding peripersonal space.

Once established that lesions can selectively affect near and far space, one may go deeper into the issue and ask whether peripersonal and far space coding are sharply separated or there is a continuum between them. This issue was addressed by Cowey *et al.* (1999), using a line bisection test. Lines of different lengths were presented at six different distances from the body (25 cm, 50 cm, 100 cm, 200 cm, 300 cm, 400 cm) to patients affected by neglect. The authors assumed that the border between near and far space should be located at 100 cm. The results showed that in 5 out of 13 patients tested, there was an effect of line presentation distance, the neglect being more severe in far than in near space. The performance impairment was achieved gradually, with no significant difference between individual steps. The only distance that significantly differed from all the others was that at 400 cm. The authors concluded that there is no evidence of a clear border between near and far space. More recently Longo and Lourenco (2006) found a gradual shift in attentional bias on a line bisection task moving from near toward far space and then they went further into the attempt to verify the classical behavioural definition of near space as the space within arm reach. Their results interestingly suggest that there is a systematic relation between the extent of near space and arm length so that arm length may constitute an intrinsic metric for the representation of near space.

3. Re-mapping of Space by Tool Use

In everyday life animals and humans act upon objects. In order to interact with the environment they need to detect, locate, orient to, and reach for the objects they are interested in. All these operations can be distinguished according to the sector of space in which they occur and to the action needed for accomplishing the task.

If the object of interest is located in the immediate surroundings of the body (peripersonal space), manual reaching and grasping can be achieved without locomotion. On the contrary, if the object of interest is placed outside a direct manual

reaching (extrapersonal space) locomotion is needed for subsequent action to the object. Alternatively, the subject can use a tool to reach and grasp for far objects. In any case, we need to encode the position of the objects with respect to the position of the body and body parts, and to activate the adequate spatial map in order to act upon them in a proper way.

If the brain constructs different maps according to far and near space the question is whether 'far' and 'near' are simply derived from the absolute distance of the object from the agent's body or whether the coding of spatial positions is a more dynamic operation that can be influenced by different actions induced by the use of tools that modify the spatial relation between the body and the object. In a seminal paper, Iriki, Tanaka and Iwamura (1996) found in the monkey parietal lobe bimodal neurons that coded the schema of the hand, similar to those studied by Rizzolatti *et al.* (Fogassi *et al.*, 1996; Gentilucci *et al.*, 1988) and by Graziano *et al.* (1994). These neurons fired when a tactile stimulus was delivered to the monkey's hand and when visual objects were presented near the hand tactile receptive field. The most striking characteristic of these neurons was that their visual receptive field was modified, during a reaching movement performed with a rake, to include the entire length of the rake and to cover the expanded accessible space. In other words, in that experiment, the body schema was altered using the tool (Head and Holmes, 1911): the tool was assimilated to the hand, becoming part of the hand representation (Aglioti *et al.*, 1996; Paillard, 1993). Berti and Frassinetti (2000) showed that also in humans a sector of space previously mapped as far on the basis of the reaching distance can be re-mapped as near when the cerebral representation of body space is extended to include objects or tools used by the subject. They studied Patient P.P., who, after a right hemisphere stroke, showed a dissociation between near and far space in the manifestation of neglect. Indeed, in a line bisection task, neglect was apparent in near space, but not in far space when bisection in the far space was performed with a projection light-pen. However, when in far space bisection was performed with a stick, by which the patient could reach the line, neglect appeared and was as severe as neglect in the near space. Like in Iriki *et al.*'s monkey, the use of a stick influenced the patient's computation of space. Indeed, the data can be explained as follows: when the patient used the stick to reach for the object of interest in far space, the tool was coded as part of the patient's hand, as in monkeys, causing an expansion of the representation of the body schema. This affected the spatial relation between far space and the body. The structure of peripersonal space was then altered and peripersonal space was expanded to include the far space reachable by the tool. The reaching of 'far' space with a tool determined a switch between spatial representations, so that the representation of near space was now activated. Because near space representation was affected by the brain damaged, the re-mapping of far space as near affected patient performance in line bisection and neglect reappeared.

The capacity of using tools is, evolutionarily, one of the most important achievements for monkeys and man. By holding a stick, we can reach for objects

that are beyond the limit of our arm without using locomotion. Consequently, the relation between our body and the external objects is modified so that a far object can become near when we can reach for it, no matter what means we use, the hand or a tool.

Recently we described the opposite dissociation i.e., re-mapping of near space into far space when a sensory discontinuity between the patient's limb and the target object was introduced by using a laser pointer for bisecting lines (Neppi *et al.*, 2007). In other words, the use of a tool that prevents the contact with the target object may cause a contraction of peripersonal space and the recoding of a 'near' object as 'far'.

At this point another question we can ask is what kind of sensory input influences the re-mapping. Is it the fact that the subject *sees* the continuity/discontinuity between the body/tool and the object, or is it the fact that subject *feels* the continuity/discontinuity between the body/tool and the object that affects re-coding of space? In other words which source of information, visual or tactile-proprioceptive, about the 'body-tool' complex is critical for determining the enlargement or the contraction of peripersonal space? In a recent experiment (Neppi *et al.*, 2007) we manipulated sensory feedbacks and tool use in far and near space. We found that in order to re-map far space into near space the *presence of tactile proprioceptive* feedback is crucial (i.e., it is crucial to feel the contact between the body/tool complex and the object), whereas in order to re-map near space into far space the *absence of visual feedback* is crucial (i.e., the visual discontinuity between the body/tool complex and the object). We also found that in some patients re-mapping occurs before any sensory feedback is available and it is based on the kind of action induced by a specific tool. In these cases the use of a laser pen for accomplishing the task, inducing a pointing action, always elicited far space activation, whereas the use of a stick, inducing a reaching action, elicited near space activation, independently of the sensory feedbacks.

Therefore our results suggest that sensory feedbacks are not *necessary* for re-mapping space representations. At least in some cases, simply programming an action, depending on the intrinsic functional characteristics of the tool, activates the space that is congruent with the kind of action induced by the tool: pointing actions, which we usually employ to interact with far objects, activate far space, while reaching actions, that we use when a direct interaction with closer objects is possible, trigger near space representation.

4. Space Representation During Walking

Having shown that the activation of space representation can be modulated by actions that change the subject's effective spatial relationship to a target object, another question we asked is whether a similar re-mapping occurs also when far

space is reached not by using a tool but by locomotion (Berti *et al.*, 2002). Neglect patients and brain-damaged patients without neglect were asked to perform two bisection tasks: one was a line bisection task to be accomplished using a projection light-pen, the other was a bisection of a doorway by walking through it. The two tasks were performed at (or starting at) three different distances from the target (3, 1.5, and 0.5 m). We found three neglect patients who showed more severe neglect in far than in near space. Based on this observation, we predicted that if the representation of space is updated during walking, our neglect patients should activate an impaired representation at the beginning of each walking path and a less impaired, or unimpaired, representation toward the end of it, in the two longer distance conditions. Therefore, their walking trajectories should deviate at the beginning of each path, especially with the most distant starting point, but as they approached the doorway, with the activation of better preserved space representations, the walking trajectories should be corrected. As a consequence, the passage through the doorway (i.e., the actual displacement error) should be *similar from all three starting point conditions*. On the contrary, if space is not updated, then the first representation that is activated, at the beginning of each path, will be the one responsible for the final bisection performance. Therefore, if spatial neglect is more severe in far than in near space, we should expect to find worse performance with the starting point in far space (activation of far space) than with the starting point in near space. This is actually what we found. Neglect was more severe when the starting point was at 3 m with respect to the other two starting points. These results are in accordance with the hypothesis that space is not re-mapped during walking.

Our conclusions were that during locomotion, at least for short, linear, and unperturbed trajectories, space representation may not be re-mapped. In our patients the spatial position of the target object was coded at the beginning of the movement, and the error in the bisection computation was produced within the first representation that was activated.

The evidence of the present study was collected in brain-damaged subjects. Therefore, the absence of space re-mapping in the locomotion task might be due to a specific deficit in shifting from one representation to another caused by the lesion. Although we cannot infer from this negative finding (absence of re-mapping) that normal participants also do not remap space during locomotion, we would suggest a similar pattern for normals. We find it very reasonable that for relatively short distances and for unperturbed pathways, our brain constructs a single, stable representation of the spatial position of an object for the execution of a particular task during movement, instead of continuously changing the representation as the participant passes across different sectors of space. It is, however, likely that updating during walking may become necessary for distances greater than those we used, or when a rapid change in some characteristic of the target or a perturbation in the walking trajectory is introduced during the participant's walking.

5. Conclusions

Although our phenomenal experience of space is characterised by a feeling of unity, the neuropsychological and neurophysiological studies just reviewed have demonstrated that the neural systems dedicated to space representation and spatial cognition are implemented in a distributed network where discrete brain areas are devoted to the coding of the different spatial attributes of the stimulus (Berti and Rizzolatti, 2002). This network is greatly modulated by the programming of purposeful actions with different effectors (for instance, an arm reaching movement or a saccade) towards specific object locations. All these circuits compute space, but the computational constraints for programming actions with different effectors are different. Thus, space is computed repetitively for different motor purposes.

The link between space representation and motor system can be interpreted in two ways. The first view is that space is primarily 'sensorial', and the link with the motor system is secondary. According to this view the multiplicity of space representations would indicate only that the motor system requirements influence the space representation, but do not determine it. The opposite view is that space is primarily 'motor' (see Rizzolatti *et al.*, 1997). The existence of a space around the body, anchored to the body parts, and coded in circuits that control body parts movements appears to be a strong argument in favour of a motor nature of space. This is also suggested by the fact that space representation, during development, is constructed through action. There is much evidence showing that movements in space precede sensory information about space. Ecographic studies show that already before birth babies have an extremely rich goal-directed motor activity that indicates the presence of a motor representation of directions well before birth (see Butterworth and Harris, 1994). At birth, the child's movements become more and more goal-directed but obviously related to the space around the body. Because the vision is limited in depth to 20 cm, the children can acquire a representation of peripersonal space, associating the motor knowledge, developed before birth, with the appearance of both his/her hand and some new external objects, in different near positions, without the necessity to disentangle between near and far stimuli. When the conditions of the visual apparatus evolve, the infant starts to receive information from far space. Correlating visual stimuli coming from far with eye/head movements and later with body movements, the child starts to construct far space representation. The phenomenal experience that the children have, while constructing different space representations, is difficult to infer, as is the one of patients with dissociation between near and far space impairments. In any case, whatever is the subjective feeling we may have during the building up of space representations and in pathological conditions, the introspective idea that individuals have of space is that of something fixed and unitary, a 'boundless, continuous

expanse extending in all directions, within which all material things are contained' (Webster, 1974).

References

- [1] Aglioti, S., Smania, N., Manfredi, M. *et al.*, 1996, Disownership of left hand and objects related to it in a patient with right brain damage. *Neuroreport* 8(1), 293–296.
- [2] Berti, A., Frassinetti, F. 2000, When far becomes near: re-mapping of space by tool use. *Journal of Cognitive Neuroscience* 12, 415–420.
- [3] Berti, A., Rizzolatti, G. 1992, Visual processing without awareness: evidence from unilateral neglect. *Journal of Cognitive Neuroscience* 4, 345–351.
- [4] Berti, A., Rizzolatti, G. 2002, Coding near and far space. In: Karnath, H.-O., Milner, A. D. and Vallar, G. (eds), *The cognitive and neural bases of spatial neglect*. Oxford University Press, New York, pp. 119–129.
- [5] Berti, A., Smania, N., Rabuffetti, M. *et al.*, 2002, Coding of far and near space during walking in neglect patients. *Neuropsychology* 16(3), 390–399.
- [6] Bisiach, E., Perani, D., Vallar, G. *et al.*, 1986, Unilateral neglect: personal and extra-personal. *Neuropsychologia* 24(6): 759–767.
- [7] Bjoertomt, O., Cowey, A. and Walsh, V. 2002, Spatial neglect in near and far space investigated by repetitive transcranial magnetic stimulation. *Brain* 125(Pt 9), 2012–2022.
- [8] Bremmer, F., Schlack, A., Shah, N. J. *et al.*, 2001, Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. *Neuron* 29, 287–296.
- [9] Cowey, A., Small, M., and Ellis, S. 1999, No abrupt change in visual hemineglect from near to far space. *Neuropsychologia* 37: 1–6.
- [10] Cowey, A., Small, M., and Ellis, S. 1994, Left visuo-spatial neglect can be worse in far than in near space. *Neuropsychologia* 32, 1059–1066.
- [11] Fogassi, L., Gallese, V., Fadiga, L. *et al.*, 1996, Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology* 76(1), 141–157.
- [12] Gentilucci, M., Fogassi, L., Luppino, G. *et al.*, 1988, Functional organization of inferior area 6 in the macaque monkey. I. Somatotopy and the control of proximal movements. *Experimental Brain Research* 71(3), 475–490.
- [13] Graziano, M. S., Yap, G. S. and Gross, C. G. 1994, Coding of visual space by premotor neurons. *Science* 266(5187), 1054–1057.
- [14] Head, H., Holmes, G. 1911, Sensory disturbances from cerebral lesions. *Brain*, Oxford 34, 102.
- [15] Halligan, P. W., Marshall, J. C. 1991, Left neglect for near but not far space in man. *Nature* 350(6318), 498–500.
- [16] Iriki, A., Tanaka, M. and Iwamura, Y. 1996, Coding of modified body schema during tool use by macaque post-central neurons. *Neuroreport* 7, 2325–2330.
- [17] Longo, M. R., Lourenco, S. F. 2006, On the nature of near space: effects of tool use and the transition to far space. *Neuropsychologia* 44(6), 977–981.

- [18] Mountcastle, V. B., Lynch, J. C., Georgopoulos, A. *et al.*, 1975, Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space. *Journal of Neurophysiology* 38, 871–908.
- [19] Neppi-Modona, M., Rabuffetti, M, Folegatti, A. *et al.*, 2007, Bisecting lines with different tools in right brain damaged patients: the role of action programming and sensory feedback in modulating spatial remapping. *Cortex* 43(3), 397–410.
- [20] Paillard, J. 2003, The Use of Tools by Human and Non-human Primates. In A. Berthelet and J. Chavaillon (eds). Clarendon Press, Oxford, pp. 36–50.
- [21] Pizzamiglio, L., Cappa, S., Vallar, G. *et al.*, 1989, Visual neglect for far and near extra-personal space in humans. *Cortex* 25, 471–477.
- [22] Rizzolatti, G., Fogassi, L. and Gallese, V. 1997, Parietal cortex: from sight to action. *Current Opinion in Neurobiology* 7(4), 562–567.
- [23] Rizzolatti, G., Matelli, M. and Pavesi, G. 1983, Deficits in attention and movement following the removal of postarcuate (area 6) and prearcuate (area 8) cortex in macaque monkeys. *Brain* 106(Pt 3), 655–673.
- [24] Rizzolatti, G., Scandolara, C. Matelli, M. *et al.*, 1981, Afferent properties of periarculate neurons in macaque monkeys. I. Somatosensory responses. *Behavioural Brain Research* 2(2), 125–146.
- [25] Rizzolatti, G., Scandolara, C., Matelli, M. *et al.*, 1981, Afferent properties of periarculate neurons in macaque monkeys. II. Visual responses. *Behavioural Brain Research* 2(2), 147–163.
- [26] Vuilleumier, P., Valenza, N., Mayer, E. *et al.*, 1998, Near and far visual space in unilateral neglect. *Annals of Neurology* 43(3), 406–410.
- [27] Webster, N. 1974, Webster’s New Word Dictionary of the American Language. New American Library, p. 711.
- [28] Weiss, P. H., Marshall, J. C. Wunderlich, G. *et al.*, 2000, Neural consequences of acting in near versus far space: a physiological basis for clinical dissociations. *Brain* 123(Pt 12), 2531–2541.

CHAPTER 7

The Space Representations in the Brain

CLAUDIO BROZZOLI and ALESSANDRO FARNÈ
INSERM, U864, Espace et Action, Bron, F-69500 France
and
Université Claude Bernard Lyon I, Lyon, 69003
16, Avenue du Doyen Lepine
69676 Bron Cedex, France
alessandro.farne@inserm.fr

Overview

This chapter reviews several highly convergent behavioural findings that provide strong evidence in favour of the existence in humans of different representations of space. In particular, we review here a series of neuropsychological studies on patients and behavioural studies on healthy subjects that show how different sectors of space can be differentiated from each other on the basis of the way in which they reflect the processing of multisensory information coming from the space around us. These findings are consistent with the functional properties of multisensory neuronal structures coding (near and far) peripersonal space in animals. This high level of convergence ultimately favours the idea that multisensory space coding is achieved through similar multisensory interaction in both humans and non-human primates. Recent findings about the plasticity of this space processing are also presented, suggesting a dynamic role of the space representation for the planning and execution of action.

1. Introduction

When we speak about space, we speak about a special object of experience that could be better described as a way to perceive the objects around us. ‘Spatial perception’ refers to the analysis of the spatial relations between different events out

of the observer's body or between these events and the observer's body or between different events on the body itself. The motor behaviour and the sensory input provide the necessary information to build the spatial representations. Since all different sensory information arrives separately to the brain, at a certain level of the spatial processing all inputs must be functionally linked and share common reference systems. Visual input are firstly coded in retinal coordinates, somatosensory input in somatotopic coordinates and the motor feedback is given by the reciprocal position of joints. However, all events seem to occur in the same given portion of the external 'real' space. We still need to locate the objects accurately in space. That is to say, objects must occupy only one position, irrespective of how many sensory channels are in use to perceive them. The idea of an object without a position in the space, or occupying two different portions of space simultaneously, is inconceivable (Làdavias, Berti & Farnè, 2000). The integration between different inputs coming from visual, auditory and tactile receptors allows for a multisensory representation of the space that can be used to plan, perform and correct the ongoing motor behaviour.

Thus, even if we perceive the space as something continuously defined and unitary represented, as the geometrical definition of Descartes' space we are used to, this derives from the 'perceptual space', that is a biologically determined space composed by multiple representations, functionally built on the basis of the behaviour we can perform in the environment (Craighero *et al.*, 1999; Farnè *et al.*, 2005).

From a phenomenological point of view, the motor behaviour suggests three different portions of space: the personal, the reaching and the extrapersonal space. The personal space is occupied by the body itself. Its representation is built from proprioceptive and tactile information mainly, that in each moment update the central nervous system about the position of the different body parts in the space and their relative orientation. The visual input about the body might also contribute to the representation of the personal space. The extrapersonal space representation is principally based on visual and auditory inputs that convey information from the far space. Finally, within the extrapersonal space but proximal to the body there is a region of space, called the reaching space, which is functionally defined according to the distance at which an object can be reached by the subject's hand without moving his/her trunk. This phenomenological distinction is corroborated by neurophysiological studies on animals and neuropsychological human studies that will be described in what follows. Moreover, from a physiological point of view, a more subtle distinction can be made within the extrapersonal space and consists of a spatial area immediately surrounding the body parts that has been called peripersonal space. The peri-personal space is further characterised by the high degree of multisensory integration between visual and tactile information, not present for farther regions of space, and can thus be added to the classical triadic space taxonomy. Indeed, neurophysiological findings in animal studies allow such a strict definition of peripersonal space, which limits it to an area of a few centimetres around the

body, thus differentiating it from the reaching space. Neuropsychological studies on brain damaged patients provided converging evidence of such a discontinuity in the spatial representation in humans, further showing the strong modularity of space representation.

In the human and non-human primates' central nervous system, several cerebral areas receive convergent multisensory afferent inputs necessary to build up such a representation of the space. Below, we briefly review the main multisensory structures involved in the spatial representation that have been discovered firstly in non-human primates.

2. The Multisensory Bases of the Space Representation

The superior colliculus is one of the cerebral areas involved in the spatial representation on the bases of the multisensory information present in the environment. Neurophysiological studies in the cat and monkey have revealed the existence of multisensory neurons in this region and in particular, the superior colliculus shows a single neuron convergence of multisensory signals: in fact, in this area single cells discharge in response to a visual or an auditory or a tactile stimulus (Stein and Meredith, 1993). Three different sensory maps are present in the colliculus, one for each modality. It is thus able to provide the correct coordination of the head and eyes movements towards the events present in the environment, integrating the visual, auditory and somatosensory inputs. Besides the superior colliculus, which seems to be particularly involved in the multisensory coding of relatively far space (Làdavas & Farnè, 2004), one of the most relevant structures of the brain involved in the spatial perception is the parietal lobe. In such an area, multisensory integration arises and, owing to its functional properties, contributes to the definition of the region of space near to the body, the peripersonal space. This region is defined by the relation of a stimulus in the near space, coded in the visual modality and the somatosensory receptors on the body. Among the first evidence of an interaction between visual and tactile events in the body proximity, it is worth mentioning the neurophysiological work of Hyvarinen and Poranen (1974), who recorded the single neuron activity in the monkeys' parietal lobe, in particular in the area 7. They found different classes of neurons whose activity varied as a function of the presented stimulus. Interestingly, they described a class of neurons which were preferentially activated by a tactile stimulus presented on a particular body part (i.e., where the tactile receptive field of that neuron was located), but also by a visual stimulus, if presented near the same body part, in correspondence with the tactile receptive field of the neuron. This characteristic of the parietal neurons in area 7 is the neurophysiological basis of the visuo-tactile integration, which thus occurs at single-neuron level.

The first systematic study of visual-tactile neurons, however, has been conducted by Rizzolatti and colleagues (Rizzolatti *et al.*, 1981a,b) in an anterior

region of the monkey brain, namely the premotor cortex (area F4). Most of the neurons the authors recorded, responded to stimuli in one or two sensory modalities. Accordingly to the particular modality activating the neurons, they were classified as somatosensory, visual or bimodal (visual and somatosensory) neurons. Visual neurons were located rostral to the arcuate sulcus (area 8, or FEF), whereas the somatosensory and the bimodal neurons were found predominantly caudal to this sulcus (area F4). The parts of the body most represented were the hands and the mouth. The neurons located rostral to the arcuate sulcus, predominantly visual neurons, were indeed activated by stimuli presented far from the animal. The neurons found caudally to the arcuate sulcus were maximally or even exclusively activated by stimuli presented in the space immediately around the animal. These neurons were bimodal, responding also to somatosensory stimuli. According to the location of their visual responding regions the bimodal neurons were subdivided into pericutaneous (54%) and distant peripersonal neurons (46%). The former responded best to stimuli presented a few centimetres from the skin, the latter to stimuli within the animal's reaching distance. The visual responding regions were spatially related to the tactile fields. Therefore, an important property of these neurons, as other cells in different multisensory areas (see below) is that the extent of their visual receptive fields is limited in depth to a few centimetres (in most cases from ~ 5 to ~ 50 cm) out of the tactile ones. For example, a neuron might have a tactile receptive field on the palmar surface of the monkey hand and a visual receptive field 'coming out' of the tactile one by ~ 30 cm. In this case, when an object is visually presented in front of the animal hand laying palm-up in front of him on a table, the neuron will respond (by increasing the frequency of action potentials) if the object is located above the hand, within the distance of about 30 cm. If the same object is presented at further distance above the hand (e.g., 50 cm), or close to the hand (e.g., 20 cm) but on the side of its palmar surface (e.g., on the left or right instead of above it), the neuron will respond with a much lower frequency, or not at all. This means that, in order to be an appropriate stimulus for such a class of neuronal cells, not only the visual object has to be in the vicinity of the body part where the tactile receptive field is located, but also in the correct position with respect to it. In other words, it is not sufficient for a visual stimulus to be in the 'reaching space' to make visual-tactile neurons to code for it (see Figure 1). This fine selectivity is the neurophysiological basis of the distinction between two spatial representations: the near peripersonal space and the far peripersonal space.

Moreover, when the hand and the arm are moved through the visual space of the monkey, the visual receptive fields follow the body part, as they are anchored to the tactile receptive field of a specific body part. When the arm or the hand are out of the visual space of the monkey, the bimodal neurons activity is largely reduced, suggesting that the vision of the body part is more important than proprioception to code the peripersonal space associated to a hidden body part (Graziano and

Multisensory representation of peripersonal space: Convergent findings in monkey and humans

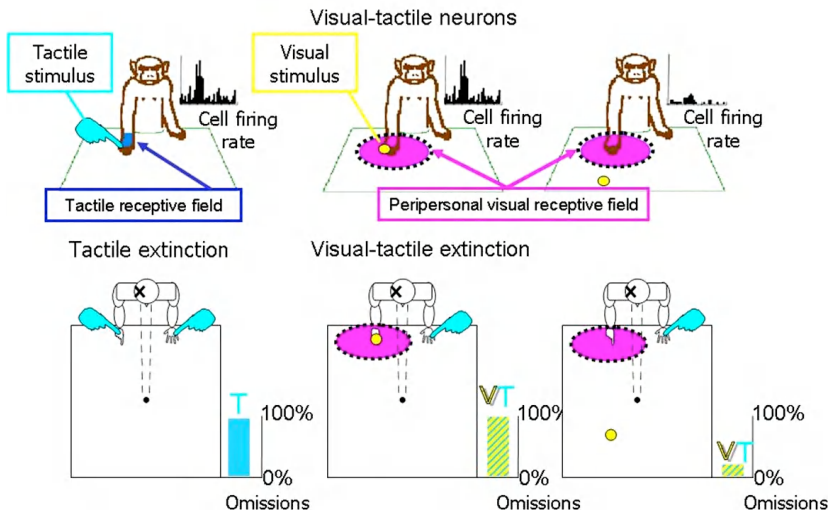


Figure 1 In agreement with the neurophysiological findings reported in the text, and illustrated in the upper row, the phenomenon of multisensory visual-tactile extinction in humans is spatially selective; extinction of a touch delivered to the contralesional hand is significantly more severe when the ipsilesional visual stimulus is presented close to (about 5 cm), than far from (about 35 cm) the patient's ipsilesional hand. This spatial selectivity suggests the existence, in the human brain, of integrated visual-tactile systems coding for near-peripersonal space, functionally analogous to the one described in monkeys.

In the upper row is shown the firing rate (histograms) of a monkey's bimodal neuron. This visuo-tactile neuron is activated by a simple tactile stimulation on a specific body part (e.g., the hand), as well as visual stimulation (the yellow dot in the picture) when it is delivered in the space close around the same body part. The same bimodal neuron is not more activated if the same visual stimulation is presented far from the body part that is out of the visual receptive field (pink area in the picture).

In the lower row, the picture shows the phenomenon of cross-modal extinction in humans. These patients classically omit the contralesional touch in case of double simultaneous stimulation (left panel). A similar phenomenon is shown when an ipsilesional visual stimulus is presented simultaneously to the tactile one in the contralesional side: the visual stimulation extinguishes the tactile one (central panel). However, if the same visual stimulation is delivered far from the body-part, that is out of the functionally defined peripersonal space, patients' tactile detection improves in the contralesional side. These findings suggest that a common mechanism may be shared by monkeys and humans in the building of a representation of the peripersonal space on the basis of the visuo-tactile integration (modified from Maravita & Iriki, 2004 & Farnè *et al.*, 2005).

Gross 1993; 1995). More recently, other researches have confirmed and developed these findings, showing that the integration of visual and tactile information is the basis to build peripersonal space representations that are not only related to the hand, but also to other body parts, such as the face. Indeed, also for the face it is possible to distinguish a region of peripersonal space from the farther one. In particular, some neurons in the ventral intraparietal area (VIP) have visuo-tactile receptive fields mostly localized on the animal's face and head. With the same mechanism as for other body parts, the multisensory VIP neurons may thus build a multisensory representation of the head-centred peripersonal space (Duhamel *et al.*, 1997; Avillac *et al.*, 2005).

One of the most influential researches about bimodal neurons' characteristics investigated the dynamic properties of their visual receptive fields. Iriki and colleagues (Iriki *et al.*, 1996) studied the bimodal neurons of the post-central parietal gyrus, somewhat extending into the intraparietal sulcus, that code for the peripersonal space of the hand-arm in monkeys. These authors found that the extent of the visual receptive field of the bimodal neurons is not fixed but can be expanded. Indeed, Iriki and colleagues' monkeys were trained to use a rake to reach for some food pellets placed out of the animal's hand-reaching space. In consequence of the tool-training, the visual receptive fields of the bimodal neurons coding for the hand peripersonal space mapped after the use of the rake were expanded in the direction of the tool-tip, such that the tool appeared to be included within the enlarged visual receptive field. Moreover, a few minutes after the training with the tool was paused, the visual receptive field area changed again, apparently shrinking back to the original size. These modifications were not observed if the rake was only passively held, but not used, by the animal. These findings suggest that the expansion of the visual receptive field of such visuo-tactile neurons that followed tool use was not merely due to the modification induced by the fact of holding a rake: that is, the visual and static proprioceptive information of the animal holding the rake was not enough to induce any significant change of the size of the visual receptive field. For such a change to occur, it seems necessary that the tool is actively employed to perform an action. In other words, the dynamic aspect depends on the execution of a specific motor action (Rizzolatti *et al.*, 1998). In a similar vein, Fogassi and colleagues (Fogassi *et al.*, 1996) have also found that the visual receptive fields of visuo-tactile neurons in area F4 expand when the visual stimulus velocity increases while approaching the cutaneous receptive field, a property that could be relevant for the preparation of action towards nearby stimuli. The approaching or escaping nature of the action could be partly determined by the characteristics of the visual stimulus (see Farnè *et al.*, 2003). Recently, the idea that multisensory-motor interfaces might code defensive movements received some preliminary support. Electrical stimulation of precentral multisensory areas seems to evoke complex avoidance or defensive reactions, such as withdrawal of the hand, turning of the head or lifting the hand as if to defend the side of the head

(Graziano *et al.*, 2001). It would thus be adaptive that responses possibly evoked by multisensory neurons are fast and mainly outside the control of top-down mechanisms. However, the tool-use studies that have been recently made available in the animal and human literature suggest that these multisensory-motor interfaces might also code for the execution of purposeful movements, aimed at voluntarily act on objects.

To resume, the bimodal neurons code for two different sensory modality: the tactile and the visual modality (see, for the auditory modality, Farnè & Làdavas, 2002). The main functional characteristics by which these neurons are able to integrate these two modalities are summarised below.

- The tactile and visual receptive fields roughly overlap in space.
- The visual receptive field is limited in depth to a few centimetres.
- Typically, the visual receptive field of the bimodal neurons, functionally defined as the region of space where a visual stimulus activates this particular neuron, is anchored to the tactile receptive field. Therefore, if the latter moves in the space, the visual one follows it.
- A bimodal neuron is activated by a visual stimulus if it is presented close to the correspondent tactile receptive field. Roughly, three classes of neurons can be differentiated as a function of the maximum distance at which a visual stimulus activates the neuron: (a) less than 5 cm; (b) less than 35/45 cm; (c) more than 1 metre. The most part of the bimodal neurons is included in the two first classes.
- The neuronal response evoked by visual stimulation is a function of the distance of the visual stimulus from the tactile receptive field. That is, besides the selectivity described above, the neuronal activation, in terms of spike frequency, increases in most cells as the distance of the visual stimulus from the body decreases.
- Somatotopic representation in multisensory areas. Multisensory areas of different regions, taken together, form a complete, although not ordered somatotopic representation of the body (area VIP has a detailed representation of the monkeys head/face/torso, area F4 mostly of the hand/arm/torso). Since tactile and visual receptive fields largely overlap, a multisensory map of the space around the body is provided. The most represented body parts are the face and the hand/arm.
- The spatial coordinates system is body parts centred. The visuo-tactile region position is independent of the orbital eye position. Thus, these neurons do not code the peripersonal space in retinal coordinates, but mainly in somatotopic, body parts centred coordinates, although several level of transition among the reference systems can be found in some areas of this circuit (e.g., area VIP).

These characteristics allow the bimodal neurons to build representations of the near peripersonal space.

3. Several Representations of the Space in the Human Brain

Most parts of information about the cerebral areas involved in the representations of the space in humans come from the neuropsychological studies on brain damaged patients. Neuropsychology represents a natural interface among diverse disciplines such as neuroanatomy, neurophysiology, cognitive psychology and computational modelling. The main source of information for neuropsychologists consists of the pathological behavioural phenomena that become manifest following damage to the central nervous system. By describing, examining, quantifying and classifying altered cognitive abilities in humans after brain damage, the neuropsychological approach may contribute substantially to our understanding of the normal organization of brain functions (Shallice, 1991).

As we discussed before, one of the most important cerebral structure involved in the construction of the spatial representations is the right parietal cortex. Evidence of the parietal cortex involvement in spatial representations in humans comes from the different syndromes which can follow a parietal lesion. One of the most complex spatial deficits is known as the Balint–Holmes’ syndrome, consisting in four main symptoms: gaze apraxia (inability to generate voluntary saccades), optic ataxia (inability to reach and grasp an object in peripheral vision), neglect (Balint, 1909) and a deficit in perceiving distances (Holmes, 1919). Holmes in particular, reported that patients failed to describe spatial characteristics of the objects they could still recognize, showing a deficit in localization of the objects and in the judgement of their distance from the body. Post mortem analysis of these patients, revealed a bilateral lesion involving the posterior parietal cortex.

Another syndrome following parietal cortex damage is the Gerstmann’s syndrome (Gerstmann, 1930). Characteristic components of this deficit, among others, are the finger agnosia (inability to recognize and name the different fingers) and left-right confusion. Gerstmann’s syndrome patients thus show a deficit in the recognition of body parts itself, suggesting that the parietal lobe is not only implied in peripersonal and extrapersonal space representation but also in the “personal” space.

The phenomenon that more than others provided information about the modularity of the spatial representations in humans in different sensory modalities is spatial neglect (simply neglect hereinafter) (Brozzoli *et al.*, 2006). Neglect is a complex syndrome associated to a lesion of the right inferior parietal lobule, at the parieto-temporo-occipital carrefour. Patients affected by neglect show characteristic deficits. In the acute phase they show a deviation of gaze and the head towards the ipsilesional side; they are often hemiplegic in the contralesional side of the body and they seem not perceive neither explore the contralesional side of the personal and extrapersonal space; a reduction of activity toward the contralesional hemispace is also reported. The presentation of a visual stimulus or a person speaking on the patient’s left side is not attended to and does not produce any orientation

behaviour. Thus, in everyday life these patients do not care about the left side of their body, shaving only the right side of their face, for example. In addition to this lack of attention to the contralesional side, they show defective performance also for stimuli presented in the right side (Snow & Mattingley, 2006) as well as non-spatially lateralised deficits (Husain & Rorden, 2003). Clinically the presence of the deficit is assessed by the use of tests as the cancellation, the line bisection or drawing from memory or on copy. In the first mentioned, the patient is provided with a sheet of paper where several target or distractor stimuli are printed. The middle of the sheet is aligned with the central meridian of the patient's body and he/she is asked to individuate and mark with a pen as many target stimuli as he/she can, avoiding distractors. Typically, neglect patients can mark only the stimuli placed on the right (ipsilesional) side of the sheet, omitting the contralesional ones. This kind of behaviour is present also in other tasks involving visuo-spatial perception. When asked to copy an object by drawing it, neglect patients usually report only the ipsilesional elements. In a line bisection task, patients are asked to report the middle of a line. The misperception of the left side of space biases the middle point rightwards. Thus, these patients behave as they could not perceive or attend to the left side of their body and the space out of their body.

This description of the syndrome already shows how the representation of the space in the brain is not continuous, since the left and the right representation of the space can be differentially damaged by the parietal lesion. The most interesting dissociation in terms of space representations shown by neglect patients is the disruption of either the peripersonal or the extrapersonal space representation. Several studies (Halligan & Marshall, 1991; Cowey *et al.*, 1994; Berti & Frassinetti, 2000) provided the evidence of this double dissociation. Halligan & Marshall, showed in a group of four patients the presence of neglect in the near but not in the far space, whereas Cowey and colleagues showed in a group of five patients the opposite dissociation in a line bisection task. In an elegant study, Berti and Frassinetti (2000) reported the case of a patient who presented with a severe neglect when asked to perform a line bisection task in the near space, as in the described classical procedure. However, when asked to perform the same task marking the same stimuli but placed 1 metre far from her, by the use of a laser pointer, she did not present neglect symptoms. The different behaviour this patient showed in the two spaces corroborates the evidence of the dissociation between the representations of a near and a far space. While no definite answer has yet been given to the anatomical counterpart of these behavioural dissociations in humans, the work by Rizzolatti and colleagues (1981; 1983) described above provided a neurobiological support to the distinction between peripersonal and extrapersonal space in monkeys. While neurons in area F4 responded to somatosensory and visual stimuli, provided that they were presented within monkeys' peripersonal space, those in area FEF responded when the same visual stimuli were located farther away, in the extrapersonal space. Accordingly, unilateral ablation of area F4, or

FEF provoked contralesional visual neglect for objects located, respectively, in the monkey's peripersonal, or extrapersonal space.

These findings are good evidence in favour of the presence in the brain of separated but interconnected representations for different sectors of space in humans as those described in non-humans primates. An interesting question is whether a similar mechanism underlies the construction of these representations across the two species. In this respect, the study of a neuropsychological condition called 'extinction' provided considerable insight into the behavioural characteristics of multimodal spatial representation in humans. Extinction (Loeb, 1885; Oppenheim, 1885) is a pathological sign following brain damage whereby patients may fail to perceive contralesional stimuli only under conditions of double (contra- and ipsilesional) simultaneous stimulation (Bender, 1952), thus revealing the competitive nature of this phenomenon (di Pellegrino & De Renzi, 1995; Driver, 1998; Duncan, 1980; Ward *et al.*, 1994).

A number of studies have shown that extinction can emerge when concurrent stimuli are presented in different sensory modalities, i.e., different sensory inputs delivered to the ipsi- and contra-lesional side of the patient's body. Tactile extinction, for example, can be modulated by visual and auditory events simultaneously presented in the space region near the tactile stimulation, increasing or reducing tactile perception, depending upon the spatial arrangement of the stimuli. In a series of studies, we tested whether the presentation of a visual stimulus in the right ipsilesional field could extinguish the tactile stimulus presented on the contralesional hand, which was otherwise well detected by patients when presented alone. The prediction of these studies was that if a multisensory (visuo-tactile) system processing tactile and visual stimuli near the body is in charge of coding left and right spatial representations, then delivering visual stimuli close to a body part (≤ 7 cm, i.e., in the near-peripersonal space) would be more effective in producing cross-modal visual-tactile extinction than presenting the same visual stimuli at larger distances (≥ 35 cm, i.e., in the far-peripersonal space).

The results of these studies confirmed the presence of stronger cross-modal visual-tactile extinction when visual stimuli were displayed in the near- as compared to the far-peripersonal space. These findings were taken as providing a strong neuropsychological support to the idea that the human brain represents near-peripersonal space through an integrated multisensory visuo-tactile system. Owing to this system's activity, the somatosensory representation of the ipsilesional hand may be activated by the nearby presentation of a visual stimulus, thus competing with the contralesional hand representation activated by a tactile stimulus. Since the competition is biased in favour of the ipsilesional side in extinction patients, the ipsilesional visual stimulus appears to extinguish the contralesional stimulus presented in a different modality. This would be due to the fact that the processing of the somatosensory stimulation of the contralesional hand is

disadvantaged in terms of competitive weights, bearing a comparatively weaker representation.

To assess which reference frame is used to code multisensory near-peripersonal space, a patient with left tactile extinction who was asked to cross the hands, so that the left hand was in the right hemispace and the right hand in the left hemispace (di Pellegrino *et al.*, 1997). In such a crossed-hand situation, a visual stimulus presented near the right hand (located in the left space) still extinguished tactile stimuli applied to the left hand (now located in the right hemispace). Thus, visual-tactile extinction was not modulated by the position of the hands in space, as far as the spatial correspondence between sensory modality and the stimulated hand was kept constant (i.e., visual stimulus-right hand/tactile stimulus-left hand). This finding, by showing that the visual peripersonal space remains anchored to the hand even when it is moved in another hemi-space, strongly suggests that the near-peripersonal space is at least partially coded in a hand-centred coordinate system. The pattern of results observed in the case of visual-tactile stimulation of the hand is consistent with the functional properties of the multisensory system that has been described in monkeys, further suggesting that human and non-human primates might share, at some level, similar cerebral mechanisms for near space representation.

The multisensory representation of space is anchored neither to a mere 'bodily point', nor to the body as a whole but to specific body-parts, in this case the hand. This raises the question of whether humans represent near-peripersonal space not only in relation to hands, but also to other body parts. In this respect, as described above, the neurophysiological findings revealed a somatotopic distribution of multisensory neurons' receptive fields, which are known to be mostly located on the animal hand/arm, trunk, and face. The latter neurons seem to be particularly relevant for the coding of near-peripersonal space, since a specific multimodal area of the parietal lobe (VIP) is mainly devoted to representing space near the face (Colby *et al.*, 1993; Duhamel *et al.*, 1991; 1998). On this basis, we reasoned that a multisensory mechanism, similar to that operating in the case of the hand, might also be involved in representing near-peripersonal space in relation to the human face. Therefore, we followed the same rationale to investigate whether the presentation of ipsilesional visual stimuli might modulate left tactile extinction also at the level of the face (Làdavias *et al.*, 1998). Similarly, we expected that cross-modal extinction would be stronger by presenting visual stimuli near, as compared to far from, the patients' face. This hypothesis has been assessed in a group of patients with right brain damage presenting left tactile extinction and was clearly supported. As for the hand, visual stimuli presented to the ipsilesional side produced a decrease in the detection of contralesional tactile stimuli, particularly when visual stimuli were presented near the ipsilesional cheek. In this near-peripersonal condition, patients reported only few touches of the left cheek, while these stimuli were otherwise

well perceived when delivered alone. The extinction phenomenon was much less severe when visual stimuli were delivered far from the face; in the far peripersonal condition, patients were able to report the majority of contralesional touches, thus confirming the existence of a representation of multisensory peripersonal space also relative to the humans face/head.

All together, these studies suggest that multisensory representations of space are coded within the near-peripersonal space of the face and the hand, and these representations might differ from those controlling visual information in the far-peripersonal space (Farnè & Làdavas, 2002; Làdavas, 2002).

4. Multiple Representations of Peripersonal Space

Does the modular organisation of space, which seems to operate as a general principle governing spatial perception, also apply to the representation of the near-peripersonal space? By referring to the Graziano and Gross' metaphor of near-peripersonal space as a 'gelatinous medium' surrounding the body, we asked whether this would be a unitary and homogeneous sector of space encompassing the whole body, or an ensemble of modules separately representing the space immediately adjacent to a given body part. We recently tested this unitary vs. modular representation hypothesis (Farnè *et al.*, 2005). As the two hypotheses make opposite predictions, we contrasted them directly by investigating cross-modal visual-tactile extinction in a group of right brain damaged patients. We reasoned that, if the unitary hypothesis were true, then tactile stimuli delivered on the contralesional hand would be comparably extinguished by ipsilesional visual stimuli irrespective of the stimulated body part (either the hand or the face), provided that the visual stimulus were presented near the body. Alternatively, if near-peripersonal space is represented in a modular way, then tactile stimuli delivered on the contralesional hand would be more severely extinguished when ipsilesional visual stimuli are presented near the homologous body part (i.e., the right hand), than near the non homologous body part (i.e., the right side of the face). The two hypotheses also differed with respect to the near-far modulation of cross-modal extinction since its presence in the case of stimulation of non homologous sectors would support the unitary hypothesis, whereas its absence would favour the modular organization hypothesis. The results showed a visual-tactile extinction stronger for homologous than for non homologous combinations and showed that the effect was selectively present when visual stimuli were presented near the ipsilesional side of the patients' body. In sharp contrast, when visual stimuli were presented far from the ipsilesional side of the patients' body, the amount of visual-tactile extinction obtained in homologous and non homologous combinations was comparable. By extending to this peculiar sector of space the principle of the modular space organisation, these findings support the view that different multisensory representations are coded

within the near-peripersonal space of the hand and the face. Further support to this view has been recently provided by neuroimaging findings showing a human parietal face area representing head-centred visual and tactile maps (Serenio & Huang, 2006).

5. Multisensory Representation of Peripersonal Space for Action

The neurophysiological and neuropsychological findings reviewed above converge in showing that peripersonal space is structured in far and near peripersonal space sectors, the latter being specifically coded in a multisensory, body part-centred and modular manner. So far, these considerations allow for a fine-grained description of the structure of space and of its anchoring to the body. This provides the adequate basis to ask further questions about the determinants of such a spatial structure. Specifically, we will ask the following questions: Is the extension of the peripersonal space fixed in space or can it be modified? If it can be modified, what are the conditions of such a modulation? Is a simple change of our visual body-image sufficient to dynamically re-map far space as near, or is some kind of sensori-motor activity necessary to produce this re-mapping? In what follows, we review empirical investigations of the specific manner in which space can be structured by the perceiver's own action.

We described recent neurophysiological animal studies which have examined whether the near-peripersonal space of monkeys' hands, and especially its spatial extension and location, might be modified through different kinds of sensorimotor experience. The question at stake is whether a passive change of the corporeal configuration is sufficient, or whether some goal-directed activity is needed. So far, this question has been investigated by considering the effect of tool use on the extension of the peripersonal space (Iriki *et al.*, 1996; 2001; Obayashi *et al.*, 2000). Tools enable human beings and other animals to manipulate objects that would otherwise not be reachable by hands. Acting on distant objects by means of a physical tool requires sensory information that is mainly provided by vision and touch. The expansion of the peri-hand area whereby vision and touch are integrated would render the possibility of reaching and manipulating far objects as if they were closer to the hand.

In particular, a re-coding of relatively far visual stimuli as nearer ones has been observed in monkey single-cells studies, after extensive training in using a rake to retrieve distant food, thus extending the hand's reachable space by connecting the animal's hand with objects located outside its reaching distance. A few minutes of tool-use induced an expansion of visual RF of visual-tactile neurons recorded in the parietal cortex. This rapid expansion along the tool axis seemed to incorporate the tool into the peri-hand space representation. The extended visual RF contracted

back to the pre-tool-use dimension after a short rest, even if the monkey was still passively holding the rake (Iriki *et al.*, 1996). No modification of the visual RF was ever found if the monkey was just passively holding the tool. Therefore, the tool-use related expansion of the visual RF was strictly dependent upon the active use of the rake to reach distant objects.

A similar effect of re-coding of visual stimuli located in far-peripersonal space, as if they were closer to the participants' body, has been documented behaviourally in right brain-damaged patients with tactile extinction (Farnè & Làdavas, 2000). In this study, the amount of cross-modal visual-tactile extinction was assessed by presenting visual stimuli far from the patients' ipsilesional hand, at the distal edge of a 38 cm-long rake passively held in their hand. The patients' performance was evaluated before tool-use, immediately after a five minutes period of tool-use, and after a further five to ten minutes resting period. To control for any possible effect due to directional motor activity, cross-modal extinction was also assessed immediately after a five minutes period of hand pointing movements. We found that far visual stimuli induced more contralesional tactile extinction immediately after tool-use (retrieving distant objects with the rake), than before tool-use, when they just hold the rake passively. This evidence of an expansion of the peri-hand space lasted a few minutes after tool use. After the resting period, the severity of cross-modal extinction was back to pre-tool-use levels, suggesting that the spatial extension of the hand's near-peripersonal space contracted back towards the patients' hand. Finally, no change in cross-modal extinction was found immediately after the execution of control pointing movements toward the same distant objects. Closely related evidence comes from the study of Berti and Frassinetti already recalled above, whereby neglect symptoms were limited to the near space and did not extend to the far space when the patient was asked to bisect far lines (1 metre away) with a laser-pen. Very interestingly, the authors asked the patient to perform the same task but this time by the use of a wooden stick, that was in spatial continuity with the hand of the patient. Surprisingly, when using this tool to reach the lines placed in the far space, the patient showed a rightward bisection bias, as severe as in the near space. Therefore, the near and far space are separately represented but the codification of what is near and what is far is not absolutely but functionally defined on the basis of how the body needs to interact with the objects in the space, that is the behaviour that has to be performed in the environment. In the Berti and Frassinetti's study, the use of a tool in rigid continuity with the hand in order to reach objects placed in the far space induced a re-mapping of this region as a near space region, with the consequence that the disruption of the left side representation was then evident in the far (but behaviourally near) space. This is an elegant demonstration of the incongruity between the geometrically defined and the behaviourally represented space.

More recently, Longo and Lourenco (2006) studied whether the transition between near and far space is gradual or to the contrary they are abruptly separated, using the same tool paradigm as in the Berti and Frassinetti's study, in

a group of healthy subjects. In fact, also normal subjects present a slight bias in a bisection task, showing a leftward bias in the near and a slight rightward bias in the far space. When employing the stick to reach and mark the presented lines, subjects always showed the same amount of leftward bias, independently of how far the lines were presented. Also, when using the laser pointer, participants showed a gradual shift of the bias from the left to the right, suggesting that the brain gradually code the transition from the near to the far space.

Specifically considering the role played by passive or active experience in reshaping peripersonal space, the results of a recent study (Farnè *et al.*, 2005a) were clear in showing that a relatively prolonged, but passive experience with a tool is not sufficient to induce such a dynamic re-mapping of far space as near space. Indeed, passive exposure to the proprioceptive and visual experience of wielding a rake did not alter the severity of visual-tactile extinction, which was found to be comparable to that obtained when the tool was actually absent. This favours the idea that plastic modifications of the structure of peripersonal space are not the product of passive changes in proprioceptive/kinesthetic, or visual inputs per se. An artificial extension of our reachable space by a hand-held tool would not necessarily imply a phenomenon of tool incorporation, unless the tool is used in some active way. Indeed, when cross-modal extinction was assessed equally far in space, but immediately after the active use of a long tool, we observed a significant increase of cross-modal extinction. These findings considerably extend our knowledge about dynamic tool incorporation in humans, by making clear that the plastic modifications are tightly linked to the active, purposeful use of a tool as physical extension of the body, which allows interactions with otherwise non-reachable objects.

These findings underline that the representation of space is neither static nor passive. Rather, the structure of space is specifically build up transiently in a body-related way thanks to processes of sensorimotor integration. The data just reviewed suggest that tool-use can change space representation both in normal subjects and in brain damaged patients. In particular, a passive change of the corporeal configuration (hand+tool) is not sufficient: some goal-directed activity is needed. These results raise a further question concerning the critical determinant of the extent to which peri-hand space increases. Does this depend upon the physical, absolute length of the tool, or the operative length of the tool that can be effectively used to act on objects?

In this respect, the differential amount of cross-modal extinction obtained with different tools was not determined by the absolute length of the tool, but by its operative length (Farnè *et al.*, 2005a). These results favour the notion that peri-hand space elongation is directly related to the functionally effective length of the tool, i.e., by the distance at which the operative part of the tool is located with respect to the hand. Importantly, this coheres with the aforementioned functional reshaping of spatial representation.

This functional perspective raises a last question concerning the actual shape of the expansion of the peri-hand space after tool use: does it consist in an elongation of the multisensory integrative area along the axis of the tool or in a shift of these proprieties to the tip of the tool? In other terms, does the functional expansion of the peri-hand space after tool use consist in an actual elongation of the visual-tactile integrative area along the tool axis, encompassing the whole tool axis (Farnè & Làdavas, 2000; Farnè *et al.*, 2005a,b), or in a selective incorporation of the specifically functional part, i.e., the tool-tip, due to the shift or the creation of a new integrative area at the distal edge of the used tool (Holmes *et al.*, 2004)? This issue has been addressed by assessing cross-modal visual-tactile extinction in a right-brain damaged patient while she was wielding a 60 cm long rake, before and immediately after its use to retrieve distant objects (Farnè *et al.*, 2007). At variance with previous patients studies, visual-tactile extinction was assessed near the ipsilesional hand (holding the rake handle), near the distal edge of the rake, as well as in a middle position between the hand and the distal end of the rake. Following the tool-use training, visual-tactile extinction increased both at the distal edge as well as midway between the hand and the tool-tip, no change being observed near the hand. This result, which has been recently confirmed and extended on a group-study base (Bonifazi *et al.*, 2007) suggests that after tool use the visuotactile peri-hand space is expanded to incorporate the whole tool rather than being displaced to a restricted area around the tip of the tool. In summary, neurophysiological and neuropsychological findings converge in showing that the strength of multisensory coding of peri-hand space can be modified along the axis of a tool to include its length, the re-mapping being achieved through a functional re-sizing of the peri-hand area where visual-tactile integration occurs.

This finding is coherent with the general property of near-peripersonal space recalled above: it is coded in a body part-centred manner. Here we see that its functional elongation does not amount to detaching near-peripersonal space from its bodily anchoring. Rather, the reported data suggest that the functional modulation of spatial representation involved a modification of the functional body itself: the latter remains anchored to the effector (the hand), but is elongated to include all relevant functional parts (all along the tool up to its functional part). These considerations link the coding of peripersonal space to the body schema (Head & Holmes, 1911). In this respect, Iriki's and colleagues' seminal paper (Iriki *et al.*, 1996) suggested that the tool was 'embodied', thus inducing a modification of the body schema. The body schema concept (Head & Holmes, 1911) refers to the representation of the body in the space, closely related to the action behaviour, and is usually differentiated from the body-image, a more conscious representation of the body, related to the conscious experience of visual, tactile and motor information of corporal origin (Head & Holmes, 1912; Paillard, 1991; Sirigu *et al.*, 1991). The two points of view are not mutually exclusive since, in principle, both modifications might arise as a consequence of the use of a tool. The central point

is to understand if the two concept of peripersonal space and body schema are really separable concepts. Although, it is logically conceivable a modification of one of them leaving unchanged the other, no evidence is present in literature about this issue. May the body schema and the peripersonal space be conceived of as the two faces of the same concept? The former, classically action-related, would be referred to the structure of the body in order to perform an action and the modification of this structure when the body is performing an action. The latter could also be action-oriented and referred to the multisensory space around the body, whose input could be used to perform the action.

6. Conclusion

We are used to thinking of the space around us as a whole continuum, as the geometrical Descartes' definition of space, where the objects are available for our actions. However, the space is represented in the brain in a different way: several spatial representations are built up for different regions of the real space. As we reviewed in the first part of this chapter, these different representations can be differentiated from each other on the basis of the sensory input they are built on. Thus, three main regions of space in respect to the body can be differentiated: a personal space, that is the body itself, principally based upon the somatosensory input (proprioceptive and tactile modality; but also important is the vision of the body, which contributes to the definition of this region of space); a peripersonal space region, that is the area around the body where tactile and visual information are massively integrated; finally, the extrapersonal region of space, whose representation is principally based on the "tele-sensory" modalities (vision and audition). The most important issue we described in this chapter is the evidence of the action dependent modifications of the multisensory integration. In fact, is the intention to perform a particular action, the criterion the brain adopts to code different areas as near to or far from the body. The representations of the space in the brain are thus dynamical, since they can be updated as a function of the action we are demanded to perform and the tool that we may use to reach the goal of the action.

