UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

FACOLTÀ DI INGEGNERIA BIOMEDICA

CORSO DI DOTTORATO IN INGEGNERIA BIOMEDICA

XXVII CICLO

Multiscale Modelling to Unravel the Interplay between Morphogen Gradients and Zonation in the Root Meristem of *A.thaliana*

Micol De Ruvo

SUPERVISORI
Ing. Luisa Di Paola

Università Campus Bio-Medico di Roma

Dr. Sabrina Sabatini

Università Sapienza di Roma

Contents

List of Figures

List of Tables

List of Publications

De Ruvo M., Giuliani A., Paci P., Santoni D., Di Paola L. Shedding light on protein-ligand binding by graph theory: the topological nature of allostery, *Biophysical Chemistry* 165-166, 21-29, 2012

Arrigo N., Paci P., Di Paola L., Santoni D., **De Ruvo M.**, Giuliani A., Castiglione F. Characterizing Protein Shape by a Volume Distribution Asymmetry Index, The Open *Bioinformatics Journal* 6(1), 20-27, 2012

Di Paola L., Paci P., Santoni D., **De Ruvo M.**, Giuliani A. Proteins as sponges: a statistical journey along protein structure organization principles, *Journal of Chemical Information and Modeling* 52(2), 474-482, 2012

Paci P., Di Paola L., Santoni D., **De Ruvo M.**, Giuliani A. Structural and Functional Analysis of Hemoglobin and Serum Albumin through Protein Long-range Interaction Networks, *Current Proteomics* 9(3), 160-166, 2012

Giuliani A., Di Paola L., Paci P., **De Ruvo M.**, Arcangeli C., Santoni D., Celino M. Chapter - Updating and revising "Proteins as Networks: Usefulness of Graph Theory in Protein Science", *Advances in Protein and Peptide Science*, 2012

Di Paola L., **De Ruvo M.**, Paci P., Santoni D., Giuliani A. Proteins Contact Networks: an emerging paradigm in chemistry, *Chemical Reviews* 113(3), 1598-1613, 2013

Tasdighian S., Di Paola L., **De Ruvo M.**, Paci P., Santoni D., Palumbo P., Mei G., Di Venere A., Giuliani A. Modules identification in protein structures: the topological and geometrical solutions, *Journal of Chemical Information and Modeling*, 54(1): 159-168, 2014

Shimotohno A., Sotta N., Sato T., **De Ruvo M.**, Marée A.F., Grieneisen V.A., Fujiwara T. Mathematical Modeling and Experimental Validation of the Spatial Distribution of Boron in the Root of Arabidopsis thaliana Identify High Boron Accumulation in the Tip and Predict a Distinct Root Tip Uptake Function. *Plant Cell Physiology*, 56(4):620-30, 2015

Chapter 1

Introduction

1.1 Outline/Aim of the Thesis

An intriguing challenge in developmental biology is to understand how organ development is spatially coordinated to form well-structured, patterned complex organisms in a reproducible way [44, 81].

The development of animals and plants is based on a similar logic, which can be identified at the molecular level with the control of cell fate and at the theoretical level with the generation of pattern [72]. In both kingdoms, pattern formation arises from a break of simmetry due to the spatial distribution of signalling molecules (morphogens) [88, 109, 139]: dividing cells exposed to particular concentration thresholds of a morphogen follow a developmental path of cell differentiation, which result in spatio-temporal patterns. Formation, positioning and maintenance of the boundaries between cell compartments are essential for the correct outcome of the patterning events [27].

Both in animal and plant development, a gradient of cell differentiation arises from stem cell niches, where local signals from an organizer coordinate the balance between self-renewal and the generation of daughter cells that differentiate into new tissues. However, in animals, at the completion of development, self-renewal and differentiation is strictly localized to very few stem cells maintained segregated from the rest of the tissue and guaranteeing only for local self-renewal. In contrast, in plants, meristematic

tissues are continuously implicated in the growth of the organism as a whole throughout the plant lifespan.

In addition, as opposed to animals mobile cells, in plants there is virtually no cell migration due to rigid cell walls [95, 135] and organ shape is largely generated by differences in the rate of cell division and cell differentiation mainly driven by hormones. This provides a unique opportunity for dissecting the relations between pattern formation and final organ structure. Therefore, plant biology has become a fertile field of research providing a comparative analysis of development [52].

Plant post-embryonic development is maintained by the activity of root meristems, where stem cell are localized and the transition from cell division to cell differentiation is orchestrated to generate distinct developmental zones. In the model plant *Arabidopsis thaliana* advanced molecular, genetic and genomic tools are available and root development is arguably the most tractable system, due to a number of characteristics of the *Arabidopsis* root, among which: a comparatively simple and very conserved anatomy, which allows easy detection of mutant phenotypes; the possibility to can be traced down every root cell file to a single stem cell that originated it; transparency, which allows easy visualization by means of confocal and light microscopy.

A key role for plant hormones in *Arabidopsis* root development is well established [18, 115, 32]. Specifically, the phytohormone auxin acts as a morphogen, as its asymmetric distribution within tissues sets positional information that guides cell-type specification [17, 79, 93]. Concentration thresholds of auxin create distinct cell compartments with specific developmental fates. Indeed, auxin peculiar carrier-mediated polar transport gives rise to a concentration gradient along the root longitudinal axis, that shapes developmental zones: stem cell niche, division zone and differentiation zone [55, 29, 7]. The formation of an auxin maximum in the stem cell niche is essential to maintain stem cell function [115]. At the boundary between the zones (transition boundary) auxin divisional activity is counteracted by the phytohormone cytokinin and their dynamic crosstalk is necessary to balance cell division over cell differentiation, in order to set a stable meristem size [30, 32, 31].

Molecular genetic approaches have identified many of the key signals components underlying auxin and cytokinin interaction in the *Arabidopsis* root, providing qualita-

tive insights into the activity of cytokinin on auxin distribution [32, 31]. This molecular mechanism, however, fell short of providing a satisfactory mechanistic understanding of how cytokinin actually shapes the auxin gradient, positioning and setting the transition boundary. Moreover, there are no tools available to make auxin gradients directly visible in living tissues. In order to explain the observed cell-type specific auxin distribution and how cytokinins shape auxin gradients, I adopted a systems biology approach, integrating experimental evidence with mathematical and computational modelling.

The use of mathematical models enables a simplified and formal description of the biological mechanisms at different scales (molecular, subcellular, cellular and supracellular) allowing to formulate theoretical assumptions that could guide future experiments, whose results can feedback into the model [125].

Modelling approaches are extensively used to investigate the spatial regulation of tissues: Turing's reaction-diffusion equation models are a well established framework used to explain pattern formation as a result of chemicals interaction [84, 65, 81]; morphogen gradient theory has succeded in making predictions on how a morphogenetic gradient is generated and maintained in multicellular organisms [139, 94, 53].

In recent years, several works have been focusing on modelling auxin transport, spanning from continuous -at tissue level- to discrete -at cellular level- descriptions, demonstrating that auxin carrier localization is fundamental to drive a reflux-loop mechanism able to generate a stable auxin maximum [50, 83, 6, 90]. However, while providing a detailed description of auxin transport, the link between physico-chemical (diffusivity, auxin distribution) and biological (gene expression) descriptors is still missing and none of the proposed mechanisms unveils the developmental cues that drive the emergence of meristem zonation at the cellular level.

During my PhD project, I dissected the problem both by theoretical and computational tools. Mathematical modelling is essential to define explicitly the relationship between physical entities attempting to find an analytical solution to the problem. On the other hand, computational modelling is advantageous in that it provides numerical solution to complex problems through the implementation of powerful algorithms.

Therefore, I first developed a one-dimensional analytical and theoretical description based on physico-chemical laws, in order to provide a straightforward condition for

auxin maximum formation. Within this framework, I linked microscopic (cell-based) description to macroscopic (organ-scale) dynamics through a derived auxin diffusivity parameter and auxin reaction terms.

To investigate the effect of cytokinin on auxin gradient, I then developed a two-dimensional computational model extending the framework developed by Grieneisen et al [50]. I integrated experimentally derived parameters into a spatial model at cellular resolution that simulates auxin transport within a layout resembling root geometries, tissues and zones.

1.2 Biological Background

1.2.1 A Comparative Analysis of Development in Animals and Plants

Higher animals and plants originate from a common eukariotic ancestor. The basic mechanisms of pattern formation and of cell-cell communication in development evolved independently in the two kingdoms [82], leading to major differences in morphology, which reflect the evolution of their survival strategies: as opposed to animals mobile cells, in plants there is virtually no cell migration due to rigid cell walls and organ shape is largely generated by differences in the rate of cell division and cell differentiation. This provides a unique opportunity for dissecting the relations between pattern formation and final organ structure. However, they share a similar logic of development, which can be identified at the molecular level with the control of cell fate and at the theoretical level with the generation of pattern [72].

In both lineages, the specification of tissue and shape within a developing organ originates from the establishment and coordination of cellular and intracellular asymmetries [91, 19], which are either transmitted from pre-existing patterns or initiated through symmetry-breaking events [13]. In biology, symmetry-breaking is often driven by cell polarity: the activity of signalling molecules [89, 43] - endogenous or external - leads to an asymmetric distribution of key molecules within the cell [19]. Well known examples of cell polarization are chemotaxis of motile cells, epithelial cells morphogenesis [70]

and tissue patterning in plants [116].

1.2.1.1 Cell-cell Communication

The coordination of cell polarity among cells occurs differently in animals and plants. A prevalent feature of intercellular communication in animals is cell migration: a complex sensory and locomotory system enables direct communication between neighbouring cells, based on protein-protein interactions at the cell membranes (e.g., Delta-Notch system for lateral inhibition [2] and the planar cell polarity proteins [126]). Such direct interactions are impossible for plant cells, whose relative movement is largely prohibited due to the presence of stiff cell walls, which form upon plasma membrane polarization impeding cell migration [95, 135]. Thus, endogenous mobile signals are necessary to instruct cells [77] with positional information and to provide the responses to external - environmental - stimuli, such as lights and gravity [141, 98]: this function is performed by plant hormones.

In plants, polar localized proteins are involved in various mechanisms, spanning from cell morphogenesis, asymmetric cell division, nutrient transport and the establishment of gradients for developmental patterning. Although plants lack many cell polarity processes of animal cells, flexible rearrangement of cell polarities in plant cells enables them to redefine the developmental programs according to environmental changes [33].

1.2.1.2 Building a Body: the Role of Stem Cells

The different life strategies adopted by multicellular organisms reflect the establishment of their body architecture.

Unlike animals, where the adult body shape is mostly defined during embryogenesis, plants continue to produce organs throughout their life [16], acquiring their final shape during post-embryonic development. The continuous generation of new plant organs is ensured by the presence of a pools of stem cells within regions called meristems, at the growing shoot and root apices: the shoot apical and the root apical meristems, which generate aboveground and belowground structures, respectively [115, 31, 140, 101, 52]. From meristems, stem cells progeny are recruited into developing organs. Dividing stem

cell progeny in the meristems are equivalent to the animal transitamplifying cells.

Both in the animal and plant kingdom, stem cells are defined by their ability to both renew themselves and to generate daughter cells to produce new tissues. A prominent feature of plant stem cells is their totipotency, which allows for the regeneration of an entire organism in culture [72]. Notably, animal cells can survive only in the context where the correct survival factors are present. The absence of this regulation can lead to dramatic effects, such as the formation of cancer cells [72].

Both in plants and animals, stem cells reside in specialized microenvironments, the stem cell niches, where local signals from an organizer coordinate the balance between self-renewal and the generation of daughter cells that differentiate into new tissues. Stem cell niches in the two kingdoms share structural similarities (Figure 1.1). However, while organizing signals are well known in animals (e.g., the pro-epidermal growth factor (EGF) and the Delta-like protein secreted by the Paneth cells in the intestinal crypt), the nature of these signals has not been uncovered yet in plants [52].

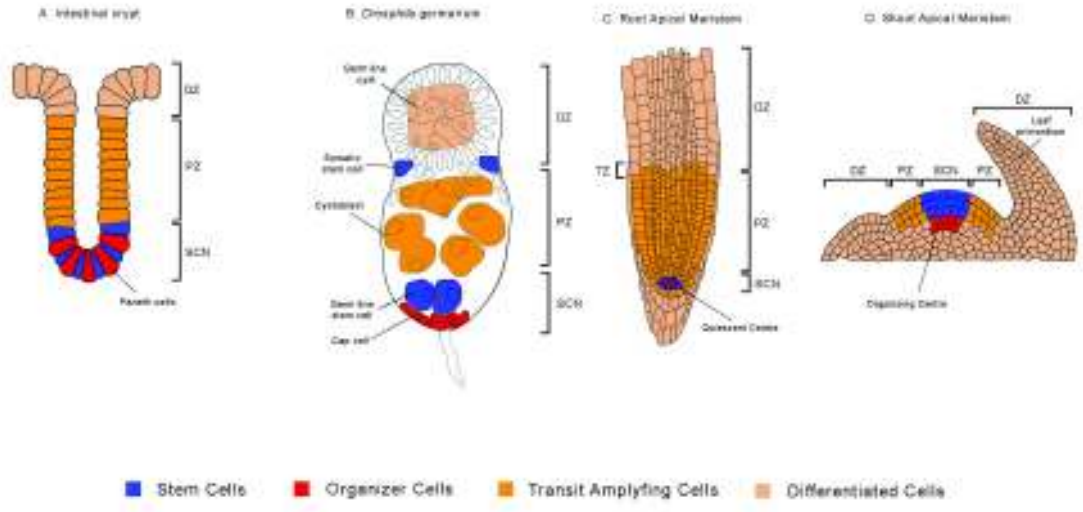


Figure 1.1: **Structural similarity between stem cell niches in animals and plants**. A) The mouse intestinal crypt. B) The *A. thaliana* root meristem. C) The *A. thaliana* shoot apical meristem. In all cases, stem cells (blue) are maintained by short-range signals that arise from specialized organizing cells (red) (Paneth cells in A, Quiescent Centre in B, Organizing Centre in C) Stem cell daughter cells undergo additional divisions in the proliferation zone (PZ) (in orange) to generate transit-amplifying cells, which eventually differentiate in the differentiation zone (DZ), generating specific tissues (A) or organs (B, C).

The structural similarity between stem cell niches in the two kingdoms has determined an increasing interest in plant stem cell research, to investigate which factors underlie the conservation of basic mechanisms [52].

1.2.1.3 Pattern Formation

Pattern formation occurs in various developmental processes, including somitogenesis, limb bud development and root development. Hence, it leads to comparisons between plant and animal development.

During development of multicellular organisms, cell fate has to be regulated in order to generate pattern of cells, tissues and organs. Both in plants and animals, the control of cell fate is based on transcriptional cascades [72]: an initial spatially specific pattern

of gene transcription is activated and then transduced into changes in gene activity during cell differentiation, leading to the specification of segment or organ identity (e.g., segmental identity in *Drosophila* and radial patterning in *Arabidopsis* flowers) and of developmental axes (e.g, dorsal-ventral axis in animal embryos and adaxial-abaxial axis in leaves) [82].

However, despite the similar use of transcription factors as master regulators of developmental pattern, plant and animal mechanisms of pattern formation are nonhomologous.

1.2.1.3.1 Hormones Both in plants and animals, hormones regulate several biological processes at different levels. The specification, maintenance and differentiation of stem cells are regulated by hormones, in a cell autonomous or non-autonomous manner: for example, in plants, the hormone auxin provides cell autonomous positional cues, which control root stem cell niche function and specification, while cytokinin provides non-autonomous signals to control stem cell differentiation [52, 115].

The most prominent difference between plant and animal hormones lies in the site of their synthesis: while animal hormones are synthesized in a specific tissue or organ due to the presence of glands, biosynthesis of plant hormones occurs in multiple cells or regions within the tissue or organ. Moreover, plant hormones are small molecules with large diffusion coefficients, so to generate signals triggering specific reactions or pathways (i.e, gene expression activation) by interacting with hormone-specific receptors. In contrast, most of animal hormones are proteins, which are larger and contain functional domains that directly perform signaling tasks, without intermediate pathways, allowing an easier diversification of the transmitted signal [134].

Importantly, hormones play a role in the coordination of development over longer distances through the establishment of concentration gradients, tipically upon diffusion over a tissue. Such diffusibile molecules are called morphogens and they act in a concentration-dependent manner directly on target cells, which interpret the information as a developmental cue, generating diverse patterns [88, 109, 139, 134]. Paradigmatic examples are the gradient of the protein Bicoid in the *Drosophila* embryo in animals and the gradient of the hormone auxin in plants. Auxin has also been de-

scribed as a morphogenic trigger since it activates a response over a threshold level [96, 9].

1.2.1.3.2 The Formation of Boundaries Positional information acts as a key mechanism to specify patterns in several developing systems [142, 143]. For instance, a plant cell displaced out of its normal position will switch to a fate specific to its new position [72].

A central feature of positional systems is the role of boundaries in setting up the positional field: cells acquire positional identities with respect to given boundaries and interpret this positional value by changing their state, leading to the specification of different cell types and shape. For instance, a signal gradient across a field of cells is converted into gene expression domains through thresholds corresponding to concentration-specific response of target genes. The boundaries between these domains lie in correspondence to cells of different types, which do not overlap, i.e., groups of cells with distinct functions are physically separated. Interactions between these cell types were predicted to form secondary signalling centres, which specify different regions, patterning the field [142]. The formation of developmental boundaries is therefore essential for the correct outcome of downstream patterning events [27]. To instruct tissue development, boundaries need to remain sharp upon cell rearrangements during cell proliferation and morphogenetic tissue movements (mechanical tension, cell adhesion/repulsion) (Figure 1.2). To ensure the completion of these complex biological processes, an extensive integration of several components of cell-to-cell signaling and gene regulatory networks takes place, whose synergic activities have an impact on cell division and growth.

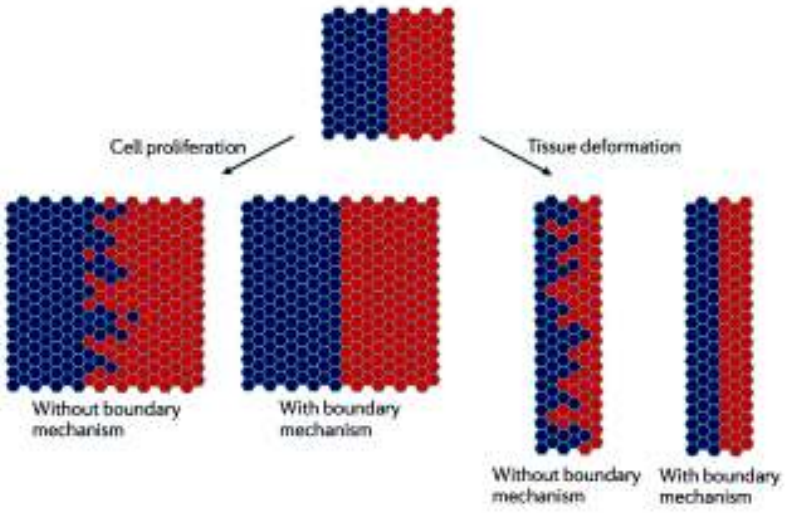


Figure 1.2: **The formation of boundaries.** Straight and sharp boundaries between two groups of cells (blue and red) when subjected to cell proliferation (left) or tissue deformation (right): boundary mechanisms are essential to define and mantain sharp interfaces. Modified from [27].

Understanding how these mechanisms are regulated is a central and open question to date. It has been suggested that the positioning and formation of boundaries is regulated by the concerted activity of selector genes, cell signalling among cells and their physical interactions: cells at the boundary act as organizers, which express a unique set of transcription factors, locally repressing cell proliferation and controlling the fate of cells; this local information is then transmitted to adjacent cells in order to control their developmental program and morphology.

In the animal kingdom, the concept of boundaries is identified with the spatial delimitation of cell identities within tissues. In particular, lineage-based boundaries correspond to the formation of compartments between cell populations in correspondence to selector gene expression domains. These boundaries act as a source for positional information in several processes, such the development of *D. Melanogaster* wing, where the graded distribution of Decapentaplegic (Dpp) protein - a member of the bone morphogenetic protein family- acts as a signalling input for anterior cells, keeping the

anterior-posterior boundary well defined.

In plants, there is evidence that boundaries between cell types act as a signalling centre [72]. Along with a mechanism similar to that occurring in the development of *Drosophila* wing, the ougrowth of leaf lamina appears to be induced by the juxtaposition of adaxial and abaxial tissue. On a different scale, boundaries play a role in the formation of new organs from the meristems [138]. In the shoot apical meristem (SAM), the establishment and maintainance of morphological meristem-to-organ boundaries occurs upon the coordination between transcriptional regulators, hormonal crosstalk and mechanical forces.

The use of live imaging techniques together with mathematical modelling will help investigating the dynamics of cell behaviour and morphogenesis during boundary development and maintainance [27, 138].

1.2.2 Plant Morphogenesis

As confirmed by etimology, morphogenesis (*morphê*: shape and *genesis*: creation) is the biological process that induces an organism to develop its shape [76]. Specifically, plant morphogenesis was defined as "the formation of shape and structure by coordination of cell shape, growth, and proliferation by mitosis" [111].

Given that in a developing tissue mechanical interactions between plant cells are inhibited, morphogenesis in plants emerges upon the coupled genetic (cell proliferation) and physical (growth) levels in a closed network [38]. Plant cells are indeed bound in a semi-rigid cell wall matrix, composed of cellulose microfibrils, other polysaccharides and proteins [111]. After each cell division, a new wall is produced along a specific axis, which will determine the position and the fate of daughter cells [123]. Tissue and organ growth is then coordinated with respect to this axis; as a result, plant cells are often polar. The mechanical strength of cell walls during growth is maintained by enzymes and other agents, while cell shape and proliferation are regulated by factors such as hormones, transcription factors and other signalling molecules, transported from cell to cell [111].

Given the complexity of the mechanisms in action, to fully understand how plant

morphogenesis is established, it is necessary to dissect the problem using biological systems that allow simple and direct observations.

1.2.3 *Arabidopsis thaliana*: a Model System

Plant development research has greatly benefited from studies on *Arabidopsis thaliana* as a model plant (a typical wild type seedling is shown in Figure 1.3), as it shows an array of advantages [124]. Among these, it is a small and fast growing plant with a short lifecycle ($\sim$ 2 months) and it has a comparatively small genome (157 Mbp) that has been fully sequenced, ensuring a high availability of genetic data [102]. Genetic tools are available to access insertion alleles for most genes (BLAST search), and genomic resources are commonly used to access mutants for these genes [23].

Figure 1.3: **Optical microscope image of *Arabidopsis* thaliana seedling at 5 days after germination.** The blue circles highlights the shoot apical meristem (top) and the root apical meristem (bottom).

The root apical meristem (RAM) of the plant *Arabidopsis thaliana* is one of the best-characterized systems in plant developmental biology [37] regarding the analysis

of development, patterning and growth. In particular, the *Arabidopsis* root meristem is advantageous due to its simple cellular organization, with cell types arranged in well-defined tissues (cell files) along the longitudinal axis. Both the number of cell files and the number of cells in each cell file are highly conserved. Therefore, the linearity of its structural and functional organization (fig 1A), together with a well known gene regulatory network, and the possibility of growing it in nonsoil media, make the RAM a suitable model system not only for wet biology techinques, but also for modelling approaches in Systems Biology, offering powerful tools to shed light on the mechanisms still poorly understood solely by means of experiments.

1.2.3.1 The *Arabidopis* Root

The root is an essential organ of the plant established during embryogenesis, generally located below the surface of the ground [103]. It is involved in a range of life-maintaining processes [55], which include adsorbing water - as well as inorganic nutrients - from the ground, anchoring the plant in the ground, as well as coping with toxic elements. In order to bear changes in nutrient and water availability, roots developed mechanisms of adaptation, such as changing their architecture by the continuous production of lateral roots and root hairs, which increase root surface area (up to 70% of the total root surface area), thus allowing for an optimised uptake of nutrients and water [96].

Root architecture can be dissected into a series of iterative modules and the radial structure of *Arabidopsis* root tissues is arranged in a set of concentric cylinders (Figure 1.4). Adult tissues differentiate in three parts, from the circumference inwards:

- The outermost layer of the root (in yellow in Figure 1.4b) is called epidermis, a mono-stratified tissue that protects the principal root and that consists of cell files with hair-bearing cells (root hair or RH cells) alternated with files of hairless cells (non-hair or NH cells) [36].

- The intermediate section includes the cortex and endodermis layers (respectively in light blue and orange in Figure 1.4b), both composed of a single cell type [72]. In particular, in the endodermal cells a substance called suberin isolates the central cylinder from water. The function of the endodermis is to direct the

flow of water and solutes across the plasma membrane and the cytoplasm before reaching the deepest parts.

- The central cylinder, called stele, includes the pericycle and the vascular tissue. The pericycle is formed by a single layer of parenchymal cells that can initiate the formation of new lateral roots [28]. The vasculature has bilateral symmetry and the vascular strands (bundles) that compose the vascular system contain two cell types: the xylem and the phloem [78]. The xylem provides support for the plant body and transport of water, minerals and phytohormones - such as cytokinin and abscisic acid - unidirectionally, from the root to the sites of photosynthesis in the leaves. From here, the end products of photosynthesis are transported through the phloem to other parts of the plant, such as roots and reproductive organs, in a bi-directional way.

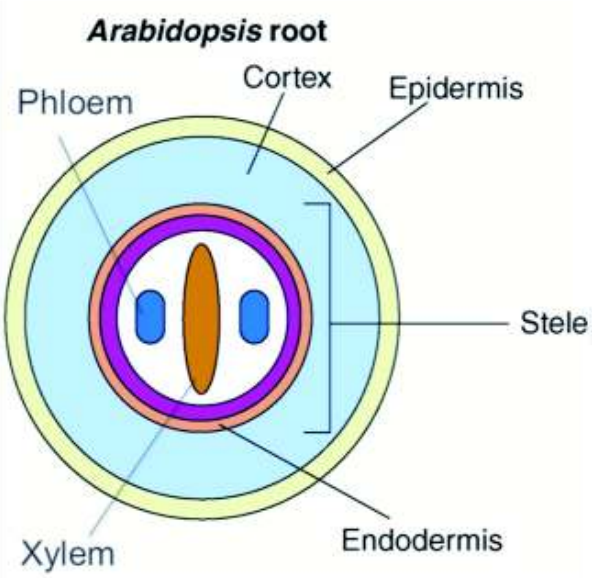


Figure 1.4: **Radial tissue organization in a transverse section of *Arabidopsis* root.** Modified from [35].

Along the longitudinal section, the *Arabidopsis* root shows a peculiar gradient of cell

differentiation, which determines the pattern of cellular events involved in plant post-embryonic development [55, 29] (Fig. 1.5). Along the apical–basal axis, cells proceed through distinct phases of cellular activities [136], shaping the root into developmental regions, depending on the morphology and state of development of the cells.

At the root apex, a zone of cell division defines the region of the root meristem. Here, in the stem cell niche (SCN) stem cells self-renew producing daughter cells that undergo additional divisions in the meristem zone (MZ) until they leave the meristem and start expanding/elongating in the elongation-differentiation zone (EDZ), with the final aim at differentiating and reaching the mature length. The boundary that encompasses the transition between dividing cells in the meristem and expanding (differentiating) cells in the different cell files is called transition zone (TZ) [31, 85, 79]; importantly, its positioning during the growth of the root determines root meristem size. This distinction in zones of distinct cellular activities is called zonation, and resembles the corresponding features of animal stem cells systems [79].

The root meristem continuously gives rise to new cells for all of the root tissues formed during postembryonic development [72]. In fact, in the *Arabidopsis* root stem cell niche (SCN), multipotent stem cells for all root tissue types surround a small group of mitotically inactive, organizing cells, the quiescent center (QC) (in red in Figure 1.5 and inset) [36, 72, 99]. The QC provides short-range cell non-autonomous signals that maintain the stem cells in an undifferentiated state and promotes its neighboring cells to continuously produce initial cells that give rise to cell files. Stem cell asymmetric division produces two daughter cells: the self-renewing daughter cell that contacts the QC maintains its stemness, whereas the other daughter cell becomes displaced and generates single-cell files that extend along the longitudinal root axis, giving rise to all tissue types [103]. Thus, the longitudinal axis of a root represents a constantly renewing gradient of cell differentiation that ensures the indeterminate growth of the root [96].

The meristem is shielded from the soil by several layers of cells at the root tip -the root cap- and a lateral single cell layer - the lateral root cap (LRC) - (in light purple in 1.5), which covers all the meristem tissues. The lateral portion of the root cap is generated from the same set of stem cells as the epidermis (in brown in Figure 1.5).

The distal part of the root cap, the columella (in light blue in Figure 1.5), has its

own set of stem cells as the vascular tissue does [36, 72]. The remaining stem cells give rise to both the cortex and endodermis [119].

Due to the lack of cell migration, the spatial relationship of cells in a file reflects their age: younger cells lie near the root tip, older cells are higher up in the root (Figure 1 and 2A).

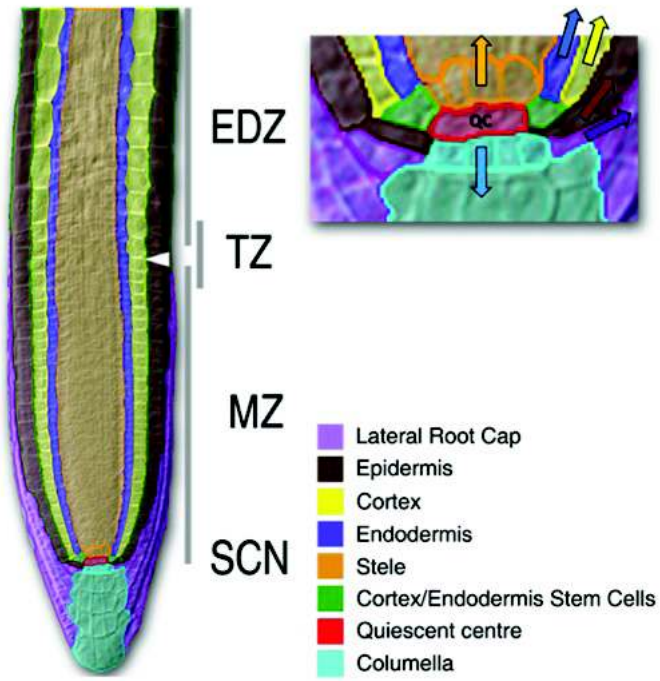


Figure 1.5: **Tissue organization in the developing root of** *Arabidopsis*. Different tissues are represented in false colors, as indicated in the legend. SCN, stem cell niche; QC, quiescent centre; MZ, division or meristem zone; EDZ, elongation/differentiation zone; TZ, transition zone. Inset. Schematic representation of the stem cell niche, with direction of stem cell division. White arrow indicates the TZ. Adapted from [99].

1.2.3.2 Root Meristem Development

The coordinated activity of the root developmental regions is crucial for the establishment of a dynamic equilibrium between dividing cells and those that differentiate, maintaining constant in time the size of the root meristem, expressed as the number of

cortex cells in a file extending from the quiescent center (light blue arrowhead in Figure 1.6) to the first elongated cortex cell (white arrowhead in Figure 1.6) [31]. In fact, the longitudinal organization of the root is a dynamic process that tends towards a steady state once the balance in cell division and differentiation rate is reached (i.e., transition zone is positioned).

In Figure 1.6 a time course analysis of cortex cell number over 8 days after germination (dag) is reported. Upon seed germination, cell division in the apical root meristem prevails over cell differentiation, favouring meristem growth until 5 days, as indicated by the different size of the meristems in Figure 1.6. After 5 dpg, cell differentiation rate increases, balancing cell division. At this stage, the establishment of a dynamic equilibrimium ensures the maintainance of a stable and constant meristem size (Figure 1.6 B).

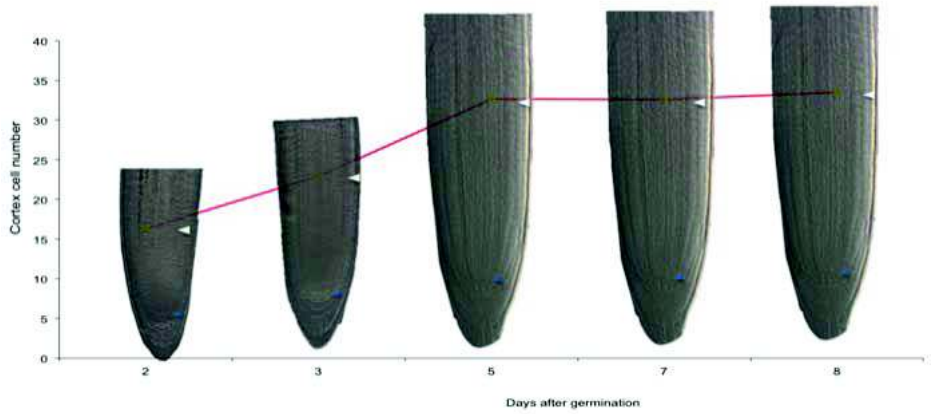


Figure 1.6: **Time course analysis of root meristem size development over time**. Root meristem size is expressed as the number of cortex cells in a file extending from the quiescent center (blue arrowhead) to the first elongated cortex cell (black arrowhead). Root meristem cell number of wild-type Col-0 plants increases until 5 days after germination, when a stable number of approximately 30 cells is established in the meristem and maintained constant in time.

The balance between the rate of cell proliferation in the meristem and the extent of cell elongation/differentiation at the transition zone ensures continuous root growth

throughout plant lifespan [114, 100, 9]. How this balance is achieved is a central question in plant developmental biology, and in the last years plant research has been focusing on elucidating the mechanism underlying this process [85, 86, 31, 30, 114].

1.2.4 The Role of Plant Hormones in Root Development

Phytohormones act as key intercellular signals throughout plant development [18]: several plant responses, such as gravitropism, tropic response and root growth, rely on signalling systems driven by hormones, even at low concentrations. The major plant hormones involved in the regulation of root growth and development include abscisic acid (ABA), brassinosteroids, ethylene, gibberellic acid (GA), cytokinins, and auxins [103, 132]. Each hormone has its specific biosynthetic and signal transduction pathway.

In this thesis, I will focus on the activity of two plant hormones, auxin and cytokinin. These phytohormones interact to regulate multiple processes during root development, including the formation of the embryonic root, the emergence of lateral roots, the control of root vascular pattern [13] and more importantly, the control of meristem size [31, 30, 86].

Indole-3-acetic acid (IAA), the most common of the naturally occurring auxins, is synthetized mainly in the aerial tissues of young leaves, but also in the meristematic region of root tips [75]. From these sites of production, auxin is redistributed to the rest of the plant through active and passive transport mechanisms. In the root, auxin plays a key role in developmental programs as well as in mediating environmental stimuli (e.g., tropic response to light and gravity) to shape the final root architecture. Auxin is known as the "root forming hormones of plants" as it is essential for the formation of both lateral roots and root hairs, as well as for the formation of the root apical meristem and the stem cell niche [128]. The coordination of IAA biosynthesis, degradation, signalling and transport ensures proper auxin concentrations [102], which leads to a differential auxin distribution between cells. A detailed analysis of auxin transport is provided in the next paragraph.

Cytokinins are purine derivatives, which promote and maintain plant cell division in culture; they also partecipate in various processes such as shoot apical meristem main-

tenances and leaf senescence [39]. In higher plants, bioactive cytokinins are isoprenoid cytokinins as isopentenyladenine (iP), trans-zeatin (tZ), cis-zeatin, and dihydrozeatin (reviewed in [11]).

The most abundant forms found in *Arabidopsis* are iP and tZ [117]. As opposed to auxin transport, the mechanism and function of cytokinin transport are poorly characterised. It is assumed that cytokinin is long-distance transported from the shoot to the root via the phloem, controlling auxin distribution in the vasculature of the root meristem [13].

1.2.4.1 Polar Auxin Transport Mediates Root Patterning

Auxin (IAA) is transported throughout the plant both at short and long range: symplastic polar transport redistributes IAA between adjacent cells, whereas phloem transport through the vasculature carries IAA flux over long distances from the shoot to the root system [98].

The peculiar distribution of auxin concentration arises from the coordination of its local biosynthesis, inactivation (conjugation and direct oxidation), diffusion through the cell cytoplasm and apoplast, and polar auxin transport (PAT) between the cytoplasm and the apoplast -crossing the plasma membrane [73, 135, 14, 5]. PAT is finely regulated by the concerted action of influx and efflux carriers and members of ATP-binding cassette protein, and its unique feature comes from the physico-chemical properties of the hormone, as originally explained by the chemiosmotic theory [47].

The chemiosmotic model for polar auxin transport, sketched in Figure 1.7, postulates a mechanistic basis for auxin movement between cells based on the ability of IAA to penetrate the membrane depending on pH [135, 102]:

$$IAAH \rightleftharpoons IAA^- + H^+ \tag{1.1}$$

where IAA^- is the conjugate base to the weak acid IAAH.

As a weak acid (pKa = 4.75), at low pHs IAA dissociation is inhibited and the acid is present prevalently in the undissociated form (IAAH); conversely, under alcaline conditions, it dissociates in the anionic, conjugate form (IAA^-). Once in the acidic

environment of the apoplast (pH 5.5), auxin in its undissociated form enters the cell across the cell wall through passive diffusion as a protonated molecule, or sustained by active importers, AUXIN RESISTANT1/LIKE AUX1 (AUX1/LAX) uptake permeases, as an ion [20]. When in the cytoplasm, IAA dissociates into the alcaline environment (pH = 7) and gets trapped inside the cell, as the anionic form is not able to passively diffuse out of the cell wall (or apoplast). In fact, only specific efflux carriers localized on the plasma membrane can lead it out of the cell [98, 144].

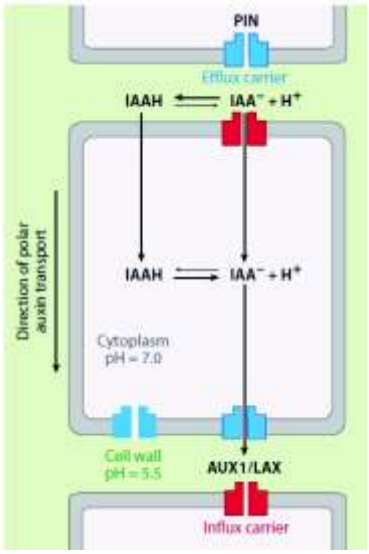


Figure 1.7: **Chemiosmotic model of polar auxin transport**. In the cell wall (grey), at a pH around 5.5, auxin is protonated (IAAH) and enters the cell with or without the aid of influx carriers AUX1/LAX (red blocks). In the alcaline cytoplasm (white), auxin is present in its anionic form (IAA^-) and can only exit through efflux transporters PINs (light blue blocks). Figure taken from [63]

The best characterized auxin efflux protein class is the PIN-FORMED (PIN) family. Importantly, the asymmetric membrane localization of PINs provides directionality to auxin transport, while the apolar carrier-mediated auxin import controls auxin concentrations and distributions within cells and tissues [6]. The allocation of PINs is a highly dynamic process that involves continuous cycles of endocytosis and recycling back to the plasma membrane. The cycling of PINs is important for their positioning

on different sides of the cell [63].

The modulation of PIN activity can independently affect meristem size, elongation rate, and final cell size [40]: a repositioning of PINs or changes in the rate of efflux would cause the relocation of auxin maxima, governing the directionality of the auxin flux, thus of the root growth [57].

The PIN protein family includes 8 members, most of which mediate auxin efflux [129]. PIN1 is highly expressed on the lower side of the membrane, leading to rootward transport of auxin [55, 8], which accumulates in the QC cells forming a maximum of concentration. From the columella, this flux is redistributed laterally to the lateral root cap and the external files by PIN3,4, and PIN7 (Fig. 1.8) [42] Here, apical-lateral PIN2 orientation redirects the flux basipetally, at the same time driving its reintroduction at the transition zone level towards the stem cell niche, thereby generating a "reflux loop", which induces an auxin maximum in the QC [7, 50]. PIN5, PIN6 and PIN8 regulate intracellular auxin homeostasis and metabolism but do not have a direct role in cell-to-cell transport.