Acknowledgements

This work was supported by the European Mobility Fellowship and the AVENIR project funding N° R05265CS.

References

- [1] Avillac, M., Deneve, S., Olivier, E. *et al.*, 2005, Reference frames for representing visual and tactile locations in parietal cortex. *Nat Neurosci.*, 8(7), 941–949.

- [2] Balint, R. 1909, Seelenlahmung des “Schauens”, optisce Ataxie, raumliche Störung der Aufmerksamkeit. *Monatsschrift für Psychiatrie und Neurologie*, 25, 51–81.
- [3] Bender, M. B. 1952, Disorders in perception. Springfield, IL: Charles C. Thomas.
- [4] Berti, A., Frassinetti, F. 2000, When far becomes near: remapping of space by tool use. *Journal of Cognitive Neuroscience*, 3, 415–420.
- [5] Bonifazi, S., Farnè, A., Rinaldesi, L. *et al.*, 2007, Dynamic size-change of peri-hand space through tool-use: spatial extension or shift of the multisensory area? *Journal of NeuroPsychology*, 1, 101–114.
- [6] Brozzoli, C., Dematte, M. L., Pavani, F. *et al.*, 2006, Neglect and extinction: within and between sensory modalities. *Restor Neurol Neurosci.*, 24(4-6), 217–232.
- [7] Colby, C. L., Duhamel, J. R. and Goldberg, M. E. 1993, Ventral intraparietal area of the macaque: anatomic location and visual response properties. *Journal of Neurophysiology*, 69, 902–914.
- [8] Cowey, A., Small, M. and Ellis, S. 1994, Left visuo-spatial neglect can be worse in far than in near space. *Neuropsychologia*, 32, 1059–1066.
- [9] Craighero, L., Fadiga, L., Rizzolatti, G. *et al.*, 1999, Action for perception: a motor-visual attentional effect. *J Exp Psychol. Hum. Percept. Perform.*, 25(6), 1673–1692.
- [10] di Pellegrino, G., De Renzi, E. 1995, An experimental investigation on the nature of extinction. *Neuropsychologia*, 33, 153–170.
- [11] di Pellegrino, G., Làdavas, E. and Farnè, A. 1997a, Seeing where your hands are. *Nature*, 388, 730.
- [12] Driver, J. 1998, The neuropsychology of spatial attention. In Pashler H. (Ed) *Attention*, pp. 297–340. Hove: Psychology Press.
- [13] Duhamel, J. R., Bremmer, F., BenHamed, S. *et al.*, 1997, Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*. 389(6653), 845–848.
- [14] Duhamel, J. R., Colby, C. L. and Goldberg, M. E. 1991, Congruent representation of visual and somatosensory space in single neurons of monkey ventral intra-parietal area (VIP). In Paillard J. (Ed) *Brain and Space*, pp. 223–236. New York: Oxford University Press.
- [15] Duhamel, J. R., Colby, C. L. and Goldberg, M. E. 1998, Ventral intraparietal area of the macaque: congruent visual and somatic response properties. *Journal of Neurophysiology*, 79, 126–136.
- [16] Duncan, J. 1980, The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, 87, 272–300.
- [17] Farnè, A., Làdavas, E. 2000, Dynamic size-change of hand peripersonal space following tool use. *Neuroreport*, 85, 1645–1649.
- [18] Farnè, A., Làdavas, E. 2002, Auditory peripersonal space in humans. *J Cogn Neurosci.*, 14(7), 1030–1043.
- [19] Farnè, A., Dematte, M. L. and Làdavas, E. 2003, Beyond the window: multisensory representation of peripersonal space across a transparent barrier. *Int J Psychophysiol.*, 50(1–2), 51–61.
- [20] Farnè, A., Dematte, M. L. and Làdavas, E. 2005, Neuropsychological evidence of modular organization of the near peripersonal space. *Neurology*, 65(11), 1754–1758.
- [21] Farnè, A., Iriki, A. and Làdavas, E. 2005, Shaping multisensory action-space with tools: evidence from patients with cross-modal extinction. *Neuropsychologia*, 43(2), 238–48.

- [22] Farnè, A., Serino, A. and Làdavas, E. 2007, Dynamic size-change of peri-hand space following tool-use: determinants and spatial characteristics revealed through cross-modal extinction. *Cortex*, 3, 436–443.
- [23] Fogassi, L., Gallese, V., Fadiga, L. *et al.*, 1996, Coding of peripersonal space in inferior premotor cortex (area F4). *J. Neurophysiol.*, 76, 141–157.
- [24] Gerstmann, J. 1930, Zur Symptomatologie der Hirnläsionen im Übergangsgebiet der unteren Parietal- und mittleren Occipital- windung. *Nervenarzt* 3, 691–695.
- [25] Graziano, M. S. A., Gross, C. G. 1993, A bimodal map of space: somatosensory receptive fields in the macaque putamen with corresponding visual receptive fields. *Exp Brain Res.*, 97(1), 96–109.
- [26] Graziano, M. S. A., Gross, C. G. 1995, The representation of extrapersonal space: a possible role for bimodal, visuo-tactile neurons. In M. S. Gazzaniga (Ed) *The Cognitive Neuroscience*, pp. 1021–1034. Cambridge, MA: MIT Press.
- [27] Graziano, M. S. A., Taylor C. S. R. and Moore T. 2001, Electrical stimulation of the bimodal, visual-tactile zone in the precentral gyrus evokes defensive movements. *Soc. Neurosci. Abs.*, 129.8.
- [28] Halligan, P. W., Marshall, J. C. 1991, Left neglect for near but not far space in man. *Nature*, 350, 498–500.
- [29] Head, H., Holmes, G. 1911, Sensory disturbances from cerebral lesions. *Brain*, 34, 102–254.
- [30] Holmes, G. 1919, Disturbances of visual space perception. *British Medical Journal*, 2, 230–233.
- [31] Holmes, N. P., Calvert, G. A., Spence, C. 2004, Extending or projecting peripersonal space with tools? Multisensory interactions highlight only the distal and proximal ends of tools. *Neurosci Lett.*, 372(1–2), 62–67.
- [32] Hyvarinen, J., Poranen, A. 1974, Function of the parietal associative area 7 as revealed from cellular discharges in alert monkeys. *Brain*, 97, 673–692.
- [33] Husain, M., Rorden, C. 2003, Non-spatially lateralized mechanisms in hemispatial neglect. *Nat Rev Neurosci.*, 4, 26–36.
- [34] Iriki, A., Tanaka, M. and Iwamura, Y. 1996, Coding of modified body schema during tool use by macaque postcentral neurones. *Neuroreport*, 7, 2325–2330.
- [35] Iriki, A., Tanaka, M., Obayashi, S. *et al.*, 2001, Self-images in the video monitor coded by monkey intraparietal neurons. *Neuroscience Research*, 40, 163–173.
- [36] Làdavas, E. 2002, Functional and dynamic properties of visual peripersonal space in humans. *Trends Cog Sci.*, 6, 17–22.
- [37] Làdavas E., Berti A. and Farnè A. 2000, Dissociation between conscious and non-conscious processing in neglect. In Rossetti Y. & Revounsoo A. (Eds.) *Beyond dissociation: interaction between dissociated implicit and explicit processing. Advances in Consciousness Research Vol. 22*, pp. 175–193. Philadelphia: John Benjamins B.V.
- [38] Làdavas, E., di Pellegrino, G., Farnè, A. *et al.*, 1998, Neuropsychological evidence of an integrated visuo-tactile representation of peripersonal space in humans. *Journal of Cognitive Neuroscience*, 10, 581–589.
- [39] Loeb, J. 1885, Die elementaren Störungen einfacher Funktionen nach oberflächlicher, umschriebener Verletzung des Grosshirns. *Pfluger Archiv*, 37, 51–56.
- [40] Longo, M. R., Lourenco, S. F. 2006, On the nature of near space: effects of tool use and the transition to far space. *Neuropsychologia*, 44(6), 977–981.

- [41] Obayashi, S., Tanaka, M. and Iriki, A. 2000, Subjective image of invisible hand coded by monkey intraparietal neurons. *Neuroreport*, 11, 3499–3505.
- [42] Oppenheim, H. 1885, Ueber eine durch eine klinisch bisher nicht verwertete Untersuchungsmethode ermittelte Form der Sensibilitätsstörung bei einseitigen Erkrankungen des Grosshirns. *Neurologisches Centralblatt*, 4, 529–533.
- [43] Paillard, J. 1991, Motor and representational framing of space. In Paillard J. (Ed.), *Brain and space*. New York: Oxford University Press.
- [44] Rizzolatti, G., Fadiga, L. 1998, Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5). *Novartis Found Symp.*, 218, 81–95; discussion 95–103.
- [45] Rizzolatti, G., Scandolara, C., Matelli, M. *et al.*, 1981, Afferent properties of periarculate neurons in macaque monkeys. I. Somatosensory responses. *Behav. Brain Res.*, 2(2), 125–146.
- [46] Rizzolatti, G., Scandolara, C., Matelli, M. *et al.*, 1981, Afferent properties of periarculate neurons in macaque monkeys. II. Visual responses. *Behav. Brain Res.*, 2(2), 147–163.
- [47] Sereno, M. I., Huang, R. S. 2006, A human parietal face area contains aligned head-centered visual and tactile maps. *Nat Neurosci.*, 9(10), 1337–1343.
- [48] Shallice, T. 1991, *Précis of From neuropsychology to mental structure*. Behavioral and Brain Science, 14, 429–469. US: Cambridge University Press.
- [49] Sirigu, A., Grafman, J., Bressler, K. *et al.*, 1991, Multiple representations contribute to body knowledge processing. Evidence from a case of autotopagnosia. *Brain*. 114 (Pt 1B), 629–642.
- [50] Snow, J. C., Mattingley, J. B. 2006, Goal-driven selective attention in patients with right hemisphere lesions: how intact is the ipsilesional field? *Brain*, 129(Pt 1), 168–181.
- [51] Stein, B. E., Meredith, M. A. 1993, *The merging of the Senses*, Cambridge, MA: MIT Press.
- [52] Ward, R., Goodrich, S. and Driver, J. 1994, Grouping reduces visual extinction: neuropsychological evidence for weight-linkage in visual selection. *Visual Cognition*, 1, 101–129.

CHAPTER 8

The Enactive Constitution of Space

CORRADO SINIGAGLIA^{*,‡} and CHIARA BROZZO^{*,†}

**Department of Philosophy, University of Milan
via Festa del Perdono 7,
I-20122 Milano, Italy*

*†St Hilda's College, University of Oxford,
Cowley Place, OX4 1DY Oxford, UK*

‡corrado.sinigaglia@unimi.it

I. Introduction

Over the last few years, an increasing intertwining between philosophical and neuroscientific research has occurred, often promoting a proper paradigm shift in the study of the basic aspects of cognition. This is particularly true for the study of the cortical motor system, which has shown to be involved in a growing number of cognitive functions that used to be typically attributed to high-order processes.

The constitution of peripersonal space is one of the most relevant and intriguing ways in which the motor system displays its cognitive function. Through the exploration of this issue, our paper aims to demonstrate the urgency of a paradigm shift towards an enactive approach to cognition.

This paper will be structured into three main sections. In the first one, a review will be given of the empirical evidence supporting the notion of peripersonal space as distinct from personal and extrapersonal space. The defining traits of peripersonal space will be shown to consist in its being *multisensory* (i.e., based on the integration of visual, tactile and proprioceptive information), *body-centred* (encoded not in retinal, but in somatic coordinates) and *motor* in nature. The second section will focus on the role that bodily movements play in constituting space. This idea is not new in philosophy, as we will prove by looking at Husserl and Poincaré. Although they provide one of the most sophisticated analyses of space constitution, we will argue that this is yet not sufficient to grasp the motor nature of peripersonal

space as *near, reachable* space. The bottom line will be that we should acknowledge the primacy of action with respect to movement in shaping space, given that the former but not the latter succeeds in accounting for the *near* versus *far* dichotomy. The third section will show that only by adopting an *enactive* approach to cognition one can fully understand the motor constitution of peripersonal space. Its dynamic plasticity, that is, its varying range will be argued to depend on the actual reach of motor goals and actions.

2. Peripersonal Space as Body-centred and Multisensory Space

Peripersonal space is usually defined as the space that encompasses objects within reach. It differs from *personal* (or *cutaneous*) space as well as from *extrapersonal* (or *far*) space, that is, the space traditionally described as that outside the body and including objects which are beyond our immediate reach and that one can get close to enough only by locomotion.

There is a large consensus that the neural circuit involved in encoding peripersonal space is mainly formed by two areas: area F4, which lies in the caudal-dorsal portion of the ventral premotor cortex, and the ventral intraparietal area (VIP). Electrical microstimulation showed that neck, mouth and arm movements are represented in the area F4 (Gentilucci *et al.*, 1988; Fogassi *et al.*, 1996a). Moreover, recordings of single neurons indicated that the majority of F4 neurons become active both during the execution of motor acts (such as reaching, orienting and facial movements) and in response to sensory stimuli (Rizzolatti *et al.*, 1981a, b); consequently these neurons have been subdivided into two groups: ‘somatosensory’ neurons and ‘somatosensory and visual’ neurons, known also as *bimodal* neurons (Fogassi *et al.*, 1992; 1996a, b; Graziano *et al.*, 1994).

Most of the F4 *somatosensory* neurons are activated by superficial tactile stimuli: a caress or the sensation of something brushing against the skin is all that is needed to trigger them. The somatosensory receptive fields of these neurons are located on the face, neck, arms and hands; these fields are fairly extensive, covering areas that extend over a number of square centimetres. The somatosensory characteristics of the *bimodal* neurons are similar to those of the pure somatosensory neurons, but they are triggered also by visual stimuli, particularly by three-dimensional objects. Most of them are susceptible to moving objects (especially objects that are moving towards the body), although some neurons respond strongly to stationary objects (Fogassi *et al.*, 1996a; Graziano *et al.*, 1997).

Besides these properties, the particularly interesting functional aspect of most F4 bimodal neurons is that they respond to visual stimuli *only* when these appear in the proximity of their tactile receptive fields — more precisely, within that specific portion of space which represents their *visual* receptive field and appears to be an

extension of their somatosensory receptive field. The shape and size of these visual receptive fields differ, with a depth ranging from just a few centimeters to 40–50 cm. For this reason, the same neuron that discharges when the experimenters brush the monkey's forearm also becomes active when they move their hand close to the animal's forearm, entering its visual receptive field. If you find this hard to believe, bring your hand close to your cheek: you will feel it before your fingers actually touch the skin. It is almost as if the personal (i.e., cutaneous) space of your cheek reaches out to embrace the visual space that surrounds it. Visual and somatic stimuli are here more than just 'equivalent'. As Alain Berthoz stated: '[Spatial-visual] [P]roximity is a form of anticipated contact with the area of the body that will be touched' (Berthoz 1997: 78). Our body uses this form of 'anticipated contact' to define its surrounding space, locating its effectors (arm, mouth, neck, etc.) and the objects that are in their visual proximity, regardless of whether they are at rest or in motion.

What is crucial for our purposes is that the visual responses of most F4 bimodal neurons are independent of the direction of gaze (Gentilucci *et al.*, 1983; Fogassi *et al.*, 1996a, b). This has been established by a series of elegant single neuron recordings testing the visual properties of F4 bimodal neurons in four different conditions. In the first one the monkey was fixating a point directly in front of it, while at the same time a visual stimulus entered the visual receptive field of the recorded neuron. The neuron discharge timing and modalities clearly show that activation started when the stimulus was at a distance of approximately 40 cm from the animal. In the second condition, the monkey's gaze was still directed at a point in front of it, but the stimulus moved to the opposite side of the fixation point, outside the visual receptive field of the recorded neuron. In this case the latter did not become active. In the third condition, the monkey was fixating a point at, say, 30 degrees to the left with respect to that of the first two conditions. The visual stimulus was approaching the animal through a physical path that was identical to that of the first condition. Although the trajectory followed by the stimulus was totally different in terms of retinal coordinates with respect to that followed in the first condition, the neuron's visual response was practically the same as that recorded in the first condition. In the last condition, the monkey continued to fixate the same point as in the third condition, but the stimulus was moved to the opposite side of the fixation point, as in the second condition. If the receptive field were retinocentrically coded, the response of the neuron should be similar to that of the first condition, but in fact the results showed that the neuron did not become active (just like in the second condition).

Overall, these experiments show that the visual responses of F4 bimodal neurons do not depend on the position of the stimulus on the retina. If this were the case, when the monkey moved its gaze, the visual receptive field should have shifted accordingly, but these experiments demonstrate that this was not the case. In addition to the fact that the coordinates of the F4 visual receptive fields are not

retinocentric, more recent experiments have shown that these receptive fields do not relate to a single reference frame located in a specific part of the body such as the head or the shoulders. On the contrary, there is a manifold of visual reference frames each centred on the corresponding somatosensory field, and this makes it possible to locate the visual stimuli in the space surrounding the bodily parts to which they are linked (for more details see Rizzolatti & Sinigaglia, 2008; and Fogassi in this volume).

In order to realize the relevance of the latter point, imagine for a while that you are fixating a point on the keyboard of your computer. If you raise your eyes to the computer screen in front of you, the visual receptive fields anchored to the somatosensory receptive fields around your mouth and forearm remain in the same position as before. Now, if you turn to look at the cup of coffee on your right to pick it up, these visual receptive fields move. This does not depend on the direction of your gaze, but on the position of your head and forearm. For the sake of simplicity we have referred to just two of the many receptive fields present in your body and to only one form of movement (rotation of the trunk and the head). But you do possess many visual and somatosensory receptive fields, some of which cover the area where the cup is located. But what happens when you move your hand towards the cup? Irrespective of the direction of your gaze, the position of the cup with respect to your hand, forearm, etc. is specified by the appropriate visual and somatosensory receptive fields. Their stimulation anticipates the actual contact with your skin, so that your hand does not have to physically touch the cup to 'know' where it is. It is sufficient for your hand to be close enough to trigger these neurons through their visual receptive fields. As these fields are a three-dimensional extension of the respective somatosensory fields, the visual individuation of that cup should initiate the specific movements of your arm that propel your hand towards it just as if it were a tactile stimulus, without any need to convert the visual coordinates into any other form (which would be extremely complex and onerous).

3. The Role of Bodily Movements in Constituting Space

The idea that bodily movements play a key role in the constitution of space is certainly not a novelty, particularly not for philosophers investigating the origin of space representation *more geometrico*. We are referring, for example, to Jules-Henri Poincaré and Edmund Husserl, but of course we could have quoted many other thinkers such as Hermann von Helmholtz and Federigo Enriques. Along the lines of what a great mathematician such as René Thom maintained, we believe that Poincaré and Husserl have understood better than anyone the motor roots of space representation (see Thom, 1990).

Indeed, in the *Foundations of Geometry* (1898) as well as in *Science and Hypothesis* (1902), Poincaré states that 'a motionless being could not have acquired

[the concept of space], because not being able to correct by his movement the effects of the change of position of external objects, he would have had no reason to distinguish them from [qualitative] changes of states' (1902, 1952: 58). By pointing out the critical distinction between the change of our impressions due to an object's displacement and the change of our impressions due to its qualitative change of state, Poincaré lays the foundations of our ability to represent space in terms of our sense of movement: 'It may therefore happen that we pass from the aggregate of impressions *A* to the aggregate *B* in two different ways. First, involuntarily and without experiencing muscular sensations — which happens when it is the object that is displaced; secondly, voluntarily, and with muscular sensations — which happens when the object is motionless, but when we displace ourselves in such a way that the object has relative motion with respect to us. If this be so, the translation of the aggregate *A* to the aggregate *B* is only a change of position. It follows that sight and touch could not have given us the idea of space without the help of the "muscular sense"' (1902, 1952: 58).

As far as Husserl is concerned, not only the *Thing and Space Lectures* (1907), but also the manuscripts devoted to a 'systematic analysis of space constitution' (1916) are worth mentioning. In these texts, Husserl tries to demonstrate that not only is movement at the basis of the constitution of various sensory spaces (visual and tactile), but also that this motor constitution is by no means unitary, but presupposes a number of frames of reference (e.g., the oculomotor system, the system of head movement around the basic axis, the complete cephalomotor system, and so on) and geometrically different spaces (e.g., respectively a delimited plane space, cylindrical field of vision; Riemannian space, and so on) (Husserl 1997; on this subject let me refer to Giorello & Sinigaglia, 2007; and Sinigaglia, 2000). Indeed, Husserl brings our attention to the fact that the movements of the eyes, of the head, and of the upper body constitute different kinesthetic systems, each of which is a 'system of power [*System der Vermöglichkeit*]: its 'basic directions of modification' are determined by the null-position of the system, which accordingly has the form of a 'coordinate system of orientation' (Husserl 1997: 282).

Although Poincaré's and Husserl's considerations are worthwhile insofar as they highlight some key features of the constitution of space, still they do not enable us to grasp the most original aspect about the nature of peripersonal space as it emerges from the recent neurophysiological research. In order to obtain a better appreciation of this point, it is indeed sufficient to compare what we know about the VIP-F4 neurons underpinning the constitution of peripersonal space with what we know about the frontal eye field (FEF) and the lateral intraparietal (LIP) neurons, which in turn are motor in nature and play a role in encoding visual space. As is well known, the FEF-LIP neurons control the rapid eye movements (*saccadic*), whose function is to bring the fovea onto targets located at the periphery of the visual field. Like the VIP-F4 neurons, they respond to visual stimuli and discharge

in relation to particular types of movement. However, no further similarities exist. In fact, the LIP-FEF neurons

- (i) respond to a visual stimulus independently of the distance at which it is located;
- (ii) their visual receptive fields are retinocentrically encoded (i.e., each field has its specific position on the retina relative to the fovea);
- (iii) their motor properties concern eye movements only (see Andersen *et al.*, 1997; Colby & Goldberg, 1999).

On the other hand, the VIP-F4 neurons

- (i) are mostly bimodal and respond more strongly to three-dimensional objects than to simple luminous stimuli;
- (ii) their receptive fields are coded in somatic coordinates and anchored to various parts of the body;
- (iii) *last but not least*, the visual stimuli must appear *close* to the bodily parts to which their visual and somatosensory receptive fields are anchored.

Although being both motor in nature and as such playing a role in the constitution of space, VIP-F4 and LIP-FEF encode different kinds of space, peripersonal and extrapersonal space respectively, typically also called *near* and *far* space. The distinction between these two kinds of space has been corroborated by a series of studies on deficits following lesions of FEF and F4 in the monkey. Unilateral lesion of F4 impaired reaching movements and, what's more, produced neglect for visual and tactile stimuli appearing in the contralateral near space (Rizzolatti *et al.*, 1983; Schieber, 2000; Fogassi *et al.*, 2001). Lesion of FEF prevented the monkey from moving its eyes toward the visual stimuli presented in the contralateral far space, whereas it did not present any deficits in the contralateral near space (Rizzolatti *et al.*, 1983; Li *et al.*, 1999; Wardak *et al.*, 2004).

A similar distinction between near and far space has been observed in human patients affected by spatial neglect. In a reported case, the patients' neglect was more severe in their peripersonal space than in their extrapersonal space (Halligan & Marshall, 1991; see also Berti, Frassinetti, 2000; Berti, Rizzolatti, 2002). The opposite form of neglect was also recorded, in which the impairment of the patients' extrapersonal space was much more severe than that of their peripersonal space (Cowey *et al.*, 1994; see also Cowey *et al.*, 1999; Vuillemier *et al.*, 1998; Frassinetti *et al.*, 2001). This has been proved to hold also for patients with visuo-tactile extinction. These patients typically suffer from a right hemisphere brain damage. They can detect a single touch on their left or right hand in isolation, but if two (tactile or visual/tactile) stimuli are presented simultaneously, one to their right hand and the other to their left one, only the right stimulus can reliably be detected. It has been shown that as soon as a visual stimulus was presented *close* to the right

hand of some patients, they no longer perceived the tactile stimulus delivered to their left hand. Most interestingly, when the visual stimulus was shown *outside* the patients' peripersonal space, the visual extinction effect on their sense of touch was very weak, or absent altogether (di Pellegrino *et al.*, 1997; see also Brozzoli and Farnè in this volume).

4. Near and Far: How Action Shapes Space

Just the comparison between the functional properties of the areas VIP-F4 and LIP/FEF, as well as the hints previously given to some lesion studies in humans, clearly show that the very distinction between *near* and *far* cannot be interpreted in purely metric terms. Therefore, taking into account bodily movements is not *per se* sufficient for clarifying the nature of peripersonal space and to shed light on the distinctive trait of such space, that is what makes it a different space from both personal and extrapersonal ones. Because of this, it's no coincidence that the *near* versus *far* distinction isn't at all crucial either in Husserl's analysis or in Poincaré's (or rather, as for the latter, not at the time of *Science and Hypothesis*, 1902), as well as it's not relevant for nearly everyone who investigated the origin of space representation *more geometrico*. As a matter of fact, in spite of being *motor in nature*, peripersonal space is marked by a *dynamic plasticity* which specifies it as such and distinguishes it from any other form of space, as we will see below. Such dynamic plasticity cannot be accounted for in terms of mere bodily movements, and this forces us to realize the primacy of action, that is, the primacy of the motor goal-relatedness which identifies each basic action as such, characterizing it as something more than a sequence of bodily movements.

On the other hand, how else could one explain the multiplicity of reference systems anchored to the different bodily parts? And how could one explain the fact that F4 bimodal neurons receptive fields differ by extension and reach? The above reviewed data clearly indicate that both the multisensoriality and also the body-centredness of peripersonal space have to do with the possibility to act. And it is precisely the relation to the motor goal-relatedness of a motor act, as well as the range of such motor goal-relatedness, which enables one to grasp not only the nearness of the points belonging to the peripersonal space, but also and above all its *dynamic plasticity*, that is, the fact that the extension of one's near space is not fixed but can change. The parameters upon which this varying extension depends will be shown to crucially depend on the *variable reach of motor acts*. Therefore, in order to fully understand the constitution of peripersonal space, we need to take into account *motor acts* as opposed to mere bodily movements, that is, we have to adopt an *enactive* (instead of *motor*) perspective with respect to the genesis of space.

This point may be fully appreciated by first considering that *an increase in the speed at which the stimulus approaches has been shown to expand the receptive*

fields of F4 bimodal neurons in depth (Fogassi *et al.*, 1996a). This means that rapidly approaching stimuli are signalled while they are still at a greater distance from the body, in comparison with stimuli approaching more slowly. The advantage is quite obvious: the earlier the neuron discharges, the earlier the motor act it codes is evoked. This enables an efficient mapping of what is really *near*, thus permitting to either take advantage of an opportunity or to avoid a threat. Similar results can be found in humans. Chieffi *et al.* (1992) asked subjects to reach for and grasp a sphere approaching them, and across different trials the speed of the object was varied. When speed was higher, participants moved their forelimb earlier in time and farther than at lower speed.

Further evidence supporting the dynamic plasticity of peripersonal space is given by a series of studies on how tool use can extend the multisensory coding of peripersonal space into extrapersonal space. In a seminal experiment, Iriki *et al.* (1996; see also Ishibashi *et al.*, 2000) showed that the visual receptive fields of a monkey's parietal neurons, which code hand movements in a similar fashion to F4 neurons, can be modified by actions involving tool use. They trained monkeys to retrieve pieces of food with a small rake, and observed that, when the instrument was used repeatedly, the receptive fields anchored to the hand expanded to encompass the space around both the hand and the rake. If the animal stopped using the rake, but continued to hold it, the animal's receptive fields shrunk back to their normal extension.

Analogous results have been found in healthy and brain-damaged humans. It has been shown that reaching a visual stimulus with one's hand or with a tool produced similar interference effects: in the latter case, these effects depended on the tool but not on the hand posture, and they increased with extensive tool-use (Maravita *et al.*, 2002). Moreover, several line-bisection studies on patients with selective neglect for the hemispace close to (or far from) their body indicated that tool use might reduce or increase the neglect according to the status of the line to be bisected (reachable or out-of-reach) in relation to tool use. Such dynamical re-mapping was modulated both by the planned motor act and by tactile and visual feedback received during the execution of that act (Berti & Frassinetti, 2000; Pegna *et al.*, 2001; Ackroyd *et al.*, 2002; Neppi-Mòdona *et al.*, 2007; see also Folegatti and Berti in this volume). Finally, studies on patients with visuo-tactile extinction selectively confined to the space close to one hand showed that the severity of the extinction can be modified by tool use, which extends the reach of hand actions (Farnè & Làdavas, 2000; Maravita *et al.*, 2001). This extension has been demonstrated to be tightly related to the functionally effective length of the tool (Farnè *et al.*, 2005; see also Brozzoli and Farnè in this volume).

Taken together, these findings show that one's multisensory peripersonal space can be extended differentially by using tools, and this appears to be a further corroboration of the *enactive* character of near space, that is, near space is not only motor but also *goal-centred* in nature.

Note that this concept was touched on by Ernst Mach, who, in his *Knowledge and Error*, wrote that ‘the points of physiological space’ are nothing other than ‘the goals of various movements’ (Mach 1905, 1967: 260). However, Poincaré was the first one who fully acknowledged the multisensory and enactive nature of peripersonal space. This can be found in an essay (collected in *Science and Method*, 1908) where, for the first time, he investigated the origin of space representation by means of an analysis which was still *more geometrico*, but also inspired to the specific biological interactions between an organism and its environment rather than to the detection of physical properties of moving objects, as it was in *Science and Hypothesis*.

According to Poincaré’s view as it emerges in *Science and Method*, the relations between our body and objects surrounding us are to be construed in terms of motor acts, by which we can reach such objects:

For instance, at a moment α the presence of an object *A* is revealed to me by the sense of sight; at another moment β the presence of another object *B* is revealed by another sense, that, for instance, of hearing or of touch. I judge that this object *B* occupies the same place as the object *A*. What does this mean? [...] The impressions that have come to us from these objects have followed absolutely different paths [...] and] have nothing in common from the qualitative point of view. The representations we can form of these two objects are absolutely heterogeneous and irreducible one to the other. Only I know that, in order to reach the object *A*, I have only to extend my right arm in a certain way; even though I refrain from doing it, I represent to myself the muscular and other analogous sensations which accompany that extension, and that representation is associated with that of the object *A*. Now I know equally that I can reach the object *B* by extending my right arm in the same way, an extension accompanied by the same train of muscular sensations. And I mean nothing else but this when I say that these two objects occupy the same position. [...] And this is very important, since it is in this way that I could defend myself against the dangers with which the object *A* or the object *B* might threaten me. With each of the blows that may strike us, nature has associated one or several parries which enable us to protect ourselves against them. The same parry may answer to several blows. [...] All these parries have nothing in common with one another, except that they enable us to avoid the same blow, and it is that, and nothing but that, we mean when we say that they are movements ending in the same point of the space. Similarly, these objects, of which we say that they occupy the same point in space, have nothing in common, except that the same parry can enable us to defend ourselves against them (Poincaré 1908, 1952: 101–102).

Insofar as it is the space resulting from the mutual ‘co-ordination’ of ‘the multiplicity of [possible] parries’ (Poincaré 1908, 1952: 104), peripersonal space binds together different sensory modalities, thus localizing visual or tactile stimuli in terms of our potential motor acts. Therefore space cannot be represented *per se*

somewhere in the brain; its constitution depends on the activity of neural circuits whose primary function is to organize motor acts which, albeit through different effectors (hands, mouth, eyes, etc.), ensure interaction with the surroundings, detecting possible threats and opportunities.

Given that this constitution is not just a conquest by the *individual*, but by the *species*, its ‘traces’ can be seen in the newborn and even in the foetus. It has been shown that foetuses engage in various goal-directed motor activities in the womb: for example, in the sixth month of gestation they’re able to put their thumb in their mouth to suck it (Butterworth & Harris, 1994). More recently, Zoia *et al.* (2007) measured the kinematics of hand movements of 22-weeks-old foetuses. The results showed that the spatial and temporal characteristics of foetal movements were by no means uncoordinated: on the contrary, they displayed kinematic patterns that depend on the goal of different motor acts. After birth, the child’s movements are increasingly goal-directed and clearly referred to the space around his/her body. The optical condition is congruent with the motor situation. As the crystalline lens is not completely operational at that age, the focal distance is more or less fixed and the baby can only see clearly objects that are within a distance of 20 cm. In this way, he/she acquires a representation of his/her peripersonal space without having to distinguish whether a visual stimulus is ‘near’ or ‘far’.

Taken together, these findings allow one to speculate that during prenatal development specific connections may develop between the motor (and somatosensory) centres controlling goal-directed behaviour and brain regions that will become recipient of visual inputs after birth. Such connectivity could provide functional templates (e.g., specific spatio-temporal patterns of neural firing) to areas of the brain that, once activated, would be ready to specifically encode visual stimuli in terms of potential motor acts (e.g., reaching, avoiding). In other words, neonates and infants, by means of specific connectivity developed during the late phase of gestation between motor and “to-become-visual” regions of the brain, would be ready to map the surrounding space as reachable space, and would be endowed with the neural resources enabling their interactions with objects around them, characterizing post-natal life since its very beginning (see Gallese *et al.*, 2009).

5. Concluding Remarks

At this point, we should have gained a better understanding of how the motor system displays its cognitive function through the constitution of peripersonal space. As has been argued at great length, the primacy of action with respect to movement provides the key for an account of the *near* versus *far* dichotomy. The role of bodily movements won’t suffice for us to understand the nature of near space, but is nonetheless worth mentioning, as it already reveals some important features of space constitution. For instance, we’ve seen that Husserl’s reflection highlights the

existence of various sensory spaces (visual and tactile), as well as the fact that the motor constitution of space presupposes a number of frames of reference.

Still, there is a lot more to the distinctive character of peripersonal space than what its *motor* nature can reveal. As a matter of fact, we've shown that peripersonal space is marked by its *dynamic plasticity*, that is, its varying range. Since the latter has been argued to depend on the actual reach of motor goals and actions, the nature of peripersonal space can only be fully appreciated and understood by adopting an *enactive* approach to cognition — enactive in that it is rooted in the goal-centredness of action.

In the light of this concept, the results of many experiments reviewed above will be much clearer. The identification with specific *motor goals* disambiguates similar movements insofar as they are part of different motor acts. This should be evident by reflecting on the fact that, in the previously mentioned experiments, similar movements were more or less effective in expanding peripersonal space depending on whether they constituted actions or not, and, if they did, on how far those actions could actually reach. In Iriki and coworkers' (1996) experiment, the monkey's peripersonal space was only increased if the rake was used *for reaching something*, and not when it was passively held. Furthermore, in Farnè and coworkers' (2005) experiment, it was the *functionally effective length* of the tool that mattered to the extension of the patient's peripersonal space, that is, the latter coincided with the actual reach of the patient's action.

We can therefore appreciate to what extent the constitution of space in terms of peripersonal and extrapersonal space requires the adoption of an enactive perspective. The way of constructing space that lies at the basis of the distinction between peripersonal and extrapersonal space indeed provides a mapping of space in terms of motor action.

References

- [1] Ackroyd, K., Riddoch, M. J., Humphreys, G. *et al.*, 2002, Widening the sphere of influence: Using a tool to extend extrapersonal visual space in patient with severe neglect. *Neurocase*, 8, 1–12.
- [2] Andersen, R. A., Snyder, A. L., Bradley, D. C. *et al.*, 1997, Multimodal representation of space in the posterior parietal cortex and its use in planning movements, *Annual Review of Neuroscience*, 20, 303–330.
- [3] Berti, A., Frassinetti, F. 2000, When far becomes near: re-mapping of space by tool use, *Journal of Cognitive Neuroscience*, 12, 415–420.
- [4] Berti, A., Rizzolatti, G. 2002, Coding near and far space. In Karnath, H.-O., Milner, A. D. and Vallar G. (Eds.), *The Cognitive and Neural Bases of Spatial Neglect* (pp. 119–129). Oxford University Press, Oxford.
- [5] Berthoz, A. 1997, *The Sense of Movement*. Odile Jacob (Ed.), Paris.
- [6] Brozzoli, C., Farnè, A. (in this volume), *The space representations in the brain*.

- [7] Butterworth, G., Harris, M. 1994, *Principles of Developmental Psychology*, Lawrence Erlbaum Associates, Hove, East Sussex (UK).
- [8] Chieffi, S., Fogassi, L., Gallese, V. *et al.*, 1992, Prehension movements directed to approaching objects: influence of stimulus velocity on the transport and the grasp components, *Neuropsychologia*, 30, 877–897.
- [9] Colby, C. L., Goldberg, M. E. 1999, Space and attention in parietal cortex, *Annual Review of Neuroscience*, 22, 319–349.
- [10] Cowey, A., Small, M., Ellis, S. 1994, Left visuo-spatial neglect can be worse in far than near space, *Neuropsychologia*, 32, 1059–1066.
- [11] Cowey, A., Small, M. and Ellis, S. 1999, No abrupt change in visual hemineglect from near to far space, *Neuropsychologia*, 37, 1–6.
- [12] di Pellegrino, G., Làdavas, E., and Farnè, A. 1997, Seeing where your hands are, *Nature*, 388, 730.
- [13] Farnè, A., Iriki, A., and Làdavas, E. 2005, Shaping multi-sensory action space with tools: evidence from patients with cross-modal extinction, *Neuropsychologia*, 43(2), 238–248.
- [14] Farnè, A., Làdavas, E. 2000, Dynamic size-change of hand peripersonal space following tool use, *Neuroreport*, 11, 1–5.
- [15] Fogassi, L. (this volume), Space coding in the cerebral cortex.
- [16] Fogassi, L., Gallese, V., Buccino, G. *et al.*, 2001, Cortical mechanism for the visual guidance of hand grasping movements in the monkey: A reversible inactivation study, *Brain*, 124, 571–586.
- [17] Fogassi, L., Gallese, V., di Pellegrino, G. *et al.*, 1992, Space coding by premotor cortex, *Experimental Brain Research*, 89, 686–690.
- [18] Fogassi, L., Gallese, V., Fadiga, L. *et al.*, 1996a, Coding of peripersonal space in inferior premotor cortex (F4), *Journal of Neurophysiology*, 76, 141–157.
- [19] Fogassi, L., Gallese, V., Fadiga, L. *et al.*, 1996b, Space coding in inferior premotor cortex (area F4): facts and speculations. In Lacquaniti, F., Viviani, P. (Eds.), *Neural Bases of Motor Behaviour* (pp. 99–120). Kluwer, Dordrecht.
- [20] Folegatti, A., Berti, A. (this volume), Action and space representation.
- [21] Frassinetti, F., Rossi, M. and Làdavas, E. 2001, Passive limb movements improve visual neglect, *Neuropsychologia*, 39, 725–733.
- [22] Gallese, V., Fadiga, L., Fogassi, L. *et al.*, 1996, Action recognition in the premotor cortex, *Brain*, 119, 593–609.
- [23] Gallese, V., Rochat, M., Cossu, G. *et al.*, 2009, Motor Cognition and Its Role in Phylogeny and Ontogeny of Action Understanding, *Developmental Psychology*, 45, 103–113.
- [24] Gentilucci, M., Fogassi, L., Luppino, G. *et al.*, 1988, Functional organization of inferior area 6 in the macaque monkey: I. Somatotopy and the control of proximal movements, *Experimental Brain Research*, 71, 475–490.
- [25] Gentilucci, M., Scandolara, C., Pigarev I. N. *et al.*, 1983, Visual responses in the postarcuate cortex (area 6) of the monkey that are independent of eye position, *Experimental Brain Research*, 50, 464–468.
- [26] Giorello, G., Sinigaglia, C. 2007, Space and Movement. Husserl’s Geometry of Visual Field. In Boi, L. Kerszberg, P. and Patras, D. (Eds.), *Rediscovering Phenomenology*.

- Phenomenological Essays concerning Mathematical Beings, Physical Reality, Perception and Consciousness (pp. 103–123). Kluwer, Dordrecht.
- [27] Graziano, M. S. A., Gross, C. G. 1994, Mapping space with neurons, *Current Directions in Psychological Science*, 3(5), 164–167.
- [28] Graziano, M. S. A., Hu, X. and Gross, C. G. 1997, Visuospatial properties of ventral premotor cortex, *Journal of Neurophysiology*, 77, 2268–2292.
- [29] Halligan, P. W., Marshall, J. C. 1991, Left neglect for near but not far space in man, *Nature*, 350, 498–500.
- [30] Husserl, E. (1907, 1916), *Thing and Space. Lectures 1907* (Edmund Husserl Collected Works VII). Eng. Trans. Dordrecht, Kluwer, 1997.
- [31] Iriki, A., Tanaka, M. and Iwamura, Y. 1996, Coding of modified body schema during tool use by macaque post-central neurons, *Neuroreport*, 7, 2325–2330.
- [32] Ishibashi, H., Hihara, S. and Iriki, A. 2000, Acquisition and development of monkey tool-use: Behavioural and kinematic analyses, *Canadian Journal of Physiology and Pharmacology*, 78, 1–9.
- [33] Li, C. S., Mazzone, P. and Andersen, R. A. 1999, Effect of reversible inactivation of macaque lateral intraparietal area on visual and memory saccades, *Journal of Neurophysiology*, 81, 1827–1838.
- [34] Mach, E. 1905, *Knowledge and Error. Sketches on the Psychology of Enquiry*, Engl. Trans. Reidel, Dordrecht, 1967.
- [35] Maravita, A., Husain, M., Clarke, K. *et al.*, 2001, Reaching with a tool extends visual and tactile interactions into far space: Evidence from cross-modal extinction. *Neuropsychologia*, 39, 580–585.
- [36] Maravita, A., Spence, C., Kennet, S. *et al.*, 2002, Tool use changes multimodal spatial interactions between vision and touch in normal humans, *Cognition*, 83, 25–34.
- [37] Neppi-Mòdona, M., Rabuffetti, M., Folegatti, A. *et al.*, 2007, Bisecting lines with different tools in right brain damaged patients: the role of action programming and sensory feedback in modulating spatial remapping, *Cortex*, 43(3), 397–410.
- [38] Pegna, A. J., Petit, L., Caldara-Schnetzer, A. S. *et al.*, 2001, So near yet so far. Neglect in far space depends on tool use, *Annals of Neurology*, 50, 820–822.
- [39] Poincaré, J.-H. 1898, *On the foundations of geometry*. Translated by T. J. McCormack. *The Monist*, 9, 1–43.
- [40] Poincaré, J.-H. 1902, *Science and Hypothesis*, Engl. Trans. New York, Dover, 1952.
- [41] Poincaré, J.-H. 1908, *Science and Method*, Engl. Trans. New York, Dover, 1952.
- [42] Rizzolatti, G., Fadiga, L., Gallese, V. *et al.*, 1996, Premotor cortex and the recognition of motor actions, *Cognitive Brain Research*, 3, 131–141.
- [43] Rizzolatti, G., Matelli, M. and Pavesi, G. 1983, Deficits in attention and movement following the removal of postarcuate (area 6) and prearcuate (area 8) cortex in macaque monkeys, *Brain*, 106, 655–673.
- [44] Rizzolatti, G., Scandolara, C., Gentilucci, M. *et al.*, 1981a, Afferent properties of periarculate neurons in macaque monkeys. I. Somatosensory responses, *Experimental Brain Research*, 2, 125–146.
- [45] Rizzolatti, G., Scandolara, C., Matelli, M. *et al.*, 1981b, Afferent properties of periarculate neurons in macaque monkeys. II. Visual responses, *Experimental Brain Research*, 2, 147–163.

- [46] Rizzolatti, G., Sinigaglia, C. 2007, Mirror neurons and motor intentionality, *Functional Neurology* 2007, 22(4), 205–210.
- [47] Rizzolatti, G., Sinigaglia, C. 2008, *Mirrors in the Brain. How our Minds Share Actions and Emotions*, London/New York, Oxford University Press.
- [48] Schieber, M. 2000, Inactivation of the ventral premotor cortex biases the laterality of motoric choices, *Experimental Brain Research*, 130, 497–507.
- [49] Sinigaglia, C. 2000, *La seduzione dello spazio. Geometria e filosofia nel primo Husserl*, Unicopli, Milano.
- [50] Thom, R. 1990, *Apologie du logos*, Hachette, Paris.
- [51] Vuilleumier, P., Valenza, N., Mayer, E. *et al.*, 1998, Near and far space in unilateral neglect, *Annals of Neurology*, 43, 406–410.
- [52] Wardak, C., Olivier, E., Duhamel, J. R. 2004, A deficit in covert attention after parietal cortex inactivation in the monkey, *Neuron*, 42, 501–508.
- [53] Zoia, S., Blason, L., D’Ottavio, G. *et al.*, 2007, Evidence of early development of action planning in the human foetus: A kinematic study. *Experimental Brain Research*, 176, 217–226.

PART 3

**Geometrical Methods in the
Biological Sciences**

This page is intentionally left blank

CHAPTER 9

Causes and Symmetries in Natural Sciences: The Continuum and the Discrete in Mathematical Modelling

FRANCIS BAILLY¹

*Laboratoire de Physique des Solides, CNRS, Meudon, France
bailly@cnrs-bellevue.fr*

GIUSEPPE LONGO

*LIENS, CNRS – ENS and CREA, Paris
ENS, 45 rue d’ulm, 75005 Paris, France
longo@di.ens.fr*

I. Introduction

How do we make sense of physical phenomena? The answer is far from being univocal, particularly because the whole history of physics has set, at the centre of the intelligibility of phenomena, changing notions of *cause*, from Aristotle’s rich classification, to which we will return, to Galileo’s (too strong?) simplification and their modern understanding in terms of ‘structural relationships’ or the replacement of these notions by structural relationships. It is then an issue of the stability of the structures in question, of their invariants and symmetries (Weyl, 1927 and 1952; van Fraassen, 1994); to the point of the attempt to completely dispel the notion of cause, following a great and still open debate, in favour, for instance, of *probability correlations* [in quantum physics, see, for example, (Anandan, 2002)].

The situation is even more complex in biology, where the ‘reduction’ to one or another of the current physico-mathematical theories is far from being

¹Francis Bailly passed away on February 5, 2009.

accomplished [see (Bailly, Longo, 2006)]. From our point of view, the difficulties in doing this reside as much within the specificities of the causal regimes of physical theories — which, moreover, differ amongst themselves — as in the richness specific to the dynamics of living phenomena. Our approach, as presented in (Bailly, Longo, 2006), has attempted to highlight certain aspects, such as the intertwining and coupling of levels of organisation, which are strongly related to the phenomena of autopoiesis, of ago-antagonistic effects, of the hybrid causalities often mentioned in the theoretical reflections in biology [see (Varela, 1989; Rosen, 1991; Stewart, 2002; Bernard-Weil, 2002; Bailly, Longo, 2003)].

We will now return to some aspects of the construction of scientific objectivity, as explication of a theoretical web of relationships. And we will mostly speak of causal relationships, since causal links are fundamental structures of intelligibility. Our approach will again be centred upon symmetries and invariances, because they enable causes to manifest themselves, namely by the constraints they impose. In a strong sense, they thus present themselves as conditions of possibility for the construction of mathematical or physical objectivity.

Now, if mathematics is constitutive of physical objectivity and if it makes phenomena intelligible, its own ‘internal structure’, that of the continuum, for example, as opposed to the discrete, contributes to physical and biological determination and structures their causal links. To put it in other words, mathematical structures are, on the one hand, the result of a *historical formation of meaning*, where history should be understood as the constitutive process from our phylogenetic history to the construction of intersubjectivity and of knowledge within our human communities. But, on the other hand, mathematics is also *constitutive of the meaning of the physical world*, since we make reality intelligible via mathematics. Particularly, it organises regularities and correlates phenomena which, otherwise, would make no sense to us. The thesis outlined in (Longo, 2007) and which we further develop here, is that the mathematics of continua and discrete mathematics, the latter characteristic of computer modelling, propose different intelligibilities both for physical and living phenomena, particularly for that which concerns causal determinations and relationships as well as their associated symmetries/asymmetries.

In a final section, we will attempt to address the field of biology by questioning ourselves about the operational relevance and status of the concepts thus under consideration. But in this text, we will first propose to illustrate, in the case of physics, the situation which we have just summarily described. This will enable us to ‘enframe’ physical causality and to compare it to computational models and to biology.

2. Causal Structures and Symmetries, in Physics

The representation usually associated to physical causality is oriented (asymmetric): an originary cause generates a consecutive effect. Physical theory is supposed

to be able to express and measure this relationship. Thus, in the classic expression $\mathbf{F} = m\mathbf{a}$, we consider the force \mathbf{F} to ‘cause’ the acceleration \mathbf{a} of the body of mass m and it would seem downright incongruous, despite the presence of the equality sign, to consider that acceleration, conversely, may be at the origin of a force relating to mass. Yet, since the advent of the theory of General Relativity, this representation found itself to be questioned in favour of a much more balanced interactive representation (a ‘reticulated’ representation, one may say): thus, the energy-momentum tensor doubtlessly ‘causes’ the deformation of space, but, reciprocally, the curvature of a space may be considered as field source. Finally, it is the whole of the manifesting network of interactions which is to be analysed from the angle of geometry or from that, more physical, of the distribution of energy-momentum. It is that an essential conceptual step has been made: to the expression of an isolated physical ‘law’ (expressing the causality at hand) has been substituted a general principle of relativity (a principle of symmetry) and the latter re-establishes an effective equivalence (interactive determinations) where there appeared to be an order (from cause to effect).

Here is an organising role of mathematical determination, a ‘set of rules’ and a reading which is abstract, but rich in physical meaning. Causes become interactions and these interactions themselves constitute the fabric of the universe; deform this fabric and the interactions appear to be modified, intervene upon the interactions themselves and it is the fabric which will be modified.

We will first of all distinguish between *determinations* and *causes* as such. For instance, we will see the symmetries proposed within a theoretical framework as related to the determinations which enable causes to find expression and to act; in this they are more general than the causes and are logically situated as ‘prior’ despite having been established, historically, ‘afterwards’ (the analysis of the force of gravitation, as cause of an acceleration, preceded Newton’s equation).

Let’s then specify that which we mean by ‘determination’ in physics, enabling us to return to the causal relationships which we will examine extensively. For us, all these notions are the *result* of a construction of knowledge: by proposing a theory, we organise reality mathematically (formally) and thus constitute (determine) a phenomenal level as well as the objectivity and the very ‘object’ of physics. We will therefore address first of all the ‘objective and formal determinations’, particular to a theory.

More specifically, once given the theoretical framework, we may consider that:

D.1 The objective determinations are given by the *invariants* relative to the symmetries of the theory at hand.

D.2 The formal determinations correspond to the set of *rules and equations* relative to the system at hand.

To return to our example, when we represent the dynamic by means of Newton's equation, we have a formal determination based upon a representation of causal relationships, which we will call 'efficient' (the force 'causes' acceleration). However, when having recourse to Hamilton's equations we still have a formal determination, but one which refers to a different organisation of principles (based on energy conservation, typically). It is still different with the optimality of the Lagrangian action, which refers to the minimality of an action associated to a trajectory. In this case, we have, for classical dynamics, three different mathematical characterisations of the events; and it is only with the advent of the notion of 'gauge invariant' (that is, of 'relativity principles') that these distinct formal determinations have been unified under an overreaching *objective determination*, related to the corresponding symmetries and invariants (manifested by transformation groups, such as the Galileo, Lorentz–Poincaré or Lie groups). A single objective determination then, for instance, the movement of a mobile with a certain mass, may account for (result from!) distinct formal determinations, based upon the concepts of force, of energy conservation and of geodesics, respectively. In the first case, the invariant is a property (mass), in the second, it is a state (energy), in the third it is question of the criticality of a geodesic (action, energy multiplied by time). If the final results of the mobile's dynamic may thus be the same, on the other hand the equations leading to them may take quite different forms unifying only under the even larger constraint of objective determinations (relating, in our example, to a mass in movement).

It is in fact the physical *objects* themselves which are the consequence of — given by — these determinations. More specifically, the physical objects are theoretically characterised by that which we designate, rather commonly, as properties and accessible states:

O.1 Properties (mass, charge, spin, other field sources...),

O.2 Accessible states, potential or actual (position, moments, quantum numbers, field intensity...),

being understood that their specific values essentially depend on empirical measurement. To highlight as simply as possible the difference we make between property and state, by means of their invariance characteristics, we may say that properties (which characterise an object) do not change when the states of the object change; conversely, if the properties change, it is the object itself which is modified.

These objective determinations thus constitute in a way the referential framework, at a given moment in time, to which are related *experience, observation and theory*, enabling to interpret and to correlate the ones to the others. In themselves, and as we have just indicated, they thus do not completely characterise the objects they construct, but constrain — among other things by extricating

invariants — properties and behaviours. Thus, for instance, they impose the fact that there is a mass (sensitive to the gravitational field), but without nevertheless fixing the magnitude of this mass or, as we shall see, the manifestation of fields such as the electro-magnetic field. We are therefore facing properties which we may qualify as ‘categorical’ and qualitative, but without necessarily specifying the associated quantities which quantitatively characterise the object in direct relationship to the measurement. This is also the case for that which we call accessible states: their structure is qualitatively characterised, but the fact that the system quantitatively attains such or such of these theoretically determined possible states depends on empirical factors.

Why distinguish here between properties and accessible states? It would enable us to understand as *cause*, in the traditional sense (which after Aristotle we will call ‘efficient cause’), all which affects (can modify) states; while we may consider that in the traditional approach, *the invariants of efficient causal reduction* are constituted by the set of properties. However, these very properties participate to a causality, which we shall relate to ‘material’ causality.

So let’s attempt to refine the analysis, not only by distinguishing between different types of ‘causes’ but also by trying to affect the distinct elements of objectivity. Let’s agree that, relatively to the effect of an object upon another:

C.1 The **material cause** is associated to the set of *properties*;

C.2 The **efficient cause** is correlated to the variation of one or more *states*.

We can recognise here a revitalisation of Aristotle’s classification, so dear to René Thom. In fact, if we want to maintain a parallel with Aristotelian categorisation, let’s observe that we have called formal determination that which the modern interpretation of the philosopher would designate as “formal cause” (that is, that which corresponds to the set of theoretical constraints which define and measure the effects of other causes — laws, rules, theories,...).² In our approach, it is the determinations, formal and objective, which produce the specification of objects, by means of the notions of properties and states (of which the structures and variations participate to the material and efficient causes, respectively). With regard to the causes, we will preserve the Aristotelian terminology, although material causes may be classified as ‘material structures’. Indeed, a change in properties changes an

²In the debate with I. Prigogine concerning determinism, R. Thom highlights the role of structural stability, even within the framework of highly unstable dynamics (the forms are maintained, all the while being deformed). It is the equations of the dynamic which determine their possible evolutions (as formal causes — determinations, for us). On the other hand, Prigogine highlights the play between locally stable structures and global systems where small, amplified fluctuations induce the choice of one of these evolutions. While preserving his new view on Aristotle’s finesse, but in a different way than R. Thom, we do not attribute to these different notions of causality an ontological hierarchy of the platonic type, where the formal determinations (causes) *ontologically* precede the other causes.

object, as we mentioned earlier, but, at the same time, it induces — it causes! — a change of states. For example, a change in mass or charge, *in an equation*, modifies the values of the acceleration or of the electrical field.