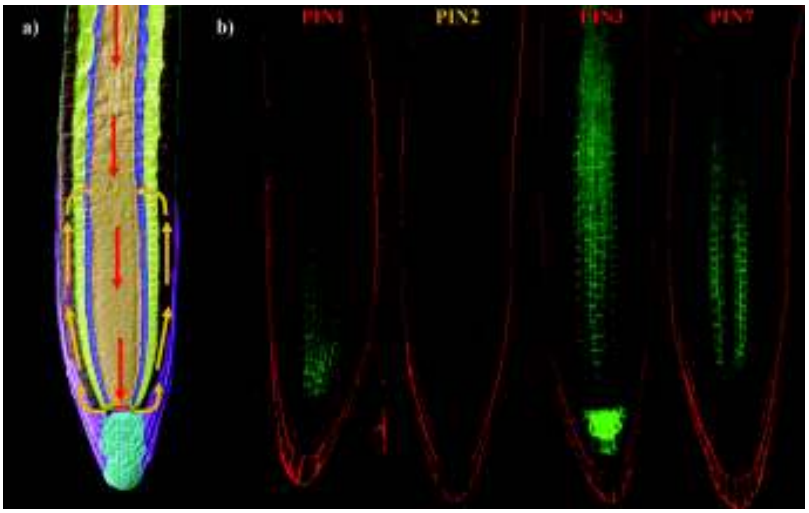


Figure 1.8: **Auxin transport routes (a) and PIN expression (b) in the *Arabidopsis* root**. a) Inverse fountain mechanism for auxin flux: in the vasculature, PIN1,3,7 (red arrows) direct auxin towards the QC cells. Here PIN2 (yellow arrow) redistributes auxin flux upward through the external files and inward towards the central files at the transition zone. b) Confocal images of PIN1,2,3,7 expression. Green Fluorescence Protein (GFP) indicates PIN localization on the plasma membrane.

Early IAA indirect measurements have demonstrated that the generation of auxin local maximum and its perception are essential to maintain stem cell function and to instruct a correct patterning of the root [115]. Actually, the local control of auxin levels generates local concentration gradients and maxima, which are crucial to set up and keep a root primordium.

1.2.4.2 The Dynamic Interplay Between Auxin and Cytokinin Sets Meristem Size

A morphogenetic role for cytokinin has not been proven, as it does not interact with auxin in the specification of the quiescent center (QC) and the regulation of cell division [113]. However, cytokinin plays a pivotal role in the regulation of root meristem size as it affects cell differentiation rate: exogenous application of cytokinins to wild type roots shifts the dynamic equilibrium between cell division and cell differentiation towards differentiation, leading to a progressive decrease in meristem size; on the other hand,

exogenous application of small amounts of auxin leads to an increase in meristem size (due to an increase in the cell division rate).

Thus, the root meristem balance is mainly mediated by auxin and cytokinin antagonistic interaction [30, 86, 32]. During the past few years, plant research has been focusing on uncovering the molecular mechanisms through which cytokinin and auxin interact to modulate the trade-off between cell division and differentiation, thus root growth and development.

Dello Ioio and coworkers [31, 30, 32] have previously shown that the cell differentiation and division balance necessary to control root meristem size and growth is the result of the interaction between cytokinin and auxin [31, 30, 32]. Specifically, cytokinin activates a type-B primary ARABIDOPSIS RESPONSE REGULATORS (ARR1) that act as transcription factors and mediate the cytokinin-dependent cell diffentiation input. In the vascular tissue of the transition zone, the transcription factor ARR1 activates the gene SHY2, a repressor of auxin signalling, which negatively regulates PIN genes. This mechanism suggested that cytokinin may promote cell differentiation by triggering auxin redistribution. They also showed that, conversely, auxin mediates degradation of the SHY2 protein, sustaining the activity of the PIN genes and fueling cell division. Thus, auxin/cytokinin antagonistic interaction defines the position of the transition zone and sets meristem size by controlling auxin redistribution.

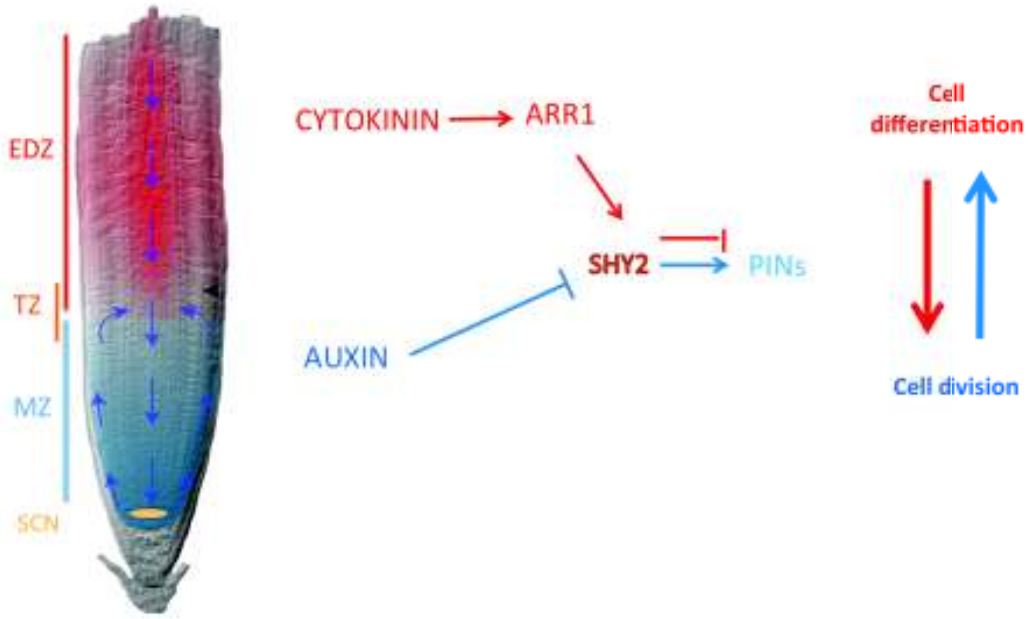


Figure 1.9: **The antagonistic interplay between auxin and cytokinin in the *Arabidopsis* root**. False colors represent domain of activity of auxin (blue) and cytokinin (red). At the transition zone (TZ), cytokinin (CK) induces cell differentiation through upregulation of SHY2, mediated by the CK-responsive transcription factor ARR1. SHY2 in turn represses PIN genes, thus limiting auxin transport. On the other hand, auxin (IAA) promotes SHY2 degradation, thus sustaining PIN activity and cell division in the meristem or division zone (MZ). Modified from [31, 30]

1.3 Mathematical Background

1.3.1 Partial Differential Equations

The most basic theories in physics and engineering often translate into partial differential equations.

Definition. A partial differential equation (PDE) is a differential equation containing partial derivatives of the dependent variable.

A partial differential equation (PDE) for the function $u(x_1, x_2, ..., x_n)$ is an equation of the form:

$$F(x, y, ..., u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}, ...) = 0 \tag{1.2}$$

where $x, y, \cdots$ are the independent variables and u is the unknown function of these variables. The solution of a differential equation is the function u that satisfies the identity. For problems defined in a given spatial domain, among the infinite number of linearly independent solutions, we choose the solution that satisfies the prescribed contraints.

There are a number of properties that lead to the classification of PDEs into families of similar equations. The two main properties are order and linearity:

- The order of a PDE is the order of the highest derivative.
- A PDE is linear if it is linear in the unknown function and all its derivatives with coefficients depending only on the independent variables.

In this Thesis, I will focus on linear second order PDEs of the general form:

$$A(x, y)u_{xx} + B(x, y)u_{xy} + C(x, y)u_{yy} + D(x, y)u_x + E(x, y)u_y + F(x, y)u = G \tag{1.3}$$

where all the coefficients A through F are real functions of the independent variables x, y.

If $G = 0$, the PDE in 1.3 is called homogeneous, as the equation does not contain a term independent of the unknown function and its derivatives. The most general solution of the linear PDE is the sum of any particular solution and the most general solution obtained for the homogeneous equation (called the complementary solution).

1.3.1.1 Classification

Second order PDEs can be classified into three types of field equations: hyperbolic, parabolic and elliptic.

If we define a discriminant $\triangle(x, y)$ such as at the point (x_0, y_0): $\triangle(x_0, y_0) = B^2(x_0, y_0) - 4A(x_0, y_0)C(x_0, y_0)$, a PDE is defined as:

- Hyperbolic equation if $\triangle(x_0, y_0) > 0$

  The hyperbolic equation tends to amplify any noise in the data. The prototypical model hyperbolic equation is the wave equation. In one dimension, this is $u_{tt} - u_{xx} = 0$

- Parabolic equations if $\triangle(x_0, y_0) = 0$

  Parabolic equations tend to keep pre-existing noise in the data, but without amplifying it. A classical example of a parabolic PDE is the one-dimensional heat equation: $u_t = ku_{xx}$

- Elliptic equations if $\triangle(x_0, y_0) < 0$

  In elliptic equations, each point of the domain describes the evolution towards an equilibrium condition, smoothing any perturbation.

While the parabolic and hyperbolic equations are associated with time-dependent problems (initial value problems), elliptic equation are time-independent (boundary value problems) [108].

Fot the specific topic of this work, the parabolic form is suited to model auxin transport. Therefore, in the following paragraph I will discuss a specific case of parabolic equation.

1.3.1.2 The Advection-Diffusion-Reaction Equation

The advection-diffusion-reaction partial differential equation provides a useful mathematical model in a wide range of applications both in natural sciences and engineering [4], such as the transport of air, the adsorption of pollutants in soil and modelling of biological systems. In many of these applications, the unknown variables in the governing partial differential equation represent physical quantities [87].

The advection-diffusion-reaction equation is a second-order PDE that decribes three processes involved:

- advection, due to movement of materials from one region to another;

- diffusion, due to movement of materials from a region of high concentration to a region of low concentration;
- reaction, due to decay, adsorption and reaction of substances with other components.

In general, one-dimensional linear advection with constant diffusion takes the form:

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}(au) = \frac{\partial}{\partial x}\left(D\frac{\partial u}{\partial x}\right) + f(u) \tag{1.4}$$

where $u(x,t)$ is the unknown quantity, $a(x,t)$ is the velocity of the medium (also called convection field) in the x direction, $D(x,t)$ is the diffusion coefficient and $f(u)$ is the reaction term.

This equation requires two boundary conditions and one initial condition. In the next paragraph I will discuss the choice of these conditions.

1.3.1.3 Boundary and Initial Conditions

Defining a domain Γ, its boundary $\partial\Gamma$, the coordinates n and s - normal (outward) and along the boundary respectively- and functions f and g on the boundary, the boundary conditions used to solve PDEs belong to three types:

- Dirichlet conditions with $u = f$ on $\partial\Gamma$. Dirichlet conditions can only be applied if the solution is known on the boundary and if f is analytic. These are frequently used for the flow (velocity) into a domain.
- Neumann conditions with $\partial u/\partial n = f$ or $\partial u/\partial s = g$ on $\partial\Gamma$.
- Mixed (Robin) conditions $\partial u/\partial n + ku = f$,$k > 0$, on $\partial\Gamma$. Mixed boundary conditions indicate that different boundary conditions are used on different parts of the domain boundary.

1.3.1.4 Numerical Methods

In many applications, finding an analytical solution to the governing partial differential equation is not trivial. In fact, the constitutive equations of mathematical models do not

always yield analytical or closed-form solutions. The difficulties in deriving analytical solutions often lead to the use of numerical methods to approximate the solution to the problem [92].

In the engineering field, a common way to carry out an approximation of the solution relies on the decomposition into elementary components, i.e. simple systems whose behaviour is known (discrete problems). A discrete problem can be solved even when the number of simple systems is large, using powerful software. Infact, the purpose of discretization is to obtain a problem that can be solved by a finite procedure.

To solve PDEs, different discretization methods have been developed. Discretization of any PDE consists on converting an infinite-dimensional solution and the differential operators that act on it to a finite-dimensional one, generally in terms of grid functions and difference operators to which simple algebraic techniques can be applied to produce approximate solutions to the PDE [80].

A graphical schematization of the most common methods (finite difference and finite elements) is depicted in Figure 1.10.

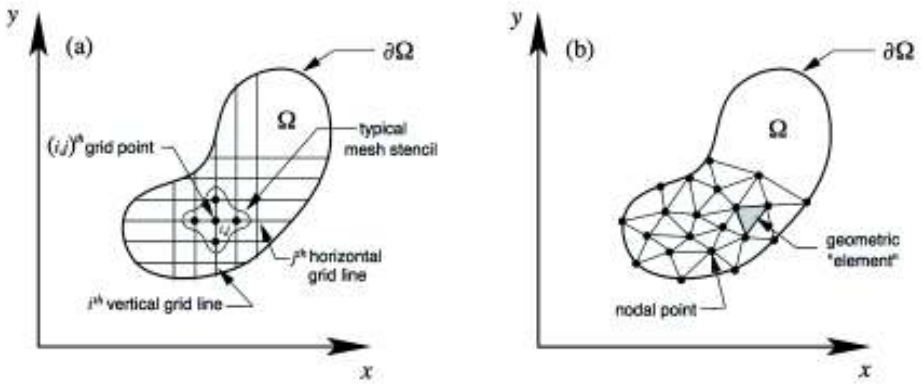


Figure 1.10: **Spatial discretization methods**. a) finite difference, b) finite elements. From [80].

Finite Difference Method (FDM) The finite difference approximation for derivatives is widely used to solve differential equations. The principle of finite difference methods consists on approximating the differential operator by replacing the derivates

in the equation with differential quotients. The solution domain is first partitioned in space and time into a grid, as indicated in Fig. 1.10a, then a system of algebraic equations for grid-point values of the solutions is computed at each space or time point as approximations to the true solution of the PDE at the discrete set of points. The error between the numerical solution and the exact solution is determined by the error introduced passing from a differential to a difference operator (discretization error) [58].

According to the way the time derivative is approximated, explicit and implicit schemes are employed.

Finite Element Method (FEM) The finite element method is a numerical technique used to find approximate solutions to boundary value problems for differential equations. The problem domain is subdivided into small regions, often of triangular shape, defining a mesh (as depicted in Figure 1.10b). The mesh is characterized by nodes on which the polynomial is used to approximate the solution. The element is then defined as the sum of this approximation plus the subregion on which it applies. Once the element equations have been determined, the elements are assembled to form the entire domain Ω. Finite-element methods are usually less efficient than finite difference methods.

1.3.2 The Use of PDEs in Pattern formation: Reaction-diffusion Equations

The idea that spatial patterns arise from the interaction between two diffusible substances (morphogens), in an initially uniform concentration field, was pionereed by Turing in 1952 [131]. He suggested that, under certain conditions, chemicals can react and diffuse in such a way that steady-state patterns emerge. Reaction–diffusion models consist of two components: the first is a set of biochemical reactions which produce, transform or remove chemical species; the second is the diffusion process, which follows Fick's laws of diffusion [110].

The key feature of Turing's framework is the role of autocatalysis in association with lateral inhibition: chemical reactions as local phenomena, while thermal diffusion

provides for the long-range process required for patterning.

The general form of Turing reaction-diffusion system of PDEs is

$$\frac{\partial \overrightarrow{c}}{\partial \mathrm{t}} = D\triangle \overrightarrow{c} + \overrightarrow{R}(c) \tag{1.5}$$

where c is a vector of morphogen concentrations, Δ is the Laplace operator, R is a function of reaction kinetics, and D is a diagonal matrix of positive constant diffusion coefficients. system of nonlinear partial differential equations for the chemical concentrations.

Following the Turing model, Gierer and Meinhardt further showed that two main features are needed for pattern formation: local self-enhancement and long-range (global scale) inhibition [46]. The local activation selects the cells to differentiate, whereas the longe-range inhibition suppresses the activation of neighbouring cells.

Given the two activator A and inhibitor B, acting on a two dimensional domain, the system 1.5 takes the form:

$$\begin{cases} \frac{\partial A}{\partial \mathrm{t}} = D_A\triangle A + F(A,B) \\ \frac{\partial B}{\partial \mathrm{t}} = D_B\triangle B + G(A,B) \end{cases} \tag{1.6}$$

where F and G are the non-linear reaction kinetics.

Both morphogens can diffuse away from where they are originally produced in the tissue. The activator is self-enhancing (or autocatalytic): a small increase in the concentration of A in the initial field causes a further increase in the activator itself. The activator also promotes the production of the inhibitor so that if the activator peaks grow, inhibitor peaks would grow in response. Therefore, self-enhancement of A must be complemented by the action of a fast diffusing antagonist B.

Long-range inhibition occurs if B diffuses much faster than its controlling activator A, i.e. $D_B \gg D_A$. This process results in a periodic pattern of activator and inhibitor peaks and can assume the most diverse shape - such as spots, stripes, spirals - depending on the dynamics of the reaction and on the pattern wavelength [21, 84].

If at the initial condition two morphogens diffuse and react with each other, there can be six stable states:

1. stationary and uniform;
2. uniform and oscillating;
3. stationary waves with extremely short wavelength;
4. oscillatory cases with extremely short wavelength;
5. oscillatory cases with finite wavelength;
6. stationary waves with finite wavelength (Turing pattern).

Representative two-dimensional Turing patterns are shown in Figure 1.11.

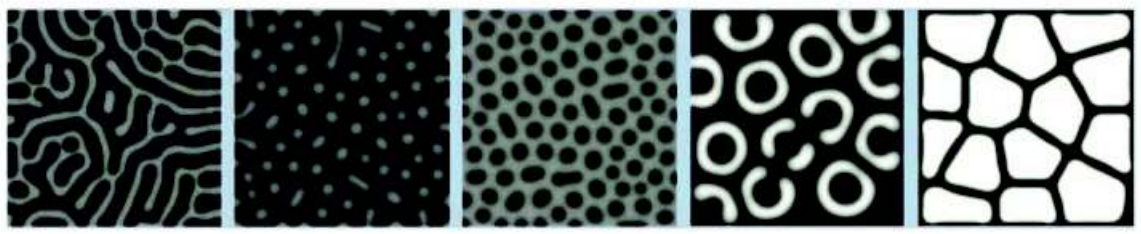


Figure 1.11: **Two-dimensional patterns generated by Turing model**. Diverse patterns can be obtained varying the range of parameter values in a Turing-like reaction-diffusion equation. Modified from [65].

A Turing pattern is defined as a nonlinear wave that is maintained by the dynamic equilibrium of the system. Its wavelength is determined by interactions between molecules and their rates of diffusion. A major feature of Turing patterns is their emergence in the absence of pre-existing positional information [65].

Turing-like reaction-diffusion models have been applied to a range of patterning phenomena in biology, including the pattern on seashells, the striped pigmentation in fish, cartilage formation and regeneration in hydra [139].

1.3.3 Morphogen Gradient Theory

In multicellar organisms, a spatial pattern can arise either as a result of Turing physico-chemical mechanism or of the establishment of morphogen gradients, which provide positional information that instruct cells to acquire their fate. The mathematical formalization given by morphogen gradient theory has been pivotal to illustrate how a morphogen can provide positional information and to unravel the spatial regulation of tissues during development [53]. The morphogen gradient model has been attractive for developmental biologists in that it reduces the problem of specifying positional information to quantify differences in a single molecule.

According to the French flag model first proposed by Wolpert in 1969 [56, 64] and revisited by Crick in 1970 [109], the simplest theoretical description of morphogen gradients relies on a source-sink mechanism, where morphogen spreads from a restricted source region into the adjacent target tissue due to diffusion [53, 49, 121, 15]. To generate steady-state gradient, the morphogen has to be degraded throughout the tissue or solely in a sink region located at some distance to the source [139]. This process leads to the establishment of concentration thresholds that convey positional information to cells, evoking differential cell responses, thereby subdividing tissues into distinct cell types [56, 94, 109].

The general mathematical formalization of a morphogen produced at a source with production rate k, spreading with a constant diffusivity (D) and a linear uniform degradation term (δ) is:

$$\frac{\partial \mathrm{c}(\mathrm{x},\mathrm{t})}{\partial \mathrm{t}} = D\frac{\partial^2 \mathrm{c}(\mathrm{x},\mathrm{t})}{\partial \mathrm{x}^2} + k - \delta \cdot c(x,t) \tag{1.7}$$

The combination of constant flux from a localized source, diffusion, and uniform degradation rate results in concentration profile that decays exponentially with distance from the source (Figure 1.12) [53].