2.1. Symmetries as starting point for intelligibility

From the point of view we have just developed, may we consider that constraints of symmetry stem from causal constraints? According to our distinction and as we have just specified, symmetries emerge from the determinations (under the form of systems of equations, typically) where the causes manifest. Their greater generality thus also imposes itself through the relation to laws corresponding to the formal determinations (which, for example, take such or such expression according to the selected gauges). To put it lapidarily, using the example we will discuss below (Intermezzo): the phase's global gauge invariance *determines* the charge (a property) as a conserved quantity of the theory and its local invariance *determines* the existence of the electro-magnetic field (a state) under the form of Maxwell equations. The interactions, described by these equations, may, but only afterwards, be considered as giving us the *causes* (material or efficient) of the observed effects.

In fact, and since Galileo, that which we usually characterise as 'causes' seems to correspond mainly to efficient causes, whereas, as we have just seen, the "determinations" seem to rather present themselves as a source common to causes which would derive from them (including material and formal). This results in that we may consider, "transcendentally speaking", the determinations, the symmetries, namely, to present themselves as conditions of possibility for the causes to manifest.

Now it appears to us that the natural sciences, with the exception of the biosciences, may be part of the conceptual framework we have just drawn out, including that which corresponds to extremalisation rules (the geodesics of the Lagrangian), which appear, but wrongly so in our opinion, to confer a tinge of *finality* to the processes which they model. It is only with living phenomena that the taking into account of a sort of 'final causality' [to put it once more in Aristotelian terms, see (Rosen R., 1991; Stewart J., 2002)], that we have characterised elsewhere as a 'contingent finality' (see also 3.1 below) and as locus of "meaning" for any living phenomenon, really becomes relevant. It is this which we will attempt to examine later, in Sec. 3.

2.2. Time and causality in physics

We have thus attempted to specify, very generally, the notions of objective determination, of object and physical cause, from the notion of symmetry and, more specifically, from the notion of invariance with regard to the given symmetries.

Let's also observe that, since about a century, in physics, the laws of conservation, as formal determinations, are understood in terms of spatio-temporal

symmetries; for instance, the conservation of the angular momentum is correlative to the symmetry of rotation (it is Noether's theory which is at the origin of this great theoretical and conceptual turning-point, see the Intermezzo below).

But at this stage and before continuing, it would seem appropriate to introduce a distinction in the view of clearing some possible confusion with regard to the representation of causality and to the reasoning that one may entertain about it. We propose, so as to distinguish,³ namely in the case of efficient causality, between *objective causality* and *epistemic causality*.

Objective causality is associated, in our opinion, to a rather essential constraint, which is constitutive of physical phenomena, and which is the irreversible characteristic of the unfolding of time (that which we call the 'arrow of time'). But even in the case where temporality does not explicitly appear, it continues to underlie any change, any process as such — including that of measurement — and constitutes in this respect a foundation to any conceptualisation, observation or experience, from the moment that such a process is considered. That is to say, this time from the angle of causal analysis, that *time is constitutive of physical objectivity*.

In contrast, epistemic causality is to be considered as independent of an arrow of time. For instance, the analysis of a phase transition in function of the value of a parameter (such as temperature) does not refer to any specific temporality. It is in a way the atemporal and abstract variation of the parameter that "causes" the transition, be it in one direction (for instance, from liquid to solid) or the other (from solid to liquid). At this level, the invoked structure of causality (effect of the change of temperature on the state of the system) remains independent of the time factor, even though, at another level, it is indeed over time that the effective variation of this parameter occurs — in one direction or the other. This is also the case in the very simple example which is the law of perfect gases ($pV = RT$, where p represents pressure, V the volume, T the temperature and R Joule's constant). This law is independent of time and one may conceive of various "causes" at the origin of a variation in volume, for instance, leading, under constant temperature, to a variation in the pressure associated to the occurrence of a chemical reaction. The concomitant (and symmetrical, given the equational relationship) variations in volume and pressure may be considered as the causes — of the epistemic type — of one another (in fact, these variations may be said to be 'correlated'), in contrast to an objective causality — temporalised, this time — which would find its source in the temporal unfolding of this chemical reaction at the origin of the considered variation in volume [our distinction may possibly help to understand the discussion in (Viennot, 2003, appendix)].

³As we have already done on the occasion of the approach and the deepening of the concept of 'complexity' (Bailly, Longo, 2003).

May such a distinction between the objective and the epistemic, which seems to clearly correspond to a reality in the case of an efficient causality (associated, let's remember, to modifications in the states of a system, as we have just illustrated) still be applicable to material causality?

It does appear that for material causality, one may find examples demonstrating that such is the case, inasmuch as the properties in question have different expressions depending whether they are related to their own specific system or to an external reference. This is the case in relativity, for example, where the mass (or life-span) of particles depend on their speed with regard to the laboratory reference: in the *internal* system, the rest mass remains a characteristic property of the particle's very identity (m_0), whereas within a referential animated by a speed v with regard to the system as such, the mass takes on an epistemic character, of which the measurement is $m = m_0/(1 - v^2/c^2)^{1/2}$, where c represents the speed of light (for light itself, this also that which enables to consider that the photon's mass is null, while its energy is non-null and while Einstein's relation establishes a direct relationship between mass and energy). Likewise, the "efficient mass" which we calculate following the process of renormalisation (which, in order to eliminate infinities from the calculi of perturbation, integrates with the mass some classes of interaction) takes an epistemic character with regard to the mass itself which preserves its own objective character. In this sense, we may consider that the properties which are located at the source of material causality retain an objective character in their internal system all the while acquiring an epistemic character if we relate them to different referentials.

Because we take the arrow of time into consideration while characterizing efficient objective causality, we distinguish ourselves from certain trends in relativistic and quantum physics that exclude such an arrow, in order to preserve any relationship by symmetry. In these approaches, the causal relationships are replaced by other concepts, for instance, in quantum mechanics, by probability correlations [see (Anandan, 2002), among others]. The reason for this differentiation, beyond the elements of analysis we have just exposed, appears to be crucial in an epistemological respect: we will indeed often refer to dynamic systems (thermodynamic and of the critical type) and will also address certain aspects of biology. Now, there is no analysis of these systems, even less of living phenomena, which can be performed without taking into account the existence of an arrow of time. Particularly, there would be no phylogenesis, no ontogenesis, no death... in short; there would be no life without time, oriented and irreversible. The processes of life impose an arrow of time, be it only for the thermodynamic effects to which they participate; but it even appears unavoidable to go further, because these processes require a new way of looking at complex forms of temporality, of clocks of life with causal retroactions due to intentional aims, to expectancies and previsions, characteristic of perception and action.

To conclude, our mathematical point of view is that objective determinations are given by symmetries and an efficient cause *breaks* some of the latter, be it, from an objective viewpoint, only that symmetry which is associated to the arrow of time. Reciprocally, irreversible phenomena (bifurcations, phase changes...), which are therefore oriented in time, may be read as symmetry breakings correlated to (new) causal relationships. Symmetries and their breakings therefore remain the starting point for any theoretical intelligibility.

More specifically, we will attempt to understand some causal relationships as symmetry breakings, in a very general and abstract sense. This will, among other things, enable to lay the basis of a coherent foundational framework for the analysis of the different causal regimes proposed by continuous mathematics in comparison to those of discrete arithmetics. This will therefore consist of a mathematical view upon the constitutive role of mathematics in the construction of scientific objectivity; through this approach, we aim to grasp the importance of our digital machines in this construction, since these machines are the practical realisation of the arithmetisation of knowledge.

The final reflection regarding biology will bring us back to natural phenomena, in all their causal specificity. Of course, computerised modelling, in biology as in physics, remains a fundamental issue. It is precisely for this reason that it must be based upon a fine analysis of the different structures of relationships, particularly causal relationships, proposed within the various theoretical frameworks (physical, biological, of discrete mathematics).

2.3. Symmetry breakings and fabrics of interaction

It is thus by the means of mathematics that we organise causal links; mathematics makes intelligible and unifies, particularly via symmetries, certain phenomenal regularities, at least those of classical physics, both dynamical and relativistic systems. But mathematics also makes explicit the symmetries in relation to which probability correlations are quantum invariants.

In the dynamic and relativistic cases, the geodesic principles governing the evolution of systems apply to abstract spaces, 'manifolds' endowed with a metric where *symmetry transformations*⁴ leave equations of movement invariant⁵ (typically, the trajectories defined by Euler–Lagrange equations). It is in this sense that these theories base themselves on invariants with regard to spatio-temporal symmetries: if we understand the 'laws' of a theory to be 'the expression of a geodesic principle within a suitable space', it is these abstract geodesics which are not modified by transformations of symmetries.

⁴Objective determinations, in our language.

⁵Formal determinations, for us.

Let's return to the most classical of physical laws: the equation $F = ma$ is symmetrical, as an equation. As we observed above, it is its *asymmetrical* reading which we associate to a causal relationship: the force F *causes* the acceleration a (the equation is read, so to speak, from left to right). We thus break, conceptually, a formal symmetry, equality, in order to better understand, following Newton, a trajectory (and its cause). More specifically, the equation *formally determines* a trajectory of which F appears as the *efficient* cause (it modifies a state, all the while leaving invariant the Newtonian mass, a property). In this sense it becomes legitimate to consider that the equation contributes to the constitution of an objectivity (the trajectory of the mobile), whereas its oriented reading (and the efficient causality it thus expresses) constitutes an interpretation and refers to an epistemic regime of causality.

We therefore propose to consider that each time a physical phenomenon is presented by an (a system of) equation(s), *a breaking in the formal symmetry* (that of equality, by an oriented reading) *makes explicit an epistemic regime of causality*. Particularly, the symmetry breaking in question may be correlated to an efficient cause which intervenes within the formal framework determined by the equation.

Of course, this breaking is not necessarily unique (that by which, namely, it manifests its epistemic character). For example, as we have just evoked in the preceding paragraph, one can read causally and from an epistemic standpoint $pV = RT$ from left to right and vice versa. By reference to relativistic systems, we have already read the equation $F = ma$ (or more exactly, its relativistic equivalent), inversely, while highlighting the fact that, reciprocally, the curvature of a space may be considered as a field source. This interpretive reversal, which reorganises phenomena radically, is legitimate; indeed, in our spatial manifolds, *the transformations* (of gauges), which enable to pass from one referential to another, are supposed to leave invariant the equations of movement, and by doing so they preserve the *symmetries*, but without necessarily preserving the asymmetrical readings of the formal determinations (among which the epistemic causal reading we have just discussed).

We thus propose to consider formal interactions, organised by the asymmetrical structures of equations, but also (efficient) causes, which can be associated to possible asymmetries in the reading of these very equations. Let's observe, once more, that certain coefficients, such as the mass m in $F = ma$, are correlated to that which we have categorised as material causes (whereas acceleration is correlated to states". And, when an "external" cause (efficient or material) is added to a given determination (equations of evolution), the geodesics of the relevant space are deformed and symmetries associated to this space may be broken, following the variation of states and properties.

Now, this mathematical intelligibility, conceptual fabric of symmetries and asymmetries which correlates the regularities of the world, is constitutive of physical phenomena as well as of scientific objectivity. As we shall see, they profoundly

change if the world's proposed reading grid is rooted in continuous or in discrete mathematics. And they must subsequently be enriched, if one hopes to better conceptualise certain phenomena pertaining to life.

Intermezzo. Remarks and Technical Commentaries

Inter. I. *More on symmetries and symmetry breakings in contemporary physics*

Let's consider the previously analysed three great types of physical theories which are the relativistic, the quantum and the critical types (dynamic and thermodynamic systems).

Relativistic theories are essentially tributary of *external* symmetries (sets operating over space-time). Classical mechanics already presents these relativistic traits, with its constraints of invariance under the Galileo group (within the Euclidean space), but it is especially with classical electro-magnetism and Special Relativity that symmetries begin to play a determining role under the Lorentz–Poincaré group (group of the rotations and translations within a Minkowski space). Regarding General Relativity and cosmology, it is the group of the set of diffeomorphisms of space-time which plays the determining role. The corresponding symmetry breakings principally manifest themselves through phenomena of dissipation, or of the arrow of time.

Quantum type theories for their part mobilise essentially *internal* symmetries operating on the fibres of the corresponding fibrates: it is the gauge sets which generate the gauge invariances and which present themselves as Lie groups (continuous groups). In quantum field theory, the most important symmetry breakings (Goldstone, Higgs fields) are considered as sources of the masses of quanta.

Critical type theories constitute theories par excellence of symmetry changes (namely by breakings): it is phase transitions, spontaneous symmetry breakings (or, conversely, of the apparition of new symmetries), of which the effects are processed this time by means of the renormalisation semi-group process in order to characterise the critical exponents and to established rules of universality which constitute, in a certain way, over the classes of equivalency which they bring forth, the basis of new relativities and symmetries (very different symmetries may present critical exponents, and thus behaviours, which are identical, depending only on parameters as general as the magnitudes of the prolongation spaces or that of the order parameters).

We will also mention the fact that the unification processes, in the relativistic theories as well as in the quantum theories, or even between themselves, involve the expansion of the concerned symmetry groups (often at the same time as the spaces within which they operate).

Inter.2. From Noether's theorem and physical laws of conservation

One of the principal foundations of the role of symmetries for physics can be found in Noether's theorem, according to which any transformation in symmetry, operating upon a Lagrangian and conserving the equations of movements, associates conserved quantities. By a more precise analysis, one may observe that this theorem narrowly couples such laws of conservation — physical invariants, that is, objective determinations — to indeterminations of reference systems (space-time, fibres) by the fact of the principles of relativity and of the symmetries supposed to operate there (for instance, the impossibility to define a temporal or positional origin, an origin of phases, etc.).

One of the simplest and most spectacular cases we may evoke in this regard is that of quantum electrodynamics, for which the gauge group is the phase group $U(1)$. In this case, it is required that the form of the density of the Lagrangian remain invariant under the multiplication of the state vector by a phase term ($\exp(iL)$). The global gauge invariance (*L independent* of the position), conduces, by the application of Noether's theorem, to the conservation of a quantity which we identify to the charge and which corresponds, according to the classification which we propose, to a 'property', that is, to a material characteristic. Moreover, the local gauge invariance (*L dependent* of the position) requires, in order to re-establish the broken Lagrangian covariance, to introduce a gauge potential, where there results a gauge field which is no other than the electro-magnetic field itself, expressed by the Maxwell equations (the gauge potential corresponding to its vector potential), which corresponds itself to the source of an efficient causality. Thus, it is indeed the indetermination of any phase origin (an aspect of the referential universe) which very strongly determines the conservation of the charge (an aspect of the determination of the physical object) and, most of all, for local invariance, determines the electro-magnetic field itself, nevertheless generally interpreted as 'cause' of electro-magnetic phenomena. Here, one may also notice (but we will not proceed further in the analysis, at this stage) that it is the global gauge invariance which finds itself to be coupled with a property (the charge), whereas the local invariance is coupled with a phenomenon intervening upon the states (with an effect of efficient causality): the field. So there are two forms of invariance, with regard to spatio-temporal symmetries, which we understand as objective determinations of a property and of a state, respectively.

In fact, in theoretical work as in the quest for unification, it is indeed these properties of symmetry — these forms of indetermination of the referential universes — which play an essential heuristic role in the determination of physical phenomenality. As if we were passing from the prevalence of the representation by *efficient causality* to that of a representation by *formal determination* with its symmetries and equational invariants. It is appropriate to recall here the remark,

already quoted, by C. Chevalley in his preface to B. van Fraassen's book, where it is question of '*substituting to the concept of law, that of symmetry*'. This is emphasised by van Fraassen himself, when he writes on symmetries '*... I consider this concept to be the principle means of access to the world we construct through theories*'.

We will also note, besides of the continuous symmetries that we have mainly evoked, the important role played by discrete symmetries, as in the CPT theorem, according to which the result of the three transformations T (reversal of time), C (charge conjugation: passage from matter to anti-matter), P (parity, mirror symmetry in space) is conserved in all the interactions, whereas we know that P is broken by chirality in the weak interaction, and that CP is broken in certain cases of disintegration (which led Sakharov to see in this a reason for the weak prevalence of matter over anti-matter and therefore the existence of our universe). We may understand here the particularity of the approaches, in quantum physics, which do without the arrow of time [(Anandan, 2002), for instance]. The breakings of the CP symmetry are not taken into consideration, enabling to have no asymmetry in the T transformations, and therefore to not have an oriented time factor.

With regard to spontaneous symmetry breakings, we have also evoked phase transitions and, from the quantum viewpoint, the Goldstone fields (for the global level) and the Higgs fields (for the local level), supposed to confer to particles their masses. But it is necessary to emphasise that from a cosmological standpoint, the decoupling of the fundamental interactions between themselves (gravitational, weak, strong and electromagnetic), also constitute such breakings: they then correspond to differentiations which have enabled our material universe to evolve into its current form. Without mentioning the fact that the Big Bang itself may be considered as a very first symmetry breaking (due to quantum fluctuations) of a highly energetic void.

But it is doubtlessly with regard to living phenomena that these symmetry breakings play an eminently sensitive role. Thus, Pasteur, let's recall, who had lengthily worked on the chirality of tartrates, did not hesitate to assert: '*Life as it presents itself to us is a function of the asymmetry of the universe and a consequence of this fact*'. More recently, dynamic models involving sequences of bifurcations have also been proposed to represent processes of organisation of which living phenomena could be the locus [(Nicolis, 1986; Nicolis, Prigogine, 1989)].

3. From the Continuum to the Discrete

Differential and integral equations, as limits, but also variations and continuous deformations, are present everywhere, in the physico-mathematical analyses that we have evoked. From Leibniz and Newton to Riemann, the phenomenal continuum, with its infinity and its limits in action, is at the centre of mathematical

construction, from infinitesimal calculus to differential geometry: it constitutes the space of meaning for the equations (formal determinations) of which we have spoken, the structure underlying any spatial manifold (Riemannian). However, a discretisation, or a representation that is finite, approximated, but ‘effective’, should be possible. It is the dream, implicit in Laplace’s conjecture, which will find its continuation in the foundational philosophy of arithmetising formalisms. If, as Laplace had hoped for the solar system, to a small perturbation always responds a consequence of the same order of magnitude (except in critical situations, cases which are ‘isolated’ — topologically — such as a mountain peak, of which Laplace was well aware), then today it would be possible to organise the world by means of well delimited little cubes (corresponding to the approximation of digital rounding; to the pixels on our machines’ screen) and to proceed to the arithmetic calculi upon these discrete values (the encoding of pixels by integers, sequences of 0s and 1s), which would then provide a ‘complete’ theory (any statement concerning the future and the past would be decidable, *modulo* the concerned approximation). Indeed, arithmetical rounding, which associates a single number to all the values contained within a ‘little cube’, does not perturb the simulation of a linear or Laplacian system, because the approximation which is inherent to it is preserved (*modulo* a linear growth) over the course of the calculi, just as over the course of physical evolution. Let’s explain ourselves, because a whole philosophy of mathematical foundations and, in fact, of nature, stems from this approach, with its own view on causality and determination.

3.1. Computer science and the philosophy of arithmetics

Digital computers are in the course of changing our world, by means of the very powerful tools for knowledge they provide us with and by the image of the world they reflect. They participate to the construction of all scientific knowledge via simulation and the elaboration of data. But they are not neutral: their theory, as formal machines, dates back to the 1930s, when effective computability, a theory of functions upon integers of integral value, imposed itself as the paradigm for logico-formal deduction. Induction and recursion, arithmetic principles, are at its centre. Our arithmetic machines and their techniques for the digital encoding of language (Gödelization) thus derive from a strong vision of mathematics, of knowledge in fact, rooted upon arithmetics; the latter has been proposed as locus of certitude, and of the absolute (the integer number, ‘an absolute concept’, for Frege), as locus of the possible encoding of any form of knowledge (‘of all which is thinkable’, Frege), of geometry in particular (Hilbert, 1899), as an organising theory of time and space. And certitude would be attained without preoccupation for the revolution caused by the geometrisation of Physics within non-Euclidean continua, with their variable curvatures, those of Riemann geometry (a ‘delirium’, with regard to intuitive meaning — Frege *dixit*, 1884); without this incertitude of

determinism deprived of predictability, characteristic of the geometry of dynamic systems since Poincaré. So a philosophy of arithmetics imposed itself upon foundational reflection, all the while departing from the new physics, which will mark the twentieth century. And it proposes that we read the world *modulo*, an arithmetic encoding, the same one enabling us to construct, from the world, the basis for modern digital data.

For this reason, the analysis of the constitution of intelligibility and of meaning, as intrication of mathematics with the world, is not traditionally part of foundational analysis in mathematics. The mathematical logic of Frege and Hilbert, with the profoundness of its achievements and the force of its philosophy, has led us to believe that any foundational analysis could be reconducted to the analysis of an adequate logico-formal system, a logical system (Frege), or a finite collection of sequences of meaningless signs (Hilbertian school), of which the meta-mathematical investigation would then become an arithmetic game (following the digital encoding of any finitary formal system), perfectly removed from the world. And, since Hilbert, as we have seen, the *formal coherence* of these calculi of signs claims to provide the sole justification of these systems, even those of the geometry of physical space and of theories of continua, as they be reduced to arithmetic.

This has definitely separated mathematical foundations from the foundations of other sciences, including physics, despite the roles of construction and reciprocal specification between these two disciplines, having a common constitution of meaning. With regard to biology, the foundational interaction has been lesser, for the time being, following the lesser mathematisation of this discipline. However, the ideology of the construction of the computational model as main explicative objective has already marked the interface between mathematics and biology, all the while forgetting the strong commitment to structuring the world, implicit in its computational arithmetisation; and few discussions have attempted to correlate the foundations of the arithmetising theories to those of the theories of life [see (Longo, Tendero, 2005)]. This epistemological separation makes difficult the interdisciplinarity and applications from one discipline to another, because foundational dialogue is a condition of possibility for a thought-out interdisciplinarity, a starting point for a parallel constitution of concepts and practices and for a common formation of meaning.

3.2. Laplace, digital rounding and iteration

So let's return to the 'bifurcation' having taken place in history: on one side, the arithmetisation of the foundations of mathematics (from Frege and Hilbert, although in different frameworks), and on the other, the geometrisation of physics (Riemann and Poincaré, in particular). The two branches have been quite productive: one the one hand we have the theory of effective computability and, therefore, our extraordinary arithmetic machines and, on the other hand, two fundamental

aspects of modern physics. The first branch of the bifurcation, however, in its foundational autonomy, has continued to base itself upon Newtonian absolutes (Frege) and upon the Laplacian determination (Hilbert), the one which involves the predictability (and which has its counterpart, in meta-mathematics, in the '*non ignorabimus*', Hilbert's decidability: once a mathematical statement is well formalised, one must be able to demonstrate either it or its negation — to falsify it).

It is actually quite clear that Laplace's hypothesis explicitly and first of all aims for predictability ('any deterministic system is predictable'; that is, in a formally determined system, any statement — concerning the future/past — is decidable). However, it bases itself precisely upon this 'conservational' interpretation of perturbation evoked earlier: Laplace is very well aware that physical measurement always constitutes an interval (it is necessarily approximated), but he believes that the solutions for the world's systems of equations, approximated if necessary by means of series (of Fourier), will be 'stable' with regard to small perturbations, particularly those of which the amplitude remains below possible measurement. The perturbation which by 'almost insensible variations' could even induce quite important secular changes — in his words, should not impede the stability of the solar system. It is this which guarantees predictability: in a system which is deterministic (therefore, in principle, formally determined by means of equations), predictability is ensured by the resolvability of the system and/or the *preservation of the approximations* under certain conditions (given the values of the initial conditions, with a given approximation, it will be possible to describe the system's evolution by an approximation of the same order of magnitude). There is the conceptual (and historical) continuity, which we have already addressed, between Laplace's conjecture and the myth of the arithmetisation of the world: approximation and rounding (discretisation) do not modify the evolutions under consideration, physical and simulated.

Yet it is nowhere like this. Even within a system which is (relatively simple and) explicitly determined by equations (the symmetric structure of formal determinations: the nine equations of three bodies within their gravitational fields, for example), unpredictability arises, Poincaré has explained. What happens? Even within such a simple system, almost everywhere small perturbations may give rise to huge consequences; in fact, 'small dividers' (which tend towards 0) within the coefficients of approximating series (of Lindstedt–Fourier), amplify the slightest of variations in the initial values. Ninety years later, this phenomenon will be defined as 'sensitivity to the initial conditions' (or at the border, or 'at the limits'). Particularly, even perturbations of which the amplitude is below the threshold of possible physical measurement may, after a certain amount of time, produce measurable changes.

So, in our interpretation, a perturbation, a 'small force' which perturbs a trajectory even below that which is measurable, *breaks an aspect of the symmetry* described by the equation of the system's evolution; it is the cause (efficient or

material) of a variation in the initial conditions, which may produce observable consequences, even very important ones. Sometimes, it may even be question of a fluctuation, such as a local or momentary breaking of the symmetry internal to the system, without the influence of ‘external’ causes: in the *Intermezzo*, we have evoked the Big Bang in cosmology, as a very first symmetry breaking (due to quantum fluctuations) of a highly energetic void. Once again, it is a broken symmetry that is the *material or efficient cause* of a specific observable evolution, that of our universe.⁶

Now, the intelligibility of these phenomena, present at the centre of modern physics, is conceptually lost if we organise the world by means of the exact values that arithmetical discretisation imposes. Or, rather, and here lies our thesis, we obtain a different intelligibility. Particularly, the perturbation or fluctuation, which have their origin in efficient or material causes, and which manifest *below the proposed discrete approximation*, elude arithmetic intelligibility, or are neglected in favour of a forced stability of phenomena. And arithmetic calculus shows us the passing from one state to another by little iterated jumps of trajectories that are imperturbable, because perfectly iterable. Better, it shows us trajectories which are affected by their own intrinsic perturbation, at each increment of calculus, and always identically iterated and iterable: *the rounding-off*. A new cause, our computational invention, which, projected into the world, becomes a relevant cause (efficient or material) with regard to the properties and states of a system. Because digital rounding modifies the simulated geodesics and may even change, in certain cases and in its own way, the conservation phenomena (of energy, of moment...) by breaking the associated symmetries. We will return to this point.

In what sense then, do we obtain, when we superimpose upon the world an arithmetic grid, a ‘forced stability’, as well as evolutions and perturbations which are very specific and ‘iterable’ at will? We will understand this thanks to the digital computer, because, when this arithmetic machine is used as *model of the world*, it organises the world according to its own causal regime, its own symmetries and symmetry breakings. In fact, the digital simulation of a physical process is constitutive of a new objectivity, to be analyzed closely because very important, due to the mathematical tools which are at its centre: the arithmetic calculi and discrete topology of its digital databases and of its working memory space, exact and absolute.

⁶According to the Curie principle, ‘the symmetries of the causes can be found in the symmetries of the effects’. In the approach followed here and as we have observed, one would say that in these cases, at the level of the observable, this is not the case: to an apparently symmetrical initial situation may follow an observable evolution which does not reproduce the same symmetries, following a breaking of symmetry, initial or at the edges, of which the amplitude, initially, is below the threshold of possible measurement (therefore non-observable). In these cases, therefore, certain symmetries are not conserved when passing from (observable) causes to (observable) consequences.

It is clear that our analysis does not aim to oppose what would be an ‘ontology’ of continua to what would be an ontology of discrete mathematics (we are not defending the idea according to which the world itself would be continuous as such!). We are rather attempting to highlight the difference of the views proposed by discrete mathematics in relation to those proposed by continuous mathematics, in our efforts to make the world intelligible. It is the constructed objectivity of mathematics which changes and not, we repeat, any ontology.

Moreover, the relative incompleteness of computational simulation, which we emphasise here, goes hand in hand with the mathematical incompleteness of arithmetic formalisms which, it too, is relative (to the practice of mathematical proof, in this case). But incompleteness does in no way mean ‘uselessness’: on the contrary, we emphasise the need for a fine conceptual analysis of algorithmic methods, precisely for the essential and strong role they play today in any scientific construction.

3.3. Iteration and prediction

Computers iterate; that is their strength. From primitive recursion, at the centre of the mathematics of computability, to the software application, the program, the sub-program, re-run a thousand times, a billion times, once every nanosecond, all reiterate with absolute exactitude. For this reason, there is no randomness as such in a digital world: (pseudo-)randomness generators are small programs, perfectly iterable, which generate periodic sequences endowed with very long periods (they are functions iterated upon finite domains).

In the algorithmic theory of information, we call random any sequence of integers for which *we do not know* a generating program shorter than the sequence itself. That is, that we do *not see* any sufficient regularities within the sequence to be able to deduce a rule by which to generate it. This definition identifies with randomness the informational characteristics of a series of throws of dice or of roulette, in fact, their incompressibility. This identification, applied to algorithms, (to pseudo-)random generators for instance, leads to a confusion between an epistemic notion of randomness, specific to physics, and ‘randomness by incompetence’, (the programmer has not told us how the program is designed, usually a one-line program). And iteration reveals the trick: if we re-launch a programmed (pseudo-)randomness generator, using the same initial values, we will obtain the same sequence, exactly. On the other hand, dynamical systems give us the good (epistemic) notion of randomness: a process is random if, when we iterate it with ‘the same’ initial conditions, it does not follow, generally, the same evolution (dice, roulette... planetary systems having at least three bodies, if we wait long enough). All the difference lies in the topological signification of this notion of ‘same’ (*same* discrete values and *same* initial conditions): a digital database is discrete and *exact*; whereas physical measurement is necessarily an interval.

In short, in the mathematical universe of effective computability, there is no randomness as such, at best there is incomprehensible information (which may provide a good ‘imitation’, see below, of randomness). And one may say synthetically with regard to our approach: the time of calculation processes is subject to a ‘symmetry’ in terms of iterability (identical repeatability), which does not have an absolutely rigorous meaning in the physical world and even less so in that of living phenomena. This iterability is essential to computer science: it is at the centre of software portability, therefore, of the very idea one may have concerning software; that one may transfer it onto any adequate machine and run it and re-run it identically as often as one wants. And it works, in fact.

Of course, computers are in the world. If we come out of the discrete arithmetics internal to the machine, we may plug them upon physical randomness (epistemic — dynamic systems — or intrinsic — quantum physics, see Appendix). We may, for instance, use temporal shifts within a network (a distributed and concurrent system, see (Aceto *et al.*, 2003)] upon which humans also intervene, randomly; or using little boxes, sold in Geneva, which produce 0s and 1s following quantum ‘spin-ups/spin-downs’. But normally, if you run the simulation of the most complex of chaotic systems, a Lorentz attractor, a quadruple pendulum... and iterate with *the same initial digital data*, you will obtain the same phase portrait, the same trajectory. The same initial data, there is the problem. As we have emphasised earlier, this physical notion is conceived *modulo* possible measurement, which is always approximated, and the dynamic may be such that a variation, including below the threshold of measurement — the material or efficient cause -, (almost) always generates a different evolution. On the other hand, in a discrete state machine, ‘the same initial data’ signifies ‘exactly the same integers’. This is what leads Turing to say that his logico-arithmetic machine is a Laplacian machine [see (Turing, 1950; Longo, 2007)]. Like Laplace’s God, the digital computer, its operating system, has a complete mastery over the rules (implemented in its programs) and a perfect knowledge of (access to) its discrete universe, point by point. As for Laplace’s God, “prediction is possible” (Turing, 1950).

And so thus is the philosophy of nature implicit to any approach which confounds digital simulation with mathematical modelling, or which superimposes and identifies algorithms to the world. Discrete simulation is rather an *imitation*, if we recall the distinction, implicit in Turing, between model and imitation [see (Longo, 2007)]. Very briefly: a *physico-mathematical* model tries to propose, by means of mathematics, the *constitutive* formal determinations of the considered phenomenon; a functional imitation only produces a similar behaviour, based, generally, upon a different causal structure. In the case of continuous vs. digital modelling, the comparison between different causal regimes is at the centre of this distinction.

Turing has the huge merit of having invented the machines and of having, in 1950, when he abandoned the myth of the great digital brain, highlighted the

difference between *computational imitation*, the game which should demonstrate his machine to be indistinguishable from a woman [modulo the sole intermediary of a written interface, (Turing, 1950)], and *modelling*. In fact, in his 1952 article on morphogenesis, he presents a *model* of chemical actions-reactions which generates forms and which bases itself upon a system of differential equations: tiny variations generate the variety of forms in certain natural phenomena (Turing calls this sensitivity “exponential drift”, a quite relevant and original name! The notion of “sensitivity to the initial conditions” dates back to the 1970s). This mathematical model, which he explicitly says could be false, nevertheless attempts to make the world intelligible; the 1950 imitation tends for its part to trick the observer (and will thus be considered at the origin of Classical Artificial Intelligence).

3.4. Rules and the algorithm

Computerised simulation transforms all physical evolutions into an elaboration of digital information. Particularly, the simulation of a geodesic in a discrete universe should make a digital computation correspond to a trajectory, and make the conservation of information correspond to laws of conservation (energy, movement...). Any state or property, in short, any physical quantity, as determination of objects (in the sense of D.1), are in fact encoded by digital information; the quantity of movement is encoded using 0s/1s, just as is the intensity of a field or mass, and their evolution is a calculus approximated by these 0s/1s. Is this encoding ‘conservational’ (does it preserve that which is important)? If a physical trajectory is a geodesic, which geodesic do we associate to the calculus in its digital universe, which symmetry breaking do we associate to the rounding?

Let’s begin by recalling the generality of the geodesic principles in physics at the centre of this science since Copernicus, Kepler, and Galileo. As we observed already, any fundamental law of physics is the expression of a geodesic principle applied within the appropriate space. The work of the physicist, who organises and, by that, makes the phenomena intelligible, consists in a good measure to the search for this space (conceptual, or mathematical) and for its relevant metric.

This approach leads us to understand the slide of meaning around the concept of law, which intends to justify the identification of mathematical modelling with computational imitation. The notion of ‘law’ has a social origin: law is normative to human behaviour. The transferral of the concept as it is to physics corresponds to an ordinary metaphysics: an *a priori* (divine, if possible) which would dictate the world’s laws of evolution. Matter would then conform itself to this pre-existing and normative ontology (as mathematical laws of a platonic universe, for instance). On the other hand, the comprehension of the notion of law as *explicitation of the regularities and criticalities of a landscape*, with its mountain passes, valleys and peaks, its geodesics, inverts this and highlights the transcendental constitution at the centre of any construction of knowledge. Mathematics, the tool of formal determination,

then elaborates itself upon the phenomenal veil at the interface between us and reality, the reality which, of course, causes friction and canalises the cognitive act, but which is also organised by this same act. Laws are not ‘already there’, but are a co-constituted in the intrication between ourselves and the world: the discernability of geodesics, as formal determinations within the framework of a network of interactions, is its main result. And their mathematical processing coincides with the beginning of modern science. The normativity of physical law then becomes only cognitive (in order to construct knowledge), and not an ontology. The different forms of formal determination (laws) propose to us different causal regimes, ulterior tools for intelligibility.

The identification between algorithm and law causes us to make a backwards step: the algorithm is normative for the machine, for its calculi, exactly as God’s law governs any trajectory. The machine would not know where to go; it would be static, without its primary motor, the program. Once again, the myth of the computer-Universe (the genome, evolution, the brain ... all governed by algorithms) consists in a metaphysics and a notion of determination which precedes the science of the twentieth century and for good historical reasons: the re-centring of the foundations of mathematics upon a philosophy of the arithmetic absolute, at the fringes of the time’s great scientific turning points, as we have mentioned a few times.

This way of understanding law should highlight the very first difficulty for the computational simulation of a physical trajectory by means of a calculus. Physical law and algorithm therefore do not coincide: they do not have the same epistemological status. Law is also not an algorithm for another reason we have already evoked: the formal determination, as mathematical explanation of laws, does not imply the predictability of physical evolution. On the other hand, any algorithm, implemented within a discrete state machine, generates a predictable calculus, at least thanks to the ‘symmetry’ by repetition within time which we have mentioned earlier (identical iteration always being possible for a sequential computer).

However, we absolutely need digital simulation, a tool which is indispensable today to any sort of construction of scientific knowledge: by highlighting the differences, outside of any computational myth (the world would be like a computer), we aim to better identify that which is do-able, and afterwards to do better, in terms of simulation-imitation. Let’s attempt then to understand the evolution of a calculus in physical terms.

In the case of an isolated computer — a sequential machine — we remain within a Newtonian framework: the absoluteness of the clock and of the access to a database are its essential characteristics. The situation is more complex with modern networks: the distribution of machines in physical space and the ensuing relational time changes the situation. Certain aspects of the absoluteness of Turing machines are thrown into question [we discuss this in detail in (Longo, 2007) and, more technically, in (Aceto *et al.*, 2003)]. However, the *exactitude* of the discrete database subsists, as well as the issue of rounding, of course.

In both cases, of sequentiality and concurrence, we may nevertheless understand calculus as a geodesic within *the space (pre)determined by the program* (more specifically: by the programming environment, or all aspects of software — operating system, compilers and interpreters, programs...). Shortly, while accepting the divine *a priori* of the programmer who establishes, beforehand, the rules of the game, the notion of ‘following a geodesic’ would be defined by *following the rule* correctly. The hardware or software bug would then be the fluctuation or perturbation that causes the evolution to derail. However, this type of bug is not integrated to the theory; it is not inherent to it, contrarily to the theory of dynamic systems which integrates the notion of sensitivity to the conditions at the edges as well as measurement by interval. Moreover, hardware bugs and logical errors are very rare (therefore statistically very different to the variation due to approximation within a dynamic); they are to be avoided and, in principle, are avoidable (or they may belong to another phenomenal level, which is far from being integrated within the mathematics of effective calculus: quantum physics).

We are left with the issue of rounding, which is inherent to calculus. The use of rounding, today, can be very dynamic and mobile: one can aim to a desired approximation and the end of a calculus and increase beforehand the available decimals, up to hundreds, to end within the targeted interval, if possible. The modern approach to analysis by intervals provides a powerful theoretical framework for these processes (Edalat, 1997). Of course, the speed of the calculi is inversely proportional to the improvement of the approximation. An excess of the latter may prevent us from following any dynamic long enough.

This being said, this bound, the rounding, constitutive of the arithmetisation of the world (a quite necessary arithmetisation if one wants digital machines to perform calculi and therefore to participate in science today), modifies the causal regime and the symmetries correlated to it, as we hinted and keep demonstrating.

Firstly, let’s remove a possible confusion: the interval inherent to classical physical measurement and quantum incertitude have nothing to do with digital approximation. First, measurement as interval is a physical principle, a classical one; it is not a ‘practical’ issue: thermal fluctuation, at least, is always present above absolute zero, by principle. And, as we have already and often observed, the fluctuation or perturbation below the observable amplitude participate in the evolution of a dynamic system, which is somewhat unstable, because it can break the symmetries of the evolution, and thus be one of the causes of a specific trajectory. In computer science, a bug which manifests below the rounding is without effect. Afterwards, the analogy sometimes made — naively — between digital discretisation and that of elements of ‘length’ (time and space) induced by the Plank constant, h , is not relevant. The non-separability, the non-locality, the essential indetermination of which we speak in quantum physics are almost the opposite of the certitude of the little boxes, well *localised* and stable, well *separated* by predicates (the memory addresses), within which is distributed the digital universe.

So we now face the principal issue: rounding entails a loss of information, at each step of the calculus. It can be associated to the irreversible growth of a form of entropy, defined as neg-information. So if one encodes all determinations, formal and objective, of a physical object and process, all properties and states, in the form of digital information, the elaboration of the latter, the digital calculus, will follow a geodesic which is, normally, perturbed at each step by a loss of information. This perturbation does not correspond to any phenomenon intrinsic to the process we intend to simulate: the loss of information is not, generally, the encoding of the change in objective determination. It is a new type of symmetry breaking. Will this influence the proximity of the virtual reality to the physical phenomenon? Will it influence the quality of the imitation?

As we have already observed, arithmetic approximation does not affect the simulation of a linear or Laplacian process: just as the approximation of the measurement, the rounding, does not remove the computational geodesic (the following of the rule) from the physical geodesic. The initial loss of information is preserved, it remains of the same order of magnitude or it increases in a controlled way (technically: the extremes of the approximation intervals are preserved). This is not the case for non-linear cases. Let's consider, for illustrative purposes, one of the simplest dynamics, one that is well-known and one-dimensional: the discrete logistic equation,

$$x_{n+1} = kx_n(1 - x_n).$$

For $2 \leq k \leq 4$, this equation formally defines a sequence $\{x_i\}$ of real numbers, between 0 and 1 (a "time discrete" trajectory within a continuous space). Particularly, for $k = 4$, it generates chaotic trajectories (sensitive to the initial conditions, dense in $[0,1]$, with an infinity of periodic points...). Can we approximate any sequence of real numbers thus generated by a digital computer? Out of the question, at least in what concerns an initial value of x_0 taken from a set of measure 1 (that is, for almost any real value in $[0,1]$). Even if we choose a x_0 that can be represented exactly by a computer, at the first rounding in the course of the calculus, the digital sequence and the continuous sequence will begin to diverge. By improving the approximation/rounding of 10^{-14} to 10^{-15} , for instance, after approximately 40 iterations, the distance between the 2 sequences will start to oscillate between 0 and 1 (the greatest possible distance). Likewise if, with a rounding of 10^{-15} , we begin using values that differ from 10^{-14} (of course, if we want to, nothing would prevent us from restarting the digital machine upon the exact same values and from calculating, with the same rounding, exactly the same discrete trajectory...). The technical problem can be summed up by the observation that the dynamic is a 'shuffling' one: the boundaries of the interval are not preserved.

We therefore may not, in general, approximate, with the machine, a continuous trajectory; however, we may do the opposite. In fact, all that can be proved, in dynamic (metric) contexts we will not specify here, is the following "pursuit"

lemma [see the *Shadowing Lemma* (Pilyugin, 1999): notice the order of the logical quantification]:

For any x_0 and δ there is a ε such as that for any trajectory f , ε -approximated (or with a rounding-off not greater than ε , at each step), there exists a continuous one, g , such as g approaches f by a difference of δ , at each step.

Even when considering the lucky case where we have $\delta = \varepsilon$ (this is possible in certain cases), it comes to say that, globally, your digital sequences are not so ‘wild’: they can be approximated by a continuous sequence, or ... there are so many continuous trajectories that, if we take a discrete one, one can easily find a continuous one which is close to it. Thus, the image of an attractor on the screen provides qualitatively correct information: the digital trajectories are approximated by trajectories of continuous dynamics (determined by equations). But the reverse is not true: that is, it is not true in general, for a trajectory given by analytical means, that the computer may always approximate it. Different versions of the pursuit lemma apply to sufficiently regular chaotic systems. However, many dynamic systems do not even satisfy weak forms of this lemma [see (Sauer, 2002)]. This signifies the existence of initial values and intervals such that, within these intervals, any rounding and any other initial value cause any continuous sequence to diverge from the given discrete sequence.

What happens, in the terms of our approach, which is geometric, in nature? To understand this in detail, it would be necessary to refer to the technical analysis which the authors are also developing. In this work of reflection, which nevertheless guides the mathematical and computational analysis, let’s try to see it in a very informal manner. The first difficulty resides in the necessity to place oneself within the appropriate space, in order to better understand. In short, it is necessary to analyse the evolution of a system, such as for instance the discrete logistic function, within a space where the notion of neighbourhood, between real numbers, corresponds to digital approximation. A space which provides for such a metric is called a ‘Cantor space’. In this space, which we will not define here, two real points are close if and only if they have close binary or decimal representations (for instance, 0.199999... to infinity and 0.2 are very far from one another in the Cantor space, whereas they are identical with regard to the usual real line, this posing numerous problems from the computational standpoint, when we try to operate upon their approximations).

We then see that at each of the digital calculus’ iterations, the rounding induces a loss of information corresponding to the deterioration of the approximation around the point of the trajectory. If we measure the phenomenon in terms of isotropy of space (the points within this widening neighbourhood are ‘indistinguishable’, in a manner of speaking), this ‘grey’ zone, of isotropy, grows, thus augmenting the symmetries of space. A notion of entropy as negative information also enables one to grasp this change in symmetries as loss of information. Now, all that we have in

the machine is encoded information. Independently of what it encodes, the physical object's formally determined properties or states, all is in the form of digital information. So the objective determination, which is given by the preservation of the theoretical symmetries, radically changes: we are facing a change in symmetry that does not model a component of the evolution of the natural phenomenon, because it depends only upon the discrete structure of the simulation universe and upon the imitation of the formal physical determination by algorithms (or, when we make a philosophy of it, of the epistemological identification of *law* with *algorithm*).

So there is, in terms of symmetry, the explanation of the causal regime of which we were speaking. The discretisation, in fact the organisation of the world by means of discrete mathematics proposes a causal regime (in this case, an evolution of symmetries) which is different than that which is proposed by continuous mathematics. It is not an issue of finitary translations of a same physical world, but of scientific construction, because this world is itself co-constituted by our formal and objective determinations. When they change, its organisation and its intelligibility also change. Once more, this does not imply that the world is continuous 'in itself': we are only observing that, since Newton, Leibniz, Riemann, Poincaré — we have organised and made intelligible some physical phenomena by means of historical notions of continuity and of limit. If we want to do without them, causal organisation and intelligibility will be altered.

Another issue would also merit investigation, but we will leave it for another study. Singularities in modern physics play an essential role. We know for instance of shock situations, in non-linear systems, where the digital calculus does not come even remotely close to the critical situation. We have the continuous description; the mathematics are clear, explicative, organising for the physical phenomenon, we understand qualitatively, but the numerical calculi chaotically revolve around the singularity, without coming close to it. In fact, the current notions of limit and of singular point, which are absolutely necessary to the analysis of the phase changes, the shocks, in order to even speak of renormalisation processes in physics, are not always coherently approximable. The loss of symmetries and the change of correlated causal regime constitute our way of understanding this problem, which is specific to the digitalisation of phenomena, without referring to Laplacian myths and computational metaphysics. Computer science, a science which is now mature, deserves, from an epistemological and mathematical standpoint, more attention and a view from within which that is able to assume the force and the limits of its own methods.

4. Causalities in Biology

While focusing now on biology, we will not address the discussions concerning the biological levels of organisation, the intertwined hierarchies, the crossed

causalities, the ago-antagonistic effects, the variabilities within phenomena, the autopoietic processes, which we find in biology. Of course, all these properties will remain part of the backdrop of the approach which we propose here, but the approach will be more conceptual or, better, ‘schematic’ (in the geometric sense of simple schemata or diagrams) than thoroughly theoretical or descriptive: it will seek rather to contemplate a framework of representation enabling us to extricate heuristic categories of thought rather than to account for the effective phenomenality of life. Indeed, the ‘relationships’ which we highlight, by means of very abstract little patterns, do not necessarily correspond to ‘material relationships’ or to physical configurations; they are only organising structures of thought, which should aid the comprehension of phenomena, by proposing a conceptual framework. Moreover, is $F = ma$ — at a much more elaborate and mathematised level indeed — not a correlation which organises a phenomenon by making it intelligible? Let’s also recall that this equation has been preceded by the general concept of inertia, or even, way before Galileo, by cosmological speculations and concepts as eminently philosophical, as profound (see, for instance, the remarks of Giordano Bruno in *‘L’infinito universo e mondi’*, 1584). The physical intelligibility specific to this equation can be the object of highly differing conceptual ‘readings’: it may no longer be primitive, but derived (from the Hamiltonian, from the Lagrangian, as we have mentioned in the first part), where it may be correlated to distinct symmetry breakings, as we have also seen.

So, as we insist on emphasising from the onset, our approach remains very speculative here: it is for us the beginning of an attempt at a conceptual categorisation and schematisation which seeks to open new avenues without being sure of their directions and which will require, in order to be continued, more discussion with biologists and the sanction of a certain fecundity in the quest for a greater understanding of living phenomena. Of course, we will remain within the framework which we have determined for ourselves, that is in the geometric terms of symmetries as an analysis of physical causality; yet, we will *also* take into account here aspects which are specific to biology, related to forms of teleonomies or of anticipation and which we have already summed up with the concept of ‘contingent finality’ (see the next section).

4.1. Basic representation

Let’s consider the dynamic functioning of a centrifugal governor (or Watt governor: two weights are lifted in rotation by pressure, making a valve open so as to lower the pressure). This functioning is completely determined by the data, obtained afterwards, concerning the initial conditions and the physical laws. In this sense, its ‘behaviour’, though being well regulated and leading to a dynamic equilibrium, is determined in a univocal and oriented way (geodesics within a well and pre-established phase space).

In the case of living phenomena, the behaviour (and functioning) of an organism does not appear to be determined in the same way. What appears to be determined like this in a more or less rigid way (within a given domain, that is compatible with the organism's survival), is what we could call the aim of the functionings and behaviours, the functions to be fulfilled in order to ensure homeostasis (-rhresy); but what is not determined in such a manner, are on the one hand the possible ways to achieve this and on the other hand the adaptations and modulations which would ensure the achievement of the functions and behaviours. Specifically, to put it in the language of Physics, there are not only phase changes but also, changes in phase spaces, that is, observables and relevant variables.

We know that the mathematical and equational formalisation of this situation (as took place for Physics and as we hope, in time, could be the case for life sciences inasmuch as the adequate mathematics would be elaborated) is confronted with profound difficulties, sometimes even difficulties of principle, which the numerous and successive attempts at modelling have encountered. Also, before any reiterated attempt in this regard, it appears to be necessary to try to illustrate and to represent — in this case by means of schematics, the first abstract conceptual stage — that which appears to characterise these modes of functioning and that which we could call the 'finalities' which interpret them, in the sense in which Monod could speak of a *telenomy* of life. These finalities, of course, are neither necessary nor absolute; they rather contribute to our view of living phenomena and, most of all, they are contingent, as they are specific to living matter and relative to its contexts. In short, they could not be present (no life, no specific specie or individual); they are relevant to various levels of organisation and to their correlations, or to their intertwining and looping, particularly in the form of integration and regulation [see (Bailly, Longo, 2003)].

To make intelligible the notion of contingent finality, we will try to organise into networks the interactions between the 'material structures' and 'functions' of living matter. It is at this level in fact that telenomy manifests: for instance, when an organic structure appears to be finalised in relation to a certain function.

We therefore propose a conceptual framework, to organise knowledge by means of our recourse — temporarily at least — to a description of this sort (for a better, understanding, the reader may easily draw the simple intended diagrams):

- (1) We have a target set, constituted of several domains — the target domains, which may or may not overlap — corresponding to the functions to be ensured for the maintenance and the perdurance of the organism and its species.
- (2) We have a source set constituted of all the organic possibilities likely to be mobilised to this end (biochemical reactions, transport agents, etc), also represented by this set's domains (source domains).
- (3) We have a set of arrows, originating in the source domains to reach to the target domains (these arrows correspond to the orientations and functioning

modes aiming to ensure the functions) and which presents the following particularities:

- (i) Any target domain is reached by at least one arrow; usually, several source domains are at the origin of the arrows reaching a same target domain. An example of this situation is the conjunction of ‘oxygen metabolisms’ and of ‘glucose metabolisms’ to ensure the maintenance of muscular tissue such as the heart; in this case both source domains are respectively defined by the chemical reactions related to the specific energetic sources (availability of glucose) and by the cellular assimilation processes of the intake of oxygen obtained through breathing (availability of oxygen), the arrows corresponding for their part to the various transport and transformation systems which enable the effective transferrals from sources to targets.
- (ii) The arrows pointing to a same target domain are endowed with different widths depending on the prevalence of the usual modes of functioning (in the preceding example, we would have, in the normal case, an arrow width for the ‘oxygen metabolism’ that is much greater than that of the ‘glucose metabolism’. In the case where a dominant mode of functioning would fail (pathology), the corresponding arrow could narrow down to the benefit of another whose initial width was smaller (and this, without necessarily reaching the width of the first one: functional weakening all the while attempting to preserve the function): this mechanism would correspond to a property of *plasticity*.
- (iii) The arrows stemming from a same source domain and extending towards several target domains exist, but may be relatively rare in adult homeostatic (-rhetic) functioning. They refer mainly to potentialities, preceding ulterior actualisations or differentiations (cf. stem cells, for instance), or to other possibilities of plasticity (cerebral, for instance). On the other hand, in the case of the representation of a *genesis* (embryogenesis, namely), these arrows are dominant and play an essential role in the representation of the organic differentiations from totipotent eggs or from pluripotent stem cells. There is therefore a dynamic for the ‘topology’ and the width of the arrows over the course of development to reach the adult situation.

It could be interesting and enlightening to note here that the joining of the characteristics (i) and (iii), for the arrows, corresponds rather well to the concept of *degeneracy* such as it has been introduced by (Edelman, Tononi, 2000) with regard to cerebral functioning (that non-isomorphic structures may participate to a same functionality and that a given structure may participate to several of these functionalities), concept which returns to and generalises that of *redundancy*, but while differing somewhat from it (computer redundancy, for example,

is achieved by iteration of identical components). In this perspective, we could qualify the described situation by the characteristics in (i) of ‘*systemic*’ *degeneracy* (a same system participating to distinct functions) and the characteristics in (iii) of ‘*formal*’ *degeneracy* (non-isomorphic systems participating to a single and same function).

Let’s also point out right away that the concepts of ‘source domain’ and ‘target domain’ do not necessarily refer to ‘absolute’ categorisations, but are relative to a given functionality (or to a set of functionalities): a target domain for a functionality can very well operate as a source domain for another⁷ on the same level of organisation or between levels, thus the many possible intertwining.

Notice on the other hand that, in this approach, the environmental, feed-back or adaptation effects can be represented by variations in the widths of the arrows (‘metric’ aspect), or that the fundamental changes would rather correspond to changes in the structuring of the set of arrows (‘topological’ aspect). Moreover, pathology is likely to occur (in order of ‘seriousness’):

- either with a variation in the width of the arrows,
- either with the disappearance of certain arrows (without nevertheless a target domain being no longer concerned at all)
- or with the disappearance of the source domains (in this case, grafts and prostheses can play an “artificial” regulatory role).

We may consider that the disappearance of the target domains corresponds at best to a mutation, and in the worst of cases, to death.

To give a few ‘systematic’ examples of the functioning thus represented, we can propose the following triplets (by starting by the source domains, then arrows — in fact corresponding to functions — and finishing by target domains):

- Vascular system/circulation (transport)/local essentials (nutrients, oxygen, etc.);
- Respiratory system/breathing/oxygenation;
- Nervous system/information, command/adaptation, initiative;
- Genes/expression/proteins, regulation;
- Mitochondria/biochemical reactions/energy produced;
- Digestive system/digestion and transport/metabolism;
- Immune system/reconnaissance/tissue identity, struggle against aggressions.