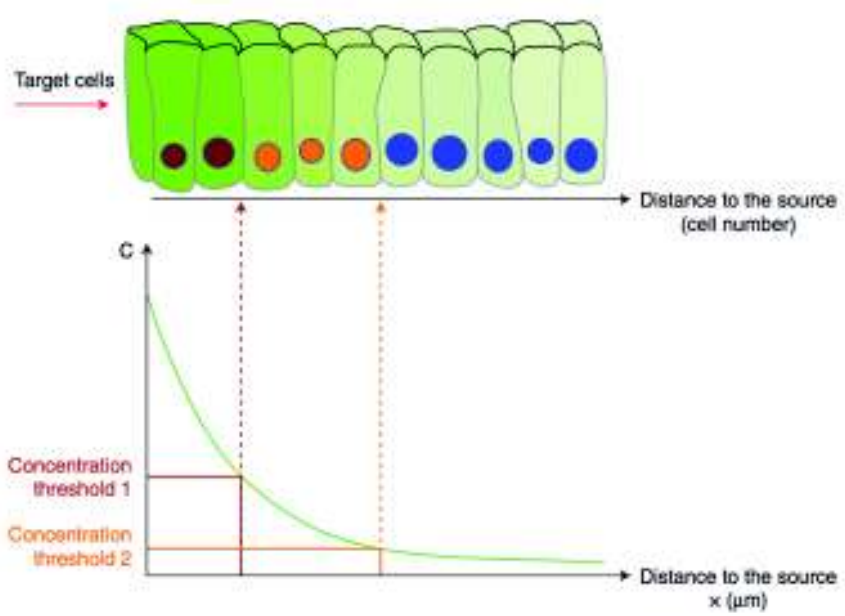


Figure 1.12: **The french flag model**. Top, Sketch of a one dimensional tissue of target cells. Cells with different colours (red, orange and blue) are exposed to different morphogen thresholds. Bottom, A continuum model (individual cells are ignored) where the concentration of the morphogen is a function of the distance from the source. The diffusion of the morphogen gives rise to a concentration gradient within the tissue. Cells exposed to morphogen concentrations exhibit different responses: below threshold 1, red; between thresholds 1 and 2, orange, beyond threshold 2, blue. These differential responses correspond to distinct cell fates and assign positional values to cells. From [94].

For a diffusive gradient, the distance over which positional information can be transferred - known as characteristic length (or decay length) - depends on the diffusivity and degradation rates. This distance can be quantified as: $\lambda = \sqrt{D/\delta}$, thus the higher the diffusivity and lower the decay rate are, the longer will be the range of the morphogen gradient. Given the maximum concentration c_0, the characteristic length indicates the distance from the location of maximum concentration at which the concentration decreases to the fraction $1/e$ of the maximum value c_0 [51, 139]. In a more general case of diffusing-reacting models, the ratio of time-scales between reaction and transport (Thiele modulus) dictates the morphogen range.

The use of computational and mathematical models has succeded in predicting how

a morphogenetic gradient is generated and maintained in a number of multicellular organisms [139]. In animals, known applications of morphogen gradients models focused on Bicoid, Decapentaplegic (dpp), Hedgehog and Wingless morphogens in *Drosophila*. These morphogens are transported from their secretory source cells without carriers [34]. Other examples include fibroblast growth factor 8 (fgf8) in mouse and chick and retinoic acid in vertebrate limb, bone morphogenetic protein (BMP) in Zebrafish [54, 133, 107]. In plants, the phytohormone auxin in *A. thaliana* is the most promising equivalent, as it alters the developmental fate of cells in a concentration-dependent manner and it regulates plant cellular responses in a dose-dependent manner [54, 88, 15, 10]. Therefore, in recent years special focus has been given to modelling auxin gradients formation.

1.3.4 Modelling Morphogen Gradients in Growing Domains: Lagrangian vs Eulerian Approach

During growth of biological systems, the morphogenetic pattern undergoes a continuous remodelling of its size and shape: either both cells and morphogen can be transported through the tissue, or cells that produce the morphogen divide over time. Thus, if the growth timescale is not negligible compared to reaction and transport timescales, the spatial pattern has to be interpreted as a dynamic process, the pattern adjusting to the growing tissue [134].

To analyze how the shape of morphogen gradients changes in response to tissue growth, it is necessary to fix a reference point for the trajectories of tissue elements (and of the morphogen inside them) during growth.

Following Crampin et al. [56], I initially derived the general expression for the evolution of a morphogen c (i.e. auxin) reacting and diffusing on a growing domain $V(t)$, passing from Eulerian coordinates $(\mathbf{x}, t)$, useful to describe the velocity of the flow at a given time and fixed position, to Lagrangian coordinates, more appropriate to observe trajectories of single material volumes that are carried about with the flow. The fluid velocity at a given time and fixed position (the Eulerian velocity) is equal to the velocity of the fluid parcel (the Lagrangian velocity) that is present at that position at that instant [105].

For an arbitrary elemental volume $V(t)$ moving with the flow due to the growth, the conservation of matter allows to write the conservation equation as [24]:

$$\frac{d}{dt}\int\limits_{V(t)} c(\mathbf{x},t)d\mathbf{x} = \int\limits_{\mathrm{V(t)}} [-\nabla\cdot\mathbf{j} + R(c)]d\mathbf{x}\ ,\ \mathrm{x}\in\mathrm{V(t)} \tag{1.8}$$

where the left-hand side denotes time change in total amounts on domain $V(t)$ and the right-hand side describes changes in total amounts due to instantaneous in- and efflux $\mathbf{j}$ and production and decay R.

Assuming Fickian diffusion in one dimension, the instantaneous flux j can be expressed as $j = -D\nabla c$ [1](according to Fick's first law), where D is the diffusion coefficient expressing the diffusivity of the morphogen molecule.

The Reynolds transport Theorem lets us express the left side of (1.8) as:

$$\frac{d}{dt}\int\limits_{V(t)} c(\mathbf{x},t)d\mathbf{x} = \int\limits_{\mathrm{V(t)}} [\frac{\partial \mathrm{c}}{\partial \mathrm{t}} + \nabla\cdot(\mathbf{a}\mathrm{c})]d\mathbf{x}, \tag{1.9}$$

where $\mathbf{a}$ is the velocity of an arbitrary but fixed volume element (velocity field of the flow) due to the domain growth.

Note that the time rate of change (total derivative) observed on a specific volume element: $d()/dt$ in the Lagrangian system, has a counterpart in the Eulerian system,

$$\begin{aligned} d()/dt &= \frac{\partial()}{\partial t}\frac{dt}{dt} + \frac{\partial()}{\partial x}\frac{dx}{dt} + \frac{\partial()}{\partial y}\frac{dy}{dt} + \frac{\partial()}{\partial z}\frac{dz}{dt} = \\ &= \partial()/\partial t + \mathbf{a}\cdot\nabla(),\ (\mathbf{a} = \{a_1, a_2, a_3\}, a_1 = dx/dt,\ a_2 = dy/dt,\ a_3 = dz/dt), \end{aligned} \tag{1.10}$$

[1]For clarity of notation, let us explain the uses of the Nabla operator (vector operator), $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$, in a 3D system of coordinates:
- if a scalar field c is defined and continuously differentiable, the gradient of c, ∇c, is a vector field $\nabla c = (\frac{\partial c}{\partial x}, \frac{\partial c}{\partial y}, \frac{\partial c}{\partial z})$;
- If a vector field a is defined and continuously differentiable, the divergence of a, $\nabla\cdot a$, is a scalar field $\nabla\cdot a = \frac{\partial a}{\partial x} + \frac{\partial a}{\partial y} + \frac{\partial a}{\partial z}$

The Laplacian ∇^2 is an operator (scalar operator) that can be mathematically defined as a combination of the divergence and gradient operators: $\nabla\cdot\nabla() = \nabla^2() = \frac{\partial^2()}{\partial x^2} + \frac{\partial^2()}{\partial y^2} + \frac{\partial^2()}{\partial z^2}$

called the material derivative.

We can think of Lagrangian coordinates as labels of volume elements: each volume element has a unique initial position (Lagrangian coordinate) and is labelled with this unique position. Hence, the material derivative of a generic variable describes the total rate of change in a control volume at a given position over time [105]. The vector $\mathbf{a}$ represents the velocity field of the flow on a three - dimensional system of coordinates.

From the above considerations, and replacing the left-hand side of (1.8) with the right-hand side of (1.9), allows to rewrite (1.8) as:

$$\int_{V(t)} [\frac{\partial \mathrm{c}}{\partial \mathrm{t}} + \nabla \cdot (\mathbf{a}\mathrm{c})] d\mathbf{x} = \int_{V(t)} [\nabla \cdot (D \nabla c) + R(c)] d\mathbf{x}, \tag{1.11}$$

and deriving the evolution equation ($V(t)$ is arbitrary):

$$\frac{\partial \mathrm{c}}{\partial \mathrm{t}} + \nabla \cdot (\mathbf{a}\mathrm{c}) = D \nabla^2 c + R(c) \tag{1.12}$$

The effect of the growth can be explained splitting the term $\nabla \cdot (\mathbf{a}\mathrm{c})$ into two components (according to the product rule of differentiation) [3]: an advection term $\mathbf{a} \cdot \nabla \mathrm{c}$, representing the transport of material around $V(t)$ at a rate determined by the (growth-induced) flow $\mathbf{a}$ (this term together with $\partial c/\partial t$ is the material derivative); a dilution term $c \nabla \cdot \mathbf{a}$, due to local volume expansion, meaning that the concentration decreases as volume increases.

Since the growth properties (hence the flow) of the constituents are locally determined, the Lagrangian description is used to follow the movement of tissue elements in time, as Lagrangian coordinates allow to me to use a growth function, which describes the trajectories (pathlines) of volume elements (e.g. cells).

Solid body translations and rotations of the domain do not affect the pattern generated by reaction and diffusion within the tissue. Thus, the coordinate system for the domain is chosen such that there is a reference point which is initially at the origin of the coordinate system and which remains at the origin during the subsequent flow [56].

Introducing Lagrangian coordinates $(\mathbf{X}, t)$, we specify the initial position $\mathbf{X}$ of an element moving with the flow $\mathbf{a}$ and we use a function Γ to describe the trajectories of

tissue elements, such that we can express $\mathbf{x} = \Gamma(\mathbf{X}, t)$ as the paths of a volume element (e.g. a cell) with initial position $\mathbf{X}$, meaning $\Gamma(\mathbf{X}, 0) = \mathbf{X}$: the coordinates of the system are specified such that there is a reference point which is initially in the origin of the system and that remains in the origin. Hence, in the absence of both solid body rotation and translation we have the boundary condition $\Gamma(0, t) = 0$.

I can always write the inverse of Γ as $X = \Lambda(x, t)$, which gives the initial location of a particle at position x at time t.

Please note that Γ is chosen freely and encodes desired growth properties such as uniformity or non - uniformity of the growth (an example is reported in Figure 1.13), keeping in mind that the two conditions $\Gamma(\mathbf{X}, 0) = \mathbf{X}$ and $\Gamma(0, t) = 0$ have to be respected.

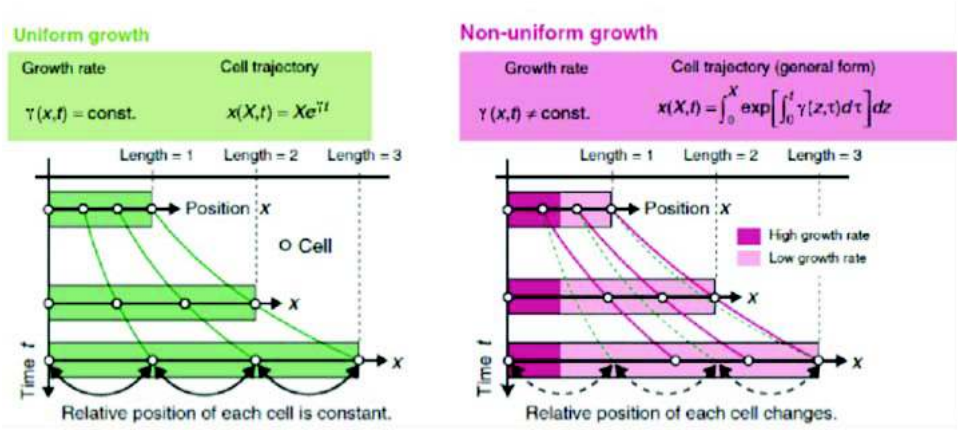


Figure 1.13: **Cell trajectories defined by $\mathbf{x}(\mathbf{X}, t) = \Gamma(\mathbf{X}, t)$ in growing tissues in the presence of uniform exponential growth (left) and non-uniform growth (right)**. Modified from [61].

Now, we need to evaluate the term due to tissue deformation $\nabla \cdot (\mathbf{a}c)$ in Lagrangian coordinates[2].

Firstly note that, given $a(\mathbf{x}, t) = \frac{d\mathbf{x}}{dt}$ and $\mathbf{x} = \Gamma(\mathbf{X}, t)$ the following holds: $a(\mathbf{X}, t) = \frac{\partial \Gamma}{\partial t}$.

[2]notation: $\Gamma_X = \frac{\partial \Gamma}{\partial X}, \Gamma_t = \frac{\partial \Gamma}{\partial t}, \Gamma_{XX} = \frac{\partial^2 \Gamma}{\partial X^2}, \Gamma_{Xt} = \frac{\partial \Gamma}{\partial X}\frac{1}{\partial t}, \Gamma_{tX} = \frac{\partial \Gamma}{\partial t}\frac{1}{\partial X}$,usually $\Gamma_{Xt} = \Gamma_{tX}$

Local tissue deformation can be expressed with the rate of strain tensor D_{ij} (in 1D case $i, j = 1$), corresponding to the local rate of volumetric expansion (contraction) $S(\mathbf{X}, t)$ [56]:

$$D_{11} = \nabla \cdot \mathrm{a}(\mathbf{X}, \mathrm{t}) = \frac{\partial}{\partial \mathrm{x}} \left(\frac{\partial \Gamma}{\partial \mathrm{t}} \right) = \frac{\partial}{\partial \mathrm{x}} \left(\Gamma_{\mathrm{t}} \right) = \frac{\Gamma_{\mathrm{tX}}}{\Gamma_{\mathrm{X}}} =: \mathrm{S}(\mathrm{X}, \mathrm{t}) \tag{1.13}$$

having applied the chain rule for composite functions[3]. It results clear that for domain growth it is required to be $S(\mathbf{X}, t)>0$.

Thus, we can rewrite $c\nabla\cdot\mathbf{a} = Sc$ and observe that the advection term $\mathbf{a} \cdot \nabla \mathrm{c}$ now represents the transport of chemical within the tissue such as there is no movement of the chemical relative to the tissue.

The evolution equation in Lagrangian coordinates becomes:

$$\frac{\partial \mathrm{c}(\mathbf{X}, t)}{\partial \mathrm{t}} + \mathbf{a} \cdot \nabla \mathrm{c} = D\nabla^2 c + R(c) - S(\mathbf{X}, t)c \tag{1.14}$$

To make the operator ∇ explicit in terms of Γ, let us rearrange single terms of equation (1.14) for 1D growth:

$$\nabla c = \frac{\partial \mathrm{c}}{\partial \mathrm{x}} = \frac{\partial c}{\partial X} \cdot \frac{1}{\Gamma_X} \tag{1.15}$$

[3]Following the notation reported in [26], if we call Γ_t a generic function $f(\mathbf{x}, t)$ and $\mathbf{x} = x(\mathbf{X}, \mathrm{t})$, meaning that x it is a known function of the two independent variable X and t, we can express $\frac{\partial \Gamma_t}{\partial X} = \frac{\partial f}{\partial X}$ as:

$$\frac{\partial f}{\partial X} = \frac{\partial f}{\partial x} \frac{\partial x}{\partial X}$$

Hence

$$\frac{\partial \Gamma_t}{\partial X} = \frac{\partial \Gamma_t}{\partial x} \frac{\partial x}{\partial X}$$

Rearranging the above expression to solve for $\frac{\partial \Gamma_t}{\partial x}$ and remembering that our known function is $\Gamma : x = \Gamma(X, t)$:

$$\frac{\partial \Gamma_t}{\partial x} = \frac{\partial \Gamma_t}{\partial X} \frac{\partial X}{\partial x} = \Gamma_{tX} \cdot \frac{\partial X}{\partial \Gamma} = \Gamma_{tX} \cdot \frac{1}{\Gamma_X} = \frac{\Gamma_{tX}}{\Gamma_X}$$

[4]

$$\nabla^2 c = \frac{\partial^2 \mathrm{c}}{\partial \mathrm{x}^2} = \frac{\partial \mathrm{c}}{\partial X} \cdot \left(-\frac{\Gamma_{XX}}{\Gamma_X^3} \right) + \frac{1}{\Gamma_X^2} \frac{\partial^2 \mathrm{c}}{\partial X^2} \tag{1.16}$$

[5]

To find the trajectories $\Gamma(\mathbf{X}, t)$ the following general link between $\Gamma(\mathbf{X}, t)$ and the strain rate $\mathrm{S}(\mathbf{X}, \mathrm{t})$ holds:

$$\Gamma(\mathbf{X}, t) = \int_0^X [e^{\int_0^t \mathbf{S}(z,\tau) d\tau)}] dz \tag{1.17}$$

Substituting (1.15) and (1.16) in (1.14):

$$\frac{\partial \mathrm{c}(\mathbf{X}, \mathrm{t})}{\partial \mathrm{t}} = D \left(-\frac{\Gamma_{XX}}{\Gamma_X^3} \frac{\partial \mathrm{c}}{\partial X} + \frac{1}{\Gamma_X^2} \frac{\partial^2 \mathrm{c}}{\partial X^2} \right) + R(c) - S(\mathbf{X}, t) c \tag{1.18}$$

for uniform growth, in which we can neglet the influence of the advection term in patterning the system. In this case, given that the strain rate is uniform across the domain, $\Gamma(\mathbf{X}, t)$ becomes a linear expression:

$$\Gamma(\mathbf{X}, t) = X \cdot (1 + \mathbf{a}t) \tag{1.19}$$

For non uniform growth

[4] If $c = c(\mathbf{x}, t)$ and $\mathbf{x} = x(\mathbf{X}, \mathrm{t})$ the chain rule let us to express $\frac{\partial c}{\partial X}$ as:

$$\frac{\partial c}{\partial \mathbf{X}} = \frac{\partial c}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial \mathbf{X}}$$

Substituting to x our known function $\Gamma : x = \Gamma(X, t)$,

$$\frac{\partial c}{\partial X} = \frac{\partial c}{\partial x} \cdot \frac{\partial x}{\partial X} = \frac{\partial c}{\partial x} \cdot \frac{\partial \Gamma}{\partial X} = \frac{\partial c}{\partial x} \cdot \Gamma_X \Rightarrow \frac{\partial c}{\partial x} = \frac{\partial c}{\partial X} \cdot \frac{1}{\Gamma_X}$$

[5] $\frac{\partial^2 \mathrm{c}}{\partial \mathrm{x}^2} = \frac{\partial}{\partial x}\left(\frac{\partial c}{\partial x}\right) = \frac{\partial}{\partial [\Gamma(X,t)]}\left(\frac{\partial c}{\partial x}\right) = \frac{\partial}{\partial X}\left(\frac{\partial c}{\partial x}\right) \cdot \frac{\partial X}{\partial [\Gamma(X,t)]} = \frac{\partial}{\partial X}\left(\frac{\partial c}{\partial X} \cdot \frac{1}{\Gamma_X}\right) \cdot \frac{\partial X}{\partial [\Gamma(X,t)]} = \left[\frac{\partial c}{\partial X} \cdot \left(\frac{\partial}{\partial X}\left(\frac{1}{\Gamma_X}\right)\right) + \frac{1}{\Gamma_X} \cdot \left(\frac{\partial}{\partial X}\left(\frac{\partial c}{\partial X}\right)\right)\right] \cdot \frac{1}{\Gamma_X} = \left[\frac{\partial c}{\partial X} \cdot \left(-\frac{1}{\Gamma_X^2} \cdot \Gamma_{XX}\right) + \frac{1}{\Gamma_X} \cdot \frac{\partial^2 c}{\partial X^2}\right] \cdot \frac{1}{\Gamma_X} = \frac{\partial \mathrm{c}}{\partial X} \cdot \left(-\frac{\Gamma_{XX}}{\Gamma_X^3}\right) + \frac{1}{\Gamma_X^2} \frac{\partial^2 \mathrm{c}}{\partial X^2}$

$$\frac{\partial \mathrm{c}(\mathbf{X},\mathrm{t})}{\partial \mathrm{t}} = D\left(-\frac{\Gamma_{XX}}{\Gamma_X^3}\frac{\partial \mathrm{c}}{\partial X} + \frac{1}{\Gamma_X^2}\frac{\partial^2 \mathrm{c}}{\partial X^2}\right) - \frac{\partial \Gamma}{\partial t}\left(\frac{1}{\Gamma_X}\frac{\partial \mathrm{c}}{\partial X}\right) + R(c) - S(\mathbf{X},t)c \quad (1.20)$$

In this case, $\Gamma(X,t)$ is a non-separable function of X and t and the convection term is not removed under a uniform transformation and may affect pattern formation in the system [56].