⁷For instance, the putting into effect of ionic equilibrium processes may constitute a source domain for the functioning of the target domain represented by a cell, itself constituting a source domain for the good functioning of the tissues to which it participates, good functioning representing one of its target domains. It would go likewise for cerebral functioning, for example, as target domain for an oxygenation and as source domain for a control or behaviour.

4.2. On contingent finality

Based upon these considerations, we can propose to call *contingent finality* the abstract structure formed,

- (1) by the triplet {source domain, arrows, target domains}
- (2) endowed with the ‘measurement’ constituted by the set of real numbers E , of the widths of the n arrows: $E = \{e_1, e_2, \dots, e_n\}$
- (3) ensuring a *structural stability* for these characterisations. We mean, by such structural stability, the conservation of the target domains in the sense that there will always be at least one arrow for which the width is non null and which points to these targets, regardless of the source domains.

Let’s go back to the preceding example and let’s attempt to compare a normal state to a pathological state. In the normal state, the ‘oxygen metabolism’ arrow has a width of e_{O1} and the ‘glucose metabolism’ arrow has a width of e_{G1} , with $e_{O1} \gg e_{G1}$ and $e_{O1} + e_{G1} = e_1$. The establishment of the pathological state is translated by a narrowing of the ‘oxygen’ arrow and the widening of the ‘glucose’ arrow; finally, we have $e_{O2} < e_{O1}$, $e_{G2} > e_{G1}$, $e_{O2} + e_{G2} = e_2 < e_1$.

The fact that the arrows do not cancel each other out and that the target domain remains translates a partial *plasticity*, whereas the decrease of the total width, on the one hand, and the internal rebalancing of the widths, on the other, demonstrates the pathological character. The influence of these two factors (total width and respective widths) could indicate that total plasticity (in the sense where we would finally have $e_2 = e_1$) does not however restore a completely ‘normal’ situation.

From a much more general point of view, we will notice that — as we have already highlighted through the approach by levels of organisation, previously considered — the same structure of ‘contingent finalisation’ thus defined, replicates itself at various levels of organisation of biolons (cell, organism, species), even if the characterisations (triplets and measurements) may differ in their specific content, according to the level. This structural likeness is doubtlessly the result of a certain form of equivalence of the objective complexities associated to these levels, as we have already noted, (Bailly, Longo, 2003).⁸

⁸Let’s recall that, according to our analysis in (Bailly, Longo, 2006) (where we had distinguished between objective complexity and epistemic complexity in Biology, also providing examples), the elements of living matter which are biolons present an objective complexity which may be considered as being infinite, with respect to any physical measure (crossing of the essential level of organization which enables to pass from inertia to life). From this point of view, and still with regard to physical complexity, the objective complexity of biological objects is comparable regardless of what these biolon-type objects are (the living cell presents an objective complexity almost identical to that of an organism such as a mammal). What is modified along life’s scale of complexity is epistemic complexity, related to the enriching structure of phenotypes, to the increasing, along evolution, levels of organization, their intertwining, the proliferations of structures and functions, the conditions of description, etc.

4.3. 'Causal' dynamics: development, maturity, aging, death

If we accept the schema we have just proposed, it proves likely to represent, thanks to the topological and 'metric' plasticity it is able to demonstrate, the great dynamic processes of which life can be the locus: the beginning of development is characterised by the prevalence of arrows which stem from a source domain to point towards several target domains which they even contribute to constituting (differentiation of tissues and of anatomical and physical systems). As the process unfolds and at the same time as the number and structure of the target domains stabilises, these arrows narrow down (some may even disappear) at the same time as the arrows originating in several source domains ending in a same target domain (functional aims) start to prevail. The set stabilises once again following development, the period of maturity.

Once the stability of the maturation achieved, the topology maintains itself 'on the whole' and aging manifests itself mainly in a 'metric' fashion (by the variation of the *measurement* of the narrowing of arrows). It is even possible that in borderline cases one may witness disappearances of arrows by cancellation of their widths, amounting, beyond metrics, to tapping onto the topological structure of the schema. And finally, to represent the death of an organism, we may agree, as suggested above, that it manifests itself by the disappearance of one or more target domains (corresponding to vital functions), in that there are no more arrows pointing towards them.

We will note that if most of an individual's target domains are oriented towards the individual's perdurance, at least one of them, corresponding to the reproductive function — is likely to produce a new source domain (child cell, fertilised egg) as origin of the reiteration of the process for a new individual. It is the set formed by the abstract joining of this particular target domain to the newly produced source domains which may constitute — at a different level — the source domain at the origin of the genesis of individuals of the level thus considered (organism for cells, species for individuals).

From the standpoint of an attempt at a more precise 'phenomenal' identification of the characteristics we have just introduced abstractly, we may consider that in the initial 'transitory regime' (time of genesis) the source domains are principally constituted by biolons [embryonic cells, individual organisms, species, as defined in (Bailly *et al.*, 1993) and further analysed in (Bailly, Longo, 2006)], the target domains being mainly constituted by orgons (cell's organites, organs and tissues to be set). The arrows correspond for their part to the phenomena of differentiation, of migration and of structuration, whereas, conversely, in the 'stationary regime' (adult organism), the source domains are mainly constituted by the constitutive orgons. Note also that the target domains would be constituted by the variety of vital functions ensuring the organism's maintenance and autonomy whilst the arrows corresponding this time to the biochemical and physical processes ensuring these

functions (integration and regulation). Such an approach suggests to propose a sort of ‘temporalized’ schema of biological functioning.

Could we refine the analysis by taking more precisely into account the nature of the fluxes which link source and target domains to one another? Namely by putting the distinction between energy and information to work? In a first approach, it appears legitimate to consider that the fluxes in the source/target direction mainly have an energetic character (transport of matter or energy), responding to fluxes which are mainly of information (gradients, divergences from the dynamic equilibrium) going in the target/source direction. The arrows are then supposed to integrate and represent both types of fluxes, their width alterable in the event of a failing of either the correlated ‘informative’ or of the ‘energetic’ character (we could, in a first approximation, take as a parameter the product of these two types of fluxes, for instance⁹). Let’s try to take an example at one of the most elementary levels, that of the cell: in this case, a particular source domain can be associated to the functioning of ionic channels enabling the ions to cross the cellular membrane and a corresponding source domain would be the stationarity (dynamic equilibrium) of the cell’s internal ionic state (homeostasis — homeorhesis) which enables it to function in the best conditions. The arrows would then correspond to the taking into account of both ‘fluxes’: on the one hand, the ‘information’ flux which would be generated by a difference in the internal ionic concentration with regard to the stationary state (gradient, difference in osmotic pressure, electric field, ...) and which would lead for instance to the opening of certain channels, and on the other hand the concomitant flux of matter (these same ions) coming from the outside in view of re-establishing homeostasis and entering through these channels.

Notice that, from a standpoint analogical to this stage, the taking into account of these two aspects (matter/energy and information) highly resembles the thermodynamic situation where the definition of free energy (of which the variations govern the system’s evolution) entails the intervention on the one hand of an enthalpy (or internal energy) and on the other hand of an entropy, these two magnitudes being associated by means of temperature.

4.4. Invariants of causal reduction in Biology

As in the case of Physics, we may question ourselves regarding the invariants of causal reduction (if they exist) specific to life and their relationships with what could

⁹If E is the matter-energy flux extending from the source domain to the target domain in order to ‘respond’ to the information flux (‘request’) going from target to source domain ‘within’ a given arrow, we could take as one of the parameters of functioning — which participates to the width of this arrow — the product $E \times F$. Thus, the failing of a flux in one or the other direction would translate as a diminution of this product, corresponding to a decrease of the width of the arrow and thus expressing an alteration of the functional process summarised by this arrow.

constitute, in the field of Biology, the determinations associated to the symmetries which we have encountered for Physics.

It seems clear that these biological invariants indeed exist and that they are constituted of sets of pure, numerical invariants (and not dimensional invariants as in Physics). It also appears that the determinations which enframe, modulate and actualise them are now rules of ‘scaling’ in function, let’s say, of the size or mass of the organisms (sorts of dilatation or scale symmetries), see also (Schmidt–Nielsen, 1984).

Thus, for instance, the average life-spans of the set of organisms do appear to scale as power 1/4 of their masses and their metabolism as power 3/4 of these masses (Peters, 1983). Likewise, on a level that is somewhat different but which is in relationship with these properties, we would recall here that mammals are characterised by an invariant average number of heartbeats or breathings (of the order of 10^9 heartbeats or of 2.5×10^8 breathings over the course of an average life), these numbers conducting to frequencies (or periods) — dimensional magnitudes, this time — which are submitted to these rules of scaling in function of the average mass of the individuals of the considered species (for instance with a power $-1/4$ of this mass for the frequencies).

But such characteristics of invariance do not manifest solely at the high level of the biological functions of evolved organisms; they can also be found at much more elementary levels such as those of cellular metabolic networks¹⁰ (Ricard, 2003; Jeong *et al.*, 2000), of which the diameter¹¹ remains invariant along the phylogenetic tree and of which the connectivity distribution, over at least 43 organisms belonging to the three categories of life forms,¹² presents the same characteristic power (2.2 approximately). As emphasised by J. Ricard, such an invariance of the network’s diameter implies that the degree of connection of nodes increases with the number of these nodes, that is, with the number of stages likely to connect them. Here again, one will notice that it is a case of numeric and not dimensional invariants.

4.5. A few comments and comparisons with physics

We see that viewed from this angle, the causality which seems to manifest in living phenomena presents similar traits as well as different traits with regard to those which we have noted in the case of Physics when only addressing inertia. Material

¹⁰We know that a network is a graph formed by nodes which are connected according to certain rules. In the case of metabolic networks, these may be constituted by metabolites or by enzymatic reactions and the connection links representing the mutual biochemical interactions.

¹¹The diameter of a metabolic network is defined by the average of the shortest paths (in terms of steps), leading from one of the network’s nodes to another.

¹²That is, archaeobacteria, bacteria, eukaryotes.

and efficient causality are manifestly present there — this having doubtlessly favoured the idea of a possible physicalistic reduction — although in a much less rigorous and structured way than for Physics (namely in connection with the plasticity and adaptability capacities). The formal determinations are relatively weakly represented there, despite the advances made in terms of various local modellings (we have mentioned the metabolic networks, but at a different level, we could evoke population dynamics, for instance, or the transport properties close to fluid dynamics). Likewise, the essential objective determinations do not really appear to have been extracted, despite the observation of variegated properties of symmetry — or of symmetry breakings — (over the course of development or in certain anatomies, for instance) and the identification of certain digital invariants. On the other hand, the dimension of ‘final causality’ (to use old categorisations) or of ‘contingent finality’ appears to play here a role which is rather important and unknown to Physics. As if the fact of finding itself in an extended critical state (therefore potentially highly unstable, although necessary to an elaborate organisation) could be compensated for the structural (momentary) stabilisation of living matter only with the introduction of these factors of telenomy/anticipation which appear to characterise it.

5. Synthesis and Conclusion

We have attempted to briefly characterise the different aspects of physical causality such as they may appear and be analysed through contemporary theories. We have emphasised the fact that symmetries and invariances constitute determinations which are even deeper than those which manifest causal laws in that they present themselves, in a way, as the conditions of possibility for the latter and as frames of reference to which they must conform.

We have also sketched out an analysis of the causality internal to the systems of effective computability, of which the symmetries and invariances obey a specific regime, rooted in the discrete arithmetic structure of databases and algorithms. The intelligibility structure proposed by these methods differ by that which is inherent, on one side, to the geometry and mathematics of the phenomenal continuum, particularly by the difference between the (modern) notion of physical law, to which we refer, and, on the other side, to the notion of algorithm. The consequences to these two aspects can be measured in terms of different causal regimes, following differences (breakings) in symmetry. Iteration, as a particular symmetry in time, is also mentioned as one of the characteristics of digital simulation, in fact as one of the strong points of computational imitation (and a starting point for effective recursion, as a mathematical theory). It is also at the centre of the particular status of predictability, even in the case of the computer implementation of

highly unstable non-linear systems, because the possibility of identically iterating a process (or of accelerating a simulation) is a form of prediction. Iterability, that is, digital calculi, also enables us to grasp the difference between the randomness of the theory of algorithmic information and the randomness of physical processes of the critical and quantum type. In the case of algorithms, randomness coincides with incompressibility. On the other hand, in the first of the physical cases (deterministic, dynamic, and thermodynamic systems), it is of an epistemic nature and it implies the non-iterability of processes; in the second (quantum physics) it is intrinsic to the theory [it is part of the objective determination, (Bailly, Longo, 2006)]. These two last cases are incompatible with the individual iterability of an individual process (which is typical of algorithmics); although it may have a statistical iterability, as in quantum mechanics.

We have then attempted to widen the causal problematic to encompass life by taking into account its specific character through what appears as a sort of finalisation of its functioning which we have attempted to conceptually systematise. This led us to refer to specific concepts, such as that of ‘contingent finality’, and to propose new representations (topologico-metrical) in order to attempt to account for it in a more or less operational fashion. Finally, we have evoked the possibility of extricating that which — through numerical constants and scaling properties — could be considered as invariants of causal reduction specific to life.

If the considerations relative to physics and to calculus base themselves upon well elaborated and mathematised theories, enabling us to refer to a *corpus* we can say to be almost completely objectivised and thus lending itself particularly well to a thorough conceptual and epistemological analyses moreover situating itself within the framework of a well established tradition it seems that for biology the situation is much more fragile in this regard. Also, with regard to life, the analyses we propose are of a much more speculative nature and require with greater necessity the theoretical and conceptual sanctions concerning experimental practices in this field. All the more so because the causal representations, in the case of life, if we seek to detail them, must take into account multiple interactions, which present themselves simultaneously all the while remaining of a highly different nature, whereas in physics, for example, the interactions may in many cases be sufficiently decoupled from one another for us to be able to approach and study them separately; even if it implies, in a second stage, the seeking of conditions and procedures for their unification. In other words and in particular, in biology, the causal representations must take account of massive retroactions, which prevent them most often from partitioning systems into weakly coupled sub-systems in order to facilitate analysis, as is done in physics, as well as to consider holistic telenomies, according to which the local organisation is dependent upon the global structure and reorganises itself according to the necessities of optimisation

or of perdurance of this structure according to criteria which are still not well known.¹³

Nevertheless, it appears that one of the points common to these disciplinary fields — and this is what we have wanted to highlight and to emphasise in this text — resides in the fact that the causal analysis, all the while remaining useful and efficient — must now be relativised and henceforth give way, for the purpose of a better comprehension of the theoretical and conceptual structures of these fields, to a more general approach relying much more upon the properties of invariance, of symmetries (and their breakings) and of conservation. These properties underly the manifestations which we tend to spontaneously (at least since the Renaissance) interpret in terms of objective causal actions. We will maybe see there the trace of a process of conceptual rehabilitation of the ‘geometrical’ (taken in a broad sense) in relation to the ‘arithmetical’,¹⁴ which is not without echoes in the most profound preoccupations of this volume.

References

- [1] Aceto, L., Longo, G. and Victor, B. eds, 2003, The difference between Sequential and Concurrent Computations. Special issue, *Mathematical Structures in Computer Science*, Cambridge University Press, 4–5.
- [2] Anandan, J. 2002, Causality, Symmetries and Quantum Mechanics. *Foundations of Physics Letters*, 15(5), 415–438.
- [3] Bailly, F. 1991, L’anneau des disciplines. *Revue Internationale de Systémique*, 5(3).
- [4] Bailly, F. 2005, Invariances, symétries et brisures de symmetries. In L. Boi, ed. *New Interactions of Mathematics with Natural Sciences*, World Scientific. [also in (Bailly, Longo, 2006)].
- [5] Bailly, F., Gaill, F. and Mosseri, R. 1993, Orgons and Biolons. In *Theoretical Biology: Phenomenological Analysis and Quantum Analogies*, *Acta Biotheoretica*, 41(3).
- [6] Bailly, F., Longo, G. 2003, Objective and Epistemic Complexity in Biology. In N.C. Singh, ed. *Invited lecture, Proceedings of the International Conference on Theoretical Neurobiology*, NBCR, New Delhi, pp. 62–79.
- [7] Bailly, F., Longo, G. 2004, Space, time and cognition. In A. Peruzzi, ed. *From The Standpoint of Mathematics and Natural Science*. Invited paper, *Mind and Causality*,

¹³We know, for example, that the genetic constraints themselves manifest ‘normally’ only in adequate extragenomic or environmental frameworks. Moreover, some phenomena, such as *local* ‘death’ — apoptosis — apparently in contrast to evolutionary survival paradigms, prove to be necessary to the *global* viability, as an integral part of ‘normal’ life and even more of the adaptation faculties of organisms in the event of the modification of the exterior environment.

¹⁴In reference to the debates of the beginning of the twentieth century concerning the foundations of mathematics, but all the while emphasising the fact that the geometrisation of physics, for its part, has never ceased to develop, probably explaining why it is in this discipline that symmetries and invariances have acquired a determining explicative and operational status rather early on.

- Benjamins, Amsterdam, pp. 149–199. [French version in *Revue de Synthèse*. Paris, 2004, also in (Bailly, Longo, 2006)].
- [8] Bailly, F., Longo, G. 2004, *Incomplétude et incertitude en Mathématiques et en Physique*. In Parrini, P. & Scarantino, L.M., eds. *Invited Paper, Il pensiero filosofico di Giulio Preti, Guerrini ed associati*, Milano, pp. 305–340. [Also in (Bailly, Longo, 2006)].
- [9] Bailly, F., Longo, G. 2006, *Mathématiques et sciences de la nature. La singularité physique du vivant*. Hermann, Paris.
- [10] Bernard-Weil, E. 2002, *Ago-antagonistic Systems*. In Mugur-Schächter, M. & Van der Merwe, A., eds. *Quantum Mechanics, Mathematics, Cognition and Action: Proposals for a Formalized Epistemology*. Kluwer, Dordrecht.
- [11] Edalat, A. 1997, *Domains for Computation in Mathematics, Physics and Exact Real Arithmetic*. *Bulletin for Symbolic Logic*, 3(4), 401–452.
- [12] Edelman, G., Tonni, G. 2000, *A Universe of Consciousness: How Matter Becomes Imagination*. Basic Books, New York.
- [13] Jeong, H., Tombor, B., Albert, R. *et al.*, 2000, *The large-scale organization of metabolic networks*. *Nature*, 407, 651.
- [14] Longo, G. 2002, *The Constructed Objectivity of Mathematics and the Cognitive Subject*. In Mugur-Schächter, M. & Van der Merwe, A., eds. *Quantum Mechanics, Mathematics, Cognition and Action: Proposals for a Formalized Epistemology*. Kluwer, Dordrecht, pp. 433–463.
- [15] Longo, G. 2005, *The reasonable effectiveness of Mathematics and its Cognitive roots*. In Boi, L. ed., *New Interactions of Mathematics with Natural Sciences*. World Scientific.
- [16] Longo, G. 2007, *Laplace, Turing and the “imitation game” impossible geometry: randomness, determinism and programs in Turing’s test*. In Epstein, R., Roberts, G. & Beber, G., eds. *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Kluwer, Dordrecht.
- [17] Nicolis, G. 1986, *Dissipative systems*. *Rep. Prog. Phys.* 49(8), 873.
- [18] Nicolis, G., Prigogine, I. 1989, *A la rencontre du complexe*. Presses Universitaires de France, Paris.
- [19] Peters, R. H. 1984, *The Ecological Implication of Body Size*. Cambridge University Press, Cambridge.
- [20] Pilyugin, S., Yu. 1999, *Shadowing in dynamical systems*. Springer, Berlin.
- [21] Ricard, J. 2003, *Emergence, organisation et causalité dans les systèmes biologiques*. In Viennot, L. & Debru, C., eds. *Enquête sur le concept de causalité*. Presses Universitaires de France, Paris.
- [22] Rosen, R. 1991, *Life Itself*. Columbia University Press, New York.
- [23] Sakharov, A. 1984, *Œuvres scientifiques. Anthopos (Economica)*, Paris.
- [24] Sauer, T. 2003, *Shadowing breakdown and large errors in dynamical simulations of physical systems*. Preprint, George Mason University.
- [25] Schmidt-Nielsen, K. 1984, *Scaling: Why is Animal Size so Important?* Cambridge University Press, Cambridge.
- [26] Stewart, J. 2002, *La modélisation en biologie*. In P. Nouvel, ed. *Enquête sur le concept de modèle*. Presses Universitaires de France, Paris.

- [27] Turing, A. 1950, Computing Machines and Intelligence. *Mind*, 59(236), 433–460.
- [28] van Fraassen, B. C. 1994, *Lois et symétrie*. J. Vrin, Paris.
- [29] Varela, F. 1989, *Autonomie et connaissance*. Seuil, Paris.
- [30] Viennot, L. 2003, *Raisonnement commun en physique: relations fonctionnelles, chronologie et causalité*. In Viennot, L. & Debru, C., eds. *Enquête sur le concept de causalité*. Presses Universitaires de France, Paris.
- [31] Weyl, H. 1927, *Philosophie der Mathematik und Naturwissenschaft*. (Translated into 2d edn. *Philosophy of Mathematics and Natural Sciences* by Princeton University Press, 1949.
- [32] Weyl, H. 1952, *Symmetry*. Princeton University Press.

CHAPTER 10

Topological Invariants of Geometrical Surfaces and the Protein Folding Problem

R. A. BROGLIA

*Department of Physics, University of Milano,
via Celoria 16, 20133 Milan, Italy*

INFN, Milan Section, Milan, Italy

*The Niels Bohr Institute, University of Copenhagen,
Blegdamsvej 17, DK-2100 Copenhagen, Denmark
ricardo.brogli@unimi.it*

I. Introduction

Empty space, as well as a plasma of electrons, positrons, neutrinos, photons and nucleons thought to be the stuff of the universe right after the Big Bang, is isotropic and invariant under rotations. The spontaneous breaking of these symmetries observed e.g., in crystals, and in proteins (aperiodic crystals), is at the basis of emergent properties — that is properties not contained in the particles forming the system nor in the forces acting among them — like rigidity, electron conduction, enzymatic activity, *etc.* Because in any three-dimensional polyhedra that the sum of the number of faces (F) and vertices (V), minus the number of sides (edges (E)) equals two is a property of space itself (topological invariant of geometrical surfaces), crystals and proteins adopt the (relatively few) three-dimensional structures they do. In other words, the only “allowed” violations of empty space symmetries are those respecting the topological invariants of which $F + V - E = 2$ is an example. This fact plays an important role in the study of nature and is at the basis of the *ex-novo* design of new nanometric materials and folds (proteins with specific properties). Even if they do not exist in nature, materials — biological or not — of which building blocks (atoms, molecules, *etc.*) respect, in their spatial arrangements, the pertinent topological invariants can, in principle, be

designed and eventually produced, helping fight diseases or making our *habitat* less hostile.

2. Protein Folding Problem

Proteins are linear sequences made out of twenty different types of amino acids (primary, one-dimensional (1D) conformation) which fold in a well defined, unique, biologically active three-dimensional (3D), native, conformation in very short times (as a rule times ranging from nanoseconds to seconds). How the primary conformation codes for the native conformation is the protein folding problem. If one would solve it one would be able, among other things, to inhibit the folding, and thus the biological activity of proteins which play a central role in the vital cycle of viruses and bacteria, and thus help fight diseases.

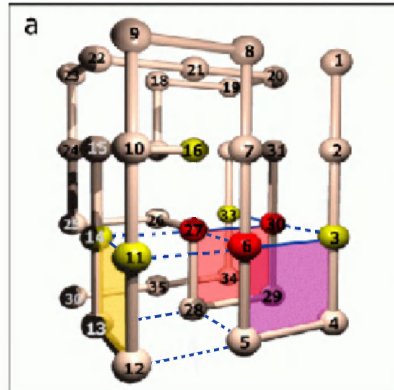
For real proteins one does not know how to solve the protein folding problem [1]. On the other hand, for a simplified model of proteins, the answer is quite simple [2].

The Lattice model is a simplified model of proteins which has proven quite useful, being simple to use in numerical simulations and, at the same time, containing many of the properties of real proteins. In it, twenty different types of amino acids are represented by beads of equal size, forced to move on the vertices of a square lattice and to interact through a 20×20 contact matrix, obtained from the analysis of the frequency of appearance of the different (native) contacts in real proteins [3].

3. Inverse Folding Problem

Because both the amino acids and the interaction are schematic one needs to design sequences which behave like proteins, that is, sequences which are good folders.

This is done by minimizing the energy with respect to amino acid sequence for a given native conformation [4]. A concrete example is shown in Figure 1, where the sequence S_{36} displayed in (b) and known as S_{36} in the literature is found to fold on short call ($\approx 1 \mu s$) on the native conformation shown in (a). It has been found that all good folders have in common few amino acids ($\approx 8\% - 10\%$) which are highly conserved, and which we call 'hot' amino acids. This condition is tantamount to saying that these sequences have an energy lying below the minimum energy of the compact conformation which random amino acid sequences can acquire. Consequently, in their folding process, good folders do not find a myriad of possible conformations but only the native one.



b SQKWLERGATRIADGDLVPVNGTYFSCKIMENVHPLA

c YPDLTKWHAMEAGKIRFSVPDACLNGEGIRQVTLNS

Figure 1 (a) The conformation of the 36-mer chosen as the native state in the design procedure. Each amino-acid residue is represented as a bead occupying a lattice site. The design tends to place the most strongly interacting amino acids in the interior of the protein where they can form most contacts. The strongest interactions are between groups *D*, *E* and *K* (cf. (b)), the last one being buried deep in the protein (amino acid in site 27). Amino acids occupying 'hot' sites (sites 6, 27, 30) have been represented by red beads, those occupying 'warm' sites (sites 3, 5, 11, 14, 16 and 28) by yellow and those occupying cold sites by light brown beads. The local elementary structures (LES) formed by the amino acid sequences $S_4^1 \equiv (3, 4, 5, 6)$, $S_4^2 \equiv (27, 28, 29, 30)$ and $S_4^3 \equiv (11, 12, 13, 14)$ and stabilized by the contacts 3–6, 27–30 and 11–14 (drawn by continuous lines) are explicitly shown by shaded areas. The contacts between the LES are shown by dashed lines. (b) Designed amino acid sequence S_{36} associated with a native energy $E_n = -16.5$. (c) Designed sequence S'_{36} associated with a native energy $E_n = -17.13$ (see Figure 2).

4. Molecular Dynamic Simulations

In Figure 2 snapshots are shown of a typical trajectory associated with the numerical simulation of the evolution of the folding of S_{36} . A hierarchical picture of the folding process emerges in which, starting from an elongated conformation (see Figure 2(a)), the basic steps are:

- (a) Formation of local elementary structures (LES), stabilized by strongly interacting, highly conserved 'hot' amino acids, very early in the folding process

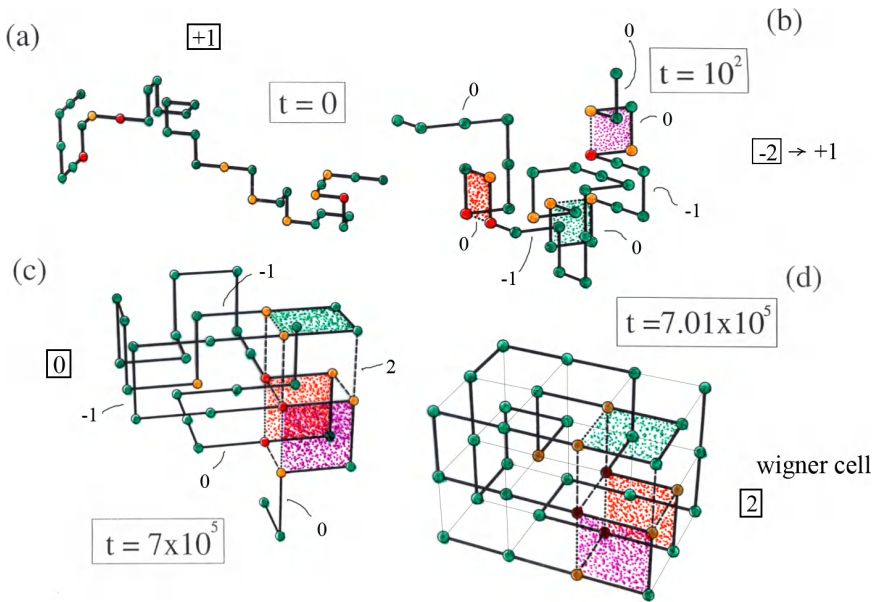


Figure 2 Snapshots of the folding of the sequence S'_{36} (see Figure 1(c)), whose energy in the native conformation is $E_n = -17.13$ carried out at $T = 0.28$. Starting from a random conformation (a), the system forms after $\approx 10^2$ MC steps (1 MCs $\approx 10^{-13}$ s) local elementary structures (LES) (b), involving three sets of four amino acids (3–6, 11–14, 17–30), whose stability is provided by the bonding indicated by dotted lines. When the LES come together to form the folding nucleus (FN) (indicated by dotted and dashed lines) after $7 \cdot 10^5$ MC steps (c), the system folds to the native conformation after only 10^3 MC steps (d). The amino acids participating in the bonding of the LES (dotted lines) are among some of the most strongly interacting amino acids, which occupy, in the native conformation (d), 'hot' and 'warm' sites indicated by red and yellow beads, respectively. The monomers number 1 and 36 of the sequence S'_{36} are indicated for each conformation (see Figure 1(c)).

- (see Figure 2(b)). From now on the folding process and the associated molecular recognition is associated with LES.
- (b) Diffusion of the LES in configuration space until they dock in the native conformation, giving rise to the (post-critical) folding nucleus (FN) (see Figure 2(c)), that is, the minimum set of native contacts which brings the system over the highest barrier of the (free) energy associated with the folding process.
 - (c) Shortly after, the remaining amino acids folds into place, the protein having reached the native conformation.

From the above results a protocol has been developed which allows reading the three-dimensional structure of a protein from the one-dimensional structure.

In fact, given the amino acid sequence and the contact matrix used to design the protein, one can first individuate the hot amino acids of the protein through multiple mutations. Knowing these sites one can design possible candidates of LES and thus of FN. In trying to compact the remaining amino acids one would find that there is a single, eventually a couple of FN leading to a low-energy compact conformation. This is the native conformation (for more details see Ref. [2]).

The extension of these results to real proteins is under way, starting with the development of methods for the identification of the LES and thus of the FN.

5. Topological Invariant Number and the Folding Process

If one closes a volume in terms of a number of plane faces (F) limited by a closed perimeter made out of straight lines which, intersecting at vertices define the edges of the surface, one would always find that the sum of the number of faces (F) plus the number of vertices (V) minus the number of edges (E) is equal to two¹ ($F + V - E = 2$). This is true for a cube as well as for a parallelepiped ($F = 6$, $V = 8$, $E = 12$ in both cases), for a pyramid made out of triangles ($F = 4$, $V = 4$, $E = 6$) or of four triangles and a squared base ($F = 5$, $V = 5$, $E = 8$). In fact, it is true for all polyhedra (from the Greek *poly* — many hedron or hedra — geometrical figure having a (specified) number of faces).

Relations of the type $F + V - E = 2$ are known as numerical topological (Greek *topos* — space, *logos* — word, reason) invariants of geometrical surfaces, being properties of (empty) space itself. This is the reason why crystals have their atoms disposed the way they are, e.g., carbon atoms in graphite are disposed in sheets made out of hexagons, with the atoms at the edges and pairs of exchanged electrons at the edges, while diamond has its atoms arranged in a face-centred cubic lattice.² But also in the case of the hollow molecule C_{60} fullerene (made out of hexagonal and pentagonal faces), the third allotropic form of carbon, and the only finite one. Also why the smallest member of the fullerene family is C_{20} containing exactly twelve pentagonal faces, the minimum needed to close a spherical space with pentagonal and hexagonal faces, still respecting $F + V - E = 2$.

This seems to be also the reason why the (folding) nuclei of proteins in their native, biologically active state, have the structure they have, a fact which is

¹The relation $F + V - E = 2$ was first found by Descartes in 1619, and 'rediscovered' by Euler in 1731. The number 2 in the above relation is known as the Descartes–Euler number $(DE)_n$ (see A.D. Aczel, *Descartes' Secret Notebook*, Broadway Books, New York (2005)).

²Quantum mechanics also 'respects' the topological invariants when it spontaneously violates rotational and translational invariance, thus leading to e.g., rigidity (example of emergent property) by correlating covalently four atoms at a time (sp^3) in diamond, and three (sp^2) in graphite.

reflected in the denomination of soft matter or aperiodic crystals for such systems (within this context cf. also [5, 6]). Topological invariants are also associated with open structures made out of edges and vertices like, for example, a chain of N beads [7]. In these systems, $F = 0$, $V = N$, $E = N - 1$, and thus $F + V - E = 1$.

Typical single-domain, monoglobular proteins are made out of $N \approx 100$ amino acids (beads) held together by $E = N - 1$ peptidic bonds. As discussed in the previous section, the folding process is controlled by highly hydrophobic, local elementary structures (LES) formed early in the folding process under the effects of the (weak) hydrophobic force. In real proteins each LES is built out of $V \approx 10$ amino acids (see e.g., [8]) held together by nine peptide bonds ($E = 9$), with $F = 0$ in keeping with the fact that they are open structures. In their elongated conformation they thus carry an invariant topological number 1. This (local) number is conserved when, very early in the folding process, the stability of the LES becomes very high ($\approx 95\% - 100\%$). Thinking in terms of S_{36} , each LES becomes a plaquette with $V = E (= 4)$ and $F = 1$ (see Figures 2 and 3). The overall topological invariant number becoming $(D.E)_n = 0$, in keeping with the fact that the polypeptides chains in between LES carry $(D.E)_n = -1$ and that the end chains carry $(D.E)_n = 0$. When the LES dock in the native conformation, they lead to a highly stable folding nucleus (FN), carrying a topological invariant number equal

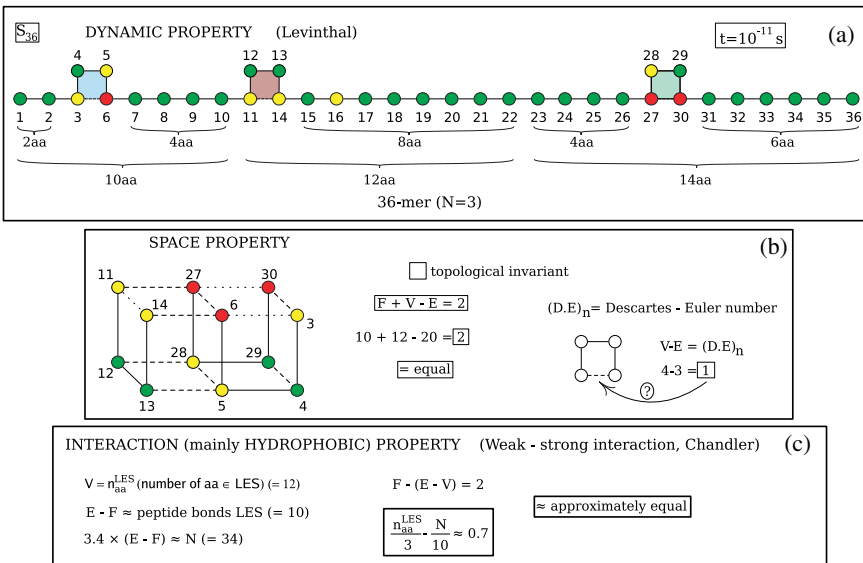


Figure 3

Table 1 The number of amino acids (aa) N , average number of aa of each LES $n_{aa}^{LES}/3$, number of aa composing each LES $N/10$ of the proteins listed in the first column are displayed in columns one to three respectively. The first three lines refer to lattice model proteins, while the last four to real proteins.

	N	$n_{aa}^{LES}/3$	$N/10$	$n_{aa}^{LES}/3 - N/10$
s_{36}^a	36	4	3.6	0.4
$s_{36} (n.l.)^a$	36	3.7	3.6	0.1
s_{48}^a	48	7	4.8	2.2
HIV-1-PR (monomer) ^b	99	9	9.9	-0.9
SH ₃ ^c	60	7.3	6.0	1.3
CI ₂ ^c	64	7.7	6.4	1.3
G ^c	56	6	5.6	0.4
average value				0.7
standard deviation				0.9

^aSee Ref. [2] and [13]

^bSee Ref. [8]

^cSee Ref. [14]

to 2, the overall $(D.E)_n$ being 0. In the last, rather fast steps, all the remaining amino acids will fall into place. For a very stable system (like lysozyme) or, at low temperature, the native conformation of proteins is stable and it becomes a (aperiodic) solid, which carries the topological invariant number 2 corresponding to the primitive cell (cube).

Summing up, the folding process is a hierarchical process: $D \rightarrow LES \rightarrow FN \rightarrow N$, where D and N indicate the denaturated and the native conformations respectively. Arguably, the most complicated and less known of all these conformations is the denaturated state [9]. It is known that this state is rich in some native contacts. It is likely that LES (see Figure 3(a)) is a better representation of the denaturated state.

Consequently, at a temperature much lower than the folding temperature (at which the (D, LES) and N states are equally populated), the (first order) transition associated with the folding of a protein can be decomposed in a sum of partial transitions: a) $LES \rightarrow FN$ (implying a variation of the $(D.E)_n$ from 1 to 0), b) $FN \rightarrow N$ ($0 \rightarrow 2$).

The properties of empty space are given a concrete meaning through the physics associated with the contact interactions acting among the amino acids. In fact, as shown in Table 1, with the help of orders of magnitude associated with the weak and the strong hydrophobic interactions, the relation $F + V - E = 2$ provides important insight into the structure of LES and thus of molecular recognition, let alone about the protein folding problem.

6. Protein Folding Inhibitor and Non-Conventional Drug Design

Because molecular recognition in the protein folding phenomenon is controlled by LES, a nonconventional³ folding inhibitor suggests itself: short peptides (p-LES) displaying a sequence identical to a LES [10]. This peptide conveniently structured and carrying an invariant number equal to 1 ($1 + 4 - 4 = 1$, plaquette) would, by attaching to its complementary LES belonging to the protein, stabilize it in an unfolded conformation. Because this state is different from the one in which the protein is stabilized by the substrate, one is talking about competitive inhibition. Very promising results for this type of inhibitors have been found in a number of situations, in particular in the case of the HIV-1-PR (see Refs. [8, 11] and Refs. therein).

In the case in which disulfide bonds are present, e.g., in the case of lysozyme, it is an open question whether one should include these bonds in the p-LES or not. In the first situation one would be talking about a closed LES, in the second an open LES as the stability of p-LES is, as a rule, less than the corresponding LES inside the (native state) protein. Because both p-LES carry a $(D.E)_n = 1$, one would need experimental input to answer this question in detail.

7. Conclusion

Topological invariants and dynamical properties associated with the folding of the lattice model designed 36-mer (see Figures 1 and 2). In what follows we refer to Figure 3:

- (a) First step in the folding process: formation of LES after only $\approx 10^2$ MCs (1 MCs $\approx 10^{-13}$ s, thus 10^2 MCs $\approx 10^{-11}$ s). Shown with red, yellow and green colours the hot, warm and cold amino acids of S_{36} [12]. The interaction stabilizing each of the three LES are shown as dotted lines. In a real protein, this interaction is to a large extent due to the weak hydrophobic interaction (proportional to the volume of the non-polar molecules). The steps $D \rightarrow \text{LES} \rightarrow \text{FN} \rightarrow \text{N}$, where D and N stand for denaturated and native conformations, summarize the hierarchical folding of proteins. Each of the three LES are evidenced in terms of sky blue, rosy brown and dark sea green areas. Also shown are the neutral chain segments (of summed length $\approx 2n_{aa}^{\text{LES}} = 8$) which each of the LES move

³Conventional inhibitors act on the active site of the folded protein either by capping it or, attaching on the surface of the protein distorting the active site structure so that the protein displays a much reduced ability to bind to the substrate.

around in the diffusive search of their complementary LES to form the (post-critical) folding nucleus (FN). Molecular recognition at this level is carried out by LES.

- (b) FN resulting from the docking of the LES (native conformation). Being the FN an energy minimizing compact conformation, and thus a viable physical conformation, it has to coincide with one of the topological invariant figures of space, and thus carry an exact topological invariant number (i.e., the boxed expression, the Descartes–Euler relation uses an equal (=) symbol). In other words, the fact that the FN has the structure of a polyhedron reflects only a property of space. Within this context, once the LES has been stabilized, it can be viewed as a plaquette with $F = 1$, $V = 4$, $E = 4$. Note that once the FN is stabilized, it can be viewed in terms of the primitive Wigner–cell (like a crystal) which, in this case carries $F = 6$, $V = 8$, $E = 12$, the topological invariant number being still 2 as for the case of the parallelepiped displayed.
- (c) The properties of empty space are given a concrete meaning through the detailed properties of the (contact) interactions. The one used in the design of S₃₆ (Miyazawa–Jernigan, [3]) was obtained from the frequency of amino acid (native) contacts of real proteins. Identifying the vertices of the polyhedra (elongated cube) with the number of amino acids n_{aa}^{LES} belonging to the FN and thus to the LES and making use of the fact that there are three LES, $n_{aa}^{LES}/3$ is the average number of amino acids building a LES. Because $E - F$ can be identified with the number of peptide bonds found in the FN (see (b)), and of the fact that $2n_{aa}^{LES}$ is the number of amino acids forming the (neutral) chains that each LES moves around in their search for the complementary LES to form the FN (see (a)), the quantity $3.4(E - F)/(3 \times 3.4) \approx 3.4(E - F)/10 \approx N/10$ gives an estimate of the order of magnitude of the number of amino acids entering in each LES, the total number of residues of the protein being N . Within the present context, the boxed relation uses an approximately equal symbol (\approx), referring to real proteins and depending on the actual interaction among amino acids. Paramount among the interactions is the hydrophobic interaction. One can distinguish between two components of this interaction: the weak (proportional to the volume) and the strong (proportional to the surface) components. In other words, the volume term dominates for systems with dimensions $\lesssim 10 \text{ \AA}$, the second term for systems larger than 10 \AA . This means that weak hydrophobicity can compact, early in the folding process, groups of the order of 10 mainly hydrophobic amino acids leading to LES. Because $N/10$ is the number of amino acids entering in each LES, it means that single-domain globular proteins must contain of the order of $N = 100$ amino acids. This in fact is the case, not only for single-domain proteins but also for each of the domains of a multi-domain protein.

A résumé of the relation $n_{aa}^{LES}/3 - N/10 = 2/3 \approx 0.7$ associated with different proteins which have been extensively studied both through (model)

molecular dynamic calculations, as well as through protein engineering, is given in Table 1.

Topological invariants and lattice model of proteins together with the empirical knowledge that folding domains of globular proteins contain about 10^2 allows to estimate the number of amino acids building each of the few (≈ 3) local elementary structures which provide molecular recognition and thus direct the folding of proteins. This number (≈ 10 aa) is consistent with the fact that the first steps of the folding process is controlled by the weak hydrophobic interaction, while that of the diffusion of LES (molecular recognition) to build the FN is a result of the strong hydrophobic interaction (both interactions crossing at volumes of ≈ 1 nm (10 \AA) of radius (and thus containing ≈ 10 aa). It is remarkable that Descartes relation $F + V - E = 2$ could say so much about the building blocks of life itself.

References

- [1] Fehrst, A. 1999, *Structure and Mechanism in Protein Science*, Freeman, New York.
- [2] Broglia, R. A., Tiana, G. 2001, Reading the three-dimensional structure of lattice model proteins from their amino acids sequence, *PROTEINS* 45, 421.
- [3] Miyazawa, S., Jernigan, R. 1985, Estimation of the effective inter-residue contact energies from protein crystal structures: quasi chemical approximation, *Macromolecules* 18, 534.
- [4] Shakhnovich, E. I. 1994, Proteins with selected sequences fold to their unique native conformation, *Phys. Rev. Lett.* 72, 3907.
- [5] Banavar, J. R., Maritan, A., Micheletti, C. *et al.*, 2002, Geometry and Physics of Proteins, *PROTEINS* 47, 315.
- [6] Micheletti, C., Banavar, J. R., Maritan, A. *et al.*, 1999, Protein Structures and Optimal Folding from a Geometrical Variational Principle, *Phys. Rev. Lett.* 82, 3372.
- [7] Munkres, J. R. 2000, *Topology*, Prentice Hall Inc., NJ.
- [8] Broglia, R. A., Tiana, G., Sutto, L. *et al.*, 2005, Design of HIV-1-PR inhibitors that do not create resistance: Blocking the folding of single monomers, *Protein Science* 14, 2668.
- [9] Shortle, D. 1996, The denatured state (the other half of the folding equation) and its role in protein stability, *FASEB J.* 10, 27-34.
- [10] Broglia, R. A., Tiana, G. and Berera, R. 2003, Resistance proof, folding inhibitor drugs, *J. Chem. Phys.* 118, 4754.
- [11] Broglia, R. A., Provasi, D., Vasile, F. *et al.*, 2006, Folding inhibitor of the HIV-1 Protease, *PROTEINS* 62, 928.
- [12] Tiana, G., Broglia, R. A., Roman, H. E. *et al.*, 1998, Folding and misfolding of designed proteinlike chains with mutations, *J. Chem. Phys.* 108, 757.
- [13] Broglia, R. A. and Tiana, G. 2001, Mechanism of folding and aggregation of proteins, in *Procs. of the International School of Physics "E. Fermi" Course CXLV*, R. A. Broglia, E. I. Shakhnovich, G. Tiana (Eds.), IOS Press, Amsterdam 69.
- [14] Sutto, L., Tiana, G. and Broglia, R. A. 2006, Sequence events in folding mechanism: beyond the Gō-model, *Protein Science* 15, 1638.

CHAPTER 11

The Geometry of Dense Packing and Biological Structures

J.-F. SADOE

*Laboratoire de Physique des Solides, CNRS
Meudon, France*

and

*Université Paris Sud, Bâtiment 510, 91405 Orsay, France
sadic@lps.u-psud.fr*

Overview

Dense structures appear in different scientific context: crystals, amorphous metals, foams and biological organizations. The scale characterising these structures go from atomic scales for amorphous metals to macroscopic scales for foams. The biology gives examples at molecular scales up to macroscopic scales. In sphere packing, the best compactness is obtained when sphere centres are on the vertices of regular tetrahedra. There is no solution in Euclidean space, but in curved space, the $\{3, 3, 5\}$ -polytope is a template for dense structures. Then larger structures are derived from this polytope, using disclinations. That needs a study of symmetries in this polytope. An efficient tool for symmetry analysis is the Hopf fibration, which reveals helicoidal symmetries. Helices and dense packing of spherical objects are two closely related problems. For instance, the Boerdijk–Coxeter helix, which is obtained as a linear packing of regular tetrahedra, is a very efficient solution to some close-packing problems. The shapes of biological helices result from various kinds of interaction forces, including steric repulsion. Thus, the search for a maximum density can lead to structures related to the Boerdijk–Coxeter helix. Examples are presented for the α -helix structure in proteins and for other examples of helical packings at different scales in biology.

I. Introduction

The fact that a physical system should obey a principle of minimum energy is always constrained by external conditions (temperature, pressure, geometrical confinement, . . .). One such condition, which is so evident that it is often forgotten, is the type of space underlying the system. The space is characterised by several quantities of a topological and metrical nature, which we are going to describe now. The importance of such knowledge comes from the fact that, in some cases, the local atomic configuration which minimises a local form of the energy, may not be compatible with the underlying space (it is not ‘space-filling’). This is called ‘geometric frustration’, when the local order cannot be propagated freely throughout the space. The simplest example is the case where the local rule consists in packing spheres as densely as possible.

In two dimensions, a local packing of discs, the flat analogs of spheres, is densely organised if centres of discs are located on vertices of equilateral triangles which can tile the plane along the six-fold symmetric hexagonal lattice; this is an unfrustrated case. In three dimensions, the local densest packing of spheres is achieved by placing their centres at a regular tetrahedron vertices. The geometric frustration reveals immediately in that the three-dimensional Euclidean space cannot be filled completely by regular tetrahedra.

Dense structures appear in different scientific context: crystals, amorphous metals, foams and biological organisations. The scale characterising these structures go from atomic scales for amorphous metals to macroscopic scales for foams. The biology gives examples at molecular scales up to macroscopic scales. In sphere packing, the best compactness is obtained when sphere centres are on the vertices of regular tetrahedra. There is no solution in Euclidean space, but in curved space, the $\{3, 3, 5\}$ -polytope [1–3], which is a very dense packing of tetrahedra, is now intensively used in order to understand local order and the propagation of local order [4, 5].

Nevertheless the $\{3, 3, 5\}$ -polytope is a finite structure containing only 120 vertices, and defined in a curved three-dimensional space, the sphere S_3 . More suitable structures are needed in order to model Euclidean structures in the three-dimensional space. A decurving procedure, using disclinations, allows to build larger polytopes, with a greater number of vertices, and so, locally less curved.

There are some interesting polytopes and structures in S_3 derived from the $\{3, 3, 5\}$ -polytope. They could be described in terms of packing of helices in order to model molecular structures. This can be applied to structures whose local order can mimic the local order of folded proteins, which could be considered as dense packing of amino-acids [6].

2. Geometry of the {3, 3, 5}-Polytope in S_3

2.1. The {3, 3, 5}-polytope

It is a tiling of S_3 by 600 regular tetrahedra, with 120 vertices and 720 edges. Let us recall the meaning of the Schläfli notation in the case of the {3, 3, 5} polytope: it denotes a regular structure such that five {3, 3} tetrahedra share a common edge. The first neighbour configuration around a vertex is an icosahedron coded in the two last indices {3, 5}, ‘the star’ or ‘the vertex figure’. Each vertex has twelve neighbours that form an icosahedral shell. Together they form a cluster of twenty tetrahedra all sharing the central vertex. If the edge length is taken as unity, the {3, 3, 5} regular polytope is inscribed on a hyper-sphere of radius equal to the golden ratio, $\tau = (1 + \sqrt{5})/2$.

2.2. The Hopf fibration of S_3

The symmetry of S_3 can be nicely described using its Hopf fibration by great circles. After a description of the Hopf fibration in continuous space S_3 , we consider the discrete case applied to the {3, 3, 5}-polytope.

The spherical space S_3 can be characterised by the equation: $x_1^2 + x_2^2 + x_3^2 + x_4^2 = r^2$ with four Cartesian coordinates x_i . It is simpler to consider it as a three-dimensional space, depending on three independent angles ω , θ and ϕ , such that

$$\begin{aligned} x_1 &= r \cos \theta \cos \phi \\ x_2 &= r \sin \theta \cos \phi \\ x_3 &= r \cos \omega \sin \phi \\ x_4 &= r \sin \omega \sin \phi. \end{aligned} \tag{1}$$

This is known as the toric coordinates of S_3 with $\theta \in [0, 2\pi]$, $\omega \in [0, 2\pi]$ and $\phi \in [0, \pi/2]$.

A fixed value of ϕ gives the equation of a surface since only two parameters (ω, θ) can vary. This surface is a torus (in the particular case $\phi = \pi/4$, it is the so called ‘spherical torus’). Now, if we further impose a linear relation between ω and θ , it remains only one independent parameter, defining a curve on the torus in S_3 , and thus the relation $\theta = \omega + \omega_0$, ω_0 being a constant value, leads to a set of great circles (radius r) in S_3 , each circle characterised by ω_0 . These circles never intersect and in some sense are parallel circles: they are called Clifford parallels. One way to imagine their configuration is to consider a flat rectangle on which

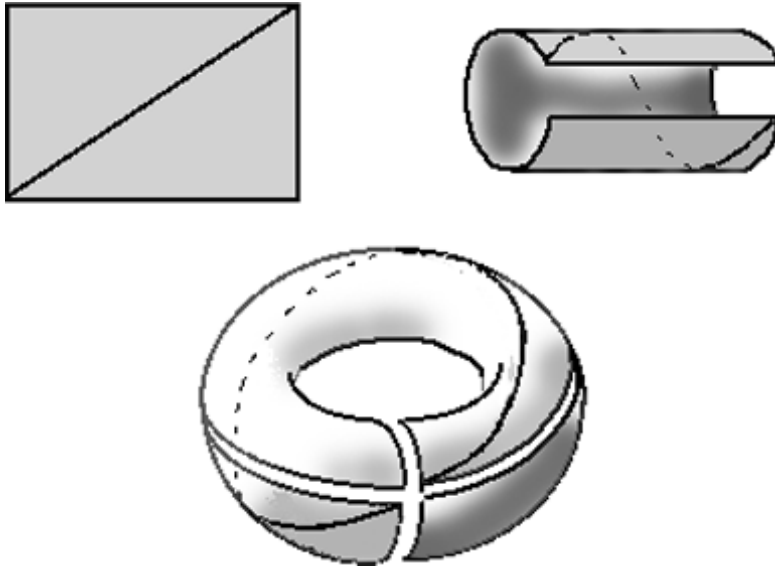


Figure 1 How a rectangle is refold in a torus and how a diagonal gives a great circle.

a family of straight lines parallel to a diagonal are drawn. Then the rectangle is folded into a torus by identification of opposite sides (Figure 1). Folding a rectangle into a torus can be done in S_3 without distortion, as the surface of a torus has an Euclidean metric, but with some elastic (metric) deformations, such a folding can be represented in the usual R_3 space by a usual torus. After this refolding, lines parallel to the diagonal form close circles running on the torus surface. Notice that there is a chirality depending on the choice of the diagonal, and that each circle is wounded one around any other. Now consider a family of torii each defined by a constant $\phi = \phi_0$. And then consider on all the torii, a family of these great circles. We have built in S_3 a set of great circles never intersecting, every one circle being at constant distance from any other. This defines a Hopf fibration of S_3 (Figure 2). Recall that a fibration is a decomposition of a space into identical sub-spaces (here the circles) so that a point in S_3 is characterised, by the fibre on which it lies and by its position on the fibre.