1.3.5 Models of Auxin Gradients

Auxin acts in gradients to trigger the developmental fate of target cells. Well known examples of auxin gradients regulate and maintain the root meristem, embryo development, the shoot apical meristem at the primordia initiation sites and the leaf vascular system (see previous section).

Unlikely most common source- or sink-driven morphogen gradients, auxin differential distribution is unique in that it is regulated by auxin metabolism (i.e., biosynthesis, conjugation/deconjugation, and degradation) as well as by PIN-directed efflux through sink tissue [10, 130, 74]. Auxin polar transport directed by its efflux carriers PINs creates a robust transcriptional auxin response gradient that peaks in the QC and gradually decreases as stem cell daughters divide in the MZ [118, 50, 17, 51]. Due to the complexity of the system and to the feedback mechanisms involved, a major issue remains to untangle which are the tuning factors that maintain a stable auxin maximum for the entire life span of the meristem, determining the patterning and the growth of the root [12].

Experimental evidence has provided qualitative insights into the different components involved in auxin distribution patterns [45]. Meristematic auxin gradients have been firstly inferred from the transcriptional response to auxin using the DR5 promoter (Figure 1.14 a). Early IAA indirect measurements through auxin-responsive reporter genes (DR5) have demonstrated that the generation of this maximum of concentration and its perception are essential to maintain stem cell function [115, 79], while auxin gradients along the meristem fuel cell proliferation in the division zone [29]. Despite its

high responsiveness to even low levels of auxin, the DR5 promoter does not completely reflect auxin distribution due to cell-specific differential responsiveness [6]. More recently, direct quantification of isolated cell types (cell-sorting) [102] (Fig 1.14b) and the development of the sensor DII-VENUS [118, 17, 137], introduced a new level of sensitivity: both methods enable the visualization of graded levels of auxin signaling intensity, suggesting the presence of cell type-specific differences in auxin distribution (Fig. 1.14c, bottom) that can be translated in maps of auxin distribution at a cellular resolution (Figure 1.14c, top)[102, 17, 6]. In particular, DII-VENUS expression is inversely correlated with auxin levels. While auxin relative levels can be inferred through these methods, measuring auxin concentration at cellular resolution is still not possible.

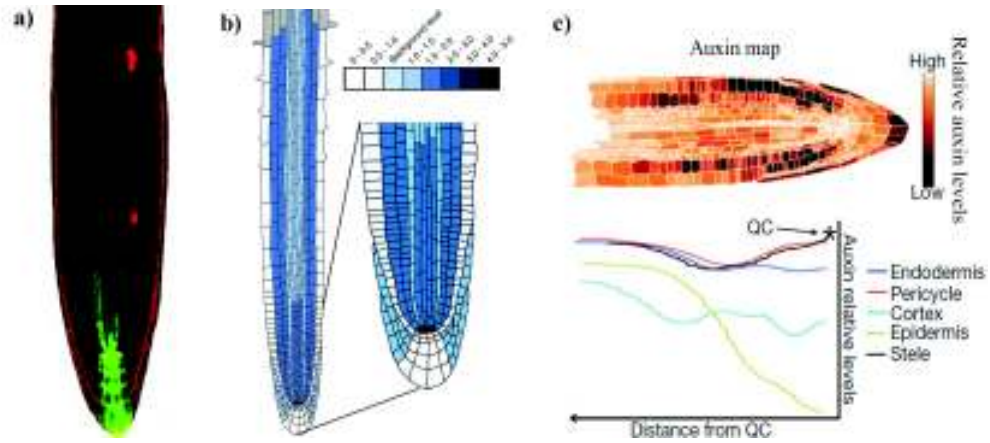


Figure 1.14: **Experimental evidence on tissue-specific auxin distribution in the root meristem**. a) DR5 expression in the root apex. A maximum of green fluorescence is correlated with a maximum of auxin. b) IAA distribution pattern in the root apex, based on cell sorting data. Auxin levels are scaled to a background level, in arbitrary units. c) Top, map of auxin distribution as resulting from DII-VENUS sensor measurements. Bottom, tissue-specific auxin gradients along the root axis. The profiles evidence the maximum in the quiescent centre in the central files of the root.

The sole interpretation of biological data, however, does not give sufficient clues to gain a quantitative understanding on the non-intuitive relations between local morphogenetic processes and global patterns [5], especially in biological systems driven by control circuits with feedback loops and complex hierarchies [106]. The unique feature

of auxin transport, which integrates the feedback emerging at cellular level with the formation of global flow and pattern at tissue level, has prompted the development of modelling works at different scales. In fact, the use of computational modelling enables a simplified yet global description of biological mechanisms taking place at different levels, allowing for the prediction and the estimation of the parameters used [68, 45, 83]. In recent years, several works have been focusing on modelling auxin transport, spanning from continuous -at tissue level- to discrete -at cellular level- descriptions. At subcellular scale, Kramer (2004) [66] first developed a computational model to investigate which process prevails in auxin transport between cells, concluding that diffusion cannot prevail at long range, where active transport is predominant. At the cellular level, Jonsson et al. (2006) [60] and Smith et al. (2006a) [122] explored the importance of PIN localization in the formation of auxin maxima. Through their analysis, they have showed that PINs polarize themselves preferentially towards cells with high concentration of auxin, promoting further auxin flux which determines the formation and sustainment of auxin maxima. The authors have further explained that this feedback mechanism can generate one- and two-dimensional periodic patterns of isolated auxin maxima, resembling patterns resulting from reaction–diffusion models [106].

One dimensional models have also addressed the interplay between auxin and cytokinin [31, 30, 89, 25]. For instance, Muraro et al. (2013) [90] focused on the role of hormonal cross-talk in meristem zonation and lateral root formation, through a subcellular and multi-cellular mathematical description.

Moving to a two-dimensional picture, Grieneisen et al. (2007) [50] developed a 2D computational model using a structured root layout, successfully predicting that tissue-specific PINs localization drives a reflux-loop mechanism, which creates a reverse fountain flux centered on a stable auxin maximum in the QC, as shown in Figure 1.15a; Band et al. (2014) [6] recently developed a computational model based on actual root cell geometries and both AUX/LAX influx carriers distribution and PIN localization, which predicts cell-type specific auxin distribution (Figure 1.15b) in good agreement with DII-VENUS auxin sensor measurements.

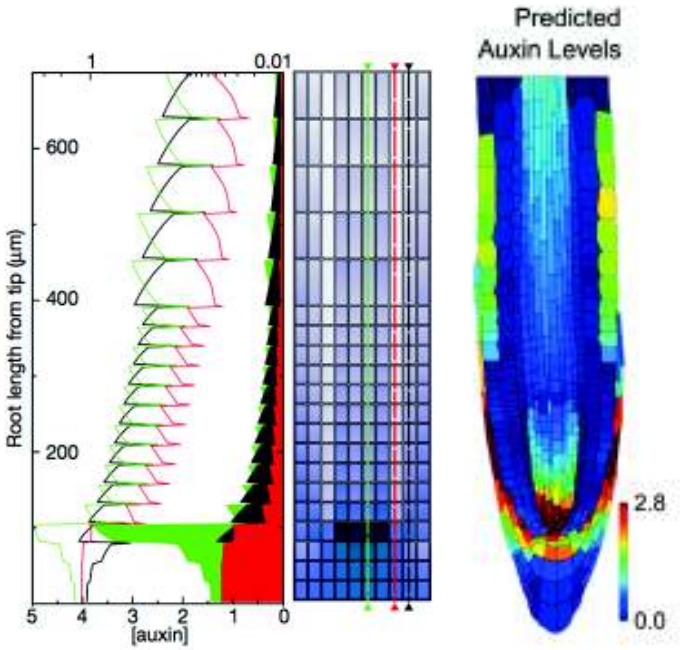


Figure 1.15: **Computational models predict auxin distribution in the root of *A. Thaliana*.** a) Steady state auxin concentration profiles (left) obtained by an *in silico* root model on a cell-structured rectangular layout (right). The model predicts the formation of a PIN-driven auxin maximum in the quiescent centre for all tissues of the root (green profile: vasculature; red profile: pericycle and endodermis; black profile: external tissue). Taken from [50]. b) Steady state auxin heat map obtained by a model based on cell geometry that predicts auxin distribution at a cellular scale and reproduces with high fidelity DII-VENUS pattern [6]. This recent model predicts high levels of auxin both in the vasculature and in the external files of the MZ, as opposed to the decaying gradient shown in a).

Further computer-based models have suggested that, together with auxin polar transport, other mechanisms are necessary for the generation of an auxin-maximum guided root patterning, such as local auxin biosynthesis and response machinery [22, 130, 74].

Chapter 2

Model description

Previous models on auxin gradient formation have clarified that auxin transport and hormonal interaction in the root meristem are required for root patterning and development. However, some fundamental questions are left unanswered. In particular, during my PhD project I seeked to understand two main processes: how cells interpret auxin morphogenetic thresholds to create a developmental response, and how the control of cytokinin shapes auxin gradients, setting a stable transition zone, thus meristem size.

Due to the complexity of the problem, we adopted a systems biology approach combining experimental tools and modelling at different scales, in order to untangle the non-intuitive relations between local morphogenetic processes and global patterns and forms [68, 106].

To first understand the link between physical processes involved in auxin transport, we worked out an analytical and theoretical description that links a cell-based description to tissue-scale dynamics on a one-dimensional domain, through the application of physical laws. The derivation of a cell-specific equivalent diffusivity allows for a direct link both between parameters and between discrete and continuous formalisms: discrete descriptions explicitly consider the role of each cell in morphogen transport and the cellular processes underlying gradient formation (recycling, intra- and extracellular degradation, receptor binding,...); instead, using a continuum approach, we allow the morphogen concentration changing continuously in space and the behaviour of single cells is neglected, to capture the essential biophysical principles that govern morphogen

transport and account for gradient shape [94]. An extensive discussion is provided in Section 2.1.

Moving forward, to investigate the cell-specific interplay between auxin and cytokinin and its effect on meristem size, we extended a previously developed two-dimensional computational model where auxin diffusion and PIN-facilitated auxin transport were integrated within a cell-structured root layout [50]. A detailed analysis on this approach is provided in Section 2.2.

2.1 Theoretical 1D model

2.1.1 Abstract

In *Arabidopsis thaliana* root meristem, the distribution of phytohormone auxin plays a pivotal role in the regulation of patterning and growth. Computational approaches have provided a good qualitative description of auxin transport, but the link between physico-chemical (diffusivity, auxin distribution) and biological (gene expression) descriptors is still missing.

In this work I focus on the physico-chemical description, approaching a multiscale analysis (cell, tissue, organ level), with the final goal of setting a framework for a quantitative assessment. I worked out a simple analytical and theoretical description that links a cell-based description to tissue-scale dynamics, through the application of physical laws, providing a straightforward condition for the formation of a stable auxin maximum in the root tip of the meristem. I first derived a multi-cellular discrete model, providing the constraints required for the observed auxin maximum. In this context, I derived an equivalent diffusivity that includes both passive and PIN facilitated diffusion and varies at each discrete position within the cell -being the cytoplasm or the apoplast- and across cells.

The equivalent diffusivity revealed to be a direct link to the properties emerging at tissue level: in the limit of a continuous description, I was able to derive a linear diffusion equation, where all transport components are embedded within the equivalent diffusivity parameter. I was eventually able to estimate an average value for the

equivalent diffusion coefficient.

Extending the analysis to the organ scale I provide further conditions for the "reflux fountain" of auxin in the meristem. As an ultimate goal, I envision that this formalism could be used as a tool for the estimation of parameters given sensor-derived auxin maps.

2.1.2 Methods

I developed a one-dimensional model, in which auxin moves through advective-diffusive transport along a file of N cells that represents, for example, a cross section through a root meristem. To capture the unique auxin physico-chemical properties (as discussed in paragraph 1.2.4.1), the model includes both passive and active auxin transport across the cell membranes, as well as auxin diffusion within the cytoplasm and the apoplast (cellular sub-structure is taken into account). A recursive formula let us link auxin concentration between non adjacent cells, travelling through the cell wall of each cell before entering a cell cytoplasm.

I computed auxin flux both across the central files (vasculature) and the peripheral files (epidermis) of the meristem, uncoupling -at the discrete level- the rootward and shootward direction of auxin transport. Evaluating the concentration entering and exiting each file, I was able to link these fluxes, obtaining a global description of auxin reflux in the root.

Evaluation of Characteristic Times

Advection and diffusion phenomena that drive auxin transport are embedded in a strongly dynamic system. In order to understand if a steady state description could reasonably hold, I ruled out characteristic times of such processes, so to compare them to the characteristic time of the growth.

Parameters for the evaluation of characteristic times are reported in Table 2.1.

Parameter description	Parameter value
Average length of *A. thaliana* root (L)	$500\,\mu m$
Auxin cytoplasmic diffusion coefficient in water (D)	$600\,\mu m^2/s$
Active transport rate (P_{PIN})	$8\,\mu m/s$

Table 2.1: Parameter set used in the 1D model to analyze characteristic times.

- Diffusive characteristic time: $\tau_D = \frac{L^2}{D} = \frac{(500)^2 \mu m^2}{600 \mu m^2/s} = 416s$
- PIN-transport time scale: $\tau_{PIN} = \frac{L}{P_{PIN}} = \frac{500 \mu m}{8 \mu m/s} \sim 63s$
- Cell cycle (mitosis) time: $\sim 24h$ [79, 90]
- Root growth rate: $g = 0.2\,\mu m/s$ [114]

When studying transport phenomena, the evaluation of *Peclet* (Pe) number enables to calculate the ratio of the advective time scale to the diffusive time scale. When $Pe \gg 1$, advection will dominate. Thus, being Peclet number: $Pe = \frac{P_{PIN} * L}{D} = \frac{8 \cdot 500}{600} \sim 6.7$, I first identified PIN polar transport as the dominating process. Then, given that the growth of the meristem spans slow time scales compared to auxin polar transport [90, 5], I was confident about performing the analysis for a pseudo steady-state snapshot, assuming the phenomenon to be steady at each time step [5].

2.1.2.1 Discrete Model

The modelled tissue consists of a single cell file (i.e, a monodimensional array of plant cells) and takes into account both plant cell geometry and carrier localization, as sketched in Figure 2.1.

Considering an idealized root cut, cells are numbered from 1 to N. The first cell is the most apical cell located at the boundary with the shoot [73] and the $N - th$ represents a cell of the root tip. A single cell from the $i - th$ row is singled out. To reproduce the internal structure of plant cells, each cell is modelled as a unit (see the

inset in Figure 2.1) composed of a cell wall (dark green region), a plasma membrane (light blue line) and a cytoplasm (light green region).

Auxin Flux in the Vasculature

I first focused on the description of auxin distribution along the vascular tissue of the root, where PINs direct auxin flux rootward, along the longitudinal axis.

I assumed PINs to be strongly expressed -higher density- on the basal side of the plasma membrane with permeability P_{PIN}, as well as a background permeability P_{BG}, supplying weak PINs expression on the apical membrane, according to the framework used in [50]. The presence of a background component represents a feedback mechanism on every cell, which leads to a systemic behaviour, giving the root the peculiar property of auto-regulation.

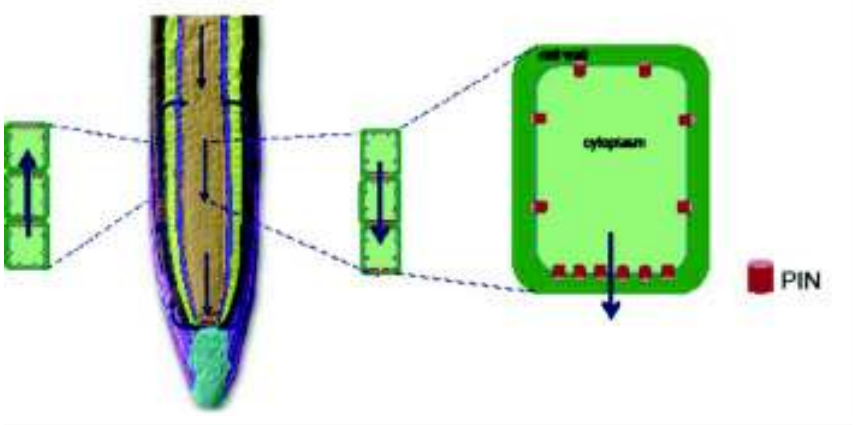


Figure 2.1: **Sketch of the 1D modelled tissue**. To build a 1D model, a file of N cells is singled out (blow up). The model takes into account plant cell structure as composed of cell wall (dark green), cell membrane (light blue) and cell cytoplasm (light green). PIN polar localization on the membrane is also taken into account when modelling the rootward flux in the vasculature (high PIN density on the basal membrane) or shootward flux (high PIN density on the apical membrane).

In correspondence to the first cell, auxin enters with a flux J_0 and a concentration c_0; neglecting local auxin production and degradation, the auxin flux entering the adjacent

cell is constantly J_0. Focusing on the single cell of the i-th row, fluxes entering and leaving the cell are those sketched in the magnified Fig. 2.2.

Considering a section encompassing z and $z + \triangle z$, it will be:

$$J_0 = -D\frac{dc^{(i)}}{dz}, \tag{2.1}$$

D is the diffusion coefficient of auxin in water.

Solving this first order linear differential equation, the concentration profile in the $i-th$ cell is:

$$c^{(i)} = -\frac{J_0}{D}z + c_1^{(i)} \tag{2.2}$$

The constant $c_1^{(i)}$ is obtained applying the boundary condition at $z = 0$:

$$c^{(i)} = c_{in}^{i}, \quad z = 0 \tag{2.3}$$

It follows:

$$c^{(i)} = c_{in}^{(i)} - \frac{J_0}{D}z \tag{2.4}$$

A direct link between auxin concentration entering and exiting the cytoplasm in correspondence to $z = l_c$, l_c being the cytoplasmic cell length, can be established, :

$$c_{out}^{(i)} = c_{in}^{(i)} - \frac{J_0}{D}l_c, \quad z = l_c \tag{2.5}$$

Thus, the flux through the cytoplasmic area of the cell, is:

$$J_0 = \frac{D}{l_c} \cdot (c_{in}^{(i)} - c_{out}^{(i)}) = P_C \cdot (c_{in}^{(i)} - c_{out}^{(i)}) \tag{2.6}$$

$P_C = D/l_c$ is the cytoplasmic permeability to auxin. Given that subcellular spatial variations in the root tip can be considered negligible [6], I supposed auxin concentration to be homogeneous within each compartment (cytoplasm and cell wall).

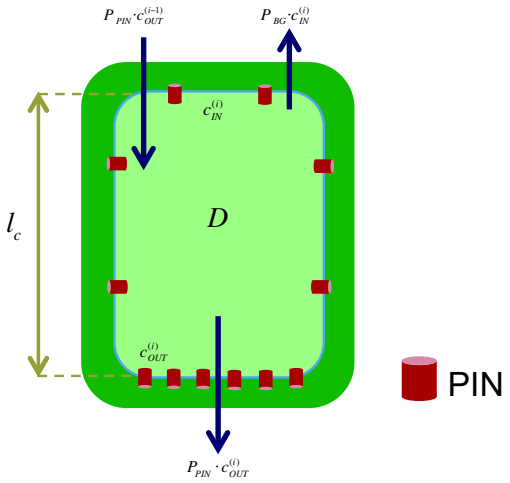


Figure 2.2: **Flux balance over a $i-th$ vascular cell.** PINs (red cylinders) are highly expressed on the basal membrane, transporting auxin outside the cell ($c_{out}^{(i)}$), towards the root tip. A background flux exits the cell, carried by apolar transporters with low permeabilty (P_{BG}).