In the present example, a fibre is characterised by the two constant parameters ϕ_0 (the torus) and ω_0 (the line on the torus) and then a point is positioned on the fibre by the parameter ω . As there are two angular parameters to define a fibre, they could be taken as angular coordinates of a surface, on which a point (ϕ_0, ω_0) represents a whole fibre. This surface is called the base of the fibration. In fact, in this example, this base is a sphere S_2 , $y_1^2 + y_2^2 + y_3^2 = r^2/4$ with spherical

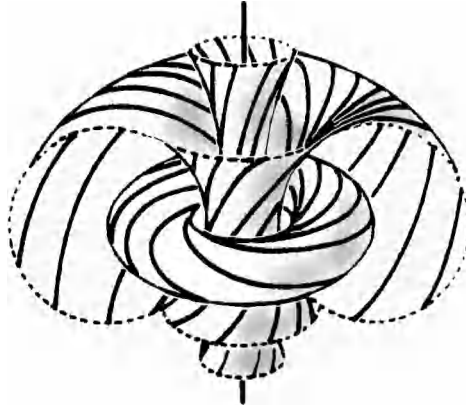


Figure 2 A representation of a Hopf fibration obtained by a stereographic projection of S_3 . In S_3 , fibres are circles all of the same diameter. On this figure, by projection effect, fibres (black lines) have different diameters.

coordinates such that

$$\begin{aligned} y_1 &= \frac{r}{2} \cos(2\phi_0) \\ y_2 &= \frac{r}{2} \cos \omega_0 \sin(2\phi_0) \\ y_3 &= \frac{r}{2} \sin \omega_0 \sin(2\phi_0) \end{aligned} \quad (2)$$

with $\omega_0 \in [0, 2\pi]$ and $2\phi_0 \in [0, \pi]$. Nevertheless, the base is not embedded in S_3 and is only a tool to represent fibres. Indeed, if it was embedded, it would then have two points intersected by a fibre: if a circle cuts a sphere, it cuts it in two points (in three-dimensional flat or curved space). This would contradict the fact that only one point on the base should characterise a given fibre. The important result of this representation of a Hopf fibration is that two points at a given distance on the base represent two parallel circles at the same distance, all distances being defined by the length of geodesic arc in spherical space. Accordingly, the distance between two fibres is given with the metric of an ordinary sphere of radius $r/2$ with spherical coordinates $2\phi_0$ and ω_0 . With this choice for the radius of the base, the distance between two fibres in S_3 is also the distance between their two representative points on the base. A torus, which is the set of fibres with the same ϕ_0 , corresponds to a circle on the base [7].

It is also possible to see the Hopf fibration as a mapping. Let us write the points on a unit radius S_3 as pair of complex numbers (u, v) such that $|u|^2 + |v|^2 = 1$. The Hopf map may then be defined as the composition of the map h_1 from S_3 to R^2

(with ∞ included) followed by an inverse stereographic projection from R^2 to S^2 :

$$\begin{aligned}
 h_1 : S_3 &\rightarrow R_3 \\
 (u, v) &\rightarrow Q = u/v \quad u, v \in C \\
 h_2 : R_2 &\rightarrow S_2 \\
 Q &\rightarrow M(x_1, x_2, x_3) \quad x_i \in R
 \end{aligned}
 \tag{3}$$

Hopf fibrations are directly related to screw symmetry operations in S_3 : fibres are the trajectories of points under the action of two rotations with axes along two fibres defined by two opposite point on the base. These two axes are in two planes completely orthogonal in the four dimensions space in which the spherical three-dimensional space is embedded.

3. The Geometry of Helices

3.1. The Boerdijk–Coxeter chain of tetrahedra

Helices and dense packing of spherical objects are two closely related problems. A very interesting geometrical figure is obtained by stacking regular tetrahedra along one direction. It is called the Boerdijk [8]–Coxeter [9] chain (B-C chain). Select one face of a tetrahedron, on which the next tetrahedron is glued, and proceed on gluing new tetrahedra, with the conditions that no more than three tetrahedra share an edge, and that edges with only one tetrahedron are more or less aligned. A chain of tetrahedra is obtained, on which external edges form three helices (Figure 3). Surprisingly, this chain is not periodic, owing to an incommensurability between the distances separating centres of neighboring tetrahedra, and the pitch of the three helices [10, 11].

There are different kinds of tetrahedral edges corresponding to the number of tetrahedra sharing a giving edge: Those which appear most parallel to the axis of the B-C chain belong to only one tetrahedron. They will be called hereafter type- $\{3\}$. Edges sharing two tetrahedra are called type- $\{2\}$ and edges sharing three tetrahedra, type- $\{1\}$. The number corresponds to the direction of the edge in the phyllotactic representation of the helix (see below). We distinguish several families of helices made of these three types of edges. There are three type- $\{3\}$ helices, but only one type- $\{1\}$ helix and two type- $\{2\}$ helices.

It is useful to describe the B-C chain (or any helical structure resulting from close packed units) as a two-dimensional graph on a cylinder. All edges of the graph are geodesic lines on the cylinder. When the cylinder is unfolded on a flat surface, this surface is tiled with triangles. concretely, the B-C chain can be built by taking an actual sheet of paper on which a triangular tiling (with equilateral triangles) is drawn, cutting a strip three triangles-wide, folding the type- $\{2\}$ edges inwards, types- $\{3\}$ and $\{1\}$ outwards, and gluing.

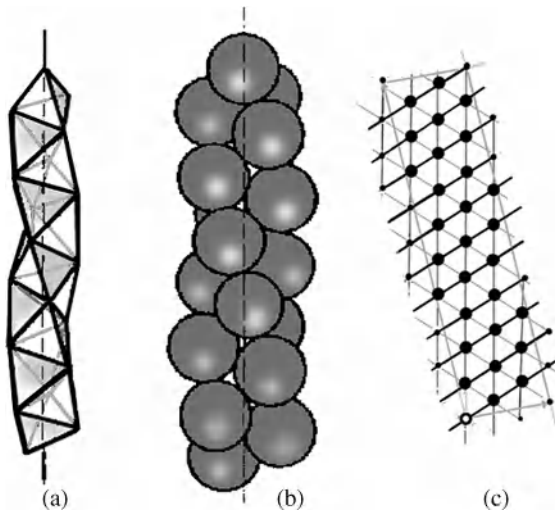


Figure 3 Boerdijk–Coxeter chain obtained from a necklace of tetrahedra. (a) A dense packing of spheres centred on the tetrahedron vertices. (b) The B-C helix (a) can be obtained by folding the edges of the triangular tiling (c), and gluing together the larger sides of the rectangle. A torus is obtained (in curved space) by identification of the smaller sides. The Coxeter helix corresponds to the black edges in (c).

3.2. Discretising the fibration for the $\{3, 3, 5\}$ polytope

Consider now the $\{3, 3, 5\}$ polytope. Since each screw symmetry gathers vertices on Hopf circles, different discretised fibrations can be drawn on this polytope. The ten-fold screw axis defines a fibration related to the B-C chain: there are twelve fibres (containing ten points each), whose Hopf map gives the twelve vertices of an icosahedron on the base S_2 . The fibres are polygons with ten vertices: edges of these decagons are edges of the tetrahedral cells. The fibres are made of type- $\{3\}$ edges.

3.2.1. Mapping the Boerdijk–Coxeter chain from S_3 to the plane

The B-C chain is related to the problem of packing spheres or tiling by regular tetrahedra, resolved by the $\{3, 3, 5\}$ polytope in curved space. Because it is impossible to tile Euclidean space with regular tetrahedra, space has to acquire a positive curvature. In curved space, the helices defined by edges on the B-C chain wind on a torus instead of a cylinder, and they form close curves. The torus can be cut and flattened into a rectangle (or a parallelogram), with identification of opposite sides (see Figure 4). Now, folding a rectangle (or a parallelogram) into a torus in curved spherical space S_3 , can be done without any metric distortions. Thus, for

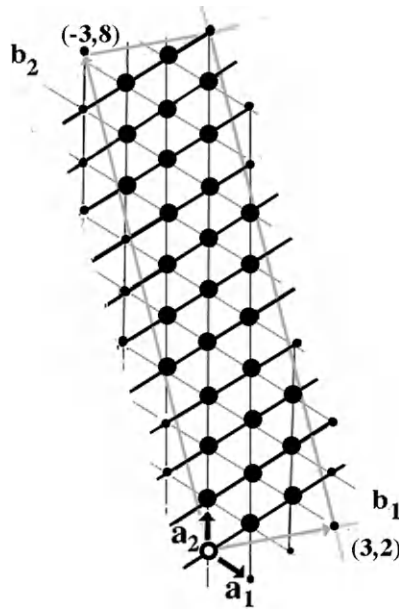


Figure 4 A flat strip leading to the Boerdijk–Coxeter chain in S_3 by identification of the two long sides of the (pseudo) rectangle.

the B-C chain, the flattened torus is tiled by triangles which are nearly equilateral (some care is needed because only type- $\{3\}$ edges are geodesics of the torus and of S_3 , the other edges are slightly distorted in the flattened torus). The flattened torus is a multiple cell of the hexagonal lattice, describing the triangular tiling, with basic vectors $\vec{b}_1 = 3\vec{a}_1 + 2\vec{a}_2$, $\vec{b}_2 = -3\vec{a}_1 + 8\vec{a}_2$ and $\vec{b}_1 \cdot \vec{b}_2 = -2$, where \vec{a}_1 and \vec{a}_2 are the unit vectors of the primitive cell (see Figure 4). With this choice, the flat representation of the cylinder is a parallelogram close to a rectangle.

3.3. The Coxeter helix

The three different types of helices observed on the B-C chain, can be easily identified on the flat strip (\vec{b}_1, \vec{b}_2) . There are three type- $\{3\}$ winding along \vec{a}_2 . With opposite sides of the strip identified, they form closed loops making one turn around the two axes of the torus, with ten edges and vertices each. In spherical space, they are geodesics, great circles of S_3 , fibres of the Hopf fibration of S_3 . Two type- $\{2\}$ helices wind along $-\vec{a}_1$; they also form close loops, making four turns around one axis and one turn around the other, with 15 vertices.

Finally, one type- $\{1\}$ helix, the Coxeter helix, winds along $\vec{a}_1 + \vec{a}_2$. It has 30 edges and vertices, and makes 11 turns around one axis and one around the other. The Coxeter helix has therefore $30/11 = 2.727272\dots$ edges per turn. The helices have opposite chiralities: If type- $\{1\}$ and $\{-3\}$ are right-handed helices, say, type- $\{2\}$ helices are left-handed.

The Coxeter helix is labelled $(3, 2, 1)$ in phyllotactic notation. This is a notation describing triangular lattices on cylinders [12, 13]. It describes, for instance, all the possible structures of composite flowers (phyllotaxis). Each vertex is labelled by a natural integer n , in order of increasing altitude on the vertical cylinder, or of increasing age in a flower. The phyllotactic notation (k, l, m) , with $k > l > m$, implies that the vertices labelled $n \pm k, n \pm l, n \pm m$ are neighbours to vertex n . Consequently, in a triangular tiling, $k = l + m$, since $n + k$ and $(n + l) + m$ label the same neighbour to vertex n . The three types of helix on the cylinder are labelled accordingly: The k helices of type- $\{k\}$ include vertices $\dots n - k, n, n + k, n + 2k, \dots$, and are the steepest ones. The m helices of type- $\{m\}$ are the flattest ones. If there is one single helix going through all the vertices, it is of type- $\{1\}$, and $m = 1$. It is labelled $(k, k - 1, 1)$. The simplest example is the B-C helix $(3, 2, 1)$. Other helices of biological interest are the α -helix $(4, 3, 1)$, the π -helix $(5, 4, 1)$ and Pauling's γ -helix (or 5.1 helix) $(6, 5, 1)$.

3.3.1. Metric properties of the Coxeter helix on an Euclidean cylinder

We can build helices in Euclidean space starting from the flat map of the helix on a torus. We make a long strip by assembling several patching units (\vec{b}_1, \vec{b}_2) joined by their smaller sides, and fold it into a cylinder.

It is easy in curved space (or on a torus) to count how many turns an helix makes around its axis, as it is a pure topological number. This is not so simple on a cylinder, as we do not know the exact angle between \vec{b}_1 and \vec{b}_2 after folding.

So, we must use coordinates in Euclidean space. The coordinates of the n th vertex A_n of an helix are given by:

$$x_n = \cos n\theta \quad y_n = \sin n\theta \quad z_n = nc. \tag{4}$$

The distances between vertex n and vertices $n + 1, n + 2$ and $n + 3$ are first neighbours distances. Then

$$\overline{A_m A_{m+n}}^2 = \overline{A_0 A_n}^2 = 2 - 2 \cos n\theta + n^2 c^2. \tag{5}$$

Since the edges $\overline{A_0 A_n}$ all have the same length, for $n = 1, 2, 3$, we find, eliminating c , an equation for $x = \cos \theta$: $3x^3 - 4x^2 - x + 2 = 0$, which factorises as $(x - 1)^2(3x + 2) = 0$. Discarding the trivial root $x = 1$, we deduce that the angle θ is given by

$$\cos \theta = -2/3, \quad \theta = 131.810^\circ. \tag{6}$$

We can also obtain the translation part of the helical motion, or pitch, $c = \sqrt{10/27}$, or $c/\overline{A_0A_1} = 1/\sqrt{10} = 0.3162$ in unit of edge length.

The number of edges per turn is given by $\xi = 2\pi/\theta$. It is $\xi = 2.7312$, close to the number $30/11$ on the torus.

3.4. The α -helix: a disclinated Coxeter helix

The α -helix is one of the important secondary structures found in proteins. The number of elementary steps of the backbone is given to be close to 3.6 units per turns. This is about one larger than $\xi = 2.7312$ for the B-C helix, so that we must increase its diameter.

Disclinations are the natural defects associated with rotational or helicoidal symmetry. In the case of helices, disclinations are characterised by an axis, which is the axis of the cylinder on which is drawn the helix, an angle $\delta\theta$ of rotation, and a vector of translation $\vec{\delta s}$ parallel to the axis. Such a wedge disclination combined with a translation is sometimes called a dispiration [14, 15]. The effect of a disclination on a cylinder is explained on the Figure 5a: the perimeter of the cylinder of radius r is changed from $2\pi r$ to $(2\pi + \delta\theta)r$, and one of the lips of the

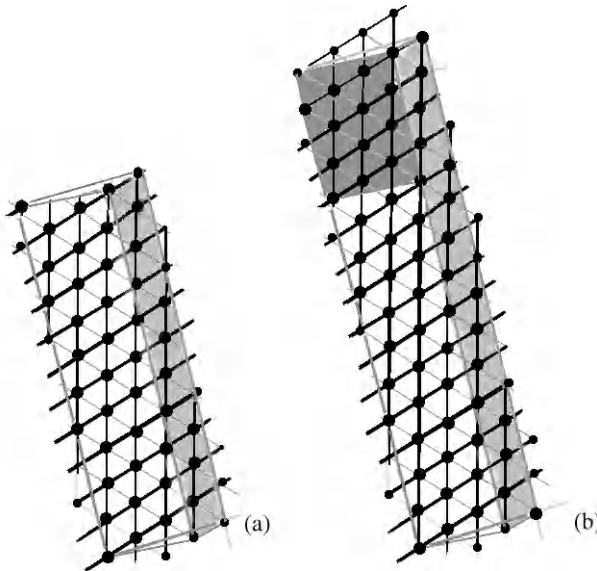


Figure 5 The strip for an α -helix, by identification of the longer sides. (a) The resulting cylinder is obtained by disclinating the cylinder supporting the B-C helix, along its axis. This adds one additional row of triangles (shaded). (b) In spherical space S_3 , in order to keep the fibration the torus has to transform by two disclinations along its two axes. A diagonal of the strip remains a fibre, which is not the case in (a).

cut cylinder is translated by $\vec{\delta s}$ before regluing. If there is a discrete geometrical structure supported by the cylinder surface, as a discrete helix, the displacement which is the combination of the rotation $\delta\theta$ and the translation $\vec{\delta s}$ must be an element of the symmetry group of the structure. If the helix is represented on a strip, the strip is sheared and its width is increased (for positive $\delta\theta$). This is the case whether the helix is drawn on a cylinder or, in curved space S_3 , on a torus. changing the parallelogram patch of the triangular tiling: its width is increased by one triangular unit and it is sheared in order to ensure identification of the longer sides. The new parallelogram is defined by vectors \vec{b}_1^α and \vec{b}_2^α : $\vec{b}_1^\alpha = 4\vec{a}_1 + 3\vec{a}_2$ and $\vec{b}_2^\alpha = -3\vec{a}_1 + 8\vec{a}_2$ (unchanged). The type-{1} helix running along edges parallel to $\vec{a}_1 + \vec{a}_2$ consists of 41 edges; it turns 11 times around one axis of the torus and once around the other, leading to a number of edges per turns $\xi = 41/11$. This helix is (4, 3, 1) in phyllotactic notation. There is an other choice, related to the torus in S_3 , describe on the Figure 5b. In this case there are 55 sites, and the type-{1} helix turns 15 times around one axis of the torus and once around the other, leading to a number of edges per turns $\xi = 55/15$.

In order to obtain the number of edges per turn ξ in Euclidean space, we use the coordinates defined in equation (1) for the B-C helix and set equal the distances between neighbours $\overline{A_0A_1} = \overline{A_0A_3} = \overline{A_0A_4}$. Note that the helix is no longer a chain of face-on-face tetrahedra, and that all other distances between vertices, notably $\overline{A_0A_2}$, are larger than $\overline{A_0A_1}$. Eliminating c , we obtain a quadratic equation for $x = \cos\theta$, $(x - 1)^2(16x^2 + 17x + 2) = 0$. The trivial roots $x = 1$ can be discarded. The root $x = -17 - \sqrt{161}$ gives a distance between non-neighbouring vertices $\overline{A_0A_2}$ smaller than $\overline{A_0A_1}$ and is incompatible with steric repulsions. The only geometrically relevant root is thus,

$$\cos\theta = (-17 + \sqrt{161})/32, \quad \theta = 97.74^\circ. \quad (7)$$

The number of edges per turn, given by $\xi = 2\pi/\theta$, is $\xi = 3.6831$, close to the rational numbers $41/11 = 3.727272\dots$ or $55/15 = 3.6666\dots$, obtained on the torus. These numbers are close to the value 3.6 observed in the α -helices, which is a two dimensional structure of proteins stabilised by hydrogen bonds, a celebrated result of L. Pauling in 1951, who "let the models fold naturally into any screw they were comfortable with" [16].

The translation parallel to the helix axis, per step, is $c = 0.3637$ or $c/\overline{A_0A_1} = 0.2347$ in units of edge length.

3.5. Other helices in proteins

Other helices that the classical α -helix are sometimes observed in proteins. All helices which cover a rolled triangular lattice are labelled $(k, k-1, 1)$ in phyllotactic notation.

- The B-C helix, (3, 2, 1) in phyllotactic notation, has hydrogen bonds represented by edges between sites i and $i + 2$ sites. Topologically, it is identical to the so-called 3_{10} -helix [17, 18]. This helix is not commonly observed in proteins as a secondary structural element. But α -helices sometimes begin or end with one single turn of a 3_{10} -helix (one hydrogen bond). There are also indications that long (3, 2, 1) helices are observed in biopolymers [19]. Hydrogen bonds in a 3_{10} -helix link the N atom of the backbone of amino acid i to the C atom of amino acid ($i + 2$).
- The next possibility is the α -helix (4, 3, 1). There are hydrogen bonds represented by triangle edges between sites i and $i + 3$, thus connecting peptide units i and $i + 4$.
- The (5, 4, 1) helix, obtained by folding a strip with one additional row of triangles compared to the α -helix is called the π -helix.
- The (6, 5, 1) helix corresponds to the Pauling 5.1-helix (or γ -helix).

Increasing k further would yield helices on very flat cylinders; the steric repulsion between side groups becomes too important and there are no proteins with $k > 6$.

Several polypeptide synthetic helices have phyllotactic structures, as was noticed by Frey–Wyssling [20].

4. Proteins as Close Packing

Proteins, could be described as a dense packing of entities representing the amino acid (AA). Therefore there is a simplified approach to proteins, in which amino acids are rigid entities. Even if a protein is a very complex object, it must follow the usual constraints. First of all it has to respect geometrical and topological rules, which govern its structure [21]. These rules are simple, but they are in competition for the influence at the level of different components of the protein and this generates the complexity of the protein world. The topology restricts strongly the local geometrical properties of complex objects like proteins. A protein structure which is a dense frustrated structure, is related to the dense {3, 3, 5} polytope and to structures derived from this polytope by introduction of defects.

In place of the {3, 3, 5} polytope, whose vertices represent centre of close packed objects, it is possible to use it dual, the {5, 3, 3} polytope. It is a packing of dodecahedra which represent the volume of the close packed objects.

4.1. Laguerre and Voronoi cells in proteins

The structure of folded proteins can be analysed by means of a quite common tool used in condensed matter physics, namely the Voronoi tessellation [22]. Given a set of points, Voronoi tessellation proceeds by determining for each point the polyhedron, called Voronoi cell, containing the portion of space closer to that point

than to all others. Done on the $\{3, 3, 5\}$ polytope this procedure gives exactly the dual, with dodecahedral Voronoi cells. The cell characteristics provide essential information on the local geometrical properties of the considered packing. That said, Voronoi decomposition does not give sizes that correspond well to the real size of the packed objects. The Voronoi decomposition increases the size of small AA (eg. *Gly*) and reduces the size of large AA (eg. *Phe*). There is, however, a modified Voronoi method, known as Laguerre decomposition [23] that takes into account the size of the packed objects.

In order to build the cell associated with a given AA, it is necessary to have a precise knowledge of its neighbourhood. This is easy for AAs which are located well inside a dense region of the bulk of the protein. In that case it is enough to know the positions of its neighbouring AA centres. But for an AA located close to the external surface or in a cavity, this becomes more difficult as, in principle, a detailed knowledge of the nature and location of the surrounding molecules is needed. This difficulty can be resolved by surrounding the protein with a model of solvent, or 'environment', whose characteristics are similar to generic proteins considered as random dense packing of equal sized spheres of average AA volume.

4.2. The protein '3chy' as an example of Laguerre tessellation

A complete Laguerre tessellation of the signal transduction protein Che Y from *Escherichia Coli* (PDB [24] code 3chy; 128 amino acids; 1 chain) is shown on Figure 6 [25]. The atomic coordinates from crystallographic X-Ray structure have been used to get the geometrical centres of every amino acid. The environment of random packing of spheres is simulated by a box of about 8000 spheres which completely bathes the protein. Spheres that overlap with the amino acids are removed, as well as spheres too far from the protein. Finally, the total number of amino acids and spheres is about 2000.

4.3. Cell statistics

Two quantities of interest give an idea of the overall geometry of the packings of AAs in proteins, the mean number of faces per cell $\langle f \rangle$, which is the mean coordination number between AAs, and the mean number of edges per faces $\langle e \rangle$, which is related to the local symmetry around bonds between neighbouring AAs. Using the Laguerre method and averaging over a set of 35 proteins we find $\langle e \rangle = 5.15$ and $\langle f \rangle = 14.17$.

In the $\{3, 3, 5\}$ polytope these quantities are exactly $\langle e \rangle = 5$ and $\langle f \rangle = 12$, but in this case, the space is curved. There is a way to estimate these quantities for a random packing of spheres in the flat space. We suppose the sphere centres on vertices of tetrahedra (a simplicial decomposition). Voronoi froth vertices are therefore tetravalent. Each such vertex belong to four Voronoi polyhedra and has

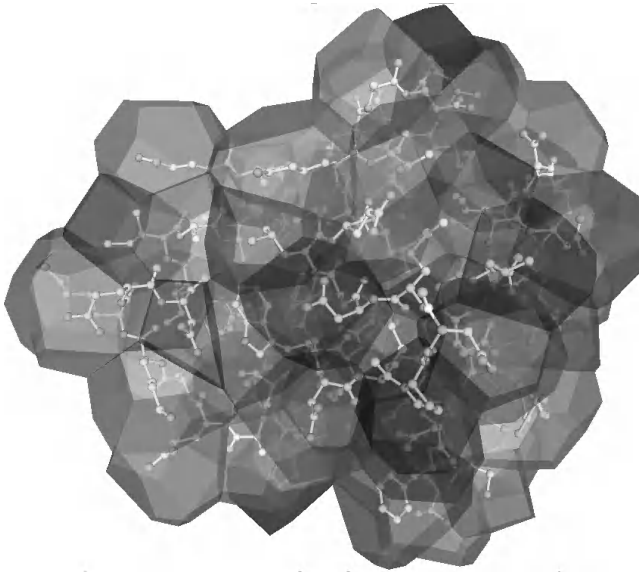


Figure 6 Laguerre tessellation of the signal transduction protein Che Y from Escherichia Coli (PDB code: 3chy). The protein main chain is shown.

three neighbours on each polyhedron. Any ring is common to two Voronoi polyhedra associated with two neighbouring sites. The ring size is equal to the number of tetrahedra sharing the two sites (sharing the edge joining the two sites in the original simplicial set).

If the tetrahedra were all regular, there would be room for about 5.1 tetrahedra around a common edge. Imagine now an ‘impossible’ structure, inside which each edge is shared by exactly $\langle p \rangle = 2\pi/\cos^{-1}(1/3) \simeq 5.104299$ tetrahedra. Such a structure has been proposed by Coxeter [26] under the name ‘statistical honeycomb’, noted $\{3, 3, \langle p \rangle\}$. It corresponds to an ideal case where frustration is as diluted as possible in space, while in disclinated structures it is concentrated along the defect lines (but in that case we get ‘possible’ structures). Using this value of $\langle p \rangle$ in order to get the approximate number of faces of the Voronoi polyhedron gives $\langle f \rangle \simeq 13.39$.

Departure from this value are explained (see [4] for references) by polydispersity of sphere packings. Large fluctuations in cell sizes decrease the coordination number $\langle f \rangle$. By contrast, cell shape anisotropy, increases $\langle f \rangle$. Since disorder cause both cell fluctuations and disorder, small values of $\langle f \rangle$ are usually observed only in crystalline structures. The value for proteins indicates that the Laguerre cells for AAs in proteins are not too anisotropic and adjusted to some extent. Anyhow these

values are still close to those of compact structures often encountered in condensed matter physics.

In the spirit of finding *ab initio* methods to obtain the protein structure knowing the sequence of its AAs, it is very useful to perform cell statistics separately for each amino-acid to try to recognise them from the geometry of their cells. On Figure 7a one provides the histograms $h(e)$ and $h(f)$, fraction of faces with e edges and fraction of cells with f faces, for the twenty amino-acids (here Cys and Cysh are not distinguished). There are particular faces which define the surface of the protein, i.e., those at the interface between an AA and a sphere of the environment. Figure 7b gives the histograms $h_S(e)$ for the number of edges of surface faces, and $h_S(f)$ for the number of surface faces per surface cells. The full histograms (regardless the AA) are given at the bottom of the figure.

On these figures it appears clearly that large AAs have a larger number of faces than small AAs. Furthermore, for surface cells, hydrophilic AAs, have a larger number of surface faces (5.75 for Glu) than hydrophobic AAs (3.13 for Leu).

The surface faces are in contact with the environment. Their statistics are especially interesting. The mean number of edges of surface faces is almost exactly $\langle e_s \rangle = 5.00$. Any finite cell packing, extracted from a disordered infinite packing should have the same statistic for its internal faces and for its surface faces, thus $\langle e_s \rangle = \langle e \rangle \simeq 5.15$. This is not the case here, indicating again some topological order in the AA packing.

There is another point appearing on these histograms which remains unexplained, but which could be fruitful in fold prediction, as it seems related to some of the AAs only. The width of the $h(f)$ distribution is very different from one AA to another; it is neither correlated to its size, nor to its hydrophobicity, as can be seen by comparing Glu and Lys or Arg and His, for instance. This probably indicates that some AAs have a local neighbourhood more uniform than others.

4.4. Proteins versus random close packed structure

At first sight, if one considers only global statistics for the number of edges per faces or for the number of faces per cells, a protein seems very similar to random close packing. If less constrained than hard sphere packing, it can be viewed as some kind of froth with cells of different sizes. Nevertheless there are interesting details which indicate that a protein has a backbone.

Analysing Voronoi cells [22, 27] in proteins, it has been observed that faces shared by two successive AA along the chain have, on average, a larger number of edges. Laguerre cells exhibit the same effect, with a number of edges for these faces $\langle e_{\text{chain}} \rangle \simeq 6.5$. For any face, the average number of edges per face is close to $\langle e \rangle = 5.1$ for isotropic cells of equal sizes. This implies that interfaces other than those separating two successive AA along the chain, are smaller than average: they have approximately five edges per face. So we can view a protein as a chain

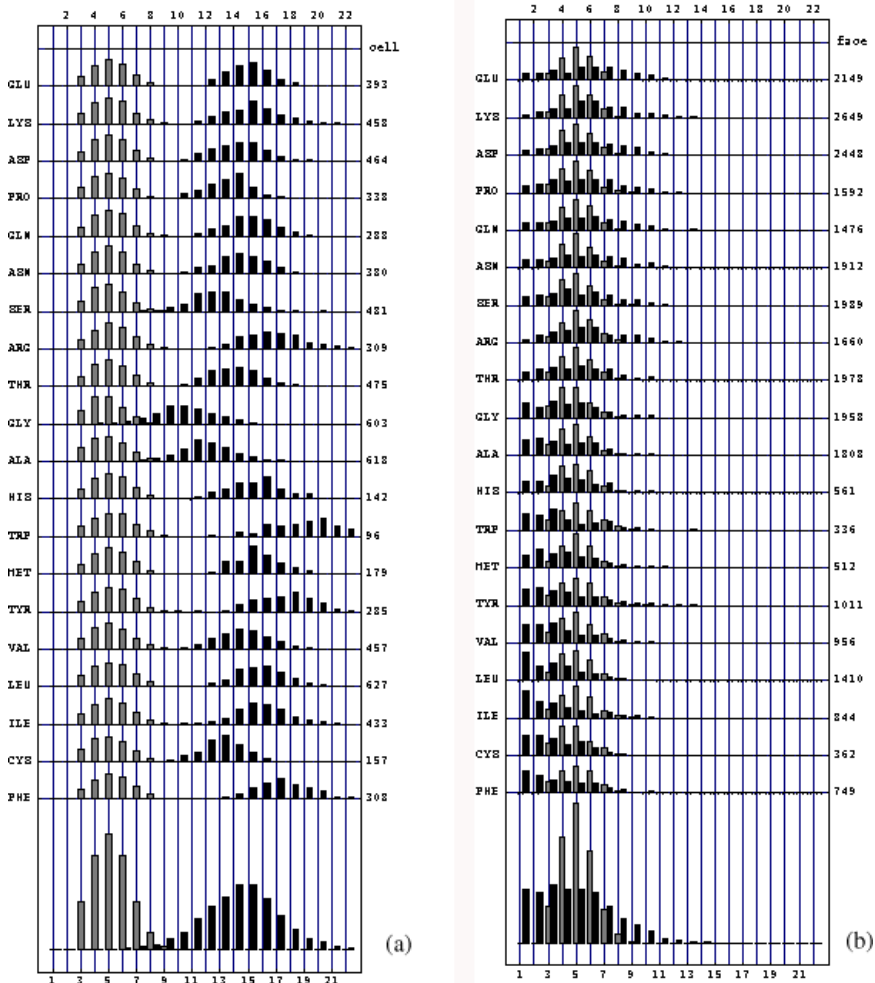


Figure 7 (a) Histograms $h(e)$, in grey, and $h(f)$, in black (fraction of faces with e edges and fraction of cells with f faces) for each amino-acid. The numbers indicated on the right are the numbers of cells involved in the statistics. The amino-acids are classified from hydrophilic (top) to hydrophobic (bottom). The full histograms (regardless the amino-acid) are given at the bottom of the figure. (b) Histograms $h_S(e)$ for faces corresponding to the surface of the protein, in grey, and $h_S(f)$, in black, of the number of surface faces for the cells at the surface of the protein. Same as in the Figure 7 save for the numbers on the right which are now the numbers of faces involved in the statistics.

of closely packed cells, slightly compressed along the chain, thus making a kind of tube tiled by faces with an average number of edges close to five.

5. Disclination Lines in Proteins

5.1. *More on disclinations*

A disclination, which is a defect involving a rotation operation, can be generated by a so-called ‘Volterra’ process, by cutting the structure along a (geodesic) line and adding (or removing) a sector of material between the two lips of the cut. It is a linear defect in three dimensions. The two lips of the sector should be equivalent under a rotation belonging to the structure symmetry group in order to get a pure topological defect confined near the apex of the cut. Wedge disclinations can be viewed as loci of curvature concentration in a three-dimensional space. If the disclination is obtained by a Volterra process in which matter has been removed (added), it is a positive (negative) disclination. Therefore, introducing negative disclinations can be used in order to flat a positively curved space like the $\{3, 3, 5\}$ or the $\{5, 3, 3\}$ polytopes. It is this structure which define the ideal structure. Disorder is necessarily associated to a mixing of positive and negative disclinations. In a Voronoi or Laguerre tessellation disclinations are easy to identify: they go through faces which are not pentagons. For instance, if we suppose a tessellation where cell faces have only four, five or six edges, positive disclinations go through four-sided faces and negative disclinations go through six-sided faces. They have to respect equilibrium rules when they intersect on cell centres: they behave like lines under tension with a tension associated to the number of edges, compare to five, of the faces they thread.

5.2. *Network of disclinations in proteins*

Consider a protein reduced to the centre (sites) of its amino-acids and then define its Voronoi or Laguerre tessellation. As given on the histograms presented Figure 7, there are faces with three, four, five, six and more edges. Nevertheless with a relaxation procedure displacing slightly sites it is possible to have only (with few exception) four, five and six sided faces. This defines a network of positive and negative disclinations. We have observed that the faces on the surface of the protein have an average of five edges. By moving the environment, it is possible to get all faces of the surfaces with five sides. This is interesting as with this condition the disclination network is entirely in the protein.

The chain of amino acids forms a zigzag of edges connecting successive sites. Following statistical results, in this model we consider that faces shared by successive sites are six-sided faces. Then the bond threading such a face can be considered as a negative disclination. These bonds make an angle at each site along the chain, so necessarily, there are positive disclinations which balance the angle of negative

disclinations. Consider three successive sites $i, i + 1, i + 2$ and suppose that i and $i + 2$ are in contact. The common face to the cells i and $i + 1$ or to the cells $i + 1$ and $i + 2$ are hexagons. These two large faces belong to the cell $i + 1$. Statistically the other faces of this cell are small faces. This is true for all the cells, so the face between i and $i + 2$ is statistically a small face. This is observed on the Figure 8,

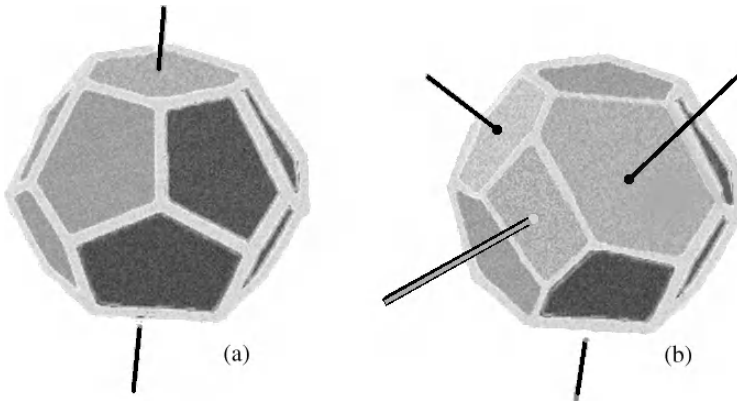


Figure 8 (a) One negative disclination threading a cell with 14 faces (two 6-gons and twelve 5-gons). On this example, a disclination is the defect which change a regular dodecahedron into this 14-face polyhedron. (b) A node of disclinations in the centre of a 14-face cell with eleven 5-gons, three 6-gons and one 4-gon. Negative disclinations (in black) are balanced at each nodes (cell centre) by a positive disclination (in grey with border) like forces of strength $+1$ and -1 .

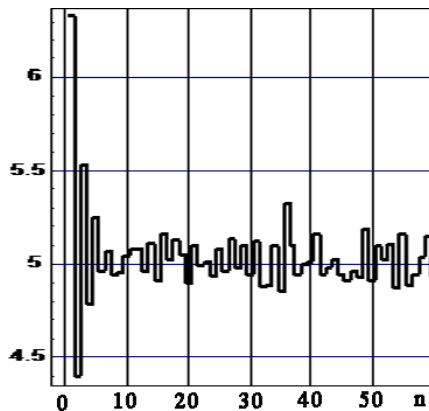


Figure 9 The mean number of edges for faces between two cells i and $i + n$. It appears, at least for n smaller than 8, that this number oscillates around the global mean value $\langle e \rangle \simeq 5.1$.

showing oscillation of the number of edges per face depending on the parity of n for contact between i and $i + n$.

6. Conclusions

Bio-polymers are, at a first level of organisation, one-dimensional sequences like any polymer chains; but, at a second level, various interactions impose organised structures in space. The different types of structures in proteins: primary, secondary or tertiary are related to close-packed structures in one, two or three dimensions. The α -helix, considered here as a two-dimensional close-packing, a triangular lattice rolled on a cylinder, is an essential step in protein folding. This is, incidentally, one of the reasons for the success of the hydrophobic cluster analysis (H.C.A.), which predicts the folding pattern of several proteins [28]. Viewing the α -helix as a two-dimensional structure makes clear the geometrical and topological constraints required. For instance, the number of residues observed per turn is imposed by the choice of a geometry of the triangular packing on the cylinder. Chiralities are also well described from this point of view.

In three dimensions, a description using tetrahedral packing is the good way to take care of the tendency of proteins to form dense random close packing structure. A packing of regular tetrahedra is undistorted and defect-free only in curved space: This is the $\{3, 3, 5\}$ polytope, and its the B-C chain. The B-C chain can be put into Euclidean space, and extended, if necessary. With small distortions of the tetrahedra, it can also coexist, tightly packed, with other B-C chains. This finely tuned local geometry with minimal distortion is extendable to longer helices without increasing the distortion.

While the exact geometry of biological helices may appear complicated, their topology is determined simply and directly by steric considerations.

Acknowledgements

Here, we have presented works which have been done in collaboration with different persons and have benefited from helpful discussions. I would like to thank Borislav Angelov, Isabelle Caillebaut, Rémi Jullien, Jean-Paul Mornon and Nicolas Rivier, together with their colleagues and students for these collaborations and discussions.

References

- [1] Coxeter, H. S. M. 1973, Regular polytopes. 3rd edn. New York: Dover Publications.
- [2] Coxeter, H. S. M. 1974, Regular complex polytopes. Cambridge: Cambridge University Press.

- [3] Lord, E. A., Ranganathan, S. 2001, Sphere packing, helices and the polytope {3, 3, 5}. *Eur. Phys. J. D.* 15, 335–343.
- [4] Sadoc, J.-F., Mosseri, R. 1999, Geometrical frustration. Cambridge: Cambridge University Press.
- [5] Sadoc, J.-F., Mosseri, R. 1997, Frustration géométrique. Paris: Eyrolles.
- [6] Soyer, A., Chomilier, J., Mornon, J. P. *et al.*, 2000, Voronoï tessellation reveals the condensed matter character of folded proteins. *Phys. Rev. Lett.* 85, 3532–3535.
- [7] Sadoc, J.-F. 2000, Toroidal DNA: topology, geometry and electrostatics. *Inter. Jour. of Modern Physics B.* 14, 737–749.
- [8] Boerdijk, A. H. 1952, Some remarks concerning close-packing of equal spheres. *Philips Res. Rep.* 7, 303–313.
- [9] Coxeter, H. S. M. 1985, The simplicial helix and the equation $\tan m\theta = n \tan \theta$. *Canad. Math. Bull.* 28, 385–393.
- [10] Nicolis, S., Mosseri, R. and Sadoc, J.-F. 1986, Polytopes with tangled disclinations. *Europhys. Lett.* 1, 571.
- [11] Sadoc, J.-F., Rivier, N. 1999, Boerdijk-Coxeter helix and biological helices. *Eur. Phys. J. B.* 12, 309–318.
- [12] Pittet, N., Boltenhagen, P., Rivier, N. *et al.*, 1996, Structural transitions in ordered, cylindrical foams. *Europhys. Lett.* 35, 547–552.
- [13] Erickson, R. O. 1973, Tubular packing of spheres in biological fine structure. *Science.* 181, 705–716.
- [14] Harris, W. F. 1977, Disclinations. *Sci. Am.* 237, 130–136, 138–142, 144–145.
- [15] Harris, W. F., Chandler, H. D. and Hepburn, H. R. 1976, *South African Journal of Science.* 72, 25–26.
- [16] Crick, F. 1988, What mad pursuit: a personal view of scientific discovery. Alfred P. Sloan Foundation Series. New York: Basic Books, p. 58.
- [17] Miick, S. M., Martinez, G. V., Fiori, W. R. *et al.*, 1992, Short alanine-based peptides may form 3_{10} -helices and not α -helices in aqueous solution. *Nature.* 359, 653–655.
- [18] Zhang, L., Hermans, J. 1994, 3_{10} helix versus α -helix: a molecular dynamics study of conformational preferences of aib and alanine. *J. Am. Chem. Soc.* 116, 11915–11921.
- [19] Pal, L., Basu, G. 1999, Novel protein structural motifs containing two-turn and longer 3_{10} -helices. *Protein Engineering.* 12, 811–814.
- [20] Frey-Wyssling, A. 1954, Divergence in helical polypeptide chains and in phyllotaxis. *Nature.* 173, 596.
- [21] Baker, D. 2000, A surprising simplicity to protein folding. *Nature.* 405, 39–42.
- [22] Angelov, B., Sadoc, J.-F., Jullien, R. *et al.*, 2002, Nonatomic solvent-driven Voronoï tessellation of proteins: an open tool to analyze protein folds. *Proteins: structure, function, and bioinformatics.* 49, 446–456.
- [23] Sadoc, J.-F., Jullien, R. and Rivier, N. 2003, The Laguerre polyhedral decomposition: application to protein folds. *Eur. Phys. J. B.* 33, 355–363.
- [24] Berman, H. M., Westbrook, J., Feng, Z. *et al.*, 2000, The Protein Data Bank. *Nucleic Acids Research.* 28, 235–242. See also the Protein Data Bank on the web site <http://www.rcsb.org/pdb/>.

- [25] Dupuis, F., Sadoc, J.-F., Jullien, R. *et al.*, 2005, Voro3D: 3D voronoi tessellations applied to protein structures. *Bioinformatics*. 21, 1715–1716. [Originally Online] 2004. [Accessed 24th June 2004]. See also the software VORO3D on <http://www.lmcp.jussieu.fr/dupuis/>.
- [26] Coxeter, H. S. M. 1958, Close-packing and so forth. *Illinois J. Math.* 2, 746–758.
- [27] Dupuis, F., Sadoc, J. F. and Mornon, J. P., 2004, Protein secondary structure assignment through voronoi tessellation. *Proteins: structure, function, and bioinformatics*. 55, 519–528.
- [28] Callebaut, I., Labesse, G., Durand, P. *et al.*, 1997, Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol. Life Sci.* 53, 621–625.

This page is intentionally left blank

CHAPTER 12

When Topology and Biology Meet ‘For Life’: The Interactions Between Topological Forms and Biological Functions

LUCIANO BOI

*École des Hautes Études en Sciences Sociales,
Centre de Mathématiques
Postal address: EHESS-CAMS,
790, avenue de France, 75013 Paris, France
boi@ehess.fr*

Overview

This study is aimed at showing that differential geometry and topological knots theory can be used notably to modelling three-dimensional structures of DNA and protein-DNA complexes. Our goal is twofold: firstly, we want to show that certain topological deformations associated to the supramolecular structures during the cell cycle take part in the dynamics of chromatin, the organisation of chromosome and therefore the cell’s metabolism; secondly, we try to illustrate the way in which these deformations might modulate the action of many different regulatory systems, ensuring in particular the transition of this action from a local-target mechanisms to global functional processes. We shall argue that these interactions between topological changes and dynamic processes constitute a deep and largely unexplored meeting point for mathematics and biology. Further, we suggest that certain geometric properties and topological patterns work like dynamic principles, which are involved in the organisation and growth of living systems. Moreover, these properties and patterns display intricate biological plasticity and complexity on every scale from very large (i.e., the organism) to very small (i.e., the molecule), and also contribute to the multilayer ordering of biological regulation and activity.

I. Remarks on the Unlinking of DNA Molecule and the Chromosome Segregation *in vivo*

Our understanding of fundamental epigenetic phenomena make it necessary to acknowledge that the true carrier of genetic information is the chromosome rather than just DNA. Indeed, the chromatin structure appears to harbour metastable key features determining the interpretation of genetic information. This organisation layer need to be understood and integrated with the organisation layer of the genome sequence to model genetic networks properly. In this manner, we will advance our global understanding of the way the information contained in the genome is interpreted by the cell. This article addresses the question of DNA structure and chromatin dynamics. It is mainly aimed at describing some aspects of the way in which, first, the two strands of DNA must be continuously unlinked during replication, and second, the chromatin is topologically condensed within the cells of organisms with nuclei.

Three families of huge ATP-powered enzymes — helicases, type II topoisomerases, and condensins — contribute to the orderly unlinking of DNA and to the chromosome segregation *in vivo*. In this process, two steps seem to be very fundamental.

- (1) For replication to occur, the DNA must be decondensed. Helicases unwind DNA creating (+) supercoils and precatenanes which are rapidly removed by topoisomerases. Type-2 topoisomerases actively remove all DNA entanglements.
- (2) Then, the organised recompactation by condensins and supercoiling are essential for chromosome partitioning. The chromosome must, indeed, be folded into topological domains. Besides, chromosome needs to be topologically remodelled in order that the genetic events and cellular process may be performed. Finally, chromatin structure and chromosome conformation are dynamic and complex and exert profound control over gene expression and other fundamental cellular activities.

The fundamental determinants of many biological phenomena are now known to be geometrically organised and topologically constrained. And the biologically most important molecules or macromolecules (proteins, RNA and DNA) are comprised of linear chains of building blocks, which yet assemble themselves according to some mathematical rules that are highly non-linear and extremely systems-dynamically complex. So, even if each protein or RNA and DNA molecule commonly folds into a specific structure that depends sensitively on its sequence; however, when we take into account large ensembles of these macromolecules and their associated systems, then we observe that they depend rather on the action of some large-range correlation structures. From a systematic statistical-mechanics

point of view analysis of DNA sequences, these large-range correlations can be interpreted in terms of the signature of the hierarchical, structural and dynamical organisation of chromatin in relation with DNA replication, gene expression and cell division. At any rate, DNA in living systems is topologically constrained, so its structure also depends on how it is constrained (as we shall show thoroughly).

Think, for example, of the local but extremely important problem of the fate of cells, of the partitioning of chromosome and the unlinking of DNA during replication. The two strands of DNA must be continuously and accurately unlinked during replication. The topoisomerases that accomplish this might instead be expected to entangle and knot chromosomes because of the enormous DNA concentration *in vivo*. In fact many factors have been identified that solve this problem and contribute to the orderly unlinking of DNA.

A major contributor to chromosome partitioning is the condensation of daughter DNA upon itself soon after replication. It has been shown that DNA condensation is due primarily to supercoiling that is introduced by topoisomerases and maintained by SMC (structural maintenance of chromosome) of proteins, often called condensins. However, it is yet scarcely known how these proteins cause the orderly long-range folding of DNA. Another factor promoting chromosome partitioning is that the type-2 topoisomerases of all organisms do not just speed up the approach to topological equilibrium, but actually change the equilibrium position. They actively remove all DNA entanglements. This requires that all topoisomerases sense the global conformation of DNA even though they interact with DNA only locally (see Rocca [92] and Cozzarelli [39]).

According to a hypothesis proposed by Holmes and coworkers [60] and by Strick *et al.* [99], topoisomerases might achieve this operation because they are like Maxwell's demon and by positioning themselves at sharp bends in DNA carry out net disentanglement of DNA. All the enzymes that play critical roles in DNA unlinking and chromosome segregation — helicases, topoisomerases and condensins — are motor proteins. They use the energy of ATP hydrolysis to move large pieces of DNA over long distances. Helicases seems to convert the energy of ATP hydrolysis into unwinding DNA.

1.1. Topological operations and biological functions

There are many mathematically challenging problems related to molecular and macromolecular sequences and structures. We mention two of these problems.

- (1) The shapes of proteins may be described using differential geometry and topology. More precisely, recent studies show that the topology of the transition state (the rate-limiting event in the folding reaction is the formation of a conformation in a set known as the transition-state ensemble) is determined by a set

of interactions involving a small number of key residues and, in addition, that the topology of the transition state is closer to that of the native than to that of any other fold in the protein universe. In other words, we have to link the folding process to the topological organisation in the transition states for protein folding (see Chapters 10 and 11 in this book).

- (2) The topological constraints on DNA commonly involve the regulation of its linking number by the transient cutting by enzymes. The activities of DNA, including gene expression and replication, depend sensitively, even not only, on the linking number imposed, which is a topological invariant. This topological invariant can be decomposed into the sum of two geometric invariants, whose analysis involves integral geometry. We will return to this point in detail shortly.

The general idea lying behind these puzzles is that the double-helix structure of DNA is a geometrical entity, more precisely a topological configuration. It turns out that this topological configuration is itself a manifestation of linking and knotting. DNA within the cell is a very long molecule with a remarkably complex topology. Topological properties of DNA are defined as those that can be changed only by breakage and reunion of the backbone. Moreover, it should be underlined that the complex topology of DNA is essential for the life of organisms. In particular, it is needed for the process known as DNA replication, whereby a replica of the DNA is made and one copy is passed on to each daughter cell. The most direct evidence for the vital role played by DNA topology is provided by the results of attempts to change the topology of DNA inside cells. The topology of DNA *in vivo* is set by a remarkable group of enzymes called *topoisomerases*. In short, these enzymes essentially promote the passage of DNA segments through each other until a stable state is achieved. This stability is thus made possible thanks to a conformational flexibility of the double-helix, and the continuous remodelling of nuclear structures is as well required for cell activity to be performed.

There are three important topological properties of DNA: (i) The linking number between two strands of the double helix, (ii) The interlocking of separate DNA rings into what are called catenanes, (iii) and knotting. Physical and phenomenological (observed) properties are: (i) As the number of crossing in a knot or catenane increases, the number of possible isomers grows exponentially. (ii) The linking number of DNA in all organisms is less than the energetically most stable value in unconstrained (relaxed) DNA. This puts the DNA under stress, which causes it to buckle and coil in a regular way called negative (–) supercoiling. The (–) indicates that the linking number is less than in the relaxed state. (iii) The name supercoiling arises because it is the coiling of a molecule, which is itself formed by the coiling of two strands about each other. Although supercoiling is, strictly speaking, a geometric property, it is a consequence of a topological one, the linking number difference between supercoiled and relaxed DNA.

1.2. Some useful topological notions

At this point, it is useful to give, first, some basic definition of these topological concepts, next, a more in-depth discussion of the notion of *linking number*, which is central to our scope here. We start by giving the analytical formula for the linking number of a pair of knotted curves. (Note that this formula goes back to Gauss — who gave it in 1833! — in his work on electromagnetism theory, which led him to compute inductance in a system of two linked circular wires). The linking number of a pair of knots is a combinatorial topological invariant (it is an integer number). Moreover, one can now show that this number is invariant under *Reidemeister moves*.

Recall briefly what these moves are (for further mathematical details, see Boi [23]). They apply to pairs of equivalent links. First, we need the mathematical definitions of a *link* and of *equivalence of links*.

Definition 1.1. A link L of m components is a subset of S^3 , or of R^3 , that consists of m disjoint, piecewise linear, simple closed curves. A link of one component is a knot.

Definition 1.2. Links L_1 and L_2 in S^3 are equivalent if there is an orientation-preserving piecewise linear homeomorphism $h : S^3 \rightarrow S^3$ such that $h(L_1) = (L_2)$.

Any two diagrams of equivalent links L_1 and L_2 are related by a sequence of Reidemeister moves and an orientation-preserving homeomorphism of the plane. (A *link diagram* of L is the image of L in R^2 together with 'over and under' information at the crossings. Of course, a crossing is a point of intersection of the projections of two line segments of L ; the 'over and under' information refers to the relative heights above R^2 of the two inverse images of a crossing.) The Reidemeister moves are of three types; each replaces a simple configuration of arcs and crossings in a disc by another configuration. A move of type I inserts or deletes a 'kink' in the diagram; moves of type III preserve the number of crossings. Any homeomorphism of the plane must preserve, obviously, all crossing information. In other words, and following a theorem by Reidemeister (1927), all changes of link or knot diagrams can be obtained by performing, repeatedly, if necessary, three basic motions applied just to small portions of the diagrams near crossings, along with simple deformations in the plane, called *plane isotopies*, which do not change any of the crossings of diagrams. Rephrased in more general mathematical terms, this means there exist $h_t : S^3 \rightarrow S^3$ for $t \in [0, 1]$ so that $h_0 = 1$ and $h_1 = h$ and $(x, t) \rightarrow (h_t, x, t)$ is a piecewise linear homeomorphism of $S^3 \times [0, 1]$ to itself. Thus certainly the *whole* of S^3 can be continuously deformed, using the homeomorphism h_t at time t , to move L_1 to L_2 . We may conclude, from the above description, that a link or a knot invariant may be thought of a quantity that remains unchanged when we apply any one of the previous Reidemeister moves to a regular diagram.

Moreover, it turns out that if one link diagram for an oriented link is changed into another diagram for an oriented link by any Reidemeister move, the linking number does not change. This is true in the special cases of moves type I and type II. A very important conclusion we can draw is that the absolute values of the linking numbers of two equivalent oriented links will be equal. Such a difference can account for either exactly the same sets of left-handed and right-handed crossings where components meet or an exchange of those types of crossings. So linking numbers are either equal or negatives of each other. Therefore, linking number is an invariant of unoriented links of two components. We can sum up the previous statements by using the following theorem:

Theorem 1.3. *If two equivalent (unoriented) links of two components are each oriented in any way, then the absolute values of their linking numbers will be equal.*

The linking coefficient can be generalised for the case of p - and q -dimensional manifolds in R_{p+q+1} . The expression for the parameterised curves $\gamma_1(t)$ and $\gamma_2(t)$ with radius-vectors $r_1(t)$, $r_2(t)$ is given by the following formula

$$Lk(\gamma_1, \gamma_2) = 1/4\pi \int_{\gamma_1} \int_{\gamma_2} (r_1 - r_2, dr_1, dr_2) / |r_1 - r_2|^3. \quad (1)$$

The linking coefficient allows us to distinguish some two-component links.

Example 1.4. Let us consider the trivial two-component link and enumerate its components in an arbitrary way. Obviously, their linking coefficient is zero. For the Hopf link, the linking coefficient equals ± 1 depending on the orientation of the components. Hence, the Hopf link is not trivial.

Example 1.5. For any two components of the Borromean rings, the linking coefficient equals zero; each component of this link is a trivial knot. However, the Borromean rings are not isotopic to the trivial three-component link.

2. Topological and Dynamical Aspects of DNA Structure and the Spatial Organisation of the Chromosome

The stable structures of a DNA molecule are those conformations that minimise a conformational energy subject to the constancy of the topological conditions. This phenomenon gives rise to a range of variational problems. Experiments show that the stable structures of proteins minimise energy (see Chapter 10 in this book). Thus, in order to predict protein structures from sequences one must solve an optimisation problem. This is actually very difficult to do, because there may be many thousands of degrees of freedom within a single molecule, so its configuration space is high dimensional.

Despite its immense length, the linear sequence map of the human genome is an incomplete description of our genetic information. This is because information on genome function and gene regulation is also encoded in the way that DNA sequence is folded up with proteins within chromosomes and within the nucleus. This information cannot be portrayed in the DNA sequence alone. In the nucleus, individual chromosomes occupy discrete territories. So examining the spatial organisation of human chromosomes and genes in the nucleus appear to be very important. It seems that this organisation is changed, for example, during development and in certain diseases. Consequently, the way the human chromosome is topologically organised might influence how abnormal chromosomes are formed (for a more detailed account of this topic, see Boi [19]).

Using whole chromosome painting probes and fluorescence in situ hybridisation (FISH), a territorial organisation of interphase chromosomes has been demonstrated. Chromosome territories have irregular shapes and occupy discrete nuclear positions with little overlap. In general, gene-rich chromosome is located more in the nuclear interior while gene-poor chromosome territories are located at the nuclear periphery. In agreement with this, non-transcribed sequences were predominantly found at the nuclear periphery while active genes and gene-rich regions tended to localise on chromosome surfaces exposed to the nuclear interior or on loops extending from the territories.

These experimental findings support the concept of a *functional nuclear space*, the interchromosomal domain compartment (ICD). According to the ICD model, the interface between chromosome territories is more easily accessible to large nuclear complexes than regions within the territory. More recently, it has been proposed that chromosome territories are further organised into 1 Mb domains, extending the more accessible space to open intra-chromosomal regions surrounded by denser chromatin domains. Using high-resolution light microscopy, an apparent bead-like structure of chromatin can be visualised in which ~ 1 Mb domains of chromatin are more densely packed into an approximately spherical sub-compartment structure with dimensions of 300–4000 nm (see T. Cremer and C. Cremer [40]).

These domains are thought to be formed by a specific folding of the 30 nm chromatin fibre, to which the chain of nucleosomes associates under physiological salt concentrations. The different models that have been proposed are shown in the Figure 1. The radial-loop models propose small loops of roughly 100 kb arranged in rosettes, while the random-walk/giant loop model proposes large loops of chromatin back-folded to an underlying structure. In the chromonema model, the compaction of the 30 nm fibre is achieved by its folding into 60 to 80 nm fibres that undergo additional folding to 100 to 130 nm chromonema fibres. These dense highly compacted chromatin regions (localised at the nuclear periphery around the nucleolus and at the centromeres) are often referred to as heterochromatin as opposed to the less dense euchromatin. Heterochromatin has been described as containing increased DNA methylation at cytosines, specific histone modification

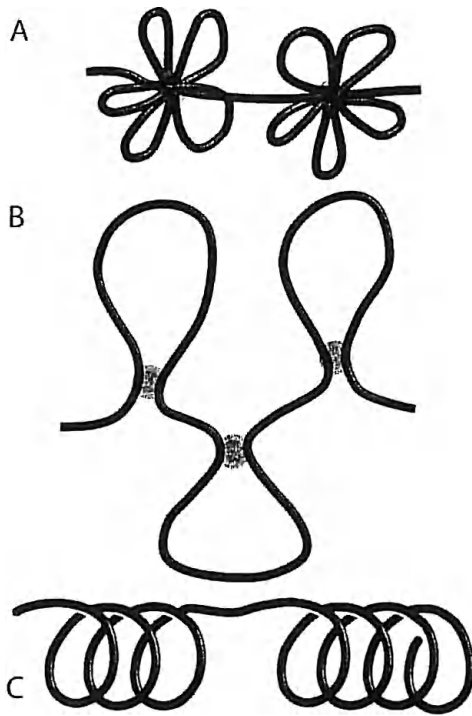


Figure 1 Models of the folding at the 30 nm fibre chromatin.

patterns like methylation of lysine 9 on histone H3 and histone hypoacetylation, binding of heterochromatin protein 1 (HP1), interaction with non-coding RNA and activities of the RNAi-mediated silencing machinery. The relation of dense heterochromatic state with a biologically inactive chromatin conformation has led to the concept that the biological activity of chromatin is regulated via its accessibility to proteins factors and co-factors.

2.1. Geometry of the double-helix and conformational modifications of chromatin

The information content of a DNA molecule is embodied in its sequence of paired nucleotide bases and is independent of how the molecule is twisted, tangled or knotted. In other words: twisting, coiling and knotting operations are able to enhance (increase) or to decrease (reduce) the structural and physiological functions of the genome and the cell nucleus. In the past decade, it has become clear that the topological form of a DNA molecule, the structural modifications of the

chromatin and the spatial architecture of the chromosome exert an important influence on the way in which DNA acts within the cell. Moreover, these three levels of organisation of the most fundamental nuclear components seem to be deeply related. Also, their functions are controlled by the action of different complexes of regulatory factors and co-factors, which may affect locally and globally the metabolism and physiology of cells. Among these different families of proteins' regulatory complexes, the remodellers of chromatin structure play a fundamental role in replication and repair of DNA sequences and in the transcriptional activities of the entire genome.

Let us first consider the basic level of DNA structure and coiling. Enzymes topoisomerases, which convert DNA from one topological form to another, appear to have a profound role in the central genetic events of DNA replication, transcription and recombination. It is a long-standing problem in biology to understand the mechanisms responsible for the knotting and unknotting of DNA molecules. Large amounts of DNA are wound up and packed into the average cell. DNA molecule is an incredibly long polymer whereas the cell's nucleus has a very thin spatial volume. This obviously means that the embedding of the DNA into chromatin within the cell core is exceedingly complicated; therefore, many complex structural modifications, topological deformations and regulatory networks interactions must work together in order to perform the proper packing of DNA into several folding-levels of chromatin, as well as to ensure the stability of the genome.

Next, it may be useful to describe the important connection between knot theory and molecular biology and, in more general terms, between topological transformations and biological processes. For many years, molecular biologists have known that the spatial conformation of DNA knots is a phenomenon involved in living matter. Indeed, macroscopic and microscopic knots and links are ubiquitous objects carrying a tremendous amount of precious information on the emergence of new forms in nature and the functions of organisms. Furthermore, knotting and unknotting are 'universal' scale-invariant principles underlying these phenomena. In particular, some topological contortions of the double-helix molecule, as well as some spatial distortions (bending, twisting...) carried out by those proteins that bind to a large variety of DNA sites, are essential to many biological processes.

It is worth of noticing that differential geometry and knot theory can be used to describe and explain the three-dimensional structure of DNA and protein-DNA complexes. Biologists devise experiments on circular DNA, which elucidate three-dimensional molecular conformation (helical twist, supercoiling, etc.) and the action of various important life-sustaining enzymes (topoisomerases and recombinases). These experiments are often performed on circular DNA molecules, in which changes in the geometric (curvature, whirling, twisting and

supercoiling) or topological (knotting and linking) state of DNA can be directly observed.

The link between the structure of the DNA double-helix and some differential geometrical concepts appear very highlighting in the ‘White’s formula’ (J. White [109]) relating the linking, twisting and writing properties of a space curve. In order to make clear the meaning of this fruitful relationship between geometry and biology, let’s start with a rigorous formulation of the ‘Jordan Curve Theorem’, which constitutes a mathematical prerequisite of White’s formula (for further mathematical details, see Massey [78]). It is well-known that a simple, closed, continuous (or if you like smooth, or piecewise smooth, or even piecewise linear) curve separates the plane R^2 into two parts with the property that it is impossible to get from one part to the other by means of a continuous path avoiding the given curve. The same conclusion (as for a simple, closed, continuous curve) holds for any ‘complete’ curve in R^2 , i.e. a simple, continuous, unboundedly extended, non-closed curve both of those ends go off to infinity, without nontrivial limit points in the finite plane. This principle generalises in the obvious way to n -dimensional space: a closed hypersurface in R^n separates it into two parts.

There is however another less obvious generalisation of this principle, having its most familiar manifestation in three-dimensional space R^3 . Consider two continuous (or smooth) simple closed curves (loops) in R^3 which do not intersect:

$$\begin{aligned}\gamma_1(t) &= (x_{11}(t), x_{21}(t), x_{31}(t)), & \gamma_1(t + 2\pi) &= \gamma_1(t) \\ \gamma_2(\tau) &= (x_{12}(t), x_{22}(t), x_{33}(t)), & \gamma_2(t + 2\pi) &= \gamma_2(t).\end{aligned}\quad (2)$$

Consider a ‘singular disc’ D_i bounded by the curve γ_i , i.e. a continuous map of the unit disc into R^3 : $x_i^\alpha(r, \alpha)$, $i = 1, 2$, $\alpha = 1, 2, 3$, where $0 \leq r \leq 1$, $0 \leq \phi \leq 2\pi$, sending the boundary of the unit disc onto γ_i :

$$x_i^\alpha(r, \phi)|_{r=1} = x_i^\alpha(\phi), \quad \alpha = 1, 2, 3 \quad (3)$$

where $\phi = t$ for $i = 1$, and $\phi = \tau$ for $i = 2$. So we have the following

Definition 2.1. Two curves γ_1 and γ_2 in R^3 are said to be *nontrivially linked* if the curve γ_2 meets every singular disc D_1 with boundary γ_1 (or, equivalently, if the curve γ_1 meets every singular disc D_2 with boundary γ_2).

Some examples are shown in Figure 2. In n -dimensional space R^n certain pairs of closed surfaces may be linked, namely sub-manifolds of dimensions p and q where $p + q = n - 1$. In particular a closed curve in R^2 may be linked with a pair of points (a “zero-dimensional surface”) — this is just the original principle that a simple closed curve separates the plane.

Gauss introduced an invariant of a link consisting of two simple closed curves γ_1, γ_2 in R^3 , namely the signed number of turns of one of the curves around the

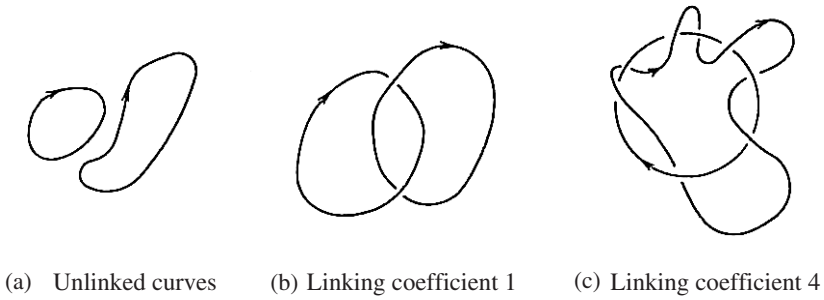


Figure 2

other, *the linking coefficient or linking number* $\{\gamma_1, \gamma_2\}$ of the link. His formula for this is

$$\begin{aligned}
 N &= \{\gamma_1, \gamma_2\} \\
 &= 1/4\pi \int_{\gamma_1} \int_{\gamma_2} ([d\gamma_1(t), d\gamma_2(t)], \gamma_1 - \gamma_2) / |\gamma_1(t) - \gamma_2(t)|^3, \quad (4)
 \end{aligned}$$

where $[,]$ denotes the vector (or cross) product of vectors in R^3 and $(,)$ the Euclidean scalar product. Thus this integral always has an integer value N . If we take one of the curves to be the z -axis in R^3 and the other to lie in the (x, y) -plane, then the previous formula (4) gives the net number of turns of the plane curve around the z -axis. It is interesting to note that the linking coefficient N may be zero even though the curves are nontrivially linked (Figure 3). Thus its having non-zero value represents only a sufficient condition for nontrivial linkage of the loops.

Let's now explain the White's formula (we follow closely L.H. Kauffman [61]). Let C be a space curve with a unit normal framing v, v^\perp and unit tangent t

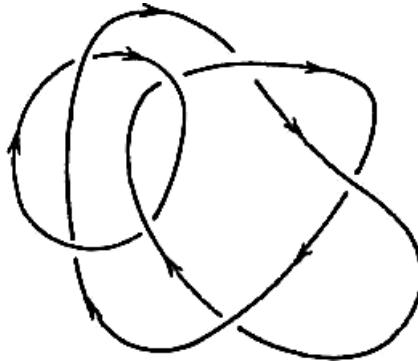


Figure 3 The linking coefficient = 0, yet the curves are non-trivially linked.

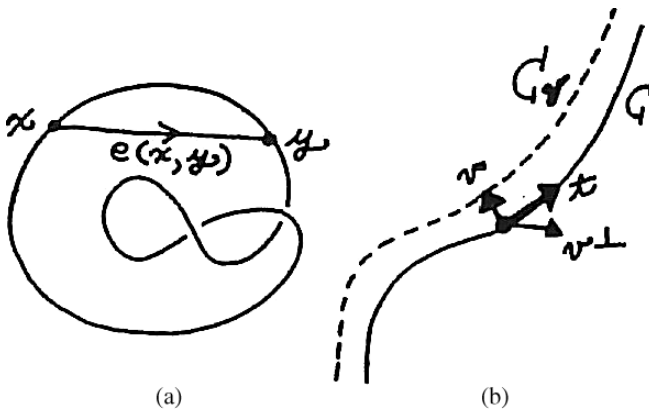


Figure 4

(v and v^\perp are perpendicular to each other and to t , forming a differentiable varying frame, $\langle v, v^\perp, t \rangle$, at each point of C .) Let C_v be the curve traced out by the tip of ϵv for $0 < \epsilon \ll 1$. Let $Lk = Lk(C, C_v)$ be the ‘linking number’ of C with this displacement C_v . Define the *total twist*, Tw , of the framed curve C by the formula

$$Tw = 1/2\pi \int v^\perp \cdot dv. \tag{5}$$

Given $(x, y) \in C \times C$, let $e(x, y) = (y - x)/|y - x|$ for $x \neq y$ and note that $e(x, y) \rightarrow t/|t|$ (for t the unit tangent vector to C at x) as x approaches y . This makes e well-defined on all of $C \times C$. Thus we have $e: C \times C \rightarrow S^2$. Let $d\Sigma$ denote the area element on S^2 and define the (spatial) *writhe* of the curve C by the formula

$$Wr = 1/4\pi \int_{C \times C} e^* d\Sigma = 1/4\pi \int_{z \in S^2} Cr(z) dz. \tag{6}$$

Here $Cr(z) = \sum_{p \in e^{-1}(z)} J(p)$ where $J(p) = \pm 1$ according to the sign of the Jacobian of e . It is easy to see, from this description, that the writhe coincides with the flat writhe (sum of crossing signs) for a curve that is (like a knot diagram) nearly embedded in a single plane.

With these definitions, White’s theorem reads

$$Lk = Tw + Wr.$$

This equation is fully valid for differentiable curves in three-dimensional space. Note that the writhe only depends upon the curve itself. It is independent of the framing. By combining two quantities (twist and writhe) that depend upon metric

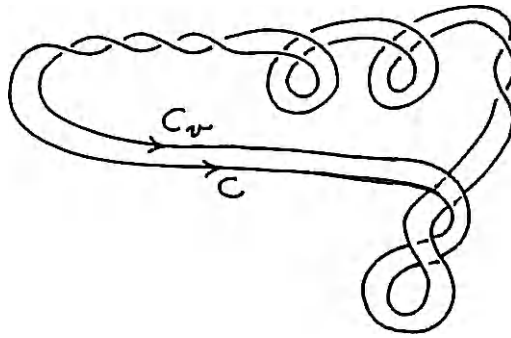


Figure 5



Figure 6

considerations, we obtain the linking number — a topological invariant of the pair (C, C_v) .

The planar version of White's theorem is worth discussing. Here we have C and C_v forming a pair of parallel curves as in the Figure 5. The twisting occurs between two curves, and is calculated as the sum of $\pm(1/2)$ for each crossing of one curve with the other, in the form as pictured in Figure 6.

The linking number is a mathematical quantity existing in two, three and also higher dimensions, topologically invariant by deformations, which tells us a great deal about the structural properties and qualitative behaviours of DNA during the cell cycle. First, it is closely related to the number of time that the two sugar-phosphate chains of DNA wrap around, or are 'linked with', one another. Here take DNA in its stress-free, relaxed state as the reference point for counting Lk , where hence $Lk = 0$. Now consider the simple model of a circular DNA with the values: $Tw = +3$, $Wr = 0$, $Lk = +3$. Thus, $Lk = +3$ tells us that the DNA has three more double-helical turns than it would have in a relaxed, open-circular form. In general, Lk measures the total excess or deficit of double-helix turns in the molecule. Note, in particular, that Lk can only be an integer, because the DNA can only join to itself by some integral number of turns. Before we pursue (in section 4) the analysis of the importance of the linking number and its relationship with the supercoiling process, we shall make few remarks about the properties of enzymes topoisomerases.

3. The Topological Role of Topoisomerases

Enzymes topoisomerases, which change the linking number of the DNA strands, appear to have a profound role in the central genetic events of DNA replication, transcription and recombination. It is a long-standing problem in biology to understand the mechanisms responsible for the knotting and unknotting of DNA molecules. Large amounts of DNA are wound up and packed into the average cell. In fact, there is enough DNA in a two-metre human body to stretch from the earth to the sun and back fifty times! This of course means that the embedding of the DNA in the cell is exceedingly complicated. The DNA in the cell knots and unknots, ties and unties itself according to a definite scheme. Knots and links appear during replication and recombination. Certain topoisomerases, which behave like topological entities in living organisms, are responsible for the knotting and unknotting. They are able to cut a strand of DNA at a particular point, grasp another strand, pass it through the opening and then close the opening. In other words, these enzymes replace over-crossing by under-crossing.

Consequently, the tying of knots in rings of DNA is one of the capabilities of these enzymes. The genetic material of many organisms has the form of a ring made up either of one strand of DNA or of two strands twisted in a double helix. The ring can assume a number of topological configurations. The conversion of the DNA ring from one configuration to another is catalyzed by the topoisomerases. Consider, for example, a single-strand DNA rings from a virus known as bacteriophage, which infects bacteria (Figure 7a). What one observes of the rings, after they were exposed a topoisomerase from the bacterium *Escherichia coli*, is then that, by cutting the DNA strand, passing a segment of the ring through the break and rejoining the cut ends, the enzyme has tied a knot in each ring (Figure 7b). In fact, the process of breaking, passage and resealing is essential to the action of all topoisomerases. Some of the enzymes, designated type I, cut a single strand of DNA; others, designated type II, cut both strands of a double helix.

4. The Relationship between the Linking Number and Supercoiling of DNA Molecule

Supercoiling of a double-strand DNA ring deforms the ring into a more twisted and compact shape. The shape of a DNA ring is strongly affected by the number of times one strand goes around the other; the quantity is called the *linking number* Lk (see the previous section). This is a topological quantity, hence, it cannot be altered while the strands are intact regardless of how the ring is pulled or twisted. If the strands are cut, however, and then rotated about each other in the direction opposite to that of the twist of the helix, the helix unwinds. When the cut ends are rejoined, the number of rotations that have been made decreases the linking

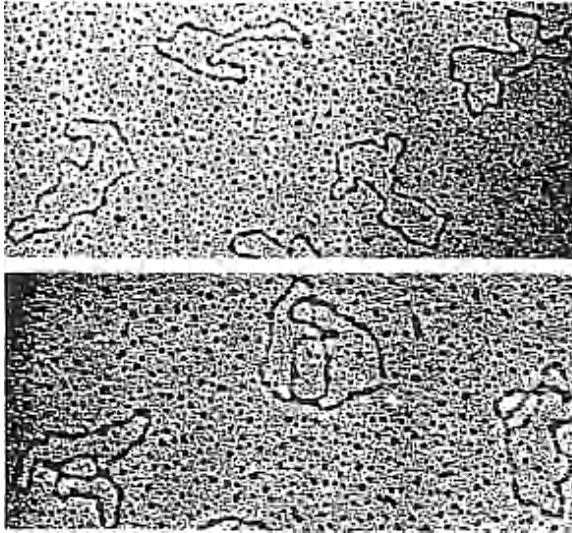


Figure 7 (a) Top and (b) bottom. The knotting of DNA rings after they were exposed a topoisomerase bacterium *Escherichia Coli*.

number. The strands of DNA in a linear molecule revolve once every 10.5 base pairs because that configuration puts the least strain on the double helix. A ring in which the ratio of base pairs to linking number is 10.5 is said to be relaxed. Increasing or decreasing the ratio strains the double helix, which responds by supercoiling. Reducing the linking number causes negative supercoiling; raising the linking number leads to positive supercoiling. The upper electron micrograph shows a relaxed DNA ring from a bacterial virus called *PM2* (Figure 8, top). The micrograph below shows a negatively supercoiled DNA ring from a bacterial virus called *PM2* (Figure 8, bottom).

We already said that, essentially, a molecule of DNA may be thought of as two linear strands intertwined in the form of a double helix with a linear axis. A molecule of DNA may also take the form of a ring, and so it can become tangled or knotted. Further, a piece of DNA can break temporarily. While in this broken state the structure of the DNA may undergo a physical change, and the DNA will finally recombine. In fact a single enzyme, a topoisomerase of type I, can facilitate this whole process, from the original splicing to the recombination. The process of recombination involves some interesting topological changes in the substrate. It is worth noting that knowledge of the topology of the substrate and product(s) can be used to compute the Jones polynomials of other products (see Murasugi [81] and Kauffman [61]). For instance, a cut in a double-strand DNA, due to a topoisomerase, allows a double-strand DNA to pass through it and recombine.

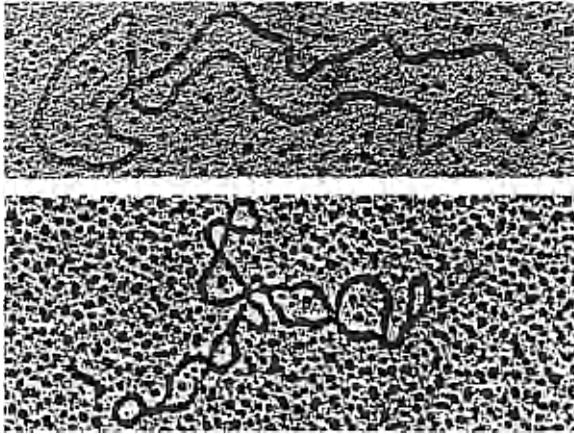


Figure 8

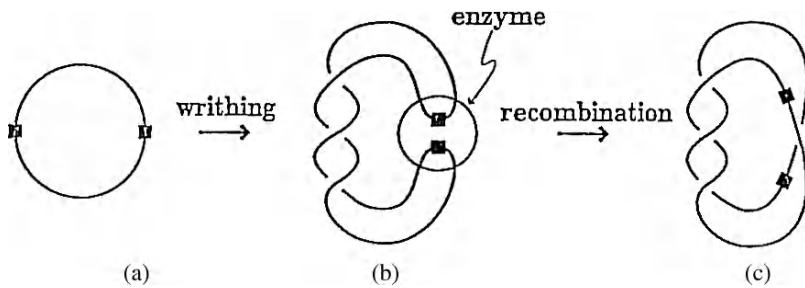


Figure 9 The writhing process of the DNA molecule.

Finding such a topoisomerase is relatively straightforward, since these enzymes occur in organisms small and large, from bacteria to the body of the reader of this article. The effect of a certain topoisomerase (called a *recombinase*) is usually called a *site-specific recombination*. Before the action of the recombinase, the DNA molecule is called a *substrate*; after the recombination it is called a *product*. The process of going from the DNA molecule to a state in which two parts of the DNA molecule have been drawn together, is said to be the *writhing process*.

The double-helix structure of DNA is a geometrical entity, or more precisely, a topological configuration. This topological configuration is itself a manifestation of linking or knotting. Further, it has been shown that when a topoisomerase causes DNA to change its form, the process is very similar to what happens locally in the skein diagrams. Therefore, for the geometrical entity — knotted or linked — the linking number is an important concept, while the action of the topoisomerase

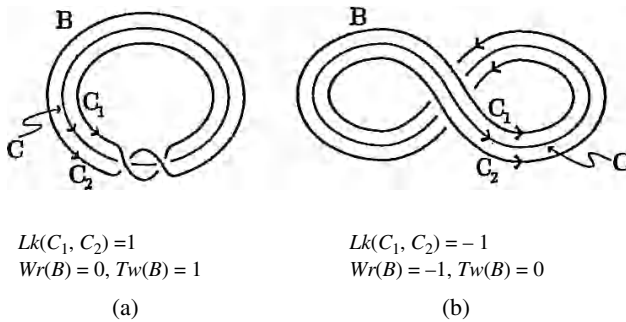


Figure 10 (a) and (b). Supercoiling process in the DNA molecule.

is related to the new skein invariants. In fact, the linking number Lk between C_1 and C_2 (where C_1 and C_2 are the two backbone curves that form the boundaries of the ribbon B and that represent the closed DNA strands) is an invariant, and its changes have a very important effect on the structure of the DNA molecule. For example, it is known that if we reduce the linking number of a double-strand DNA molecule, the DNA molecule will twist and coil, that is, it will supercoil. In other words, the linking number difference is a measure of supercoiling process (Figure 10).

Interestingly enough, in the case of a link formed from C_1 and C_2 , the linking number defined on the DNA molecule in biology, and the linking number computed from the mathematical knot theory turns out to be the same. Moreover, the untying mechanisms used in cells bear an uncanny resemblance to the simplest mathematical method for generating the new polynomial invariants. The number of twists of the ribbon B along the axis C is called the *twisting number*, and is denoted by $Tw(B)$. The *writhe*, $Wr(B)$, can be defined as the *average value of the sum of the signs of the crossing points, averaged over all projections*. These numbers, $Wr(B)$ and $Tw(B)$ are invariants. They are not, however, invariants of the knot (or link) obtained from the DNA molecule, but differential geometrical invariants of the ribbon B as a surface in space. The three invariants mentioned above are related by the following basic formula:

$$Lk(C_1, C_2) = Tw(B) + Wr(B).$$

The application of this expression to DNA molecules is an explanation of its propensity to supercoil. The remarkable fact about this result is that two geometric quantities that may change under deformations of the curves add up to a topological quantity, which is invariant under such deformations. Moreover, the linking number has a very important topological property: it is unchanged under any continuous deformation of the pair of curves, no matter how the double-strand ring is pulled or twisted, so long as the two strands remain unbroken. Topological

properties of DNA are defined as those that can be changed only by the breaking and rejoining of the backbone. There are three important topological properties of DNA: (1) the linking number between the strands of the double helix, (2) the interlocking of separate DNA rings into what are called catenanes, (3) and knotting.

Note that as the number of crossings in a knot or catenane increases, the number of possible isomers grows exponentially. The linking number of DNA in all organisms is less than the energetically most stable value in unconstrained (relaxed) DNA. This puts the DNA under mechanical stress that causes it to buckle and coil in a regular way called negative *supercoiling*. The name supercoiling derives from the fact that it is the coiling of a molecule that is itself formed by the coiling of two strands around each other. Although, strictly speaking, supercoiling is a geometric property, it is a consequence of a topological one, the linking number difference between supercoiled and relaxed DNA.

4.1. Topological complexity of DNA and its biological meaning

Let's start this section by emphasising an important point, namely, that the complex topology of DNA is essential for the life of all organisms. In particular, it is needed for the process known as DNA replication, whereby a replica of the DNA is made and one copy is passed on to each daughter cell. The most direct evidence for the vital role played by DNA topology is provided by the results of attempts to change the topology of DNA inside cells. Two related questions arise immediately from the recognition that DNA topology is essential for life: How did the complex topology of DNA evolve, and why is it so important for cells? DNA is the only molecule in cells that has a complex topology.

The evolution of proteins has taken a different course. Proteins also naturally subdivide into domains and thus local knots or links could readily occur, but they do rarely, although different types of pseudo-knots have been recently observed in proteins patterns. Besides, no knots, catenanes, or supercoiling have been found so far in RNA, polysaccharides, or lipids. Type I topoisomerases of the DNA molecule, which cut one strand at a time, can carry out several topological operations. By cutting one strand of a supercoiled DNA ring the type I enzyme can put the ring into the relaxed state. It can tie a single-strand ring into a knot. The knot is tied when the simple-strand ring crosses over itself. If the two loops formed in this way are pulled together, the enzyme can cut one loop and pass the other loop through the opening. When the break is sealed, the ring is sealed in a knot. The type I enzyme can also interlock two single-strand ring. If the rings have complementary base sequences, a double-helix results. Although the operations seem quite different, each requires that a strand be broken, a segment of DNA be passed through the break and the break be resealed.

With the evolution of type I topoisomerases, compaction by nucleosomes could occur and the size of DNA could grow to about 10^5 kb. However, as DNA grew in length, the problems of accidental knotting within domains and catenation of separate domains and segregation of the products of DNA replication became acute. These problems were solved by the evolution of the type II topoisomerases, which promote the passage of duplex DNA through transient double strand breaks. A type II topoisomerase could have evolved from a type I topoisomerase by the development of an interaction between two copies of a type I enzyme. Further increases in DNA size required only the evolution of successively higher orders of DNA compaction. So we are faced, once again, with a genuine topological problem, which we shall address in the next sections.

4.2. The structural flexibility of biomolecules. DNA compaction by successive order of coiling

It must be stressed that, essentially thanks to its topological properties, DNA is a very malleable, deformable molecule, being able to recombine through a series of stages. Very likely this property of flexibility or deformability is one of the most important properties the DNA molecule, which also might distinguish the living matter from the inanimate one. Moreover, this flexibility influences in a fundamental way the biological functions of the double-helix. In fact, the molecule can freely move about, although under certain chemical and geometrical constraints, in the space of the cell's nucleus and transform itself into several shapes without losing a certain structural stability and energetic optimal state. This movement is twofold: the three-dimensional two-strands helical structure of DNA molecule can extend and compact.

(i) The extended (unfolded) conformation of DNA, which put it under tension as if the molecule was subjected to shear (cut off) one dynamic force, seems to be especially required for DNA replication. By this process, each of the two strands of DNA is used as a template for the formation of a complementary DNA strand. The original strands therefore may remain intact through many cell generations.

(ii) DNA compaction inside cells occurs by successive orders of *coiling*. One can show that a DNA double helix is compacted in about four successive steps. Only the first step of nucleosome formation is quite well understood. In this step, DNA coils twice in a left-handed helical fashion around a set of proteins called histones. The nucleosomes are then coiled successively to give the final forms, called a chromosome. In the phases of this process (the recombination), the knot type of DNA molecule is actually changed. The whole process, from the original splicing to the recombination, is the result of the effect of a single enzyme/catalyst called a topoisomerase. To be more precise, all these nuclear processes that occur during an entire cell cycle need to be properly and continuously orchestrate by

a family of topoisomerases each of one have a specific task although severely interconnected.

5. More about Topoisomerases and their Mathematical Abilities and Biological Functions

The term topoisomerase is relatively easy to explain. Chemically, two molecules with same chemical composition but different structures are called *isomers*. It follows that two DNA molecules with the same sequence of base pairs but different linking numbers are also isomers. Due to the difference in linking numbers, “topologically” they are inequivalent. In other words, topoisomerases are those enzymes that cause the linking number to change.

Topoisomerases are essential to allowing DNA replication. Once replication is completed, the newly synthesised molecule must be disentangled from its parent. The replication of circular DNA molecules gives rise to two linked circular molecules, but the replication of whole chromosomes leaves the cell with highly entangled chromatids. If the cell does not disentangle the freshly replicated pairs of sister chromatids, they will fragment under the pull of the mitotic spindle. Disentanglement is achieved thanks to topoisomerases. Thus, one can say that topoisomerases are the cell’s tools for managing the topologies of their genomes. This means in particular that the relation between the topological form and the biological function of DNA molecule must be at the core of the nuclear organisation of cell.

The process of mutation due to topoisomerases can be described in simple terms as follow: First a strand of the DNA is cut at one place, then a segment of DNA passes through this cut, and finally the DNA reconnects itself. So surgery cutting and self-recognition are always two consecutive closed related operations. There exist two examples of the action of a topoisomerase on a DNA molecule (see Wang [106] and Rocca [92] for a detailed account):

(1) The simple strand has a single cut due to a topoisomerase and the DNA passes through it and recombines; this is called a type I topoisomerases. However, topoisomerase I is also capable of tying complex knot; in fact, the enzyme produces every knot theoretically possible. Thus, the requirement for excess enzyme to form complex knots suggests a role for topoisomerase I in contorting the DNA in addition to promoting strand passage. For example, it has been shown that *E. coli* topoisomerase I can catenate circles in nicked, duplex DNA. Furthermore, it is able to tie a knot by inverting either a (+)- or a (-)-node during strand passage. In other words, the DNA must fold to provide an inversion node, at which topoisomerase I passes one DNA segment through a transiently introduced break in the other DNA strand and thereby inverts the topological sign of the node. The inversion node divides the ring into two domains, and there must be at least two nodes between

these two domains. Intradomain nodes alone never lead to knots, although they can contribute. Even a complicated knot can be tied by just one strand passage.

(2) A cut in a double-strand DNA, due again to a topoisomerase, allows a double-strand DNA to pass through it and recombine, this is as expected called a type II topoisomerases. In other words, type II topoisomerases are essential enzymes that pass one DNA through another and thereby remove DNA entanglements. They make a transient double-stranded break in a gate segment (*G* segment) that allows passage by another segment (*T* segment) of the same or another DNA molecule. Thus, these enzymes have the potential to convert real DNA molecules into phantom chains that freely pass through themselves to generate an equilibrium distribution of knots, catenanes, and supercoils. In fact, they reduce the fraction of knotted and catenated circular DNA below thermodynamic equilibrium values. To do that, enzymes use the energy of ATP hydrolysis. Active topology simplification by topoisomerases II has an important biological consequence. It helps explain how topoisomerases can remove all DNA entanglements under the crowded cellular conditions that favour the opposite outcome. Recently, a new model was designed to explain this surprising finding, in which eukaryotic and prokaryotic topoisomerases bend DNA sharply upon binding. In this model, bending is a local, geometrical manipulation that varies the curvature of the molecule's strand site, which is responsible for the change of the global, topological configuration of circular DNA. The challenge, however, is to ask oneself whether there may be some kind of global process which could be responsible for the topological compaction of chromatin in the chromosome. In the next section, we give some hints on this problem.

6. Tangles, Knotting, and DNA Recombination: the Close Link between Topological 'Information' Acting on Supramolecular Forms and Biological Processes

Thus, as we just saw, a molecule of DNA may also take the form of a ring, and so it can become tangled or knotted. Further, a piece of DNA can break temporarily. While in this broken state the structure of the DNA may undergo a physical change, and the DNA will finally recombine. A single enzyme can facilitate this whole process, from the original splicing to the recombination. The process of recombination involves some interesting topological changes in the substrate. It is worth noting that knowledge of the topology of the substrate and product(s) can be used to compute the Jones polynomials of other products. For instance, a cut in a double-strand DNA, due to a topoisomerase, allows a double-strand DNA to pass through it and recombine. Finding such a topoisomerase is relatively straightforward, since these enzymes occur in organisms small and large, from bacteria to complex organisms. The effect of a certain topoisomerase (called *recombinase*) is

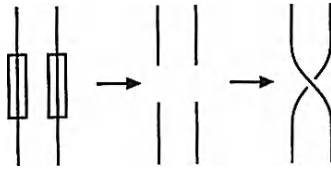


Figure 11 A possible site-specific recombination.

usually called a *site-specific recombination*, which is a process whereby an enzyme attaches to two specific sites on two strands of DNA, called recombination sites, each of which corresponds to a particular sequence of base pairs that the enzyme recognises. After lining the sites up, the enzyme cuts the two strands open and recombine the four ends in some manner. In Figure 11, we show one of the simplest actions.

Let us give some more technical details. A site-specific recombination is a local operation (see Sumners [100]). The effect of the recombinase on a DNA molecule is either to move a piece of this DNA molecule to another position within itself or to import a foreign piece of DNA molecule into it. The result is that the gene transmutes itself. The exact process of a site-specific recombination is fairly easy to understand. Firstly, two points of the same or different DNA molecules are drawn together, either by a recombinase or by a random (thermal) motion, or even possibly both. The recombinase then sets to work, causing the DNA molecule to be cut open at two points on the parts that have been drawn together. The loose ends are then recombined by the recombinase in a different combination than the original DNA molecule. Before the action of the recombinase, the DNA molecule is called a *substrate*; after the recombination it is called a *product*. The product can be a knot, an unknot, or a two-component link. The process of going from the DNA molecule to a state in which two parts of the DNA molecule have been drawn together, is said to be the *writhing process*. When at this stage the recombinase combines with the substrate, the resultant combined complex is called a *synaptic complex* (see Figure 13). Within the synaptic complex, we can assign local orientation to the respective, relatively small part of the DNA molecule (or molecules) on which the recombinase acts within a circle.

A few words about some notation we have just used are in order. First let us define mathematically what the object *tangle* is.¹ On the sphere S^2 — the

¹The British mathematician John H. Conway introduced the concept of a tangle at the beginning of the 1970s, in its attempt to give a complete table of knots [1970]. Using this variation on a knot, a new class of knots could be defined: algebraic knots. By studying this class of knots, various local problems were able to be solved, which led to a further jump in the level of understanding of knot theory. However, since there are knots that are not algebraic, the complete classification of knots could not be realised. Nevertheless, the introduction of this new research approach has had a significant impact on knot theory,

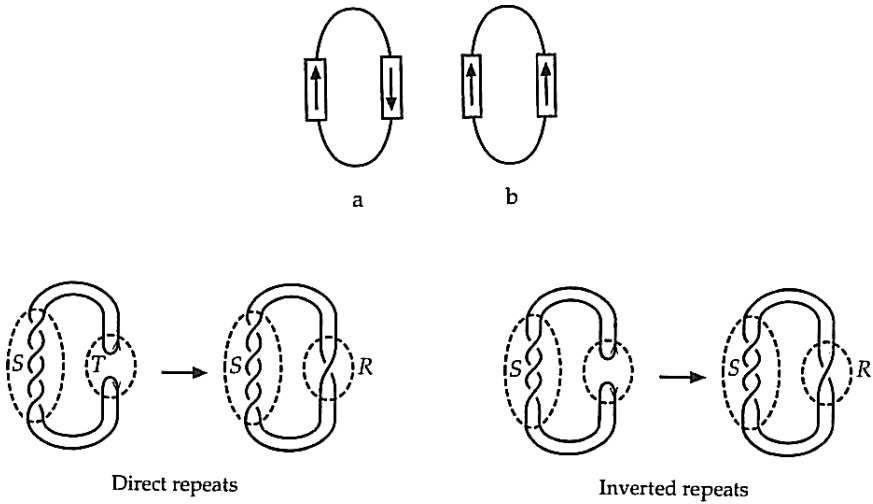


Figure 12 (a) *Direct repeats* and (b) *inverted repeats*. Suppose we have a single circular DNA molecule that contains a copy of each of the two recombination sites necessary for the reaction. Then, when the enzyme acts on this molecule, the result can be analysed to determine the effect of the enzyme. We can choose an orientation for the site. When a pair of sites is utilised in an enzyme action, we pick the orientations of the two sites so that they will match when the enzyme pulls the two sites together. When both sites appear on the same circular DNA molecule, these orientations can either point in the same direction as we traverse the molecule, in which case we say that the two have *direct repeats*, or their orientation can point in opposite directions as we traverse the molecule, this case being known as *inverted repeats*.

surface (boundary) of the three-ball B^3 — place $2n$ points. A (n, n) -tangle T is formed by attaching, within B^3 , to these points n curves, none of which would intersect each other, as illustrated in the Figure 14. (Note that the curves should be polygonal.) Suppose that we fix four points on the sphere S^2 — say, north-east, north-west, south-east, south-west — to which we attach their coordinates that lie in the yz -plane. By attaching the end points of two polygonal curves in B^3 to these four points, we can form a tangle. So, if we project this tangle onto the yz -plane, as in the case of a knot, we have what may be called a *regular diagram* of the tangle (Figure 14(f)). The knot (or link) obtained by connecting the points north-west and north-east, south-west and south-east by simple curves outside B^3 is called the *numerator* and is denoted by $N(T)$. Similarly, we may connect the points north-west and south-west, north-east and south-east by simple curves outside B^3 , and

particularly on the investigation of 2-bridge knots (or links), which are a special kind of algebraic knot obtained from trivial tangles.

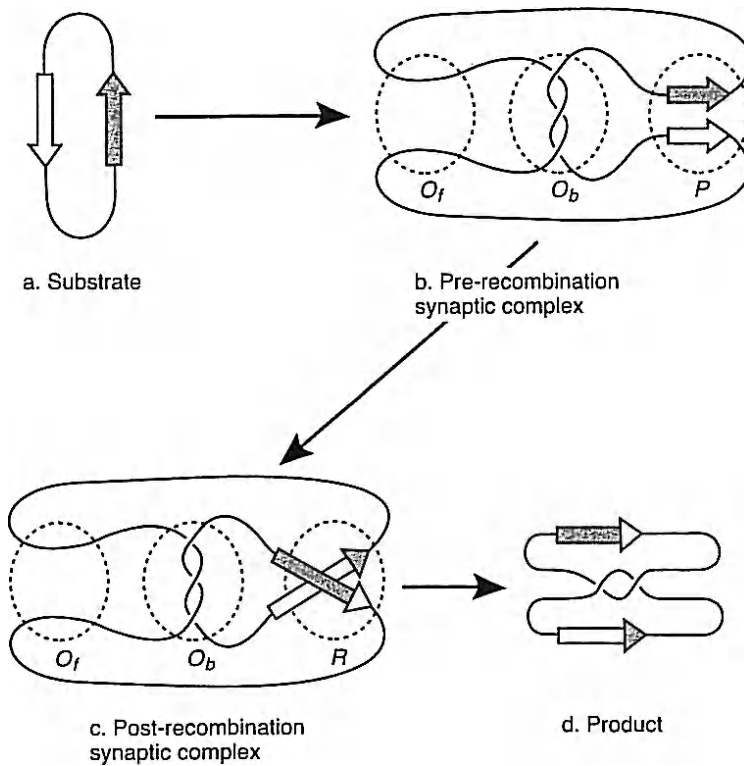


Figure 13 (Adapted from Adams [2000, 1] and Sumners [1992, 100]). Recombination with inverted repeats. Steps of the synaptic complex: (a) the *substrate*; (b) the *pre-recombination synaptic complex*; here S denotes s the *substrate tangle*, which is unchanged by the enzyme, and T stand for the *site tangle*, where the enzyme acts; (c) the *post-recombination synaptic process*, thereby the enzyme replaces the site tangle T with the *recombination tangle* R ; (d) the *product* of the recombination, which can be either a knot or a link; according to the above notation, its formula is $N(T + R)$, where T and R are enzymes determined constants independent of the variable geometry of the substrate S .

the subsequent knot (or link) is called the *denominator* and is denoted by $D(T)$. (For further details on tangle theory, we refer the reader to Sumners [100] and Murasugi [81]).

Let $N(Q)$ denote the knot or link obtained by connecting the top two strands of a *rational tangle* Q to each other and the bottom two strands of Q to each other. Let $Q + V$ denotes the rational tangle obtained by adding the two tangles Q and V together. In this notation, the facts that the substrate comes from the tangles S and T and the product from the tangles T and R can be written in two equations

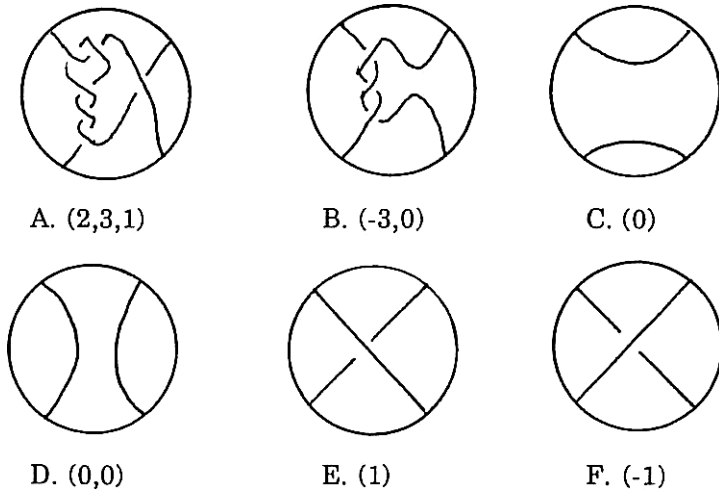


Figure 14

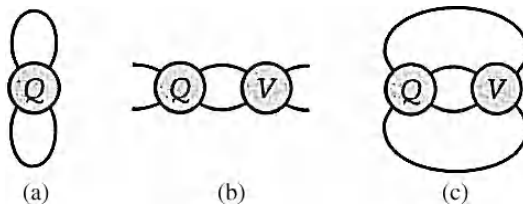


Figure 15 (a) $N(Q)$. (b) $Q + V$. (c) $N(Q + V)$.

in the three unknowns, S , T , and R :

$$N(S + T) = \text{substrate}$$

$$N(T + R) = \text{product.}$$

Since we have more variables than we have equations, we can never hope to determine all three of S , T , and R from knowing the knotting of the substrate and the product. If we happen to know one of the three, however, we should be able to determine the other two.

The rational tangles are characterised topologically by values in the extended rational numbers $\mathbf{Q}^* = \mathbf{Q} \cup \{1/0 = \infty\}$. An element in \mathbf{Q}^* has the form β/α where $\alpha \in \mathbf{N} \cup \{0\}$, (\mathbf{N} is the natural numbers), and $\beta \in \mathbf{Z}$ with $gcd(\alpha, \beta) = 1$. Rational tangles themselves are obtained by iterating operations similar to the recombination process itself. The inverse of a tangle is obtained by turning it 180°

around the left-top to right-bottom diagonal axis. Rational numbers correspond to tangles via the continued fraction expansion. Since two rational tangles are topologically equivalent if and only if they receive the same fraction in \mathbf{Q}^* , it is likely to calculate possibilities for site-specific recombination in this category. Here we have an arena in which molecular enzymes-driven manipulations, knot theoretic operations and the biologically relevant topological information carried out by a knot or link act in a cooperative manner. This brings us directly to the central question of this study: what is the nature of the topological information carried out by a knot or link? For biology this information manifests itself in the dynamics of a recombinant process, or in the organisation of the constituents of a cell, as we shall explain in the next section dedicated to the problem of chromatin folding and supercoiling.

According to the above description, the following mathematical propositions (or results) follows:

Proposition 6.1. *Almost all the products obtained by the site-specific recombination of trivial knots substrates are rational knots (or links), i.e., two-bridges knots (or links).*

Proposition 6.2. *The part of the synaptic complex acted on by an enzyme (recombinase), mathematically within the three-ball, is a (2,2)-tangle.*

Therefore, the product is just the replacement of one (2,2)-tangle by another (2,2)-tangle. Thus, for example, a (2,2)-tangle within the circle T may be replaced by a tangle R to form a product. Mathematically, it is perfectly reasonable to consider S to be a (2,2)-tangle in T . The numerator of the sum of S and R is then the product. So the following “equation” holds: $N(S + R) = P$ (the product). Further, we may divide the substrate into the external tangle S and the internal tangle E , since the substrate is the numerator of the sum of S and E . Again, we have a quasi-equation holding: $N(S + E) = S$ (the substrate).

There is an important mathematical assumption one can make, which is yet supported by biological observation. Namely, that tangles T and R do not depend on the tangle S . They depend only on the enzyme that is acting, and not on the knottedness of the molecule it acts on. One example of a topoisomerase is the enzyme *Tn3 resolvase* (see Benjamin *et al.* [14]). This enzyme acts on a particular duplex cyclic DNA molecule with direct repeats. Once its has matched up the two sites, it replaces the T tangle with a single R tangle and releases the molecule. Once in a while, however, it will repeat the tangle replacement a second time before releasing the molecule. Even more rarely, it can repeat the tangle replacement a number of times, yielding even more complicated molecules. In a series of experiments, biochemists established what products resulted when the enzyme acted, and determined the following equations, where we use the notation

for rational knots:

$$N(T + S) = N(1) \quad (\text{the unknot})$$

$$N(T + R) = N(2) \quad (\text{the Hopf link})$$

$$N(T + R + R) = N(211) \quad (\text{the figure-eight knot})$$

$$N(T + R + R + R) = N(11111) \quad (\text{the Whitehead link}).$$

From this set of equations, Summers [1992] proved that $S = (-3, 0)$ and $R = (1)$. Moreover, he proved that it should then be the case that $N(S + R + R + R + R) = N(12111)$ (the 6_2 knot). This last knot has been observed as a product in many recombination processes.

7. Condensation of the Double-Helix Molecule into the Chromatin, and the Role of Supercoiling

One of the most striking phenomena that reveal the profound interdependence between topological problems and biological processes is that of the compaction of chromatin into the chromosome within the cell nucleus. Its explanation is one of most challenging task of biology today. Here we are faced with a genuine problem of differential topology. What kind of deformations does the double-strands linear DNA molecule undergo in order that it condenses into an extremely compact form, corresponding to the metaphase of the chromosome? Though the answer to this question is far from being clear or complete, however, some aspects have been elucidated very recently.

- (1) The key distinguishing characteristic of the eukaryotic genome is its tight packaging into chromatin, a hierarchically organised complex of DNA and histone and non-histone proteins. How genome operates in the chromatin context is a central question in the molecular genetics of eukaryotes. The chromatin packaging consists of different levels of organisation. Every level of chromatin organisation, from nucleosome to higher-order structure up to its intranuclear localisation, can contribute to the regulation of gene expression, as well as affect other functions of the genome, such as replication and repair. Concerning gene expression, chromatin is important not only because of the accessibility problem it poses for the transcription apparatus, but also due to the phenomenon of chromatin memory, that is, the apparent ability of alternative chromatin states to be maintained through many cell divisions. This phenomenon is believed to be involved in the mechanism of epigenetic inheritance, an important concept of developmental biology.
- (2) Supercoiling is one of the three fundamental aspects of DNA compaction; the other two are conformational flexibility and intrinsic DNA curvature. For

example, the problem of DNA compaction in *E. coli* can be putted in the following words: the DNA must be compacted more than a thousand-fold in the cell, yet it still needs to be available to be transcribed. (Recall that the length of a typical bacterial operon — usually about three genes — is about as long as the entire bacterial cell, if it is stretched out in its B-DNA double-helical conformation!). In order this compaction to be achieved, some kind of anisotropic flexibility or ‘bendability’ of DNA, which is very much sequence-specific, and is different from the structural ‘rigidity’ of DNA, is required. Whereas *persistence length* of DNA is relatively non-specific, and just has to do with its overall ‘rigidity’ (on average, DNA has a persistence length of about 44 nm, which is quite a bit longer than proteins — one way to thinking about this is that proteins tend to fold up into little spheres, or ‘blobs’, and DNA is a bit more rigid), *anisotropic flexibility* is a measure of a particular sequence to be deformed by a protein (or some other external forces). Some sequences are both isotropically flexible and ‘bendable’ — for example, the TATA motifs. Perhaps one of the best examples of this is the binding site for the Integration Host Factor (IHF): there are certain base pairs that are highly distorted upon binding of this protein. It is quite impressive that this protein induces a bend of 180 degrees into a DNA helix. In other words, the curvature, say k , at each sequence of the two strands of DNA helix must be very sharp in order the DNA double helix may assume its extremely compact form. So the relationship between (geometric) curvature and conformational (or topological) flexibility appear to be crucial in the understanding of the biological activity of cells.

- (3) Indeed, when one consider that the DNA must be compacted more than a thousand fold in the cell, it is probably not surprising that almost any protein that binds to DNA will bend it. Moreover, since the total curvature K of an entire DNA double-helix segment depends on the torsional stress which applies to DNA strands, and, accordingly, these strands form a twisted curve, i.e., a curve of double curvature in the three-dimensional space of the cell nucleus, DNA double-helix must coil many times in a very ordered way to form chromatin structure; otherwise, if the chromosome of a human cell were in the form of a random coil, they would not fit within the nucleus. The DNA double helix coils first by overwinding or underwinding of the duplex. The supercoiled form of a circular DNA molecule is much more compact than the other possible conformations, i.e., nicked and linear. In its supercoiled form, DNA molecule minimises to the highest the space volume it occupies in the nucleus. Supercoils condense DNA and promote the disentanglement of topological domains.
- (4) Today we know that DNA is topologically polymorphic (see Lesliet *et al.* [69]). The overwound or underwound double-helix can assume exotic forms known as *plectonemes*, like the braided structures of a tangled telephone cord, or *solenoids*, similar to the winding of a magnetic coil. (i) Plectonemically supercoiled DNA is unrestrained and frequently branched, while toroidal

- supercoils is restrained by proteins and it is more compact. (ii) DNA can be either positive or negatively supercoiled. In particular, eukaryotic DNA is negatively supercoiled in and around genes, and it is transiently negatively supercoiled behind RNA polymerase during transcription. (iii) Negative supercoiling favours DNA-histone association and the formation of nucleosomes, the first step in packaging DNA. Because the solenoidal DNA wrapping around a nucleosome core creates about two negative supercoils, it is understandable that the DNA that fulfills this topological prerequisite will more easily form nucleosome. (iv) These tertiary structures have an important effect on the molecule's secondary structure and eventually its functions. For example, supercoiling induces destabilisation of certain DNA sequences and allows the extrusion of cruciform or even the transcriptional activation of eukaryotic promoters. Another essential process, DNA transcription, can both generate and be regulated by supercoiling.
- (5) During replication, the chromosomes need to be partitioned and the two strands of DNA must be continuously unlinked during replication. The topoisomerases that accomplish this might instead be expected to entangle and knot chromosomes because of the huge DNA concentration *in vivo*. There are actually several factors that solve this problem and contribute to the orderly unlinking of DNA. A major contributor to chromosome partitioning is the condensation of daughter DNA upon itself soon after replication. DNA condensation is due primarily to supercoiling. Another factor promoting chromosome partitioning is that the type II topoisomerases of all organisms do not just speed up the approach to topological equilibrium, but actually change the equilibrium position. They actively remove all DNA entanglements. This requires that topoisomerases sense the global conformation of DNA even though they interact with DNA only locally. In fact, topoisomerases achieve this because, by positioning themselves at sharp bends in DNA, they carry out net disentanglement of DNA (they act, in a way, like Maxwell's demon). An equal partner to the topoisomerases in chromosome segregation is the helicases. They seem to convert the energy of ATP hydrolysis into unwinding DNA. All the enzymes that play critical roles in DNA unlinking and chromosome segregation, topoisomerases, helicases, and condensins, are motor proteins. They use the energy of ATP hydrolysis to move large pieces of DNA over long distances.
- (6) The previous discussion can be summed up by saying that supercoiling has three essential roles. (i) First, (–) supercoiling promotes the unwinding of DNA and thereby the myriad processes that depends on helix opening. (ii) The second essential role of supercoiling is in DNA replication. For replication to be completed, the linking number of the DNA, Lk , must be reduced from its vast (+) value to exactly zero. In bacteria, DNA gyrase introduces (–) supercoils and thereby removes parental Lk . (iii) The third essential role of supercoiling is conformational. DNA manifests the difference between the relaxed and naturally

occurring values of Lk by winding up into supercoils. These supercoils condense DNA and promote the disentanglement of topological domains. This can be accomplished equally well by (-) or (+) supercoiling. Let us still underline two important facts. First, the promotion of decatenation by supercoiling has also been directly demonstrated *in vivo*. Second, the volume occupied by a supercoiled molecule is much more smaller than that of a relaxed DNA. This difference in volume is due mostly to the formation of superhelical branches. Indeed, supercoiled DNA branches and bends itself into a ball. The decrease in chromosomal volume by supercoiling reduce the probability that the septum will pass through the chromosome during cell division.