Extending the balance to the whole $i-th$ cell, the net flux will result from two contributions: flux directed towards increasing z values, due to auxin transported from the adjacent cell above, and flux directed upwards, due to the reflux dependent on the background permeability in the cell:

$$J_0 = P_{PIN} \cdot c_{out}^{(i-1)} - P_{BG} \cdot c_{in}^{(i)} = P_{PIN} \cdot c_{out}^{(i)} - P_{BG} \cdot c_{in}^{(i+1)} \tag{2.7}$$

The flux due to PINs can be roughly described as the product of the specific permeability for a single PIN molecule and the superficial density of PINs; this way, the ratio between the background P_{BG} and PIN concentration in the polarized region P_{PIN} can be defined as the *polarization factor* $\alpha = P_{PIN}/P_{BG}$, indicating the overexpression of PINs in polarization regions with respect to the background, due to low superficial density of other membrane regions as well as to a passive permeability (leakage),occurring upen a shift in the equilibrium towards the lipophilic form of auxin within the cell wall [50] .

It can be easily derived from eqs. 2.5 and 2.7:

$$c_{in}^{(i+1)} = \frac{P_{PIN}}{P_{BG}} \cdot c_{in}^{(i)} - \frac{J_0}{P_{BG}} \cdot \left(1 + \frac{P_{PIN}}{P_C}\right) = \alpha \cdot c_{in}^{(i)} - J_0 \cdot \alpha \cdot \beta \tag{2.8}$$

where $\beta = \left(\frac{1}{P_{PIN}} + \frac{1}{P_C}\right)$.

Therefore, the concentration of auxin in cells depends on that of adjacent cells, through the feedback mechanism of PIN efflux. Applying eq. 2.8 in a recursive fashion, it is possible to link the concentration of non-adjacent cells. For instance, the concentration of a generic cell belonging to the $j-th$ row depends on that in the cell of the $k-th$ row ($j > k$):

$$c_{in}^{(i)} = \alpha^{j-k} \cdot c_{in}^{(k)} - J_0 \cdot \beta \cdot \sum_{m=1}^{j-k} \alpha^m \tag{2.9}$$

The balance for the first row cells (at the boundary with the shoot) is:

$$J_0 = P \cdot c_0 - P_{BG} \cdot c_{in}^{(1)} \tag{2.10}$$

being P the cell wall permeability and choosing the boundary conditions as J_0 and c_0 at the cell wall of the first row cells.

Thus the auxin amount entering the firts row of cells is:

$$c_{in}^{(1)} = \frac{P \cdot c_0}{P_{BG}} - \frac{J_0}{P_{BG}} \tag{2.11}$$

Choosing $k = 1$, eq.2.9 becomes[1]:

$$c_{in}^{(i)} = \alpha^{j-1} \cdot c_{in}^{(1)} - J_0 \cdot \beta \cdot \sum_{m=1}^{j-1} \alpha^m = \frac{\alpha^j}{P_{PIN}} \cdot [P \cdot c_0 - J_0] - \frac{J_0 \cdot \alpha \cdot \beta}{1-\alpha} \cdot [1 - \alpha^{j-1}] \tag{2.12}$$

Once the auxin is transported by PINs into the cell wall, a balance at the frontier between the cell membrane and the wall (Figure 2.3) is established:

[1]Exploiting the properties of the geometric finite series, it is possible to write:$\sum_{j=m}^{n} \alpha^m = \frac{\alpha^m - \alpha^{n+1}}{1-\alpha}$

$$P_{PIN} \cdot c^{i}_{out} = P \cdot c^{(i)}_{W} \tag{2.13}$$

being $c^{(i)}_{W}$ the concentration at the entrance of the wall of the $i-th$ cell.

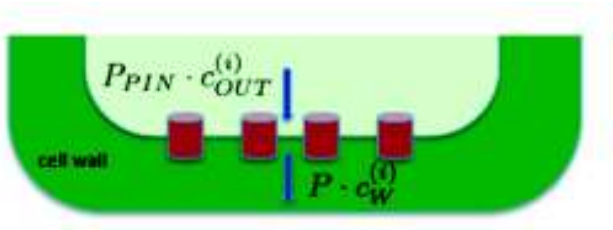


Figure 2.3: **Flux balance at the cell wall interface.** Auxin concentration in the cell wall depends on auxin concentration carried out the cytoplasm by PINs.

This expression holds prescribing auxin concentration at the membrane of the $i-th$ cell to be zero, due to the negligible permeability of the undissociated form of IAA in the cytoplasm, because of the pH unbalanced conditions between cell wall and cytoplasm:

$$c^{(i)}_{W} = \frac{P_{PIN}}{P} \cdot c^{(i)}_{out} \tag{2.14}$$

Auxin Flux in the Columella

Columella cells are located below the QC and are surrounded by a layer of root cap. Auxin concentration in these cells can be derived in a similar way as in the vascular tissue.

The concentration value of the first columella cell – at the boundary with the QC in the vasculature - corresponds to the concentration found in the last cell of the vascular tissue. Moreover, in these cells, PINs are expressed on all sides of the cell, thus one can state that the concentration exiting from a cell is the same as the concentration entering the cell below:

$$c^{(i)}_{in} = c^{(i-1)}_{out} \tag{2.15}$$

Within the cell, the gradient due to diffusion can be obtained as for vascular cells, i.e,

equation 2.5. The concentration profile for all columella cells (N_{col}) can be then easily derived.

Auxin Flux in the Epidermis

In the external files of the root meristem, PINs are localised on the apical membrane of the cell, directing auxin flux shootward. Given the derived flux across vascular cells J_0, the flux entering the epidermis can be obtained applying the flux balance to $N-th$ cell of the vasculature (i.e., QC cells). The coordinate system is oriented in the shootward direction, in accordance with the direction of auxin transport in the epidermis. QC cells are located at the boundary with columella cells, which redistribute auxin in all directions according to PIN apolar expression in these cells. Given the symmetry of the root, and given that in our schematized geometry the columella encompasses the entire width of the root (see Figure 2.5), as boundary conditions for auxin flux entering the first epidermal cell we set the same flux that flows out of QC cells, which we called J_f, and for the equivalent auxin concentration we set $c_f/2$ for each of the two epidermal files (i.e., the half of auxin concentration value in the QC).

Following the same approach to evaluate the auxin concentration profile in the vasculature, I derived the auxin concentration profile for epidermial cells.

As in the epidermal tissue the strength of PIN permeability (basipetal transport) has been measured as lower than in the vascular cells (acropetal transport), I set:

$$P_{PIN_{epi}} = P_{PIN} \cdot 1/2 \tag{2.16}$$

Given the concentration within the cytoplasm -c_{in} and c_{out} - and in the cell wall c_W for central and external files and columella cells, auxin longitudinal graded distribution in the root meristem can be fully described.

2.1.2.2 Continuous Model

Once derived the concentration profiles both for the vasculature and the epidermis, I sought to match the discrete description of the auxin transport with a continuous picture: my formulation leads to an equivalent diffusion D_{eq}, which includes all transport

terms that may be used in continuous mass balance equations. Neglecting the auxin consumption/production rate within the tissue, the flux J_0 through any cell is constant, and for the $i-th$ cell:

$$J_0 = \frac{D_{eq}^{(i)}}{l_c + 2 \cdot l_w} \cdot (c_{in}^{(i+1)} - c_{in}^{(i)}) \tag{2.17}$$

being l_w the cell wall thickness; thus, it is derived:

$$D_{eq}^{(i)} = \frac{D_{eq}^{0}}{\alpha^i} \tag{2.18}$$

where:

$$D_{eq}^{0} = \frac{(l_c + 2 \cdot l_w) \cdot J_0}{\left[\frac{P \cdot c_0}{P_{PIN}} - J_0 \cdot \left(\frac{1}{P_{PIN}} + \frac{\beta}{\alpha - 1}\right)\right] \cdot (\alpha - 1)} = \frac{(l_c + 2 \cdot l_w) \cdot J_0}{\left[\frac{\gamma}{P_{PIN}} - \left(\frac{1}{P_{PIN}} + \frac{\beta}{\alpha - 1}\right)\right] \cdot (\alpha - 1)} \tag{2.19}$$

the part of the equivalent diffusivity not depending on the position.
This equation can be recast reminding $\beta = \left(\frac{1}{P_{PIN}} + \frac{1}{P_{BG}}\right)$; thus:

$$D_{eq}^{0} = \frac{D \cdot P_{PIN} \cdot (l_c + 2 \cdot l_w)}{D \cdot [\gamma \cdot (\alpha - 1) - \alpha] - P_{PIN} \cdot l_c} \tag{2.20}$$

Thus, the cytoplasmic diffusivity D is transformed into an equivalent diffusivity, that can be expressed as:

$$D_{eq}^{(i)} = \frac{D \cdot P_{PIN} \cdot (l_c + 2 \cdot l_w)}{D \cdot [\gamma \cdot (\alpha - 1) - \alpha] - P_{PIN} \cdot l_c} \cdot \frac{1}{\alpha^i} \tag{2.21}$$

The equivalent diffusivity depends on position and even its initial value ($i = 0$) D_{eq}^{0} is a nonlinear function of the cytoplasm diffusivity D; specifically, it has a hyperbolic form, that reaches an asymptote when $D/lc \gg P_{PIN}$:

$$D_{eq}^{0,lim} = \frac{P_{PIN} \cdot (l_c + 2 \cdot l_w)}{[\gamma \cdot (\alpha - 1) - \alpha]} \tag{2.22}$$

In order to link the discrete description to a continuous approach, I rearranged the

expression derived for $D_{eq}^{(i)}$: the discrete indices i and were turned into a unit of length x belonging to the range $0 \leq x \leq l_r$, where l_r represents the length of the root, calculated as $l_r = n_c \cdot (l_c + 2 \cdot l_w)$, where n_c stands for the number of cells in the modelled 1D tissue. The continuous approximation has already been proven to be valid in a similar study, given the small differences in concentration between adjacent cells (transport is fast on the cell scale) [5].

The resulting expression becomes:

$$D_{eq}(x) = \frac{D \cdot P_{PIN} \cdot n_c \cdot (l_c + 2 \cdot l_w)}{D \cdot [\gamma \cdot (\alpha - 1) - \alpha] - P_{PIN} \cdot l_c} \cdot \frac{1}{\alpha^x} \tag{2.23}$$

The full auxin transport equation in a continuous description is of the form of a 1D diffusion-advection partial differential equation:

$$\frac{\partial \mathrm{a(x,t)}}{\partial \mathrm{t}} = \nabla(D(x,t) \cdot \nabla a(x,t)) - \nabla(P_{PIN} \cdot a) \tag{2.24}$$

, the advective term given by PIN-directed unidirectional transport of auxin, directed rootward (i.e, from the right to the left of the chosen domain).

Now, it can be observed that in a continuous description, the contribute of the background permeability across each cell membrane would be neglected. The derived $D_{eq}(x)$ instead includes both P_{PIN} and P_{BG} terms. Thus, replacing $D_{eq}(x)$ in equation 2.24, along the x direction:

$$\frac{\partial \mathrm{a(x,t)}}{\partial \mathrm{t}} = \nabla(D_{eq}(x) \cdot \nabla a(x,t)) = \frac{\partial D_{eq}(x)}{\partial x} \cdot \frac{\partial a(x,t)}{\partial x} + D_{eq}(x) \cdot \frac{\partial^2 a(x,t)}{\partial x^2} \tag{2.25}$$

One step further, in order to derive a uniform value for the parameter $D_{eq}(x)$, we averaged $D_{eq}(i)$ over the number of cells N:

$$\overline{D_{eq}} = \frac{1}{N} \sum_{i=1}^{N} D_{eq}(i) = \frac{D_{eq}^0}{N} \sum_{i=1}^{N} \frac{1}{\alpha^i} \tag{2.26}$$

The geometric series converges to the sum $S = \frac{1}{\alpha-1}$ [2]. Thus we can approximate:

$$\overline{D_{eq}} = \frac{D^0_{eq}}{N} \cdot \frac{1}{\alpha - 1} \tag{2.27}$$

Thus, using the derived $\overline{D_{eq}}$, auxin transport equation can be written as:

$$\frac{\partial \mathrm{a(x,t)}}{\partial \mathrm{t}} = \overline{D_{eq}} \cdot \frac{\partial^2 a(x,t)}{\partial x^2} \tag{2.28}$$

2.1.2.3 Continuous Model - Growing Domain

The evaluation of characteristic times reported in Section 2.1.2 revealed a neglibigle influence of the growth on the system. However, during root growth, cells transiently 'move through' the different zones due to the root tip moving deeper into the soil. I therefore investigated the effect of the root growth in a growing domain. To this aim, I applied the theoretical formalism presented in 1.3.4, considering a one dimensional root tissue where cell movement is directed towards the root tip (i.e., hypothetically down to the soil).

Starting from the conservation of matter on a time varying domain, i.e. equation 1.8, the conservation equation of auxin can be written as:

$$\frac{d}{dt} \int\limits_{V(t)} a(\mathbf{x},t) d\mathbf{x} = \int\limits_{\mathrm{V(t)}} [-(\nabla \cdot \mathbf{j}) + \mathbf{R(a)}] \mathrm{d}\mathbf{x} \ , \ \mathrm{x} \in \Omega(\mathrm{t}) \tag{2.29}$$

where the left-hand side denotes total amount of auxin (a) varying over time, whereas the right-hand side describes changes in the total amount due to instantaneous influx and efflux $\mathbf{j}$, and to reaction terms in $R(a)$. Following the discrete formalism, I will first neglet any reaction occurring.

Considering that changes in auxin concentration results both from a diffusion and an advective term (see eq. 2.24), the instantaneous flux j accounts for these two contributes: $j = -D \cdot \nabla a - P_{PIN} \cdot a$.

[2] The common ratio of this series is $\frac{1}{\alpha}$, thus we need to ensure that $\|\frac{1}{\alpha}\| < 1$. Given that α is always > 1, the series converges

Therefore, following the formalism described in, the evolution equation for changes in auxin concentration over time and space is of the form:

$$\int_{V(t)} [\frac{\partial \mathrm{a}}{\partial \mathrm{t}} + \nabla \cdot (\mathbf{v}a)] d\mathbf{x} = \int_{V(t)} [\nabla \cdot (D \nabla a - P_{PIN} \cdot a) + R(a)] d\mathbf{x}, \tag{2.30}$$

, $\mathbf{v}$ is the velocity of the material flow due to the growth.

Integrating the previous equation (assuming $V(t)$ arbitrary) and reminding equation 1.14, the rate of changes in auxin concentration in a growing domain, can be expressed as:

$$\begin{cases} \frac{\partial \mathrm{a}(\mathbf{x},\mathrm{t})}{\partial \mathrm{t}} = D \cdot \frac{\partial^2 \mathrm{a}(\mathbf{x},\mathrm{t})}{\partial \mathrm{x}^2} - P_{PIN} \cdot \frac{\partial a(\mathbf{x},t)}{\partial x} - a(\mathbf{x},t) \frac{\partial \mathbf{v}}{\partial x} + R(a) \\ R(a) = b - \delta_a \cdot a(\mathbf{x},t) \end{cases} \tag{2.31}$$

, accounting for uniform basal auxin biosynthesis b and decay δ_a.

To observe how the shape of this gradient is influenced by growth, it is useful to use the Lagrangian coordinates system.

Thus, using expressions derived in (1.13) and (1.15,1.16) equation (2.31) becomes:

$$\frac{\partial \mathrm{a}(\mathbf{X},\mathrm{t})}{\partial \mathrm{t}} = D \left(-\frac{\Gamma_{XX}}{\Gamma_X^3} \frac{\partial \mathrm{a}}{\partial X} + \frac{1}{\Gamma_X^2} \frac{\partial^2 \mathrm{a}}{\partial X^2} \right) - P_{PIN} \cdot \left(\frac{1}{\Gamma_X} \frac{\partial a}{\partial X} \right) - S(\mathbf{X},t) \cdot a(\mathbf{X},t) + b - \delta_a \cdot a(\mathbf{X},t) \tag{2.32}$$

for uniform growth, and:

$$\begin{aligned} \frac{\partial \mathrm{a}(\mathbf{X},\mathrm{t})}{\partial \mathrm{t}} \quad = \quad & D \left(-\frac{\Gamma_{XX}}{\Gamma_X^3} \frac{\partial \mathrm{a}}{\partial X} + \frac{1}{\Gamma_X^2} \frac{\partial^2 \mathrm{a}}{\partial X^2} \right) - P_{PIN} \cdot \left(\frac{1}{\Gamma_X} \frac{\partial a}{\partial X} \right) - \\ & - \frac{\partial \Gamma}{\partial t} \left(\frac{1}{\Gamma_X} \frac{\partial \mathrm{a}}{\partial X} \right) - S(\mathbf{X},t) \cdot a(\mathbf{X},t) + b - \delta_a \cdot a(\mathbf{X},t) \end{aligned} \tag{2.33}$$

for non uniform growth.

Replacing auxin cytoplasmic diffusion by the derived equivalent diffusion, equation 2.33 can be simplified in:

$$\frac{\partial \mathrm{a}(\mathbf{X},\mathrm{t})}{\partial \mathrm{t}} = \bar{D_{eq}}\left(-\frac{\Gamma_{XX}}{\Gamma_X^3}\frac{\partial \mathrm{a}}{\partial X} + \frac{1}{\Gamma_X^2}\frac{\partial^2 \mathrm{a}}{\partial X^2}\right) - \frac{\partial \Gamma}{\partial t}\left(\frac{1}{\Gamma_X}\frac{\partial a}{\partial X}\right) - S(\mathbf{X},t)\cdot a(\mathbf{X},t) + b - \delta_a \cdot a(\mathbf{X},t) \tag{2.34}$$

In a realistic case, a non constant growth rate for cells belonging to the different zones of the root should be considered, according to the strain rate profile reported in Figure 2.4.

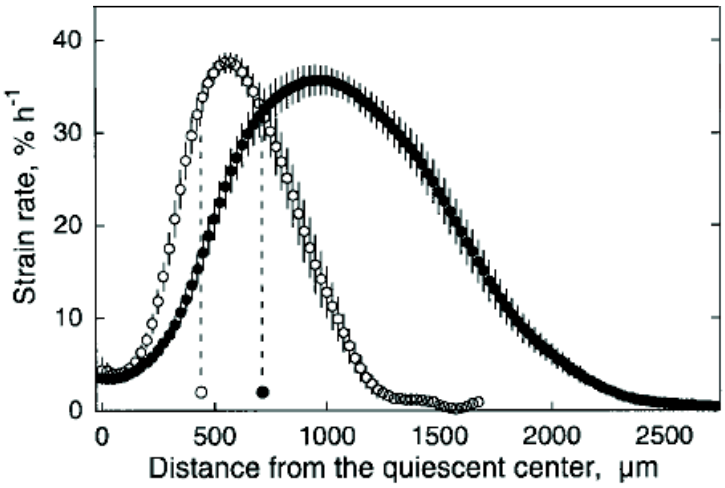


Figure 2.4: **Strain rate spatial profile for *A.thaliana* roots** calculated at day 6 (○) and 10 (●) after germination. Circles on the x axis and the vertical dashed lines mark the basal terminus of the meristem Taken from [114].

Given that in the model the length of the root is taken till $\sim 500\,\mu m$ from the QC, to obtain a function that could describe strain rate profile in figure 2.4, the expression for the strain rate was chosen as follows:

$$\mathbf{S}(\mathbf{X},\mathbf{t}) = \frac{a}{1+e^{-bX\cdot t}} \tag{2.35}$$

, with a and b positive constants.

2.1.2.4 From a Microscopic to a Macroscopic Balance

In order to integrate my systems biology approach, I linked the description derived for auxin flux along the central and the external files of the meristem, taking into account the feedback component emerging from the reverse fountain mechanism, which drives auxin reflux [50].

Adopting a macroscopic approach, I extended the microscopic balance on auxin flux over a control volume representing the meristem. To this end, I built a simplified geometry of the meristem, considering it as composed of different reactors or "building blocks", as depicted in Fig. 2.5:

In this view, auxin production and degradation were taken into account. Together with external flux coming from the shoot (J_{ext}) and auxin flux entering the vasculature (J_0), an auxin source in QC and columella cells is included, in accordance with recent findings [75, 85]. Moreover, as a first assumption, I hypothesized the molecule to be degraded at the boundary between the division and differentiation zone, where auxin enters the system due to the reflux. This assumption is necessary to close the balance within the root.