8. Topological Models for Chromosome Compaction; the Mathematical Concepts of 'Linking Number', 'Twist', 'Writhe', and their Biological Meaning

It seems clear that supercoiling play a fundamental role in the condensation of the double helix and that this condensation is responsible for DNA unlinking and chromosome partitioning. Supercoiling results from topological strain and the contortion of DNA by proteins, notably the nucleosomal histone octet and the structural maintenance of chromosomes (SMC) proteins. There are three ways, actually experimentally observed *in vivo*, in which condensation of chromosome by supercoiling occurs, and to each of them corresponds a topological model for explaining the compaction of chromosomes in the cell's nucleus.

- (i) (-) Supercoiling by gyrase compacts the chromosomes such that random passages by topoisomerase IV disentangle them. In particular, topoisomerase IV is responsible for decatenation of DNA.
- (ii) With the second type of condensation *via* supercoiling, that is by core histones. DNA is compacted in independent successive stages such that the total compaction is the product of compaction in each stage. The first stage of this compaction is *via* solenoidal wrapping of DNA in the nucleosome. Although the compaction achieved is modest, the nucleosome provides a fundamental structure for genome organisation and function. The structure of a nucleosome reveals a scaffolding that forces the DNA to adopt ordered solenoidal supercoils.
- (iii) The third type of compaction *cum* supercoiling, that by condensin, is needed for the formation of mitotic chromosomes from the open interphase forms.

Recently, it has been experimentally demonstrated that condensin was required for both the assembly and maintenance of these chromosomes. It is very worth of note that a mere *local* interpretation of these results isn't a satisfactory explanation, because a local overwinding of DNA would have no effect on condensation; nor

could a tight wrapping around condensin greatly compact DNA, because there is no more than one condensin molecule per 10 kb of DNA. Fortunately, there is a third possible explanation for the (+) supercoiling that is compatible with its physiological role. Condensin is so large, reaching out perhaps 1,000 Å, that it could torque the DNA between its reach. Thus, condensin could introduce (+) supercoiling by effecting global 'writhe' (as schematised in Figure 16). Strong evidence for this was provided by the finding that incubation of condensin and a type-2 topoisomerase with plasmid DNA forms chiral DNA knots. These knots were almost exclusively (+), as expected if condensin introduces a regular (+) writhe.

Let us explain more in detail, first mathematically then biologically, the central concepts of *writhe* and *twist*. Our aim is to show that their properties are closely linked. We need to start by recalling that the linking number is a mathematically quantity associated with two closed oriented curves. To define it, the simplest

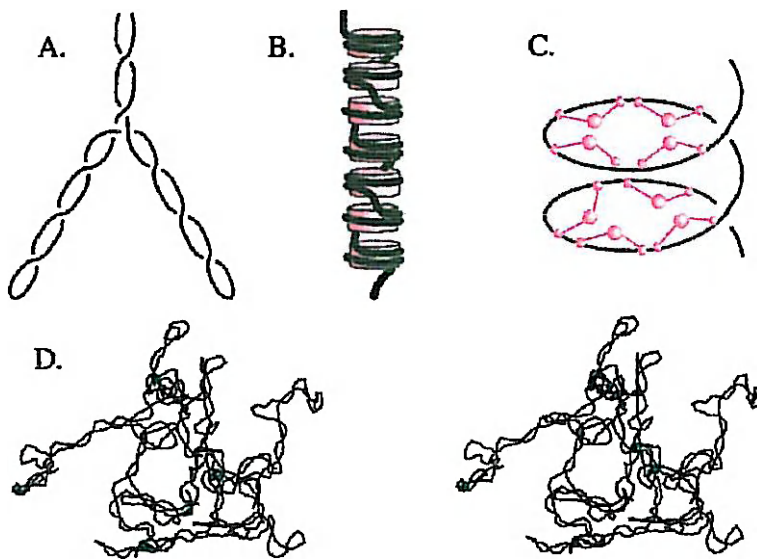


Figure 16 Comparison of four types of DNA compaction by supercoiling. (A) Free (-) supercoils twist DNA into a right-handed plectonemic superhelix. (B) Wrapping around the histone octamer compacts DNA by forming left-handed solenoidal supercoils. (C) SMC proteins, such as *Xenopus* 135 condensin (schematised as red ball and stalk structures), effect global DNA writhe by forming large (+) solenoidal supercoils. (D) Stereo image of a 25-kilobase (kb) (-) supercoiled DNA generated by a Metropolis Monte Carlo simulation. A and C represent approximately 2 kb of DNA (700 nm) at 200,000-fold magnification, whereas B is only 1.5 kb of DNA (500 nm) but at 4-fold greater magnification. D is at 100,000-fold magnification.

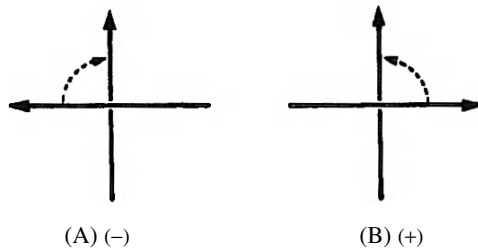


Figure 17 Sign convention for the crossing of two curves in a modified projection. The arrows indicate the orientation of the two crossing curves. To determine the sign of the crossing, the arrow on top is rotated by an angle less than 180° onto the arrow on the bottom. If the rotation required is clockwise as in A, the crossing is given a (-) sign. If the rotation required is counterclockwise as in B, the crossing is given a (+) sign.

manner is to use the so-called modified projection method. We designate the two curves C and A . These two curves when viewed from a distant point will appear to be projected into a plane perpendicular to the line of sight, except that the relative overlay of crossing segments is clearly observable. Such a view gives a modified projection of the pair of curves. In any such projection, there may be a number of crossings. To each such crossing is attached a number ± 1 , depending on the sign convention in Figure 17. Adding all the signed numbers of a given projection and dividing by two gives the linking number, $Lk(C, A)$ of C with A . Examples are shown in Figure 18.

The rigorous definition of the linking number applies to an oriented link. Recall first the following

Definition 8.1. A link L of m components is a subset of S^3 , or of R^3 , that consists of m disjoint, piecewise linear, simple closed curves. A link of one component is a knot.

Definition 8.2. Suppose L is a two-component oriented link with components L_1 and L_2 . The linking number $Lk(L_1, L_2)$ of L_1 and L_2 is half the sum of the signs, in a diagram for L , of the crossings at which one strand is from L_1 and the other is from L_2 .

Note at once that this is well defined, for any two diagrams for L are related by a sequence of Reidemeister moves, and it is easy to see that the above definition is not changed by such a move. The linking number is thus an invariant of oriented two-component links. To be equivalent, two such links must certainly have the same kinking number. The definition given of linking number is symmetric:

$$Lk(L_1, L_2) = Lk(L_2, L_1).$$

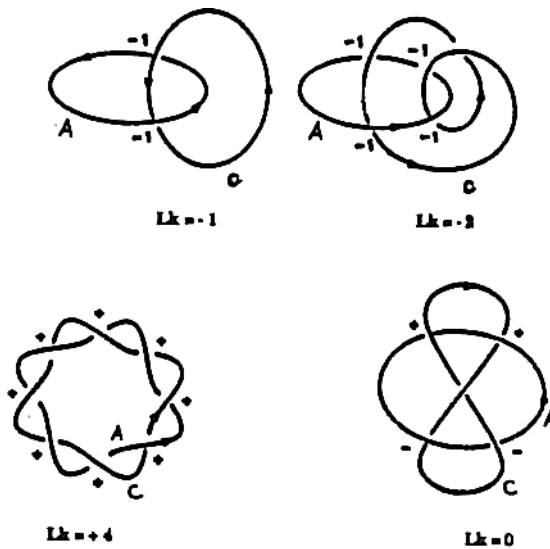


Figure 18 Examples of pairs of curves with various linking numbers, using the convention described in Figure 17.

This definition of linking number is convenient for many purposes, but should not obscure the fact that linking numbers embody some elementary homology theory. Suppose that K is a knot in S^3 . Then K has a regular neighbourhood N that is a solid torus. (Technically, the regular neighbourhood is the simplicial neighbourhood of K in the second derived subdivision of a triangulation of S^3 in which K is a subcomplex.) The *exterior* X of K is the closure of $S^3 - N$. Thus X is a connected three-manifold, with boundary ∂X that is a torus. This X has the same homotopy type as $S^3 - K$, $X \cap N = \partial X = \partial N$ and $X \cup N = S^3$. For the present, we don't need to go more in-depth into this subject (we refer the interested reader to Rolfsen [95], Lickorish [70] and Boi [24].)

Let us rather underline that the linking number has many important properties, two of which are especially important for DNA. First, it is unchanged under any continuous deformation of the pair of curves so long as no break is made in either curve. Second, it is independent of the view for which one computes it. For DNA the linking number is defined to be the linking number of the two backbone curves. However, since either backbone curves may be deformed into the axis curve A without passing through the other, the linking number of a DNA may equally be defined as the linking number of a backbone curve and the axis, $Lk(C, A)$.

The definition of writhe is similar to that of linking. However, it is a property of a single curve, in this case the axis A . In any modified projection of A there may be a number of crossings. To each such crossing is attached a signed number ± 1

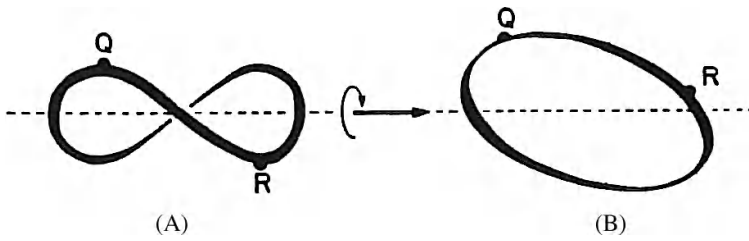


Figure 19 Illustration of the dependence of projected writhing number on projection. The axis of the same non-planar closed DNA is shown in two different projections obtained by rotating the molecule about the dashed line. The points Q and R on the axis help illustrate the rotation. The segment QR crosses in front in part A but is in the upper rear in part B . The projected writhing number is part A is -1 and 0 in part B .

as in the case of linking number. Unlike linking number, the projected writhing number may depend on the projection. This is illustrated in Figure 19 in which a figure-eight in one projection becomes an oval curve in another. In one case the projected writhe is -1 , in the other 0 . The writhing number or writhe, Wr , of the curve A is defined to be the average over all possible projections of the projected writhing number. In other words, such an average value is determined by utilising integrals. We take the integral of the signed crossover numbers, integrating over all vantage points (that is, the points perpendicular to the axis of S^3 , or those points where our eye is in the plane) on the unit sphere, and then divide by the integral of one, integrating over the unit sphere.

$$\begin{aligned} \text{Average value} &= \int \text{signed crossover number } dA / \int dA \\ &= \int \text{signed crossover number } dA / \int 4\pi \end{aligned}$$

since is just the surface area of the unit sphere.

If the axis A lies in a plane except for a few places at which it crosses itself, the writhing number Wr is the total of the signed numbers attached to the self-crossings. Figure 20 illustrates the approximate writhe of some tightly coiled DNA axes. An important fact about the writhing number is that during a self-passage of

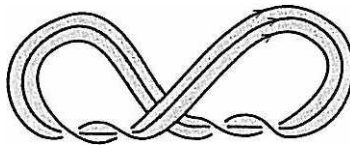


Figure 20

the curve A it must change by two. We also point out that the writhing number of a curve A is independent of the orientation chosen along the curve.

Summarising and considering a ribbon modelling cyclic duplex DNA, we shall say that the writhe of the ribbon, denoted $Wr(R)$, measures how much the axis of the ribbon is contorted in space.

We next define the twist of a DNA. For closed DNA the twist will usually refer to the twist of one of the backbone curves C about the axis curve A . This will be denoted $Tw(C, A)$ or simply Tw . We already defined twist in section 2 by using vector analysis. Let us slightly reformulate the mathematical property of twisting by showing the following picture.

Any local cross-section of a DNA perpendicular to the axis A contains a unique point a of the axis and a unique point c of the backbone curve C (see Figure 21). We denote by \mathbf{v}_{ac} a unit vector along the line joining a to c . As the DNA is traversed since the curve C winds helically about A , the vector \mathbf{v}_{ac} turns about A . Tw is a measure of this turning. As the point a moves along A , the vector \mathbf{v}_{ac} changes. The infinitesimal change in \mathbf{v}_{ac} , denoted $d\mathbf{v}_{ac}$, will have a component tangent to the axis and a component perpendicular to the axis. Tw is the measure of the total perpendicular component of the change of \mathbf{v}_{ac} as the point a traverses the entire length of the DNA. This is given by the line integral

$$Tw = 1/2\pi \int_A d\mathbf{v}_{ac} \cdot T \times \mathbf{v}_{ac},$$

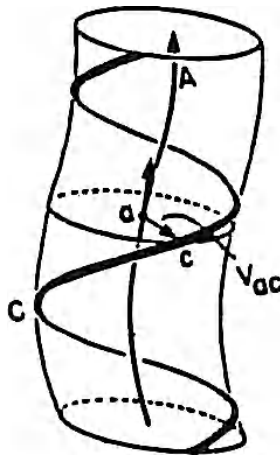


Figure 21 Cross-section of a DNA. The plane perpendicular to the DNA axis A intersects the axis in the point a and the backbone curve C in the point c . The unit vector along the line joining a to c is denoted \mathbf{v}_{ac} . Note that as the intersection plane moves along the DNA, this vector turns about the axis.

where T is the unit tangent vector along the curve A . When A is a straight line or planar, dv_{ac} is always perpendicular to A , so that in this case Tw is simply the number of times that v_{ac} winds about the axis. Examples are shown in Figure 19. It can be easily demonstrated that Tw is positive if the winding is right-handed and negative if left-handed. Furthermore, if the DNA is closed then the initial and final positions of v_{ac} are the same. Thus if the DNA is closed (and in the circular or ribbon model) and its axis planar, Tw must necessarily be an integer. However, if the axis is supercoiled this is not usually the case. A portion of a supercoiled DNA is shown in Figure 22. Here the axis A itself is a helix, so that the helically winding C becomes a superhelix. For such an example, Tw is the number of times that C winds about A plus a term $\lambda \sin \gamma$ which depends on the geometry of the helix A . The term λ is the number of times that A winds about its own straight line axis and γ is its pitch angle.

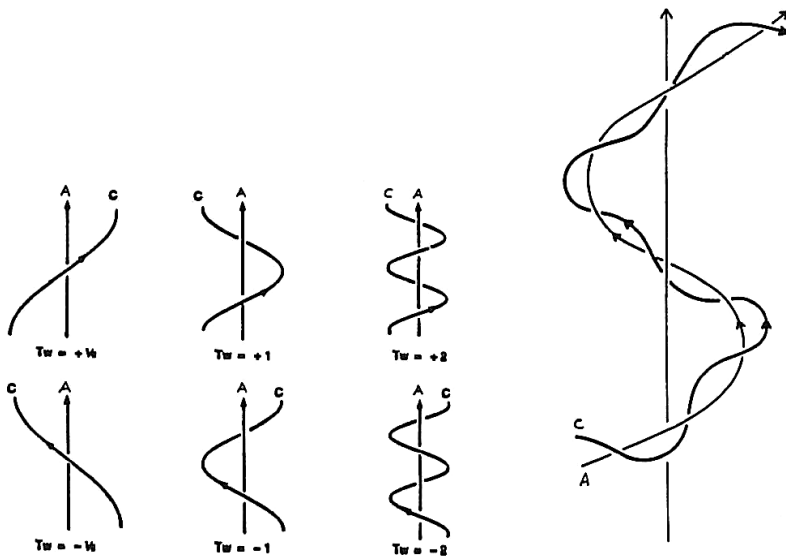


Figure 22 Examples of pairs of curves C and A with different values of twist. The first six are simple examples in which the axis A is a straight line and the twist is the number of times that C winds about A , being positive for right-handed twist and negative for left-handed twist. The last example is one in which the axis A is a helix winding around a linear axis, and the curve C is a superhelix winding about A . In this case the twist of C about A is the number of times that C winds about A (in this case approximately 3.5) plus $\lambda \sin \gamma$, where λ is the number of times that A winds about the linear axis and γ is the pitch angle of the helix A . Here λ is approximately 1.5 and γ is approximately 40° . Thus, Tw equals approximately 4.46.

Let us consider again the concept of *writhe*, which can simply be thought of in terms of the number of times the rubber rod crosses over itself. The crucial point about Wr is that it is a measure of the shape of the DNA as a three-dimensional curve through space. As already noted, one can count the number of crossovers of the DNA in a single view in order to estimate Wr . All we need to do to get Wr accurately is to count the number of crossovers that can be seen in many different randomly chosen views of the structure, and then take the average of all of these to get the actual value of Wr . This is not a hard concept to grasp, if we think of taking a large number of snapshots of the DNA as it tumbles randomly through space, due to the thermal motion. In practice, however, this may not be such a straightforward procedure, for in some views there can be many crossovers, some of which will cancel each other out.

Now the diagrams in Figure 23 are drawn for positive Lk , i.e. for overwound DNA. You may recall that DNA in living cells is normally not overwound, but rather is underwound, and so its value of Lk is negative. Therefore, one can provide a corresponding set of pictures for negatively supercoiled, underwound DNA. For example, the twist is now left-handed and counterclockwise. As a result, all the values of the Tw , Wr , and Lk numbers will be conversed, namely negative. This means that the DNA crosses over itself in two different fashions: a right-handed and a left-handed. In fact, the handedness of the crossovers in any interwound supercoil enables you to say definitely whether the DNA is underwound or overwound, simply by looking at a picture.

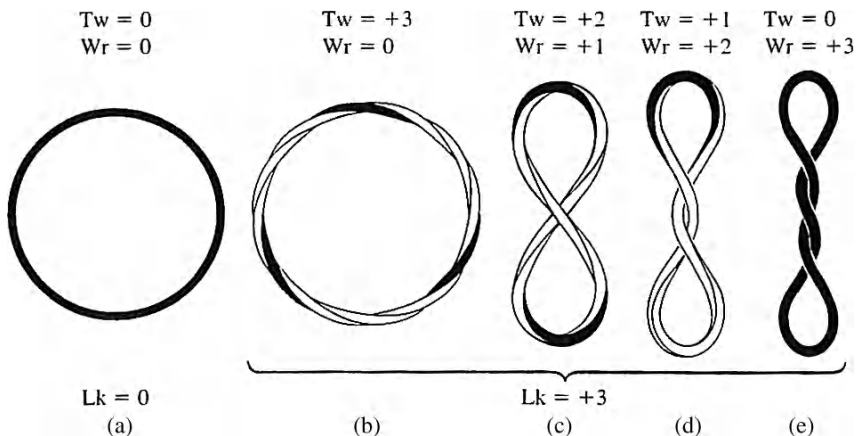


Figure 23 Five closely related circular DNA molecules: (a) and (b) show open circles, while (c), (d) and (e) show interwound supercoils. The DNA in its stress-free, relaxed form is drawn as a rubber rod of square cross-section, with one face black.

As already mentioned, there are two general classes of supercoil, known as interwound and toroidal. The circular DNA (that is, with the ends of the molecule fixed) consists of a series of open spirals that wind around an imaginary ring, or toroid; this kind of supercoiling is known as ‘toroidal’. But the circular can also wind above and below itself several times, and this kind of supercoiling is called ‘interwound’. In practice, real DNA supercoils may contain portions of both the toroidal and interwound geometries. Thus, where certain parts of the DNA are highly curved, on account of either the base sequence or due to wrapping around a protein, one may find toroidal structures, since the DNA in a toroidal supercoil is highly curved throughout. Alternatively, if such curved portions of the DNA are not very long, they may locate themselves at the two strongly curved end-loops of an interwound supercoil, as shown at the top and bottom in Figure 24. Sometimes the interwound and toroidal geometries may occur together, as in the looped-linear DNA which is shown schematically in Figure 25. On a small scale, within any loop, the coiling is toroidal on account of the wrapping of DNA around protein spools; but on a large scale, over the full length of any loop, the structure is interwound. You often see this kind of arrangement in telephone cords, if people habitually rotate the handset. In general, supercoiled DNA has the shapes seen in Figure 24 because it either has more turns of twist, or fewer turns of twist, than the underlying, relaxed, right-handed double helix from which it is made. DNA with more than the natural

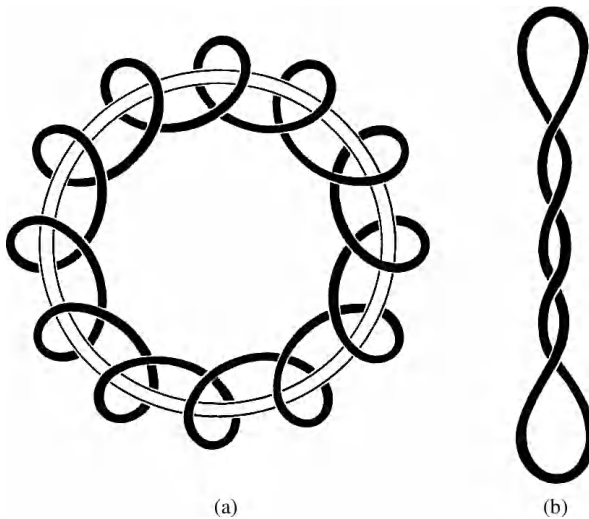


Figure 24 Two general varieties of DNA supercoil. In (a), the DNA coils into a series of spirals about an imaginary toroid or ring (shown here by open lines); and so this kind of wrapping is known as ‘toroidal’. In (b), the DNA crosses over and under itself repeatedly; and so this kind of wrapping is known as ‘interwound’.

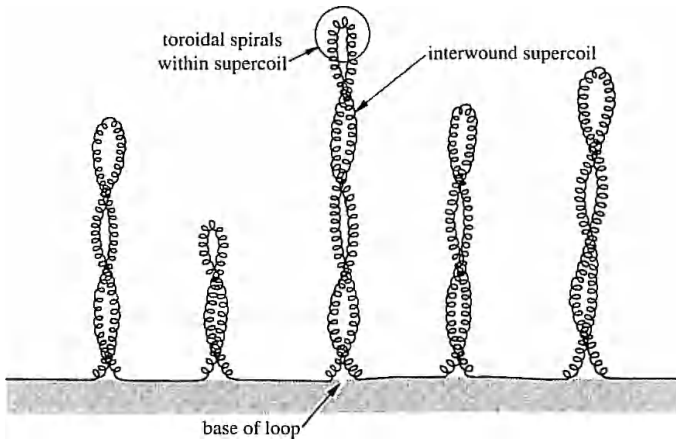


Figure 25 The division of a long, linear DNA molecule into loops generates end-restraint at the base of every loop, if the two ends are attached to some support of 'scaffold'. This kind of looped-linear arrangement is thought to be typical of the chromosomal DNA found in higher organisms.

number of turns is known as *overwound*, while DNA with fewer than the natural number of turns is known as *underwound*.

We have now described two different kinds of supercoiling for DNA — toroidal and interwound. But what are the relative stabilities of these two forms? In other words, when will a DNA molecule be interwound, and when will it be toroidal? The interwound shape is usually very stable, and most underwound or overwound DNA molecules will naturally adopt an interwound shape, in the absence of other forces. But the proteins that associate with DNA in living cells can sometimes change the situation dramatically, and favour the toroidal over the interwound form by wrapping the DNA around themselves (see next section for further details on this topic). Note, however, that the preferred interwound structure of DNA molecules in cells is somewhat similar to the idealised shape in Figure 23(e) (but with a linking number Lk of the opposite sense, which means that these DNA molecules are underwound, with Lk negative), since $Wr = 0.9 Lk$, and $Tw = 0.1 Lk$. In other words, the DNA which has been underwound finds it more favourable energetically to cross over itself repeatedly, than to alter its twist.

For example, consider the cork which has been inserted between the two turns of ribbon shown in Figure 26(c). This cork represents a typical protein 'spool' around which the DNA can wrap, and around which it does wrap in a left-handed sense in the chromosomes of most higher organisms on Earth. If the DNA or ribbon in Figure 26(c) were to be cut free from the two blocks at either end, it would stay wrapped around the 'sticky' protein spool; whereas if it were cut free in the

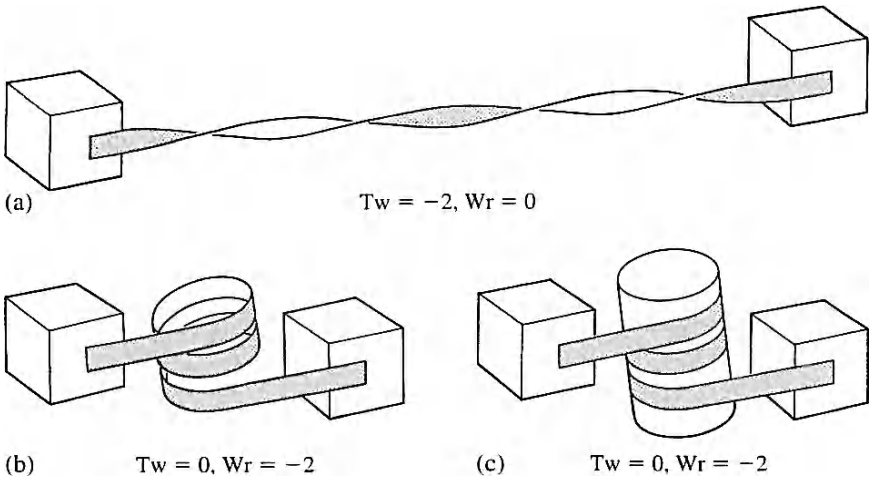


Figure 26 A highly twisted ribbon will collapse spontaneously into part of a toroidal supercoil. In (a), the two ends of the ribbon are held apart by their attachment to blocks, so that $Tw = -2$. In (b), the blocks move together so that the ribbon can collapse to $Wr = -2$. In (c), a cork or protein spool stabilises the shape of the ribbon shown in (b).

absence of a spool, as in Figure 26(b), it would immediately spring back into a straight configuration. When we isolate DNA in the laboratory in pure form from any kind of cell or cells, at some point in the procedure we must strip off the proteins around which the DNA was originally wrapped, without breaking either of its two double-helical strands. In other words, we must remove the cork from the arrangement shown in Figure 26(c), without cutting the DNA free from either of its two end-blocks. Naturally the 'naked' DNA will first spring out to the highly twisted form shown in Figure 26(a), and then it can collapse into an interwound supercoil as shown in Figure 23(e), because it has lost the curvature which stabilised the toroidal form. Therefore, we can expect to see highly interwound supercoils in the preparations of pure DNA which we make from living cells, after removal of various proteins. Incidentally, this is why DNA supercoils in Nature are usually underwound rather than overwound: the DNA always coils around proteins in the cell nucleus in the form of a left-handed toroidal spiral, giving negative Lk . In the next section, we will be especially concerned with some important topological and biological properties of supercoiling.

To conclude the present description, let us point out that most of the missing turns in any negatively supercoiled DNA molecule are stored in the form of writhe Wr , whether by crossovers in an interwound supercoil, or by flat spirals in a toroidal supercoil. How, then, can supercoiling produce a reduction of twist Tw by one or two turns, as is needed for a polymerase protein to unwind DNA in various

locations? Clearly, the DNA must be able to vibrate or fluctuate in solution, as a kind of Brownian movement, from shapes with high writhe to shapes with high twist. For example, an interwound supercoil might vibrate from the shape shown in Figure 23(e) to any of the shapes shown in Figure 23(d), (c), or (b), in order to generate twist. Similarly, a toroidal supercoil might vibrate from the shape shown in Figure 26(b) to that shown in Figure 26(a). Unfortunately, we have few direct experimental data today, which might indicate how DNA molecule fluctuates in solution over a large scale. We know, through probing for single-stranded regions using enzymes and chemicals, that negatively supercoiled DNA vibrates much more efficiently than relaxed DNA to yield negative T_w ; and we know also that many genes require negative supercoiling in order to be transcribed by RNA polymerase; but we do not know how DNA changes its shape over a large scale, to produce vibrations that lead to the generation of twist. Perhaps these involve changes in the local shape of the DNA from a right-handed supercoil to a plane curve, or from a plane curve to a left-handed supercoil. But all we have today are a great many lines of indirect evidence to suggest what might be going on. Furthermore, our indirect data are limited to observations about bacterial genes, because the genes in higher organisms are so poorly understood that one cannot draw any firm conclusions about how they work.

9. A Mathematical Model for Explaining the Folding of Chromatin Fibre During Interphase

In the nucleus of eukaryotic cells, the three-dimensional organisation of the genome takes the form of a nucleoprotein complex called chromatin. This organisation not only compacts the DNA but also plays a critical role in regulating interactions with the DNA during its metabolism. This packaging of our genome, the basic building block of which is the nucleosome, provides a whole repertoire of information in addition to that furnished by the genetic code. This mitotically stable information is not inherited genetically and is termed epigenetic. One of the challenges in chromatin research is to understand how epigenetic states are established, inherited, controlled and modified so as to guarantee that their integrity is maintained while preserving the possibility of flexibility. In other words, the aim is to understand the temporal and spatial dynamics of chromatin organisation, during the cell cycle, in response to different stimuli and in different cell type.

Beyond the level of the nucleosome, the chromatin is compacted into higher structures which delimit specialised nuclear domains such as regions of heterochromatin and euchromatin. Heterochromatin is defined as the regions of chromatin that do not change their condensation state during the cell cycle and represents the majority of the genome of higher eukaryotes. Heterochromatin principally comprises repeated non-coding DNA sequences, its characteristics generally contrast

with those of euchromatin. One essential characteristic of the heterochromatin regions, which has been highly conserved during evolution, is the presence of hypoacetylated histones (H3 and H4). Apart from its repression of transcription, heterochromatin function remains unknown.

Of paramount importance to the understanding of gene expression and biological regulation, is the mechanism which drives and controls the packaging of DNA and its organisation within the chromatin structure. The lowest level of organisation is the nucleosome, in which two superhelical turns of DNA (a total of 165 base pairs) are wound around the outside of a histone octamer. Nucleosomes are connected to one another by short stretches of linker DNA. During chromatin assembly on nascent DNA, acetylated histones H3 and H4 are sequestered by the DNA first, histones H2A and H2B follow, and, finally, H1 binds, stabilising chromatin folding within the irregular 30 nm fibre. At the next level of organisation the string of nucleosomes is folded into a fibre about 30 nm in diameter, and these fibres are then further folded into higher-order structures. More precisely, during the progressive assembly of chromatin, DNA is compacted, nucleosome formation leads to a sevenfold compaction of DNA, and the subsequent formation of the 30 nm fibre contributes a further sevenfold compaction. These four successive steps of compactations represent the major topological constraints of DNA in eukaryotic nucleus. At levels of structure beyond the nucleosome the fundamental mechanisms of folding are still unknown. We know that the 11 nm nucleosome units (the first level of packing of DNA) coil into a 30 nm solenoid structure which is stabilised by H1 histone. The solenoid forms loops that attach to a scaffold of non-histone protein, which leads to the chromatin supercoiling during condensation within metaphase chromatids. This intermediary and possibly crucial level of compaction of complexes DNA-proteins into the final form, a mitotic chromosome, is very scarcely understood.

Among the different hypothetical models that have been proposed over the last years for the folding of the chromatin fibre during interphase, the so-called radial-loop model seems to us the most suitable for explaining the formation of the 30 nm solenoid structure. Let us suggest, specifically, a theoretical model by applying methods and techniques from algebraic geometry and specifically from the classification theory of compact connected two-manifolds, which has been one of the most important and far-reaching mathematical results of the twentieth century. We start with the following theorem.

Theorem 9.1. *Let M be a closed, simply-connected orientable manifold. M can be expressed as a union*

$$M = D \cup D' \cup \bigcup_{i=1}^n S_i$$

of polyhedral two-cells with disjoint interiors, such that (1) for each i , each of the sets $S_i \cap D$ and $S_i \cap D'$ is the union of two disjoint arcs and (2) $D \cap D'$ is the union of $2n$ disjoint arcs.

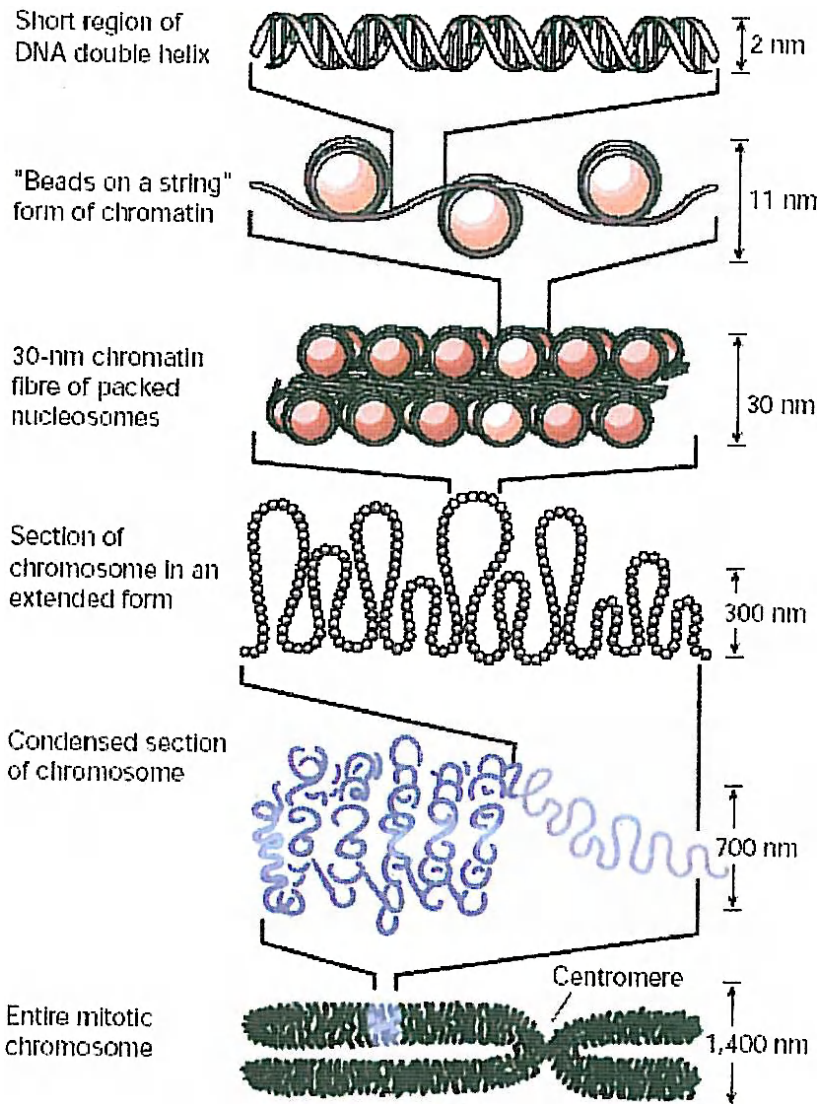


Figure 27 Chromatin supercoiling during condensation: each metaphase chromatid is 700 nm wide. The first level of packing of DNA results in an 11 nm diameter fibre. The 11 nm nucleosome units coil into a 30 nm solenoid structure, which is stabilised by H1 histone. The solenoid forms loops that attach to a scaffold of non-histone protein.

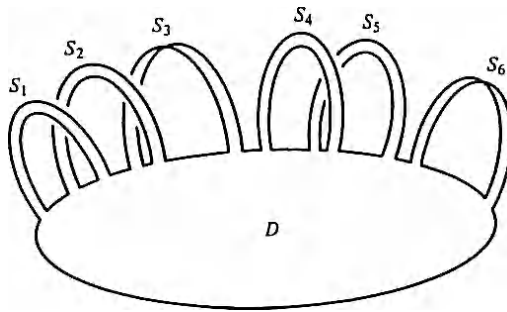


Figure 28

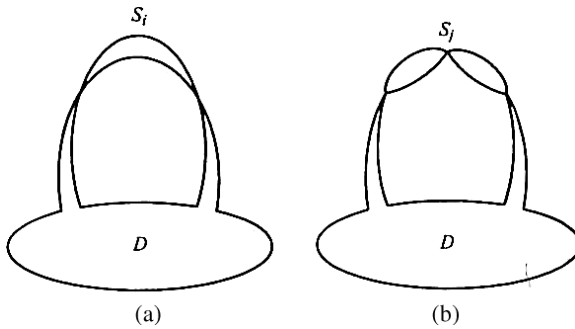


Figure 29

The sets S_i will be called *strips*, and M' will be called a *two-cell with strips*. Evidently, such an M' can always be imbedded in R^3 , and thus can be described by a figure such as Figure 28. Under the conditions of Theorem 1 boundary $\partial M'$ must be a one-sphere, but aside from this, the strips S_i may be attached to $\partial M'$ at any set of disjoint arcs. If $S_i \cup D$ is an annulus, then S_i will be called *annular* (relative to D , of course) and if $S_i \cup D$ is a Möbius band, then S_i will be called *twisted*. Thus, in Figure 28, S_3 and S_6 are twisted, and the rest of the strips are not. Note that in investigating the topology of M' , we do not care whether the sets $S_i \cup D$ are knotted. Note also that indicating “multiples twists” would contribute nothing to the generality of the figure. For example, in Figure 29(a) on the left a double twist gives an annulus, and in Figure 29(b) a triple twist gives a Möbius band.

We shall now simplify this representation of M in various ways.

- (1) Suppose that S_j is a twisted strip, so that $S_j \cup D$ is a Möbius band. Let $J = \partial(S_j \cup D)$, so that J is a polygon. As in Figure 30, let $P, Q, R,$ and T be points of J , not lying in any set S_j ; let PT be the arc in J , between P and T , that

intersects ∂S_j ; supposing that P, Q, R and T appear in the stated order on J ; and supposing that the arcs $PQ \subset PT$ and $RT \subset PT$ intersect no set S_j . We assert that there is a piecelinear homeomorphism (PLH):

$$\begin{aligned} h : M &\Leftrightarrow M, & J &\Leftrightarrow J, & D \cup S_i &\Leftrightarrow D \cup S_i \\ P &\rightarrow P, & T &\rightarrow T, & QT &\Leftrightarrow RT, \end{aligned}$$

such that $h|(J - PT)$ is the identity. Consider the two-cell $D, h(D'), S_i$, and $h(S_j)(j \neq i)$. These have all the properties stated above for D, D', S_i , and $S_j(j \neq i)$. The operation which replaces the old system of two-cells by the new will be called *operation α* . We now renumber the two-cells S_i in such a way that S_1, S_2, \dots, S_k are annular, and $S_{k+1}, S_{k+2}, \dots, S_n$ are twisted.

Lemma 9.2. *In the conclusion of Theorem 9.1, we can choose the 2-cells in such a way that (a) the intersections $S_i \cap D(i > k)$ lie in disjoint arcs in ∂D and (b) $\bigcup_{i>k} (S_i \cap D)$ lies in an arc in ∂D which intersects no annular strip S_j .*

- (2) If we have no annular strips, then we proceed to step (3) below. If we have an annular strip S_i , then there must be another annular strip S_j which is “linked with S_i on ∂D ,” as indicated in Figure 30. (If not, $\partial M' = \partial D'$ would not be connected.) The set $D \cup S_i \cup S_j$ is then a handle.

Recall that by a *handle* we mean a space obtained by deleting from a torus the interior of a two-cell. Figure 31(a) shows what a handle looks like. A *two-sphere with n holes* is a space obtained by deleting from a two-sphere the interiors of n disjoint two-cells. If a handle is attached to the boundary of each of the holes, the resulting space is a *two-sphere with n handles*, as shown in Figure 31(b). A *projective plane* is a space defined in such a way that each pair of antipodal points of the circle are supposed to be identified. A *sphere with n cross-caps* is a space

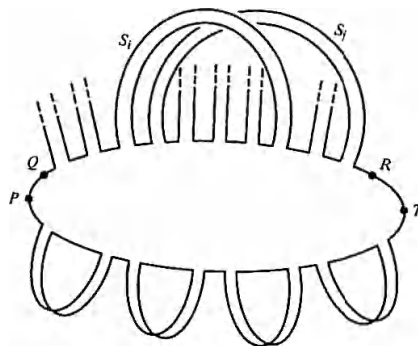


Figure 30

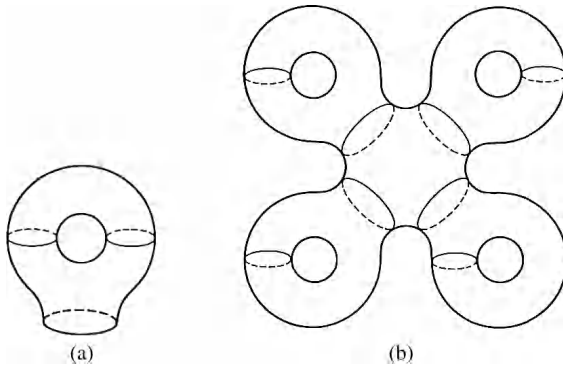


Figure 31

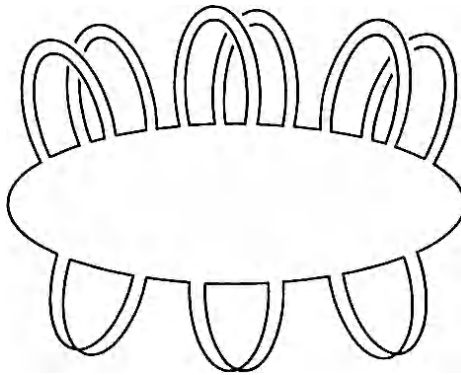


Figure 32

obtained by starting with a sphere with n holes and then attaching a Möbius band to the boundary of each of the holes.

By operation β , closely analogous to α , we slide the strip $S_r (r \leq k, r \neq i, j)$ along the arc PT , so as to get a situation in which $(S_i \cup S_j) \cap D$ lies in an arc in ∂D which intersects no set $S_r (r \neq i, j)$. We do this for each such handle. The figure now looks like Figure 32.

- (3) Let $m = n - k$ be the number of the twisted strips S_i , and suppose that $m > 2$. Consider the first three of the twisted strips (starting in some direction from the annular strips) as shown in Figure 33. By two operations of the type α , we slide PQ along $\partial(D \cup S_s)$ so as to move it onto $P'Q' \subset \text{Int } AB \subset \partial D$; and we slide RT along $\partial(D \cup S_s)$ onto $R'T' \subset \text{Int } AB$. The figure now looks like Figure 34. It is easy to check that the new strips $S'_r, \cup S'_t$ is a handle.

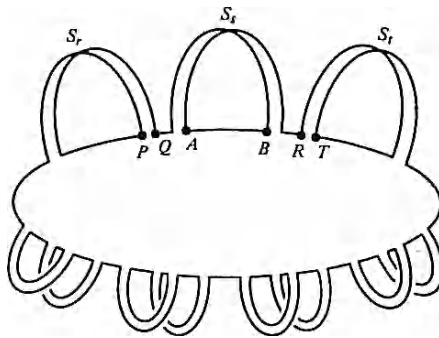


Figure 33

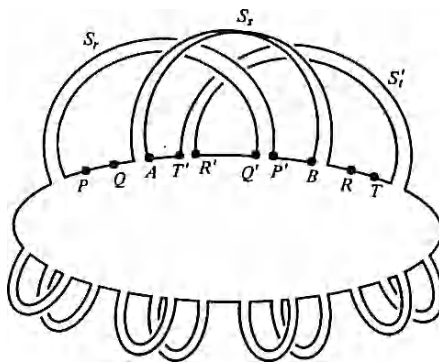


Figure 34

By another application of β , we move $S_s \cap D$ to the right of $(S'_r \cup S'_s) \cap D$ in Figure 34. Thus we have introduced a new handle into the figure, and reduced the number of twisted strips by two. Therefore we may assume, in Theorem 9.1, that the number m of twisted strips is ≤ 2 . M is orientable if and only if $m = 0$ at the final stage. To each linked pair of annular strips, and to each twisted strip, we add a two-cell lying in D , as indicated by the dotted arcs in Figure 35. This gives a set $\{H_i\}$ of handles ($h \geq 0$) and a set $\{B_j\}$ of m Möbius bands ($0 \leq m \leq 2$).

Consider the set $N = Cl[M - (\bigcup H_i \cup \bigcup B_j)]$. N is the union of two two-cells D_1 and D_2 , with $D_1 \subset D$ and ∂D_2 , it follows that N is a sphere with holes. Thus we have proved the following

Theorem 9.3. *Let M be a compact connected two-manifold. Then M is a two-sphere with h handles and m cross-caps ($h \geq 0, 0 \leq m \leq 2$).*

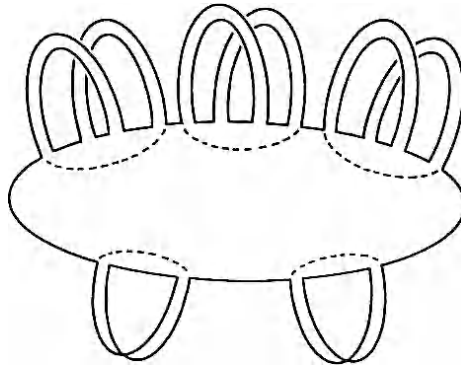


Figure 35

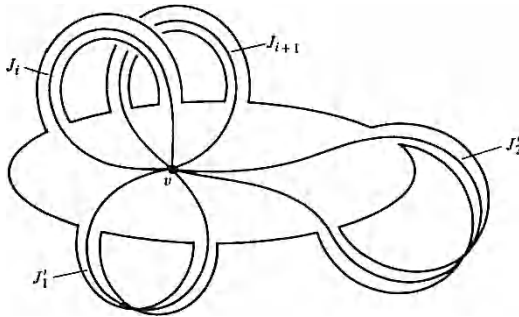


Figure 36

We now define a new open cell decomposition of M , as follows. As indicated in Figure 36, we choose a point v of $\text{Int } D$, and we define a collection $\{J_i, J'_j\}$ of polyhedral 1-spheres (one J_j for each annular strip, and one J'_j for each twisted strip) such that each of them “runs from v through the corresponding strip, and then returns to v ,” and such that each two of the sets in $\{J_i, J'_j\}$ intersect at v and nowhere else. This gives an open cell-decomposition C' of M , with one vertex v , $2h$ edges $J_i - \{v\}$, m edges $J'_j - \{v\}$, and one 2-face $C^2 = M - [\cup J_j \cup \cup J'_j]$.

Thus we have:

Theorem 9.4. *Let M be a 2-sphere with h handles and m cross-caps ($h \geq 0$, $0 \leq m \leq 2$). Then*

$$\chi(M) = 2 - (2h + m).$$

Proof. $V - E + F = 1 - (2h + m) + 1.$

10. Biological Justifications for the above model

One way to show the relevance of the topological model we sketched above is to investigate the spatial organisation and functional compartmentalisation of chromosomes, and the nucleus itself, within the quest to understand how the expression of complex genomes is regulated. Inside the higher eukaryotic chromosome, DNA is folded through DNA-protein interactions into multiple levels of organisation. At the highest level, these yield a compaction ratio of more than 20 000: 1 in terms of the ratio of linear B-form DNA to the length of the fully compacted metaphase chromosome. While the extent of compaction within mitotic chromosomes is well known, less appreciated is the fact that compaction remains extremely high within interphase nuclei. The bulk of genomic DNA in interphase is likely to be packaged within large-scale structures well above the level of the 30 nm chromatin fibre (for further details see Widom [116]).

For technical reasons, most research into chromosome structure has focused on the structure of maximally condensed, metaphase chromosomes. An experimental approach based largely on unfolding chromosome structure through extraction of chromosomal proteins has led to a radial loop model of chromosome structure. In this model, structural proteins, which are resistant to high salt and detergent extraction, anchor the bases of 30 nm chromatin fibre loops (~20–200 kb long) to a chromosome “scaffold”, which itself may be helically coiled. Specific SAR/MAR DNA sequences (scaffold attachment regions or matrix attachment regions) are hypothesised to form the bases of these loops, attached to specific proteins which are predicted to make up the chromosome scaffold. Specific sequences are found remaining at the axial core in extracted human metaphase chromosomes, but it is not clear whether the same SAR/MAR sequences are attached to an underlying scaffold in both mitotic and interphase chromosomes. Experiments using fluorescence *in situ* hybridisation (FISH) on cell nuclei has led to a giant-loop, random walk model for interphase chromosomes, based on statistical analysis of the mean separation between two chromosome sites, as a function of genomic distance.

Ideally, any model of large-scale chromatin folding would unify mitotic and interphase chromosome structure and predict the structural transitions accompanying cell-cycle-driven chromosome condensation/decondensation. The radial-loop, helical-coil model of mitotic chromosome structure (Figure 37a) has been extended to interphase chromosomes. However, this has required postulating a particular loop geometry that might, under special circumstances, give rise to a fibre with an elliptical 60–90 nm cross-section (Manuelidis [76]). An alternative model proposes a successive, helical coiling of 10 nm chromatin fibre into 30–50 nm tubes, and of these into 200 nm diameter tubes, which coil into *c.* 600 nm metaphase chromatids (Sedat and Manuelidis [96]). Finally, a folded chromonema model is based on *in vivo* light microscopy combined with TEM ultrastructural analysis of folding intermediates during the transition into an out of mitosis. In this model,

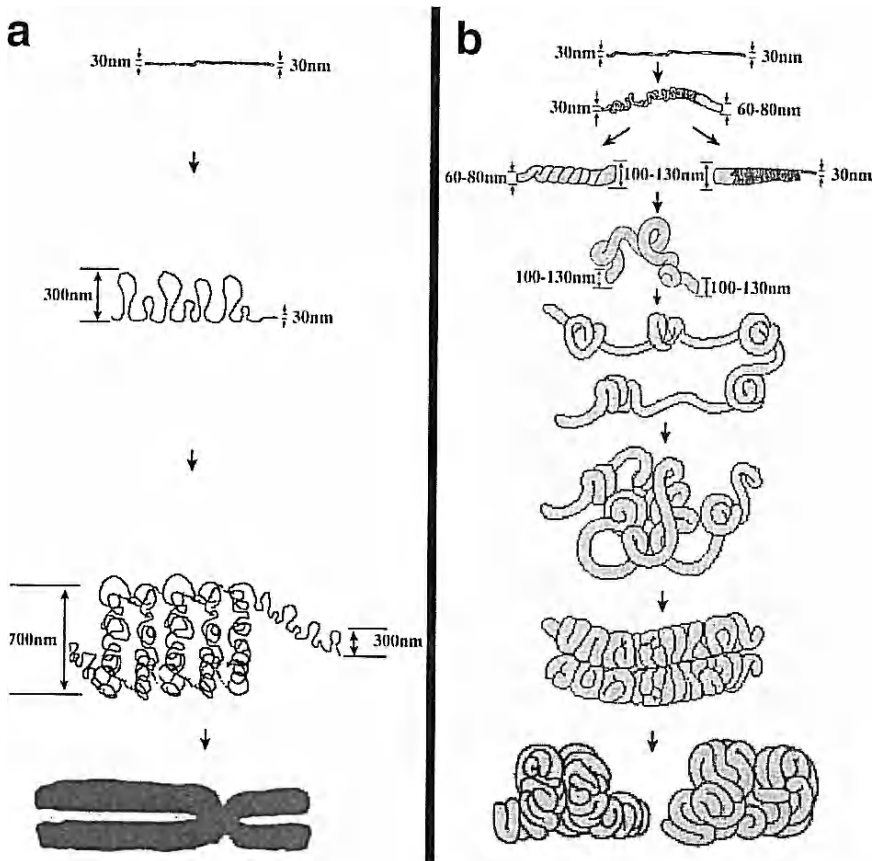


Figure 37 (a) Radial-loop model for mitotic chromosome structure. A looping of the 30 nm fibre gives rise to a 300 nm structure in which 50–100 kb looped DNA attaches at the base of the loop to a chromosomal scaffold. This structure coils helically to form the metaphase chromosome. (b) Chromosome model of interphase chromatin structure. Progressive levels of coiling of the 30 nm fibre into 60–80 nm and 100–130 nm fibres are depicted. Chromonema fibres kink and coil to form regions of more dispersed or compact chromatin. Extended chromonema fibres predominate in G_1 while more compact structures become abundant during cell-cycle progression. Chromonema folding culminates with the formation of the G_2 chromatid, which coils to form the compact metaphase chromosome.

10 and 30 nm chromatin fibre fold to form a c. 100 nm diameter chromonema fibre, which then folds into a 200–300 nm diameter prophase chromatid, which itself coils to form the metaphase chromosome (Figure 37b) (Belmont and Bruce [9]). It is still unclear how these structural models of mitotic and interphase

chromosome structure integrate with the underlying biochemistry responsible for chromosome condensation. The two chief protein components of the mitotic chromosome scaffold, topoisomerase II α and SCII, have more clearly identified DNA topological activities than structural roles. SCII is a component of the mitotic condensing complex, which recently has been demonstrated to have the ability to introduce positive supercoils into DNA in the presence of topoisomerase II in a stoichiometric manner. SCII also shows a non-ATP-dependent enhancement of re-annealing of complementarity DNA strands (see Section 1 for more details on this topic).

More specifically, the geometrical model we suggested in the previous section might fit well with the three-dimensional packing process of chromatin, first, into a 300 nm extended scaffold-associated form, followed by a 700 nm condensed scaffold-associated form. In fact, the condensation of metaphase chromosome results from several orders of folding and coiling of 30 nm chromatin fibres. For example, electron micrographs of histone-depleted metaphase chromosome from HeLa cells reveal long loops of DNA anchored to a chromosome scaffold composed of non-histone proteins. This scaffold has the shape of the metaphase chromosome and persists even when the DNA is digested by nucleases. As depicted schematically in Figure 38, megabase long loops of the 30 nm chromatin fibre are thought to associate with the flexible chromosome scaffold, yielding an extended form characteristic of chromosome during interphase. Coiling of the scaffold into a helix and further packing of this helical structure produces the highly condensed structure characteristic of metaphase chromosome.

Furthermore, *in situ* hybridisation experiments with several different fluorescent-labeled probes to DNA interphase cells support the loop model shown in Figure 37. In these experiments, some probe sequences separated by millions of base pairs in linear DNA appeared reproducibly very close to each other in interphase nuclei from different cells (Figure 39). These closely spaced probe sites are postulated to lie close to specific sequences in the DNA, called scaffold-associated regions (SARs) or matrix-attachment regions (MARs), that are bound to the chromosome scaffold. SARs have been mapped by digesting histone-depleted chromosome with restriction enzymes and then recovering the fragments that are bound to scaffold proteins. In general, SARs are found between transcription units. In other words, genes are located primarily within chromatin loops, which are attached at their bases to a chromosome scaffold. Experiments with transgenic mice indicate that some cases SARs are required for transcription of neighbouring genes. In *Drosophila*, some SARs can insulate transcription units from each other, so that proteins regulating transcription of one gene do not influence the transcription of a neighbouring gene separated by a SAR. Individual interphase chromosomes, which are less condensed than metaphase chromosomes, cannot be resolved by standard microscopy or electron microscopy. Nonetheless, the chromatin of interphase cells is associated with extended scaffold and is further organised into specific domains.

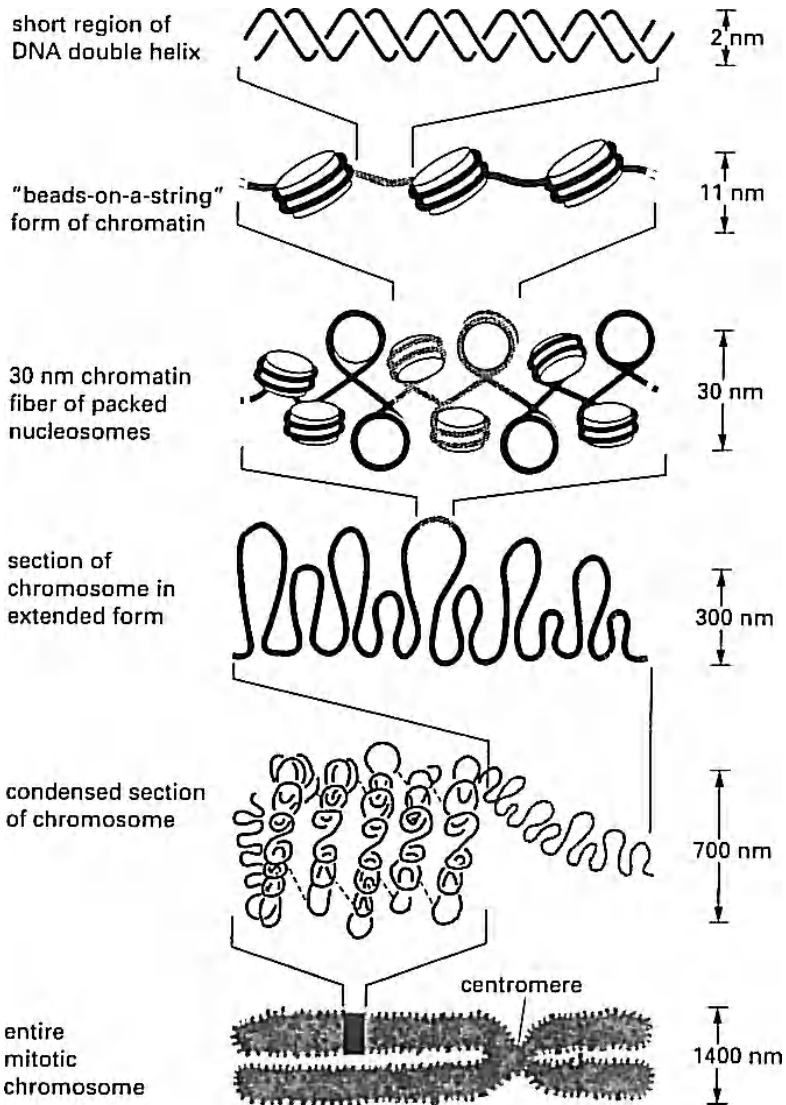


Figure 38 Model for the seven-orders packing of chromatin and the chromosome scaffold in metaphase chromosome. In interphase chromosomes, long stretches of 30 nm chromatin loop out from extended scaffolds. In metaphase chromosomes, the scaffold is folded into a helix and further packed into a highly compacted structure, whose precise geometry has not been determined. [Adapted from Lodish *et al.* (2000, p. 326).]

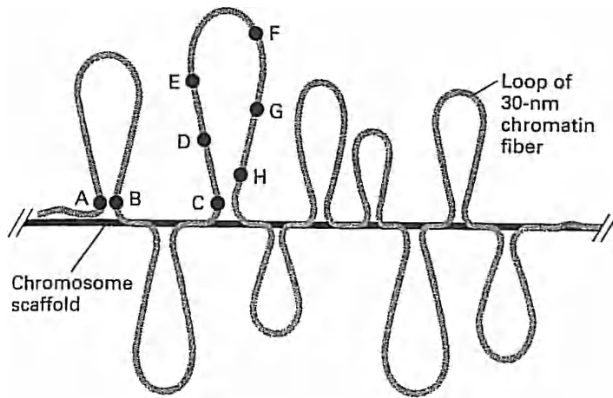


Figure 39 Experimental demonstration of chromatin loops in interphase chromosomes. In situ hybridisation of interphase cells was carried out with several different fluorescently labeled probes specific for sequences separated by known distances in linear, cloned DNA. Lettered circles represent probes. Measurement of the distances between different hybridised probes, which could be distinguished by their colour, showed that some sequences (e.g., A, B, and C), separated from each other by millions of base pairs, appear located near each other within nuclei. For some sets of sequences, the measured distances in nuclei between one probe (e.g., C) and sequences successively farther away initially appear to increase (e.g., D, E, and F) and then appear to decrease (e.g., G and H). The measured distances between probes are consistent with loops ranging in size from one to four million base pairs. [Adapted from Lodish *et al.* (2000, p. 327).]

II. Open Mathematical Questions, Biological Implications, and Some Suggestions for the Future Research

In this article we stressed the participation of topoisomerases in nearly all cellular processes involving DNA. Because the enzymes affect the topology and organisation of intracellular DNA, the primary effects of inactivating a topoisomerase are also likely to generate far-reaching ripples. The regulation of the cellular levels of the enzymes themselves and the association of the enzymes with other cellular proteins are closely tied to the cellular functions of the enzymes. One major cellular function of the topoisomerases is to prevent excessive supercoiling of intracellular DNA. However, supercoiling is sometimes utilised *in vivo* to drive a particular region of intracellular DNA into a conformation suitable for a particular process. Initiation of DNA replication, for example, often requires that the DNA be in a negatively supercoiled state. Indeed, replication is the best-known process that generates supercoils in intracellular DNA. The involvement of various topoisomerases in the removal of positive supercoils generated by replication is generally in accordance with their known *in vitro* specificities. Namely,

eukaryotic DNA topoisomerases I and II, and bacterial DNA topoisomerase IV, can efficiently remove supercoils of either sort; bacterial DNA topoisomerases I and III, and eukaryotic DNA topoisomerases III, can remove negative supercoils, but not positive supercoils, unless a single-stranded region is present in the DNA. Bacterial gyrase is unique in its ability to convert positive to negative supercoils; depending on how fast the positive supercoils are generated and how fast they are converted to negative supercoils, gyrase can either prevent accumulation of positive supercoils in an intracellular DNA segment or keep the segment in a negatively supercoiled state.

The DNA topoisomerases presumably co-evolved with the formation of very long and/or ring-shaped DNA molecule. To solve a variety of problems that are rooted in the double-helix structure of DNA, nature has created not one but three distinct enzymes. In eukaryotes, members of all three subfamilies of DNA topoisomerases have been found in the same cells; in bacteria, four members from two subfamilies participate in nearly all-cellular transactions of DNA. The past decade saw much progress in the study of the DNA topoisomerases, but many questions remain. The key to answering to them may lie in the elucidation of interactions between the DNA topoisomerases and other cellular proteins. Complexes between these enzymes and transcription factors and chromosomal proteins illustrate new avenues yet to be fully explored. Furthermore, whereas the information available on topoisomerases-DNA interactions is substantial, that on interactions in the context of chromatin is still scarce; whether eukaryotic DNA topoisomerase II has a structural role in the organisation of interphase and/or metaphase chromosomes, for example, is yet to be settled.

By the 1970s, it became clear that — although the informational content of the genetic code was embodied in a linear array of bases — it was the three-dimensional structure and the topological condensation in the chromatin-like assembly of the DNA double helix in the chromosomes that ultimately would govern its physiological functions in the cells. This is very likely the crucial point. As an illustration of this point, in perhaps the most striking biological example of ‘forms dictates function’, the two complementary parental strands of DNA must separate during semi-conservative replication in order to act as the templates for each of the two newly synthesised daughter strands. This discovery leads to the realisation that the structure of DNA, while elegant, burdened the cell with previously unimagined topological problems. Although these topological problems were originally recognised only for circular molecules, because of the long length of chromosomal DNA, we now know that they apply to linear genomes as well.

The key for finding the solution of these problems seems to lie in the following issues: (1) In the conformational, organisational and biological roles of the topoisomerases that, because of their extreme structural and functional complexity, still remains in part to be elucidated. (2) In the DNA supercoiling process, because it links the biological activity of DNA to its tertiary structure and not just its sequence.

All essential cellular processes seem to be related to the way in which supercoiling is realised. (3) In the three-dimensional organisation of the chromatin, which is a nucleoprotein complex and the stuff chromosomes are made of. This organisation not only compacts the DNA but also plays a fundamental role in regulating interactions with the DNA during its metabolism.

12. Conclusion

A number of theoretical and experimental new findings suggest that the secrets of life and what allows the biological growth of all organisms maybe lies in topology, namely in the fact that forms possess the capacity to convert dynamically structures and functions one into another. In fact, the topological compaction of our genome, the basic building block of which is the nucleosome (a protein-DNA structure), provides a whole repertoire of information in addition to that furnished by the genetic code. This mitotically stable information is not inherited genetically and is termed epigenetic. Epigenetic phenomena are propagated alternative states of gene expressions, and alternative states of protein folding, and they are closely linked with histone and chromatin modifications. Still more than genetics events, one could say that the comprehension of epigenetic processes essentially requires a better understanding of the role played by topological transformations.

One of the challenges in chromatin research is to understand how levels of chromosome organisation beyond the 30 nm chromatin filaments condense to form the cell metaphase chromosome. We need very likely a topological model that accounts for the several ordered transformations that are required for the dimensions of metaphase chromosomes, which are 10,000-fold shorter and 400- to 500-fold thicker than the double stranded DNA helices contained within them. Loop-like arrangement of chromatin and its stacking into a cylinder of 800 to 1000 nm in its thickness, which is in good agreement with the diameter of the metaphase chromosome, and twisting the cylinder into a superhelix would further compact it, is a model that account well for the corkscrew appearance of metaphase chromosome.

Happily cells achieve this tight packing of DNA while still maintaining the chromosome in a form that allows regulatory proteins to gain access to the DNA to turn on (or off) specific genes or to duplicate the chromosomal DNA. This means that all epigenetic states and processes have to be established, inherited, controlled and modified in such a way as to permit that their integrity is maintained while preserving the possibility of deformability. Thus, the topological plasticity of the many-levels structure of chromosomes, the chromatin dynamics and the gene's regulatory modifications are intimately interconnected processes and determinant factors of cellular and development organisation.

Condensation of genetic material appears to be a very fundamental mechanism of life. Now, since condensation realise as a kind of topological embedding of one

space, the restrained linear DNA helicoidal-like surface, into another space, the three-dimensional chromosome structure in the cell's nucleus, it seems reasonable to think that topological embeddings and transformations are dynamic processes that are essential for the maintain and the integrity of life. One demonstration of that is the fact that the exotic supercoiled forms that double helix can assume are tertiary structures which have an important effect on the molecule's secondary structure and its function. DNA and chromosome organisation must fulfill precise topological prerequisite in order to achieve certain functional processes. In particular, DNA transcription and replication can both be enhanced and regulated by topological supercoiling. It now appear clear, for example, that for replication to be completed, the linking number of the DNA, Lk , must be reduced from its vast (+) value to exactly zero. In bacteria, DNA gyrase introduces (-) supercoils and thereby removes parental Lk . Moreover, in certain cases, the severity of the phenotype can be controlled by changing the level of supercoiling in the cell.

We have thus three interrelated theoretical and experimental facts, which we would like to stress: (1) DNA condensation is a driving force for double helix unlinking and chromosome portioning, by folding, in topological domains. (2) Condensation is achieved by supercoiling, which is a topological state of macromolecules enhanced by three kinds of deformations (embeddings): twisting, writhing and knotting. If the DNA is modeled as a ribbon in three-space whose axis is not flat in the plane, we can define the *twist* of the ribbon abstractly as the integral of the incremental twist of the ribbon about the axis, integrated as we traverse the axis once; so it simply measures how much the ribbon twists about the axis from the frame of reference of the axis: it need not be an integer. The *writhe* measures how much the axis of the ribbon is contorted in space. Because (-) supercoiling in bacteria arises from a topological misalignment and not a protein corset, it has the flexibility to do work. (3) Supercoiling results from topological strain and the contortion of DNA by proteins, notably the nucleosomal histone octet and the structural maintenance of chromosomes (SMC) proteins.

To conclude, we would like to say few words on the general philosophy which underpins this work. We tried to explore new mathematical modelling in a variety of biological problems, paying a particular attention to the possibilities of a geometrical and topological description of biological systems such as that of chromatin and chromosome. We showed that its study involves the simultaneous integration of different geometrical concepts and biological components and their relationships with one another. A multilevel and integrative approach has to essentially take into account the fact that simply knowing the parts list of genes and proteins does not tell us much about how life's many biological processes work. The cellular organisation is a complex dynamic system with hundreds of thousands of bio-molecules interacting with one another to execute life's many functions. Developments in the mathematical and physical sciences will be very important for addressing complex questions in biology. In the view of these facts, one may foresee that a great

deal of the future research on the interface between mathematics and life sciences will relate to the following two fundamental issues: How did the topology of the double-helix and DNA-proteins complexes evolve and why it is so biologically important for the integrity of cells and organisms? These questions arise immediately from the crucial recognition that the topology and dynamics of DNA and macromolecular proteins complexes are essential for the maintenance and integrity of life.

References

- [1] Adams, C. C. 2000, *The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots*. W. H. Freeman, New York.
- [2] Alberts, B. 2003, DNA replication and recombination. *Nature*, 421, 431–435.
- [3] Alberts, B., Johnson, A., Lewis, J. *et al.*, 2002, *Molecular Biology of the Cell*. 4th edn. Garland Science, New York.
- [4] Banchoff, T. 1976, Self-linking numbers of space polygons. *Indiana Univ. Math.*, 25, 1171–1188.
- [5] Banchoff, T., White, J. 1975, The behaviour of the total twist and the self-linking number of a closed space curve under inversions. *Math. Scandinavica*, 36, 254–262.
- [6] Bates, A. D., Maxwell, A. 1993, *DNA Topology*. Oxford University Press, Oxford.
- [7] Becker, P. B., Längst, G. 2004, Nucleosome remodeling: one mechanism, many phenomena. *Biochimica et Biophysica Acta*, 1677, 58–63.
- [8] Bednar, J., Horowitz, R., Grigoryev, S. A. *et al.*, 1998, Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-folding and compaction of chromatin. *Proc. Natl. Acad. Sci. USA*, 95, 14173–14178.
- [9] Belmont, A. S. 2002, Mitotic chromosome scaffold structure: New approaches to an old controversy. *Proc. Natl. Acad. Sci. USA*, 99, 15855–15857.
- [10] Belmont, A. S., Bruce, K. 1994, Visualisation of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. *J. Cell Biol.*, 127, 287–294.
- [11] Belmont, A. S., Dietzel, S., Nye, A. C. *et al.*, 1999, Large-scale chromatin structure and function. *Curr. Opin. Cell Biol.*, 11, 307–326.
- [12] Ben Haim, E., Lesne, A. and Victor, J. M. 2002, Adaptive elastic properties of chromatin fiber. *Physica A*, 314, 592–599.
- [13] Benjamin, H. W., Cozzarelli, N. R. 1990, Geometric arrangements of Tn3 resolvases sites. *J. Biol. Chem.*, 265, 6441–6447.
- [14] Benjamin, H. W., Matzuk, M. M., Krasnov, M. A. *et al.*, 1985, Recombination site selection by Tn3 resolvase: topological tests of a tracking mechanism. *Cell*, 40, 147–158.
- [15] Berger, J. M. 1998, Structure of DNA topoisomerases. *Biochimica et Biophysica Acta*, 1400, 3–18.
- [16] Berger, J. M., Gamblin, S. J., Harrison, S. C. *et al.*, 1996, Structure and Mechanism of DNA Topoisomerase II. *Nature*, 379, 225–232.

- [17] Bi, X., Broach, J. R. 1997, DNA in Transcriptionally Silent Chromatin Assumes a Distinct Topology That Is Sensitive to Cell Cycle Progression. *Mol. Cell. Biol.*, 17, 7077–7087.
- [18] Boi, L. 2005, Topological knots models in physics and biology. In L. Boi, ed. *Geometries of Nature, Living Systems and Human Cognition*. World Scientific, Singapore, 203–278.
- [19] Boi, L. 2006, Mathematical knot theory. In J.-P. Francoise, G. Naber and T. S. Sun, eds. *Encyclopedia of Mathematical Physics*. Elsevier, Oxford, 399–406.
- [20] Boi, L. 2006, Topological knot theory and macroscopic physics. In Francoise, J.-P., Naber, G. and Sun, T. S., eds. *Encyclopedia of Mathematical Physics*. Elsevier, Oxford, 271–277.
- [21] Boi, L. 2007, Geometrical and topological modeling of supercoiling in supramolecular structures. *Biophysical Reviews and Letters*, 2, 287–299.
- [22] Boi, L. 2007, Sur quelques propriétés géométriques et topologiques des processus biologiques et des systèmes vivants. *Bull. Hist. Epist. Sci. Vie*, 14, 71–113.
- [23] Boi, L., 2011, Geometry of dynamical systems and topological stability: from bifurcations, chaos and fractals to dynamics in the natural and life sciences. *International Journal of Bifurcation and Chaos*, 21(1).
- [24] Boi, L., 2011, Plasticity and complexity in biology: topological organisation, regulatory proteins networks and mechanisms of genetic expression. Towards new vistas in the life sciences. In G. Terzis and R. Arp, eds. *The Concept of Information in the Biological Sciences*. The MIT Press, Cambridge, MA.
- [25] Boles, T. C., White, J. H. and Cozzarelli, N. R. 1990, The structure of plectonemically supercoiled DNA. *J. Molec. Biol.*, 213, 931–951.
- [26] Bott, R., Taubes, C. 1994, On the self-linking of knots. *J. Math. Phys.*, 35, 5247–5287.
- [27] Boy de la Tour, E., Laemmli, U. K. 1988, The metaphase scaffold is helically folded: sister chromatids have predominantly opposite helical handedness. *Cell*, 55, 937–944.
- [28] Brown, P. O., Cozzarelli, N. R. 1979, A sign inversion mechanism for enzymatic supercoiling of DNA. *Science*, 206, 1081–1083.
- [29] Buck, G., Simon, J. 1999, Thickness and crossing number of knots. *Topology and its Applications*, 91, 245–257.
- [30] Burde, G., Zieschang, H. 1985, *Knots*. Walter de Gruyter, Berlin.
- [31] Calgareanu, G. 1959, L'intégrale de Gauss et l'analyse des nœuds tridimensionnels. *Rev. Math. Pures Appl.*, 4, 5–20.
- [32] Calladine, C. R., Drew, H. R., Luisi, B. F. *et al.*, 2004, *Understanding DNA: The Molecule and How It Works*. 3rd edn. Elsevier, London.
- [33] Cerf, C. 1998, A note on tangle model for DNA recombination. *Bull. Math. Biol.*, 60, 67–78.
- [34] Cerf, C., Stasiak, A. 2000, A topological invariant to predict the three-dimensional writhe of ideal configurations of knots and links. *Proc. Natl. Acad. Sci. USA*, 97, 3795–3798.
- [35] Champoux, J. J. 2001, DNA topoisomerases: structure, function, and mechanism. *Annu. Rev. Biochem.*, 70, 369–413.

- [36] Chern, S-S. 1979, From triangles to manifolds. *Amer. Math. Monthly*, 86, 339–349.
- [37] Cluzel, P., Lebrun, A., Heller, C. *et al.*, 1996, DNA: an extensible molecule. *Science*, 271, 792–794.
- [38] Conway, J. 1970, An enumeration of knots and links and some of their related properties. In J. Leech, ed. *Computational Problems in Abstract Algebra*. Pergamon Press, Oxford, 329–358.
- [39] Cozzarelli, N. R. 1992, Evolution of DNA topology: implications for its biological roles. In Sumners, D. W. L. ed. *New scientific applications of geometry and topology. Proceedings of Symposia in Applied Mathematics. Vol. 45*. American Mathematical Society.
- [40] Cremer, T., Cremer C. 2001, Chromosome territories, nuclear architecture and gene regulation in mammalian cells, *Nat. Rev. Genet.*, 2, 292–301.
- [41] Cremer, T., Küpper, K., Dietzel, S. *et al.*, 2004, Higher order chromatin architecture in the cell nucleus: on the way from structure to function. *Biol. Cell.*, 96, 555–567.
- [42] Dean, F. B., Stasiak, A., Koller, T. *et al.*, 1985, Duplex DNA knots produced by *escherichia coli* topoisomerase I. *J. Biol. Chem.*, 260, 4975–4983.
- [43] DePamphilis, M. L. ed. 1999, *Concepts in Eukaryotic DNA Replication*. Cold Spring Harbor Laboratory Press, New York.
- [44] Edwards, S. F. 1998, Entanglements of polymers. In Whittington, S. G., Sumners, D. W. L. and Lodge, T., eds. *Topology and Geometry in Polymer Science*. Springer, New York, 1–8.
- [45] Eichler, E. E., Sankoff, D. 2003, Structural dynamics of eukaryotic chromosome evolution. *Science*, 301, 793–797.
- [46] Eils, R., Dietzel, S., Bertin, E. *et al.*, 1996, Three-dimensional reconstruction of painted human interphase chromosomes: active and inactive X chromosome territories have similar volumes but differ in shape and surface structure. *J. Cell Biol.*, 135, 1427–1434.
- [47] Elgin, S. C. R., Workman, J. L. eds. 2000, *Chromatin Structure and Gene expression*. 2nd edn. Oxford University Press, Oxford.
- [48] Ernst, C., Sumners, D. W. L. 1990, A calculus for rational tangles: applications to DNA recombination. *Math. Proc. Cambr. Phil. Soc.*, 108, 489–515.
- [49] Finch, J. T., Klug, A. 1976, Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. USA*, 73, 1897–1908.
- [50] Francis, N. J., Kingston, R. E. and Woodcock, C. L. 2004, Chromatin compaction by a polycomb group protein complex. *Science*, 306, 1574–1577.
- [51] Fuller, F. B. 1971, The writhing number of a space curve. *Proc. Natl. Acad. Sci. USA*, 68, 815–819.
- [52] Fuller, F. B. 1978, Decomposition of the linking number of a closed ribbon: a problem from molecular biology. *Proc. Natl. Acad. Sci. USA*, 75, 3557–3572.
- [53] Gasser, S. M., Amati, B. B., Cardenas, M. E. *et al.*, 1989, Studies on scaffold attachment sites and their relationship to genome function. *Int. Rev. Cytol.*, 119, 57–96.
- [54] Gasser, S. M., Laemmli, U. K. 1986, The organisation of chromatin loops: characterisation of a scaffold attachment site. *EMBO J.*, 5, 511–518.

- [55] Gavin, I., Horn, P. J. and Peterson, C. L. 2001, SWI/SNF chromatin remodeling requires changes in DNA topology. *Mol. Cell*, 7, 97–104.
- [56] Gilbert, N., Boyle, S., Fiegler, H. *et al.*, 2004, Chromatin architecture of the human genome; gene-rich domains are enriched in open chromatin fibers. *Cell*, 118, 555–566.
- [57] Görisch, S. M., Lichter, P. and Rippe, K. 2005, Mobility of multi-subunit complexes in the nucleus: accessibility and dynamics of chromatin subcompartments. *Histochem. Cell Biol.*, 123, 217–228.
- [58] Grange, T., Paques, F. 2002, Architecture du noyau et régulation transcriptionnelle. *Médecine/Sciences*, 18, 1245–1256.
- [59] Hirano, T., Mitchison, T. J. 1993, Topoisomerase II does not play a scaffolding role in the organisation of mitotic chromosomes assembled in xenopus egg extracts. *J. Cell Biol.*, 120, 601–612.
- [60] Holmes, V. F., Cozzarelli, N. R. 2000, Closing the ring: Links between SMC proteins and chromosome partitioning, condensation, and supercoiling. *Proc. Natl. Acad. Sci. USA*, 97, 1322–1324.
- [61] Kauffman, L. H. 1987, *On Knots*. Princeton University Press, Princeton.
- [62] Kauffman, L. H. 2001, *Knots and Physics*. Series on Knots and Everything Vol. 1, 3rd edn. World Scientific, Singapore.
- [63] Kawachi, A. 1996, *A Survey of Knot Theory*. Birkhäuser, Basel.
- [64] Kimura, K., Cuvier, O. and Hirano, T. 2001, Chromosome condensation by a human condensing complex in *Xenopus* egg extracts. *J. Biol. Chem.*, 276, 5417–5420.
- [65] Kireeva, N., Lakonishok, M., Kireev, I. *et al.*, 2004, Visualisation of early chromosome condensation: a hierarchical folding, axial glue model of chromosome structure. *J. Cell Biol.*, 166, 775–785.
- [66] Kornberg, R. D., Thomas, J. O. 1974, Chromatin structure: a repeating unit of histones and DNA. *Science*, 184, 865–868, 868–871.
- [67] Langer, J., Singer, D. A. 1984, Knotted elastic curves in R^3 . *J. London Math. Soc.*, 30, 512–520.
- [68] Laurie, B., Katritch, V., Sogo, J. *et al.*, 1998, Geometry and physics of catenanes applied to the study of DNA replication. *Biophysical Journal*, 74, 2815–2822.
- [69] Lesliet, A. G. W., Arnott, S., Chandrasekaran, R. *et al.*, 1980, Polymorphism of DNA double helices. *J. Mol. Biol.*, 143, 49–72.
- [70] Lickorish, W. B. R. 1981, Prime knots and tangles. *Trans. Amer. Math. Soc.*, 267, 321–332.
- [71] Lickorish, R. W. B. 1997, *An Introduction to Knot Theory*. Springer, New York.
- [72] Lodish, H., Berk, A., Zipursky, S. L. *et al.*, 2000, *Molecular Cell Biology*. 4th edn. W. H. Freeman and Company, New York.
- [73] Lutter, L. C., Judis, L. and Paretto, R. F. 1992, Effects of histone acetylation on chromatin topology *in vivo*. *Mol. Cellul. Biol.*, 12, 5004–5014.
- [74] Maggioni, F., Ricca, R. L. 2006, Writhing and coiling of closed filaments. *Proc. R. Soc. A*, 462, 3151–3166.
- [75] Mahy, N. L., Bickmore, W. A., Tumber, T. *et al.*, 2000, Linking large-scale chromatin structure with nuclear function. In Elgin, S. C. R., Workman, J. L. eds. *Chromatin Structure and Gene Expression*. 2nd edn. Oxford University Press, Oxford, 300–321.

- [76] Manuelidis, L. 1990, A view of interphase chromosomes. *Science*, 250, 1533–1538.
- [77] Marko, J. F. 1997, Supercoiled and braided DNA under tension. *Phys. Rev. E.*, 55, 1758–1772.
- [78] Massey, M. 1974, *An Introduction to Algebraic Geometry*. Springer, New York.
- [79] Milnor, J. 1950, On the total curvature of knots. *Ann. Math.*, 52, 248–257.
- [80] Moise, E. E. 1954, Invariance of the knot-types; local tame imbedding. *Ann. of Math.*, 59, 159–170.
- [81] Murasugi, K. 1996, *Knot Theory and its Applications*. Birkhäuser, Boston.
- [82] Papakyriakopoulos, C. D. 1957, On Dehn's lemma and the asphericity of knots. *Ann. of Math.*, 66, 1–26.
- [83] Pikaard, C. S. 1998, Chromosome topology—organising genes by loops and bounds. *The Plant Cell*, 10, 1229–1232.
- [84] Pohl, W. F. 1968, The self-linking number of a closed space curve. *J. Math. Mech.*, 17, 975–985.
- [85] Pohl, W. F., Roberts, G. W. 1978, Topological considerations in the theory of replication of DNA. *J. Math. Biology*, 6, 383–402.
- [86] Pohl, W. F. 1980/81, DNA and differential geometry. *The Mathematical Intelligencer*, 3, 20–27.
- [87] Postow, L., Crisona, N. J., Peter, B. J. *et al.*, 2001, Topological challenges to DNA replication: conformations at the fork. *Proc. Natl. Acad. Sci. USA*, 98, 8219–8226.
- [88] Qiu, W.-Y. 2000, Knot theory, DNA topology, and molecular symmetry breaking. In Bonchev, D. and Rouvray, D. H., eds. *Chemical Topology. Applications and Techniques*. Gordon & Breach Science Publ., Amsterdam, 175–237.
- [89] Richmond, T. J., Davey, C. A. 2003, The structure of DNA in the nucleosome core. *Nature*, 423, 145–150.
- [90] Richmond, T. J., Widom, J. 2000, Nucleosome and chromatin structure. In S. C. R. Elgin and J. L. Workman, eds. *Chromatin Structure and Gene Expression*. 2nd edn. Oxford University Press, Oxford, 1–23.
- [91] Ridgway, P., Almouzni, G. 2001, Chromatin assembly and organisation. *J. Cell Sci.*, 114, 2711–2713.
- [92] Roca, J. 1995, The mechanisms of DNA topoisomerases. *Trends Biochem. Sci.*, 20, 156–150.
- [93] Roca, J., Wang, J. C. 1996, The probabilities of supercoil removal and decatenation by yeast DNA topoisomerase II. *Genes to Cells*, 1, 17–27.
- [94] Roca, J., Berger, J. M., Harrison, S. C. *et al.*, 1996, DNA transport by a type II topoisomerase: direct evidence for a two-gate mechanism. *Proc. Natl. Acad. Sci. USA*, 93, 4057–4062.
- [95] Rolfsen, D. 1990, *Knots and Links*. Mathematics Lectures Series 7. Publish or Perish, Inc., Houston, Texas.
- [96] Sedat, J., Manuelidis, L. 1978, A direct approach to the structure of mitotic chromosomes, *Cold Spring Harb. Symp. Quant. Biol.*, 42 1978, 331–350.
- [97] Stasiak, A. 2000, DNA topology: feeling the pulse of a topoisomerase. *Current Biology*, 10, R526–R528.
- [98] Strick, T. R., Allemand, J.-F., Bensimon, D. *et al.*, 1996, The elasticity of a single supercoiled DNA molecule. *Science*, 271, 1835–1837.

- [99] Strick, T. R., Allemand, J.-F., Bensimon, D. *et al.*, 1998, Behaviour of supercoiled DNA. *Biophysical Journal*, 74, 2016–2028.
- [100] Sumners, D. W. L. 1992, Knot theory and DNA. In Sumners, D. W. L. ed. *New scientific applications of geometry and topology. Proceedings of Symposia on Applied Mathematics*. Vol. 45. Amer. Math. Soc., Providence, 39–72.
- [101] Sumners, D. W. L. 1987, The role of knot theory in DNA research. In McCrory, C. and Shifrin, T., eds. *Geometry and topology, manifolds, varieties, and knots*. New York, Marcel Dekker, 297–318.
- [102] Vinograd, J., Lebowitz, J., Radloff, R. *et al.*, 1965, The twisted circular form of polyoma viral DNA. *Proc. Natl. Acad. Sci. USA*, 53, 1104–1111.
- [103] Vologodskii, A. V. 1992, *Topology and Physics of Circular DNA*. CRC Press, Boca Raton, FL.
- [104] Vologodskii, A. V., Zhang, W., Rybenkov, V. V. *et al.*, 2001, Mechanism of topology simplification by type II DNA topoisomerases. *Proc. Nat. Acad. Sci. USA*, 98, 3045–3049.
- [105] Vrána, O., Boudny, V. and Brabec, V. 1996, Superhelical torsion controls DNA interstrand cross-linking by antitumor cis-diamminedichloroplatinum(II). *Nucleic Acids Research*, 24, 3918–3925.
- [106] Wang, J. C. 1996, DNA topoisomerases. *Ann. Rev. Biochem.*, 65, 635–692.
- [107] Wang, J. C., Liu, L. F. 1990, DNA replication: topological aspects and the roles of DNA topoisomerases. In Cozzarelli, N. R. and Wang, J. C., eds. *DNA Topology and its Biological Effects*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 321–340.
- [108] Wasserman, P. M., Wolffe, A. P. eds. 1999, *Methods in Enzymology*. Vol. 304: *Chromatin*. Academic Press, London.
- [109] White, J. H. 1969, Self-linking and the Gauss integral in higher dimensions. *Amer. J. Math.*, 91, 693–728.
- [110] White, J. H. 1989, *An Introduction to the Geometry and Topology of DNA Structure*. CRC Press, Boca Raton.
- [111] White, J. H., Bauer, W. R. 1986, Calculation of the twist and the writhe for representative models of DNA. *J. Mol. Biol.*, 189, 329–341.
- [112] White, J. H., Millett, K. C. and Cozzarelli, N. R. 1987, Description of the topological entanglement of DNA catenanes and knots by a powerful method involving strand passage and recombination. *J. Mol. Biol.*, 197, 585–603.
- [113] Whitehead, J. H. C. 1937, On double knots. *J. London Math. Soc.*, 12, 63–71.
- [114] Whitney, H. 1937, On regular closed curves in the plane. *Comp. Math.*, 4, 276–284.
- [115] Widom, J. 1998, Structure, dynamics, and function of chromatin in vitro. *Annu. Rev. Biophys. Biomol. Struct.*, 27, 285–327.
- [116] Widom, J., Klug, A. 1985, Structure of the 3000Å chromatin filament: X-ray diffraction from oriented samples. *Cell*, 43, 207–213.
- [117] Wolffe, A. 1994, Nucleosome positioning and modifications: chromatin structures that potentiate transcription. *Trends Biochem. Sci.*, 19, 240–244.
- [118] Wolffe, A. 2000, *Chromatin: Structure and Function*. Academic Press, London.
- [119] Wu, C. 1997, Chromatin Remodeling and the Control of Gene Expression. *J. Biol. Chem.*, 272, 28171–28174.

- [120] Yokota, H., van den Engh, G., Hearst, J. E. *et al.*, 1995, Evidence for the Organisation of Chromatin in Megabase Pair-sized Loops Arranged along a Random Walk Path in the Human G0/G1 Interphase Nucleus. *J. Cell Biol.*, 130, 1239–1249.
- [121] Zechiedrich, E. L., Cozzarelli, N. R. 1995, Roles of topoisomerase IV and DNA gyrase in DNA unlinking during replication in *Escherichia coli*. *Genes Dev.*, 9, 2859–2869.
- [122] Zlatanova, J., Leuba, S. H. and van Holde, K. 1998, Chromatin Fiber Structure: Morphology, Molecular Determinants, Structural Transitions. *Biophysical Journal*, 74, 2554–2566.

This page is intentionally left blank

About the Contributors

Claudio Bartocci (Ph.D. in mathematics, University of Warwick, UK, 1993) is associate professor at the University of Genova, where he teaches mathematical physics and history of mathematics. He has had visiting positions at the State University of New York at Stony Brook, the Université de Paris VII, the University of Philadelphia, and the École de Hautes Études en Sciences Sociales, Paris. Among his more recent publications are “Fourier–Mukai and Nahm Transforms in Geometry and Mathematical Physics” (co-authors: U. Bruzzo and D. Hernández Ruipérez, Birkhäuser, 2009), and “Vite Matematiche” (co-authors: R. Betti, A. Guerraggio and R. Lucchetti, Springer, 2007); he is the co-editor, together with P. Odifreddi, of “La Matematica” (Einaudi). He is presently completing a book, “Mostri Matematici” (Cortina Editore, Milano), about the conceptual development of 19th/early 20th century mathematics, with special emphasis on the role of pathological examples.

Ugo Bruzzo is Professor of Geometry in the Mathematical Physics Sector of SISSA (Trieste). Its research and teaching are in the area of algebraic and differential geometry and in the geometric methods in string and quantum field theory. Main research interests: moduli spaces of framed sheaves, instanton counting, principal and vector (Higgs) bundles on projective and Kaehler manifolds, geometric methods in string and quantum field theory, supermanifolds. Since November 1, 2003 he is the Editor-in-Chief of the Journal of Geometry and Physics. He has coauthored, with C. Bartocci and D. Hernandez Ruiperez, the books: “The Geometry of Supermanifolds” (Kluwer, 1991), and “Fourier–Mukai and Nahm Transforms in Geometry and Mathematical Physics”, Progress in Mathematics 276 (Birkhäuser, 2009). He has had visiting positions at Princeton University, Universidad of Salamanca, City University Of New York, Université Paris VII, Tata Institute of Fundamental Research Mumbai, University of Pennsylvania (Philadelphia), Institut des Hautes Etudes Scientifiques (Bures-sur-Yvette, France).

Mauro Carfora is full professor of mathematical physics at the University of Pavia in the Department of Theoretical and Nuclear Physics. His research interests concern Regge calculus, topological field theory, relativistic cosmology, and the relations between Ricci flow and theoretical physics. He has contributed to several books: “General Relativity in Gravitational Physics” (World Scientific, 1996), “Integrable Systems and Quantum Groups” (World Scientific, 1992), “The Geometry of Dynamic Triangulations” (Springer, 1997). He is the author of several publications in the fields of general relativity, quantum field theory and string theory.

Ugo Moschella graduated in Physics at Bologna University in 1985 and obtained a Ph.D. at the International School for Advanced Studies (ISAS) in Trieste in 1991. Since 2001 he is an Associate Professor of Theoretical Physics at the University of Insubria (Como, Italy). Since 2000, he is a member of the scientific board of the Italian Society of Gravitation. He is a referee for *Physical Review* and *Physical Review Letters*, *Classical and Quantum Gravity*, *Journal of Physics*, *Journal of Mathematical Physics*. Ugo Moschella’s main research interests are in Quantum Field Theory. His contributions are concerned with the infrared problem in gauge theory, quantum field theory on curved spacetimes with special attention to de Sitter and anti-de Sitter field theories, and the role of diffeomorphism groups in quantum physics. Among his several coauthored books are “Modern Cosmology” (Institute of Physics Publishing, 2002), and “Geometry and Physics of Branes” (Institute of Physics Publishing, 2003).

Jean-Pierre Luminet is a French astrophysicist, who specializes in black holes and cosmology. He works as research director for the CNRS (Centre National de la Recherche Scientifique), and is a member of the Laboratoire Univers et Théories (LUTH) of the Observatory of Paris-Meudon. Its main research interests are in astrophysics and cosmology, the topology of the universe and the history and philosophy of physical sciences. He has been awarded by the Georges Lemaître Prize (1999), and the International G. B. Lacchini Prize (2008), and the International Astronomical Union membership. He is the author of many books, among them are “Black Holes” (Cambridge University Press, 1992), and *The Wraparound Universe* (AK Press, 2008).

Leonardo Fogassi is born in La Spezia. He took a first degree in Biological Sciences at the University of Pisa in 1982. In 1989 He took a Ph.D. in

Neuroscience at the University of Parma. From 1990 to 1999 he worked as a researcher in the Institute of Human Physiology of Parma. He is currently Full Professor of Human Physiology at the University of Parma. His major research interests concerns the sensori-motor transformations operated by the cerebral cortex and the role of motor cortex in action perception. He has studied these topics by means of neurophysiological and behavioral techniques. His publications are related to the neurophysiology of space perception, coding of prehension movements, and action understanding.

Anna Berti is born on the 27th June, 1957 in Milan. She obtained a degree in Medicine and Surgery in 1983 at the University of Milan, another degree in Neurology in 1987 at the same university, and a Ph.D. in Neuroscience at the University of Parma. She is Full Professor of Neuropsychology in the Department of Psychology at the University of Turin. Her research fields are Neuropsychology and cognitive neuroscience, and her main interests are in the fields of spatial cognition, motor cognition and neurobiology of consciousness. She authored and coauthored many research articles that appeared in *Science*, *Brain*, *Cortex*, *Journal of Experimental Psychology*, *Behavioral Neurosciences*, and *Journal of Cognitive Neurosciences*. She has been a researcher in the Department of Experimental Psychology at the University of Oxford.

Alessia Folegatti is a junior researcher in the Department of Neuropsychology, University of Turin. Her main research interests are in the cognitive sciences and the neuropsychological aspects of perception (spatial cognition and memory, spatial neglect, motor cognition). She has published several articles on these topics.

Claudio Brozzoli is a junior researcher in cognitive neurosciences at the UNSERM U864 “Espace et Action” Lyon, France. He is interested mainly in the problem of multisensory interactions and the role of action in the construction of peripersonal space. He has published several articles on these topics.

Alessandro Farnè (Ph.D. in Experimental Psychology, University of Bologna, 1999). He was Assistant Professor at the University of Bologna and he is currently a senior researcher at the INSERM, Unit 864 “Espace et Action”, Lyon, France. His main research interests are in multisensory perception and action, multisensory extinction and neglect, body

representations, tool-use and cerebral plasticity, and self–other recognition. He chairs several advanced research project on these topics. He has published extensively in neuropsychology and cognitive sciences.

Corrado Sinigaglia is currently Professor of Philosophy of Science at the University of Milan. He studied at the Husserl-Archives of Leuven (1992–1993), at Ecole Normale Supérieure (1994), and at the University of Genova (1995–1999) where he obtained a Ph.D. in Philosophy of Science. His fields of research are Neuroscience, phenomenology, and philosophy of mind. He is currently working on the enactive roots of social cognition, and he develops a motor approach to intentionality. He is the author of several articles that appeared in international journals of neurosciences and philosophy of mind, and of the books: (with G. Rizzolatti) “Mirrors in the Brain” (Oxford University Press, 2008), “La Seduzione dello Spazio” (Unicopoli, 2000). He is member of the Editorial Board of several journals.

Chiara Brozzo is a junior researcher at the University of Milan. Her main research interests are in the Philosophy of Psychology and Neurosciences. She has been visiting Oxford University where he developed research on the perception of space.

Francis Bailly (1950–2009) was a physicist working as senior researcher at the CNRS, Laboratoire de Physique des Solides, Meudon. His main interests have been in the concepts of symmetry and breaking symmetries, in the interface between physics and biology, and in the theories of space-time in relativistic physics and quantum mechanics. He has also made significant contributions in the foundational mathematics and physics and in epistemological aspects of cognitive sciences. He has published extensively on all these topics.

Giuseppe Longo took a degree in Mathematics at the University of Pisa. He is currently Director of research at the Ecole Normale Supérieure in Paris. His principal fields of interest are in logic and theory of computation; denotational semantics and lambda-calculus; type theory, category theory and their applications to computer science, interfaces mathematics, physics, biology, philosophy of mathematics and cognitive sciences. He has been a visiting professor at the Oxford University, the University of California at Berkeley and the Carnegie Mellon University. He is editor of several international journals in the computer sciences, cognitive sciences and the philosophy of

science. He has been invited to lecture at 30 international conferences and at about 150 seminar talks in Universities or Research Institutions in Europe, USA and Asia. He is Member of the Academia Europea, since 1992. He has published tens of articles and two books: (with A. Asperti) “Categories, Types and Structures” (M.I.T. Press, 1991), (with F. Bailly) “Mathématiques et Sciences de la Nature. La Singularité Physique du Vivant” (Hermann, 2006).

Riccardo Broglia (Ph.D. in Physics, University of Cuyo, Argentina, 1964) is Professor of Physics at the University of Milan and Adjunct Professor in the Niels Bohr Institute, University of Copenhagen. His main research interests are in nuclear and solid physics, protein folding and the geometrical modeling of cell networks and biochemical processes and functions. He has been a Visiting Professor in many universities throughout the world: University of Tennessee (USA), University of Coimbra, State University of New York at Stony Brook, Université Catholique de Louvain, He is the Head of the Nuclear Theory Group, Department of Physics, University of Milan. He has been Associate Editor of *Nuovo Cimento* and of *Nuclear Physics News*. He directed a number of Schools of Physics and he has been a member of the Advisory Committee of several International Conferences. He has published hundreds of research articles in international journals of physics and biophysics. He has been the Editor of many proceedings of international conferences and he gave many invited talks and lectures to international meetings.

Jean-François Sadoc is Professor of Physics at the University of Orsay Paris-Sud. He conducts research in the Laboratory of Solid Physics. The main interests are in theoretical physics, soft matter physics, the protein folding and the topology of polymers. His work shows that the concept of geometrical frustration can be used to elucidate the structure and properties of non-periodic materials such as metallic glasses, quasi-crystals, amorphous semiconductors and complex liquid crystals. He is the author of numerous research articles and of the book “Geometrical Frustration” (Cambridge University Press, 1999).

Luciano Boi is Associate Professor of Geometry, Scientific Theorization and Natural Philosophy at the Ecole des Hautes Etudes en Sciences Sociales, in the Centre de Mathématiques (Paris). He studied mathematics, physics and philosophy at the Universities of Bologna, Paris and Berlin, and received

his Doctorate as well as his Habilitation from the EHESS (Paris). He has been a guest professor and a member of several universities and research institutes, including the IHES (Bures-sur-Yvette), the IAS (Princeton), the University of Cambridge (UK), Université de Montreal, the SISSA (Trieste) and the University of Padua. He has been invited lecturer in numerous international conferences. He has received several awards, such as a 1997 award from the Guggenheim Foundation (New York) and a 2000 award from the Singer–Polignac Foundation (Paris). His main research interests include various aspects of mathematics, the interactions between topology and biology, the geometrical modeling of spatial perception, as well as the history and philosophy of sciences. He is the author of numerous books and articles on these topics.

- α -helix, 230
- action, 134
- AdS invariance, 65
- algorithm, 193
- amino acid sequence, 215
- anti de Sitter universe, 37
- area F4, 112
- arrows, 200
- autopoiesis, 174
- bimodal neurons, 117
- biological processes, 251
- biological systems, 298
- biology, 181
- bodily movements, 163
- body, 107, 131
- body schema, 131
- Borromean rings, 248
- brain, 107
- brain damage, 144
- Calabi–Yau manifolds, 13
- catenanes, 246
- causal regimes, 193
- causal relationships, 174
- causes, 175
- cell, 244
- central nervous system, 144
- cerebral cortex, 122
- chimney spaces, 90
- chiralities, 239
- chromatin, 243
- chromatin fibre, 249
- chromatin folding, 268
- chromatin memory, 269
- chromonema model, 249
- chromosome, 244
- chromosome territories, 249
- Clifford parallels, 223
- cognition, 157
- cognitive neuroscience, 127
- compaction of chromosomes, 272
- condensins, 271
- configuration space, 248
- conformal structure, 24, 25
- conformational flexibility, 269
- connection, 9
- contingent finality, 199
- coordinate system, 107
- cortex, 110
- cortical areas, 130
- cortical motor system, 107, 157
- Cosmic Microwave Background, 81
- cosmic time, 60
- cosmic topology, 83, 95
- cosmological constant, 35
- cosmological models, 94
- cosmological principle, 94
- cosmology, 93
- critical dimensions, 27
- curvature, 9, 37, 95
- curvature of space, 93
- cyclic groups, 91
- dark energy, 35, 36, 97
- de Sitter, 51
- de Sitter geometry, 36
- deformability, 261
- dense structures, 222
- determinations, 175
- diagram, 247

- diffeomorphism group, 20
- differential geometry, 3
- dihedral groups, 91
- disclination, 237
- discrete logistic equation, 195
- DNA, 244
- DNA curvature, 269
- DNA methylation, 249
- DNA sequence, 249
- DNA topology, 246
- DNA transcription, 271
- Donaldson invariants, 10
- dual descriptions, 30
- dualities, 31
- duality symmetries, 32
- dynamic plasticity, 158, 163
- dynamic principles, 243

- Einstein's equations, 94
- embedding, 251
- emergent properties, 211
- enactive approach, 158
- entropy, 196
- enzymes, 244, 246
- epigenetic inheritance, 269
- epigenetic phenomena, 244
- euchromatin, 249
- Euclidean group, 38
- Euclidean spaces, 90
- Euclidean surfaces, 89
- extrapersonal space, 108, 120

- far space, 129
- fibration, 224
- fibre, 224
- fibre bundles, 8
- finite universe, 81
- flat torus, 85
- Fock space, 55
- folding process, 212
- formal determination, 193
- frame of reference, 108, 113
- Friedmann–Lemaître model, 94
- functional nuclear space, 249
- fundamental domain, 87
- fundamental group, 87

- gauge group, 184
- gauge invariance, 6
- gauge symmetry, 8
- gauge theories, 3
- gene expression, 269
- gene regulation, 249
- general relativity, 3
- genetic information, 244
- genome, 244
- geodesic principles, 181
- geometric frustration, 222
- geometrical invariants, 259
- global topology, 93
- goal-centred, 164
- gravitational theory, 93
- Green functions, 71

- helicases, 271
- helices, 226
- helicoidal symmetries, 221
- heterochromatin, 249
- hidden dimensions, 83
- histone modification, 249
- holographic principle, 44
- holomorphic functions, 75
- holonomy group, 88
- homeomorphism, 247
- homotopy groups, 86
- Hopf fibration, 221
- Hopf link, 248
- hydrophobic interaction, 220
- hyperbolic manifolds, 92
- hyperbolic spaces, 93
- hyperbolic surfaces, 89
- hypersphere, 38, 91
- hypertorus, 90

- inflation, 36
- information, 192, 195
- instanton, 9
- intelligibility, 193
- internal representation of space, 107
- invariances, 174

- invariants, 175
- isometry group, 91
- isotopies, 247

- Jones polynomials, 263

- Klein bottle, 91
- Klein–Gordon equation, 51, 55
- Klein–Gordon field, 54
- knot theory, 251
- knots, 243

- Lagrangian, 184
- Laguerre tessellation, 233
- lattice model, 220
- law, 193
- laws of conservation, 192
- Legendre’s function, 73
- lens spaces, 91
- levels of organisation, 174, 199
- Lie group, 8
- links, 247
- linking number, 246
- living phenomena, 174, 199
- Lobachevsky hyperbolic plane, 85
- locomotion, 131
- loop quantum gravity, 21
- loops, 249
- Lorentzian manifolds, 37

- mapping, 225
- mathematical structures, 174
- mathematics, 174
- Maxwell equations, 184
- Minkowski metric, 76
- Minkowski space, 3
- Minkowski spacetime, 37
- mirror symmetry, 30
- moduli space, 10
- molecular biology, 251
- motor acts, 163
- motor behaviour, 138
- motor constitution, 158
- motor cortex, 111
- motor neurons, 112
- motor representation, 116
- motor system, 112
- movements, 107
- multiconnected spaces, 81
- multimodal areas, 130
- multiply connected, 87
- multisensory integration, 138
- multisensory neurons, 139
- multisensory space, 137

- native conformation, 214
- near space, 129
- negative pressure, 35
- neglect, 120
- networks, 199
- neural systems, 134
- neurons, 110
- Noether’s theorem, 184
- non-commutative geometry, 20
- non-Euclidean space, 82
- normal analyticity condition, 71
- nucleosomes, 261

- organism, 199

- parietal cortex, 144
- parietal lobe, 110, 139
- peripersonal space, 113
- personal space, 107
- phase spaces, 199
- phenotype, 298
- plasticity, 137, 200
- plectonemes, 270
- Poincaré group, 76
- Poincaré dodecahedral spherical space, 81
- Poincaré–Euler characteristic, 86
- polyhedral groups, 91
- polyhedral spaces, 92
- polytope, 221
- predictability, 188
- primary motor cortex (F1), 112
- prism spaces, 92
- probability correlations, 181
- product, 258
- protein folding problem, 212

- protein structures, 248
- proteins, 212

- quantum chromodynamics, 10
- quantum cohomology, 13
- quantum conformal field theory, 25
- quantum electrodynamics, 10, 184
- quantum field, 68
- quantum fluctuations, 26
- quantum geometry, 19
- quantum invariants, 181
- quantum mechanics, 12
- quantum physics, 93

- radial-loop models, 249
- randomness, 190
- rational numbers, 268
- recombinase, 258, 264
- recombination, 263
- regular tetrahedra, 226
- regulation, 243
- regulatory systems, 243
- Reidemeister moves, 247
- renormalization, 26
- replication, 246
- retina, 109
- Riemann surface, 13
- Riemannian manifold, 82
- rigidity theorem, 92
- RNA polymerase, 271
- rotation group, 38

- secondary structures, 230
- sensorimotor transformation, 115
- sensory input, 138
- sequences, 248
- shape of space, 81
- simply connected, 87
- site-specific recombination, 264
- skein diagrams, 258
- slab spaces, 90
- solenoidal DNA, 271
- solenoids, 270
- somatosensory receptive fields, 158
- source domain, 202

- space, 107
- space perception, 109
- space representation, 122
- spaceforms, 88
- spacetime manifold, 3
- spatial cognition, 134
- spatial map, 131
- spatial neglect, 144
- spatial organisation, 249
- space representation, 120
- spectral condition, 76
- sphere packing, 221
- spherical objects, 226
- spherical surfaces, 89
- stability, 251
- standard model, 10
- string theory, 11
- structural flexibility, 261
- structural stability, 202
- substrate, 258
- supercoiling, 244
- superior colliculus, 139
- symmetries, 174, 175
- symmetry breakings, 183

- tangle, 264
- target domain, 200
- tetrahedron, 226
- topoisomerases, 244, 246
- topological complexity, 260
- topological deformations, 243
- topological embeddings, 298
- topological form, 250
- topological information, 268
- topological invariant, 86, 211, 246
- topological model, 272
- topological transformations, 251
- topologically equivalent, 86
- topology of space, 81
- toroidal structures, 280
- toroidal supercoils, 271
- torus, 223
- total twist, 254
- twistor programme, 4

universal covering space, 88, 92
unpredictability, 188

vacuum Einstein equations, 36
visual receptive field, 131, 164
visual space, 108
visual stimulus, 108
visual system, 109

Voronoi tessellation, 232

Weeks space, 93
White's formula, 253
Wightman function, 72
writhe, 254
writhing process, 258

NEW TRENDS IN GEOMETRY

THEIR ROLE IN THE NATURAL AND LIFE SCIENCES

This volume focuses on the interactions between mathematics, physics, biology and neuroscience by exploring new geometrical and topological modelling in these fields. Among the highlights are the central roles played by multilevel and scale-change approaches in these disciplines.

The integration of mathematics with physics, as well as molecular and cell biology and the neurosciences, will constitute the new frontier of 21st century science, where breakthroughs are more likely to span across traditional disciplines.

P749 hc

ISBN-13 978-1-84816-642-4
ISBN-10 1-84816-642-7



9 781848 166424

Imperial College Press

www.icpress.co.uk