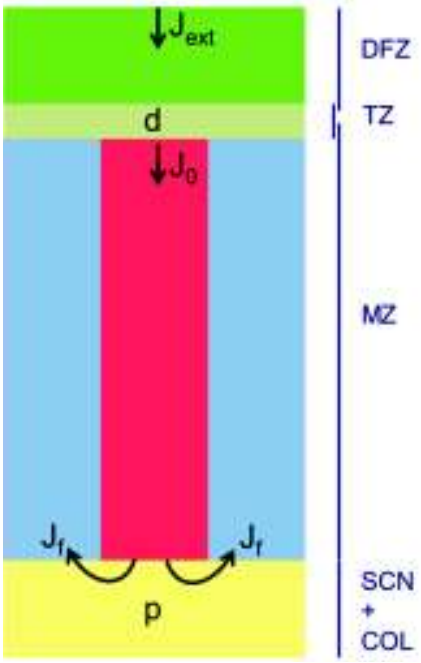


Figure 2.5: **Auxin flux balance in the *Arabidopsis* root, modelled as composed of reactors.** To extend the balance to a macroscopic picture, the root was idealized as composed of tanks. Continuous stirred tank (CSTR): differentiation zone, from where J_{ext} enters the root and SCN + columella (yellow) and transition zone (TZ, light green) with auxin generation p and consumption d, respectively; piston flow reactors (PFR): vasculature MZ (red), where auxin enters with flux J_0 and exits as J_f, which enters the epidermis MZ (light blue).

The system in figure 2.5 is then composed of:

- two piston flow regions (PFR) with no generation/depletion of auxin, i.e. vasculature (red block) and epidermis (light blue blocks), where auxin concentration varies along the y axis;
- three continuous stirred tank reactors (CSTR), two of which with generation p (QC and columella, yellow block) and depletion of auxin d (light green block), where auxin concentration is supposed to be homogeneus in space.

At the steady state, auxin flux J_{ext} entering the differentiation zone of the root (green block) is equal to the flux due to auxin transport and reaction across the regions of the meristem, indicated as a global flow g.

Assuming the section of vascular and epidermal tissue to be the same S, we can write:

$$J_{ext} \cdot 3S + g = 0 \Rightarrow g = -J_{ext} \cdot 3S \tag{2.36}$$

Flow rate balance at the boundary between columella and epidermal region is:

$$J_0 \cdot S + p = J_f \cdot 2S \tag{2.37}$$

as flux leaving the columella and entering the epidermis carries an additional contribute resulting from auxin production.

In order to keep the root as a closed system and to ensure the correct amount of auxin to flow, auxin is assumed to be degraded at the boundary between division zone and differentiation zone, i.e, in correspondence to the transition zone. Therefore, the following balance holds:

$$J_{ext} \cdot 3S + J_f \cdot 2S = J_0 \cdot S + d \tag{2.38}$$

accounting for flux coming from the top of the root, entering the epidermis and being reintroduced in the vasculature.

Substituting equation 2.37 in 2.38, I obtain:

$$J_{ext} \cdot 3S = d - p \tag{2.39}$$

To make the concentration explicit, a mass balance on the volumetric flow rate associated with auxin flow can be applied.

For each "tank" an expression for the flow rate can be written:

- At the differentiation zone: $F_{ext} = J_{ext} \cdot 3S = Q_{ext} \cdot c_{ext_0}$
- At the vasculature of MZ: $F_0 = J_0 \cdot S = Q_0 \cdot c_0$

- At the columella: $F_f = J_f \cdot 2S = Q_f \cdot c_f$. This can be written as: $F_f = F_0 + d$

I assumed that in the steady state the volumetric flow rate in the different zones of the root does not change (in turn to have a stable auxin distribution): $Q_0 = Q_{ext} = Q_f = Q$.

Thus, it follows: $c_{ext} = (d - p)/Q$.

2.1.3 Results

2.1.3.1 Limit Conditions for the Formation of an Auxin Maximum

In order to generate an auxin maximum in the QC cells, a first condition is $c_{in}^{(i+1)} > c_{in}^{(i)}$, resulting into:

$$c_{in}^{(1)} > \frac{J_0 \cdot \alpha \cdot \beta}{\alpha - 1} \tag{2.40}$$

the most restrictive condition holds for the first row ($c_{in}^{(1)}$), giving rise to the following condition for the flux J_0:

$$J_0 < P \cdot c_0 \cdot \frac{P_{PIN} - P_{BG}}{P_{PIN} \cdot \left(1 + \frac{P_{BG}}{P_C}\right)} = J_{0lim} \tag{2.41}$$

Introducing $\gamma = P \cdot c_0/J_0$ as the *enhancement factor*, i.e. the flux passing through the cell wall with respect to the incoming flux J_0, eq. 2.41 becomes:

$$\gamma > \frac{\alpha}{\alpha - 1} \cdot \left(1 + \frac{P_{BG}}{P_C}\right) = \gamma_{lim} \tag{2.42}$$

Given the eq. 2.10, the condition holding for the auxin flux through the root meristem is $J_0 < P \cdot c_0$. Therefore, recalling the condition 2.41 as more restrictive, $J_0 < P \cdot c_0$ is satisfied only if $\frac{P_{PIN} - P_{BG}}{P_{PIN} \cdot \left(1 + \frac{P_{BG}}{P_C}\right)} < 1$. It is straightforward to demonstrate that this holds always.

The equivalent diffusivity depends on position and even its initial value ($i = 0$) D_{eq}^0 is a nonlinear function of the cytoplasm diffusivity D; specifically, it has a hyperbolic form that reaches an asymptote when $D/l_c \gg P_{PIN}$:

$$D_{eq}^{0,lim} = \frac{P_{PIN} \cdot (l_c + 2 \cdot l_w)}{[\gamma \cdot (\alpha - 1) - \alpha]} \tag{2.43}$$

This term is positive (that is, the flux of auxin determines a maximum) only if $\gamma \cdot (\alpha - 1) - \alpha > 0$, that is $\gamma > \frac{\alpha}{(\alpha-1)}$; actually, this condition is less restrictive than 2.41 that can be recast as: $\gamma > \frac{\alpha}{(\alpha-1)} \cdot \left(1 + \frac{P_{BG}}{P_C}\right)$. Thus, while the condition for the maximum is ensured, $D_{eq}^{0,lim}$ is positive.

In order to D_{eq}^{0} be positive, it must stand:

$$D \cdot [\gamma \cdot (\alpha - 1) - \alpha] - P_{PIN} \cdot l_c > 0,\ P_{PIN} < P_c \cdot [\gamma \cdot (\alpha - 1) - \alpha] \tag{2.44}$$

It is straightforward to demonstrate that this condition corresponds to the auxin maximum threshold derived in equation 2.42. Dividing both members of 2.44 by P_{BG}, it is derived the following condition for the polarization factor α :

$$\alpha > \frac{\gamma}{\gamma - 1 - \frac{P_{BG}}{P_C}} = \alpha_{lim} \tag{2.45}$$

Reminding the maximum condition 2.42[3], the limit for α is always positive, and the condition 2.45 poses a lower threshold for the polarization factor as the condition for the maximum occurrence.

It can be further evaluated whether $[\gamma \cdot (\alpha - 1) - \alpha] > 1 \Rightarrow \gamma \cdot (\alpha - 1) > 1 + \alpha \Rightarrow \gamma > \frac{1+\alpha}{\alpha-1} = \frac{1}{\alpha-1} + \frac{\alpha}{\alpha-1}$

Knowing that $\gamma > \frac{\alpha}{(\alpha-1)} \cdot \left(1 + \frac{P_{BG}}{P_C}\right)$, and $\gamma > \frac{1+\alpha}{\alpha-1}$, two cases can be discussed:

1. $\frac{1+\alpha}{\alpha-1} > \frac{\alpha}{(\alpha-1)} \cdot \left(1 + \frac{P_{BG}}{P_C}\right) \Rightarrow 1 > \frac{P_{PIN}}{P_C} \Rightarrow P_{PIN} < P_C$

2. $\frac{1+\alpha}{\alpha-1} < \frac{\alpha}{(\alpha-1)} \cdot \left(1 + \frac{P_{BG}}{P_C}\right) \Rightarrow 1 < \frac{P_{PIN}}{P_C} \Rightarrow P_{PIN} > P_C$: this condition cannot hold as in 2.43 we demonstrated $\gamma \cdot (\alpha - 1) - \alpha > 0 \Rightarrow P_{PIN} < P_C$.

It can be stated that $[\gamma \cdot (\alpha - 1) - \alpha] > 1$ if $P_{PIN} < P_C$

[3]Being: $\gamma > \frac{\alpha}{(\alpha-1)} \cdot \left(1 + \frac{P_{BG}}{P_C}\right)$, since $\frac{\alpha}{(\alpha-1)} > 1$, it follows $\gamma > 1 + \frac{P_{BG}}{P_C}$

Alongside, for the epidermal files, a maximum of auxin concentration in correspondence to the first cell in the epidermal tissue of the meristem occurs if $c_{in}^{(k+1)} < c_{in}^{(k)}$:

$$c_{in}^{(k)} < \frac{J_f \cdot \alpha_{epi} \cdot \beta_{epi}}{\alpha_{epi} - 1} \tag{2.46}$$

As the most restrictive condition is for the first row, the condition for the flux J_f is:

$$J_f > P \cdot c_f \cdot \frac{PIN_{epi} - P_{BG}}{PIN_{epi} \cdot \left(1 + \frac{P_{BG}}{P_C}\right)} \tag{2.47}$$

Defining γ_{epi} as the enhancement factor for the lateral flux, equation 2.47 becomes:

$$\gamma_{epi} < \frac{\alpha_{epi}}{\alpha_{epi} - 1} \cdot \left(1 + \frac{P_{BG}}{P_C}\right) \tag{2.48}$$

To test the chosen parameter set on the limit conditions, I simulated the auxin transport 1D equation at the steady-state, in *Matlab* environment. Parameters are reported in Table 2.2.

PARAMETER DESCRIPTION		PARAMETER SYMBOL	PARAMETER VALUE
Boundary values	Auxin flux entering the vasculature	J_0	$10^{-4}\, a.u./\mu m^2 \cdot s$
	Initial auxin concentration	c_0	$1\, a.u.$
Transport parameters	Cell wall permeability	P	$1\, \mu m/s$ (in accordance with [67])
	PIN permeability in the vasculature	P_{PIN}	$8\, \mu m/s$ (in accordance with [67])
	PIN permeability in the epidermis	$P_{PIN_{epi}}$	$4\, \mu m/s$ (in accordance with [71, 67])
	Background permeability	P_{BG}	$0.7 \mu m/s$
Physical parameters	Number of cells in the meristem zone	N	30
	Number of cells in the columella	N_c	5
	Cell length (cytoplasm)	l_c	$8\, \mu m$
	Cell length (apoplast)	l_w	$1\, \mu m$

Table 2.2: **Boundary values and parameter values chosen to simulate auxin steady-state distribution in the 1D model.** Parameter values were assigned according to literature data and experimental evidence.

In Figure 2.6, the profiles for auxin graded distribution both in the vascular (red) and in the epidermal (blue) tissue are displayed. As expected, PIN-directed flux generates an auxin maximum in the QC (i.e., approximately the fifth cell from the root tip). As expected, although concentration profiles for different tissues reflect the establishment of the same graded distribution, the overall auxin concentration values are lower than auxin concentration in the vasculature - due to lower strength of PIN2 in the external files - compared to the permeability in the vasculature, resulting from PIN1,3,7 (see Table 2.2 for parameter values). The concentration profile reproduces the auxin gradient inferred from DR5 reporter (see figure 1.14 a)).

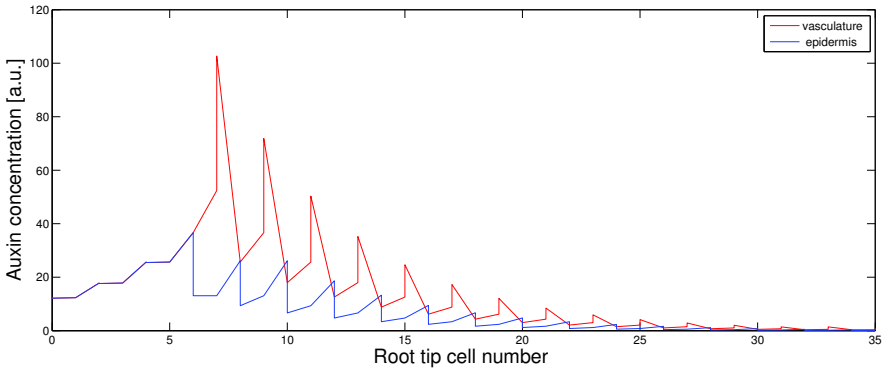


Figure 2.6: **Auxin concentration profiles in the root meristem.** The 1D model predicts the establishment of a graded distribution along the root tip (columella and meristem cells) and the formation an auxin maximum in the QC, assumed to be the $N-th$ cell of the vasculature (i.e., the $5th$ cell on the x axis located at the boundary with last columella cell). Predicted auxin values in the vasculature (red) and the epidermis (blue) show a 10-fold difference, due to the difference in PIN permeability (PIN1,3,7 permeability is twice as strong as PIN2 permeability).

The chosen parameter set reported in Table 2.1 ensures the observation of limit conditions for auxin maximum formation, i.e. the conditions 2.41,2.42,2.45 derived on fluxes, enhancement factor and polarization factor for the vasculature and in turn for epidermal files.

To test the robustness of the model towards the derived limit conditions, I tested multiple parameter sets. Importantly, when the condition on the flux (equation 2.41 and 2.47) is not respected, auxin concentration for both vascular and epidermal tissues reach negative values (Figure 2.7a), thus suggesting that auxin flux coming from the shoot is finely tuned so that high auxin fluxes would be not be allowed. Likewise, when violating the condition in eq. 2.45, e.g, when $P_{BG} \geq P_{PIN}$, auxin gradients become not informative as auxin concentration values becomes negative in the columella and subcellular profiles are (Figure 2.7b). Strikingly, breaking the condition on the polarization factor α leads to the violation of the condition on the flux J_0, which becomes higher that its limit value J_{0lim} (equal to 0 for the specific case $P_{BG} = P_{PIN}$). This result confirms the condition on α being the most restrictive.

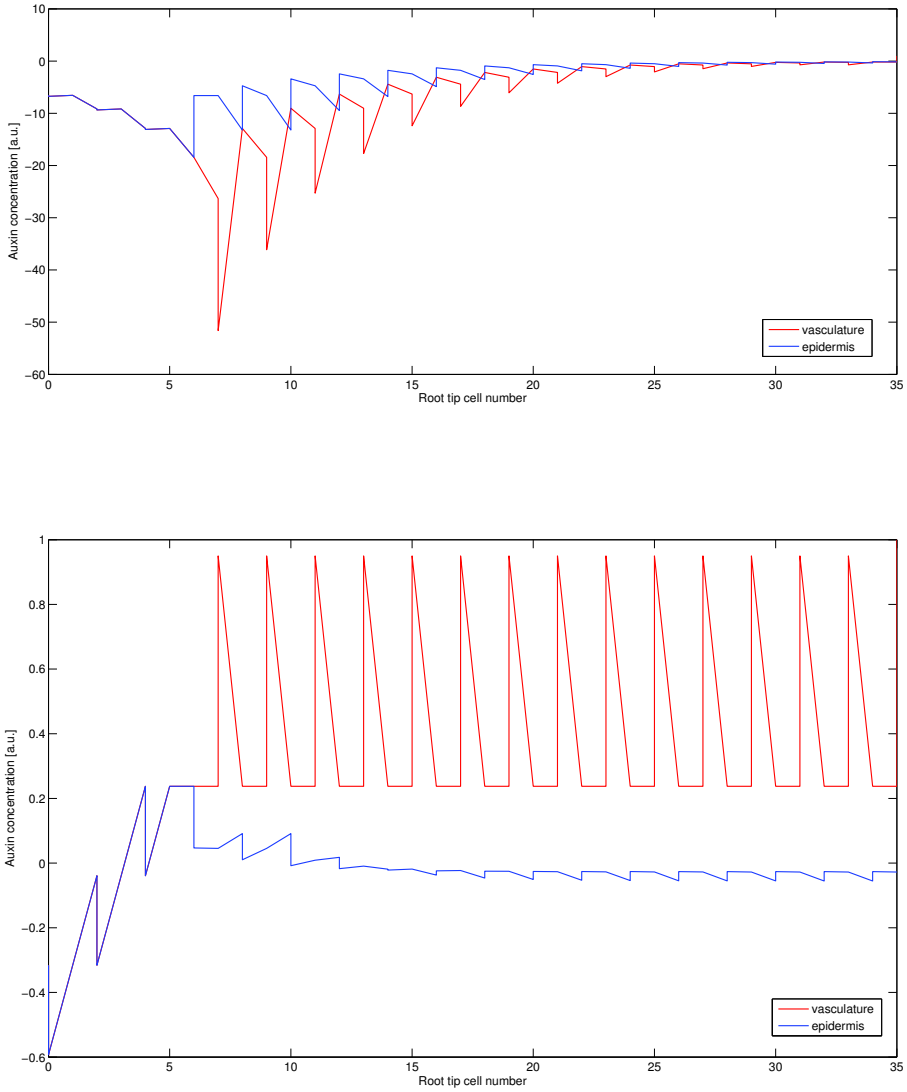


Figure 2.7: **Auxin concentration profiles when limit conditions are not respected**. a) $J_0 > J_{0lim}$; b) $\alpha < \alpha_{lim}$. These conditions would result in negative concentration values in the root meristem, which are not physically possible (a, b) and in a not informative profile in the vasculature (b), where the graded concentration would be constant in each cell.

I further investigated whether breaking limit conditions would affect the derived

equivalent diffusivity. Using the parameter set in Table 2.1, the profile of $D_{eq}^{(i)}$ has a graded exponential trend, reaching a maximum value in the QC and decaying along the meristem with a disproportion occuring between adjacent cells (Figure 2.8, top), which is negleted when translating $D_{eq}^{(i)}$ to its continuous equivalent $D_{eq}(x)$ (Figure 2.8, bottom).

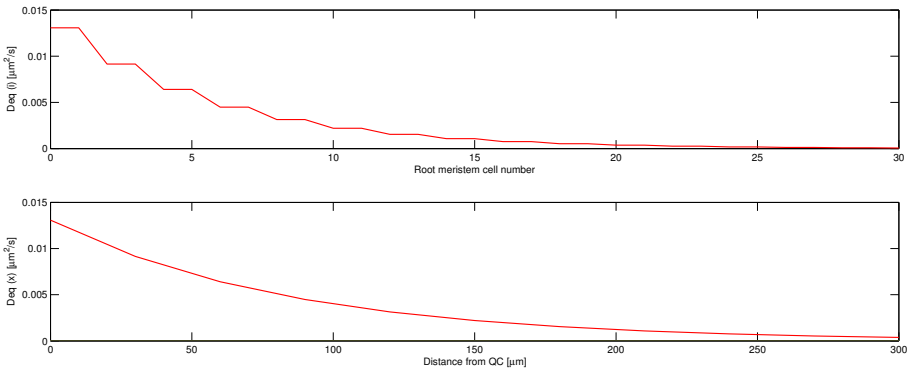


Figure 2.8: **Profile of the equivalent diffusivity $D_{eq}^{(i)}$ for root tip cells of the vasculature.** The cell-specific $D_{eq}^{(i)}$ (top) reaches its highest value in correspondence to the QC, decaying towards the end of the meristem with a disproportion between cells. In the continuous limit, the profile for the equivalent diffusivity along the tissue $D_{eq}(x)$ (bottom) is derived.

According to the developed formalism, the condition 2.42 ensures $D_{eq}^{(i)}$ to be positive. If γ becomes lower the $D_{eq}^{(i)}$ is affected indeed, reaching negative values, as shown in figure 2.9.

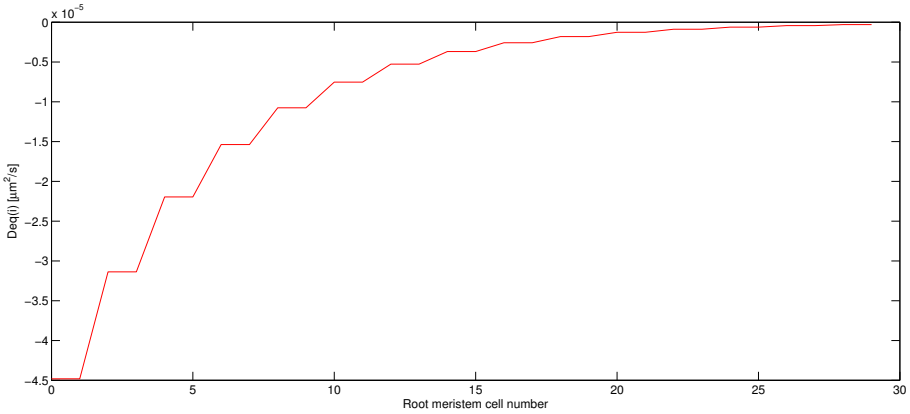


Figure 2.9: **Profile of the equivalent diffusivity $D_{eq}^{(i)}$ for root tip cells of the vasculature when limit conditions are not respected.** If the limit condition on γ (eq. 2.42) is violated, $D_{eq}^{(i)}$ would turn to negative values.

2.1.3.2 Continuous description

Through the expression derived for $\overline{D_{eq}}$, auxin graded distribution was translated into a continuous one, firstly neglecting auxin production and degradation, i.e. solving eq 2.28. Using the same parameter set in Table 2.2, auxin concentration profile along root tissue was obtained numerically solving the equation in *Mathematica* environment. As initial condition, the same value for c_0 used for the discrete model was chosen. As boundary conditions, no flux at the left side (QC) and a constant influx at the right side (at the end of the meristem length) were prescribed.

The resulting $\overline{D_{eq}}$ ($1.46 \cdot 10^{-2}$ $\mu m^2/s$) and the value for auxin diffusion in the cytoplasm D differ for four orders of magnitude. The continuous profile (Figure 2.10) still maintains the exponential trend with an auxin maximum in the QC on the same range of values as in the discrete distribution (see fig. 2.6). However, the continuos approximation leads to a sharper slope and a flat gradient after few microns from the QC. This result suggests that, in absence of auxin reaction terms, the auxin maximum is robust towards changes in the diffusion constant, whereas a low value for auxin diffusivity would determine a steep drop of auxin levels to low values after five cells from QC, i.e. the characteristic length of the gradient would be too short: the maximum concentra-

tion value would drop to the 37% of its maximum value at $37\mu m$, three times lower than the characteristic length resulting from the analytical solution (see Section A.1 in the Appendix). Moreover, the auxin gradient in the vasculature would become flat (Figure 2.10), thus no positional information would be provided for cell fate specification.

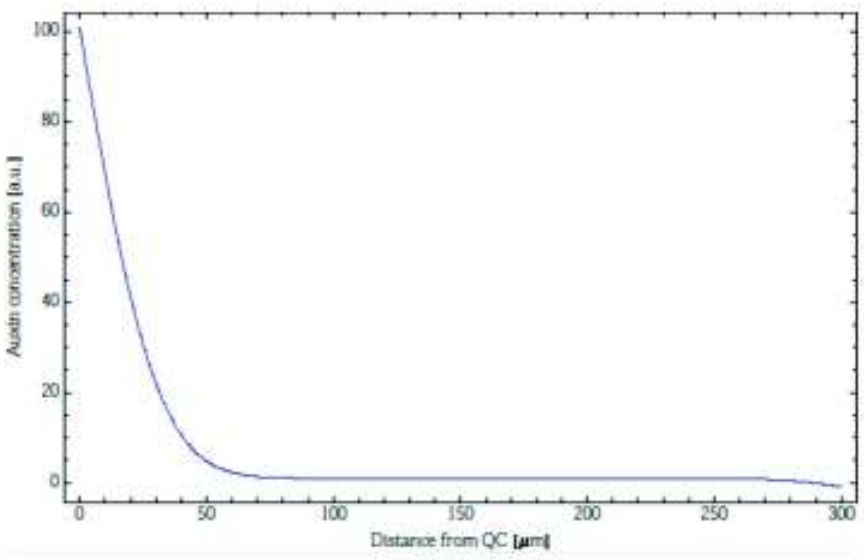


Figure 2.10: **Auxin concentration profile as derived from a continuous approximation.** As derived from the discrete formalism and shown in Figure 2.6, an auxin maximum occurs in the QC. However, in the continuous approximation the slope of the gradient is steeper (i.e., the characteristic length is shorter) and the gradient becomes flat (not informative) in the vascular cells.

2.1.3.3 Condition on the macroscopic balance

Considering equation 2.39, given that $J_{ext} \cdot 3S > 0$, it follows $d > p$, i.e. at the steady state, the overall production of auxin must be lower than the overall degradation to maintain the correct reflux in the root. However, the continuous view of the model neglects both basal production and decay in each cell of the root, as well the transversal component of auxin flux, which is essential to link auxin gradients in the vasculature and the epidermis at the TZ level.

2.2 Computational 2D model

To investigate the importance of the transversal component of auxin flux as well as to integrate the cell-specific input of cytokinin on auxin distribution in the *Arabidopsis* root meristem, a 2D model was developed. Importantly, such a model allows for the integration of new experimental data on the cytokinin-mediated regulation of auxin degradation, as obtained by microarray analysis and cell-specific gene expression in our lab.

2.2.1 Abstract

In multicellular organisms, pattern formation has been related to the graded distribution and activity of signalling molecules called morphogens. Concentration thresholds of morphogens create distinct cell compartments with specific developmental fates. Formation, positioning and maintenance of the boundaries between these cell compartments are essential for the correct outcome of the patterning events. In plants, the hormone auxin acts as a morphogen. However, little is known on how the shape of its graded distribution is generated, and how it sets the positional information that guides tissue specification and zonation during organ formation.

Combining computational modeling with a molecular and genetic approach, I provide a mechanistic understanding of how the shape of the auxin graded distribution in the root depends on the hormone cytokinin that controls both auxin transport and local auxin degradation. The model I developed predicts as an emerging property that both types of control are necessary in setting an auxin minimum, which provides positional cues for the meristematic cell to trigger a developmental switch towards differentiation. This auxin minimum is essential for positioning and maintaining the transition boundary between dividing and differentiating cells, thus setting meristem size and ensuring optimal root growth.

2.2.2 Methods

The cell-structured 2D model is based on a mathematical and computational model, developed by Grieneisen et al. [50] to perform simulations on the dynamics of auxin transport [50, 25, 51]. Grieneisen et al. [50] have provided proof of PIN-directed auxin flux as a main factor to determine the resulting flux patterns underlying the establishment of concentration profiles, although other directional transporters have an additional role in enhancing auxin reflux [6].
This concept acquires a more critical connotation in the present work, where I integrated the regulation of cytokinin on PIN expression: I embodied the regulation of cytokinin on the activity of PINs at the transition between the MZ and the DFZ [30] as a tissue- and zone- specific localization and permeability based on literature and on systematic analysis of the PIN expression pattern. In addition, to mimick the input of cytokinin on auxin degradation, I derived and introduced in the model a local auxin degradation rate, on the base of new experimental evidence found in our lab.

Both the 2D spatial layout (2.2.2.1) and auxin dynamics (2.2.2.2) were implemented using *C* programming language.

2.2.2.1 Building the 2D Root Layout

To test my hypotheses, I performed simulations on a two-dimensional (2D) grid, which represents a cross-section through the root, capturing the bilateral symmetry of the root across the xylem axis [50, 25]. Cells are represented as extended regions on a lattice characterized by appropriately scaled sizes and shapes, including surrounding cell walls. Each cell consists of multiple grid points, such that auxin concentrations may vary within cells. Each grid point is allocated to be part of cytosol, cell wall or media - cell wall is described as a separate entity, one grid point wide. The interface between cytosol and cell wall represents the cell membrane, through which auxin may permeate. This *in silico* root layout resembling the *A. thaliana* root has been already used in our previous works [50, 25, 120]. However, as I was interested in dissecting the processes occurring at a local level, I modified the spatial setting in turn to allow for a more realistic shape and cellular organization of the root tip, as well as for a detailed

cell-type specification (Fig. 2.11), based on observation of typical roots (see also A.2.1).

Therefore, I represented the root layout (Figure 2.11) longitudinally divided into a Meristem Zone (MZ), an Elongation-Differentiation Zone (DFZ), and a Transition Zone (TZ). In the model, the TZ was approximated as a straight boundary (TB) (Figure 2.11, right), given that the number of cells and cell length for different tissues was assumed to be the same.

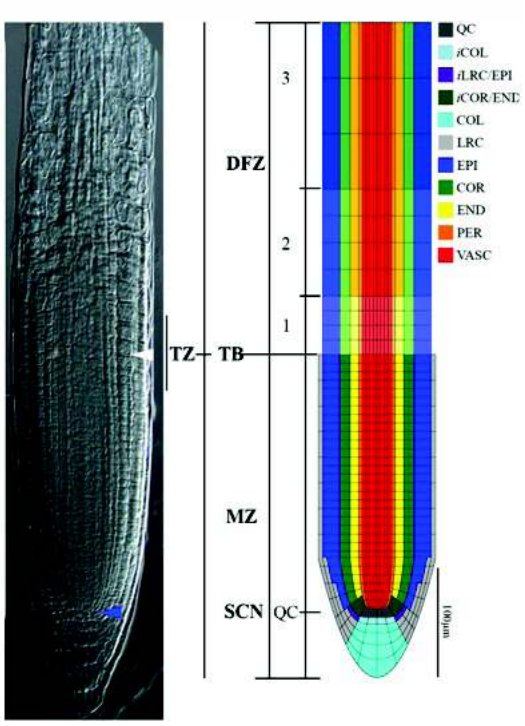


Figure 2.11: ***In silico* root layout of the 2D spatial model**. Left, optical microscope image of a real root. Blue arrowhead indicates QC, while white arrowhead indicates the transition boundary (TB) of the cortex. Right, the modelled root structured into cells, tissues and zones. SCN, stem cell niche; MZ, meristem of division zone; DFZ, differentiation zone. Cells are classified into different cell types: QC, quiescent centre; *i*LRC/EPI, lateral root cap/epidermis initial; *i*COR/END, cortex/endodermis initial; *i*COL, columella initial; COL, columella; LRC, lateral root cap; EPI, epidermis; COR, cortex; END, endodermis; PER, pericycle; VASC, vasculature. The same scale bar (100 µm) applies both to the image (left) and the model (right). Blue and white arrowheads indicate the QC and the cortex TB, respectively.

Alike the experimental procedure [101], the position of the cortex transition boundary (COR_TB) was taken as a reference for the other tissues. The number of cells till the transition boundary defines the meristem size. Root meristem size is expressed as the number of cortex cells in a file extending from the quiescent center to the first elongated cortex cell [31, 30, 99]. We thus provided the modified *in silico* root with a measured average meristem size (as reported in Table 2.3) and with a more accu-

rate measurement of cell length, which we obtained applying the tool *Cell-o-tape [41]* to confocal images of 12 roots (details are given in the Appendix, sectionA.2.1). The resulting root layout is on the same scale as real roots. Moreover, as in live roots cell length increases from MZ to DFZ, until reaching a terminal length, we incorporated differences in cell length for each zone, accounting for an additional distinction within elongation-differentiation zone, which was divided into three groups of cells of different length. In table 2.3 the numeric values for parameters are reported. Our measurements are in good agreement with previous studies [48].

MEASURED PROPERTY		**MEASURED VALUE**
Cell length (μm)	DFZ 3	59.7 ± 13.1
	DFZ 2	31.4 ± 4.7
	DFZ 1	19.4 ± 3.9
	MZ	7.5 ± 0.7
Meristem size (distance iCOR/END cell - COR_TB)(μm)		230.8 ± 20.4
Cortex cell number		30.7 ± 3.0
Root cap length (μm)		95.9 ± 4.1
Root diameter (μm)		114.8 ± 10.4

Table 2.3: **Measurement of spatial properties in real roots**. Values measured for 12 roots at 5 days after germination, SD as indicated. iCOR/END: cortex-endodermis initials; COR_TB: cortex transition boundary, taken as a reference to calculate meristem size. Total meristem length is given by the sum of meristem size plus root cap length.

Differences in width between cell files were also included. Cell width differs depending on cell type: measured lateral root cap are $5\,\mu m$, epidermal cells are $13\,\mu m$, cortical $17\,\mu m$, endodermal $8\,\mu m$, and vascular and pericycle cells $3\,\mu m$ wide.

The modelled root consists of 2576 x 355 grid points till the LRC and 2576 x 342 from the TB up. The spatial resolution of the grid ($\triangle x$) was set to 0.25, so that each grid point corresponds to 0.25 μm $\times$ 0.25 μm, giving rise to a diameter of $\sim 90\,\mu m$, in accordance with experimental measurements (Table 2.3).

2.2.2.2 Cytokinin-regulated Auxin Dynamics

The computational framework allows to solve numerically partial differential equations (PDEs) with space-dependent parameters and complex boundary conditions. The

changes in concentration results from the contribute of transport terms (simple and facilitated diffusion) and reaction terms (production and degradation), according to the conservation equation (eq. 1.8).

The auxin transport equation was solved numerically, using an Alternating Direction Implicit (ADI) method [97], implemented in previous works [50, 25, 69]. A space step $\triangle x$ was used corresponding to $0.25\,\mu m$, and a time step $\triangle t$ corresponding to $0.1s$. As boundary conditions, a fixed auxin concentration in the grid points belonging to the medium was prescribed, while at the shootward end of the simulated root a constant influx of auxin into the root was ensured, hypothesizing the vasculature of the DFZ to be connected to the upper parts of the plant. As initial condition for simulations, auxin concentration was set to 0 within the whole root tissue; as soon as auxin from the medium started to diffuse in, a pattern in auxin distribution emerged.

For the specific purpose of this work I used a static model, as I was interested in dissecting the process at the condition of equal cell division and cell differentiation rate (5 days after germination). The auxin distribution profiles shown were obtained letting each simulation to reach the steady state, i.e. a negligible difference ($\sim 10^{-7}$) in the total amount of auxin between 100 consecutive time steps was observed.

2.2.2.2.1 Transport Terms Alike the assumption taken for the 1D discrete model (see subsection 2.1.2.1), diffusion and permeability were dealt with independently for grid points belonging to cell cytoplasm, membrane or apoplast (cell wall), using parameter values from literature and direct experimental measurements (see Table 2.4). Diffusion within cells and cell walls was set according to Fick's laws. Auxin efflux and influx across cell membranes included both passive and cell-type specific carrier-mediated influx and efflux: a major component is enhanced by the presence of AUX/LAX and PIN expression, respectively. Compared to the setting used in [50], [6], the role of apolar AUX1/LAX influx carriers on auxin distribution at the root tip was introduced in the model as a parameter ($P_{AUX/LAX}$) that accounts for the expression pattern of AUX1, LAX2 and LAX3, set as in [6], while PIN permeability was changed into a space-dependent parameter, as explained in the next paragraph. A minor flux component comes from influx and spreading along all cell membranes (P_{IAAH}) and background permeability rates P_{ebg} (see Table 2.4).

Cytokinin Regulation on PIN Permeability To introduce the regulation of cytokinin on PINs mediated by ARR1 transcription factor (as discussed in paragraph 1.2.4.2), I embodied the known activity of this hormone on PIN expression as tissue- and zone- specific PIN orientation and efflux permeability: notably, cytokinin decreases PIN1, PIN2, PIN3 and PIN7 levels in the root [30, 113] and in particular at the transition zone via SHY2 [30] and data not shown. To quantify this effect I measured PIN fluorescence in all root tissues and zones of plants expressing PIN1:, PIN2:, PIN3: and PIN7:GFP translational fusions, using the software *ImageJ* (experimental details are given in Section A.2.2of the Appendix). As expected, a differential expression of PINs both between cell types and along the longitudinal axis of the root was observed. This regulation was translated into relative transport rates as tissue- and zone-specific parameters in the model, as reported in Table 2.5. The range of values for all PINs was set to be the same, as there are no means to establish their relative expressions. Importantly, total efflux results from the sum of each member contribution, as to have a shootward efflux (only directed by PIN2) lower that rootward one (resulting from PIN1,3,7 efflux along the entire length of the meristem and aided by PIN2 in the cortex

	PARAMETER DESCRIPTION	PARAMETER SYMBOL	PARAMETER VALUE
Transport terms	Influx at the apical boundary ($^{[a.u.]}/_{\mu m \cdot s}$)	J_0	2.5
	Auxin diffusion constant ($\mu m^2 \cdot s^{-1}$)	D	600 [50]
	AUX/LAX permeability ($\mu m \cdot s^{-1}$)	$P_{AUX/LAX}$	2 (in accordance with [6, 112])
	Passive influx permeability ($\mu m \cdot s^{-1}$)	P_{IAAH}	0.5 [6, 67]
	Passive efflux permeability ($\mu m \cdot s^{-1}$)	P_{ebg}	0.2
Reaction terms	Auxin decay rate (s^{-1})	δ_{IAA}	$5 \cdot 10^{-6}$ [50, 69]
	GH3-mediated auxin degradation rate (s^{-1})	δ_{GH3}	10^{-4}
	Auxin biosynthesis rate in each cell ($^{[a.u.]}/_{\mu m^2 \cdot s}$)	b	$5 \cdot 10^{-4}$ [50, 79]
	Auxin biosynthesis rate in QC cells ($^{[a.u.]}/_{\mu m^2 \cdot s}$)	b_{QC}	$5 \cdot 10^{-3}$ [6]

Table 2.4: **Parameter set used in the 2D model**. Parameter values were assigned according to literature data and experimental evidence.

of MZ), in accordance with experimental measurement and previous works [71]. Reflecting cytokinin effect, the overall PIN strength was measured as weaker the MZ than in the DFZ (as also reported in [69]).

To embed in the model the specific PIN localization on the different sides on cell membrane (inner I, outer O, lower L, upper U), I took advantage of an algorithm implemented in the code that compartmentalizes each cell into four different zones, which are oriented upward, downward, inward, or outward relative to the centroid of the cell [25]. In the model, PIN polarization along a given direction will correspond to the property acquired by membranes belonging to different zones. Figure 2.12 shows a section of the root tip where for each cell those different zonations and membrane (blow up) are coloured in green, purple, cyan, and red. Observed PIN localization was included in the model as indicated in Table 2.5.

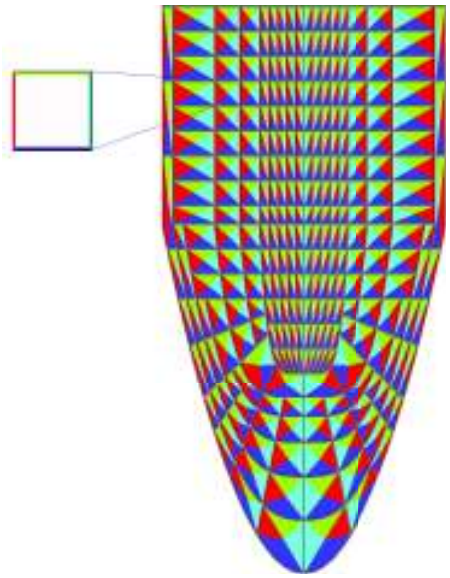


Figure 2.12: **Implementation of PIN orientation in the 2D model.** Section of the root tip and blow up indicating the upward (green), downward (purple), inward (cyan), and outward (red) part of the cell, which instructs the correspondent PIN orientation on cell membrane, as reported in Table 2.5.

Thus, following the formalism used for the discrete $1D$ model and the framework used in previous works [50, 25, 69], auxin flux across each grid point was computed as:

ZONE	TISSUE	PIN1		PIN2		PIN3		PIN7	
		LOCALIZ.	STRENGTH	LOCALIZ.	STRENGTH	LOCALIZ.	STRENGTH	LOCALIZ.	STRENGTH
2-3	EPIDERMIS	-	-	U	$PIN2_{quarter}$	-	-	-	-
	CORTEX	-	-	U	$PIN2_{third}$				
				O, I	$PIN2_{quarter}$				
	ENDODERMIS	L	$PIN1_{min}$	-	-	L	$PIN3_{sixth}$	L	$PIN7_{min}$
	PERICYCLE	L, O	$PIN1_{sixth}$			L, O	$PIN3_{half}$	L, O	$PIN7_{quarter}$
	VASCULATURE	L	$PIN1_{quarter}$			L	$PIN3_{half}$	L	$PIN7_{half}$
DFZ	EPIDERMIS	-	-	U, I	$PIN2_{half}$	-	-	-	-
1	CORTEX			U	$PIN2_{half}$				
				O, I	$PIN2_{third}$				
	ENDODERMIS	L	$PIN1_{min}$	-	-	L	$PIN3_{sixth}$	L	$PIN7_{min}$
	PERICYCLE	L, O	$PIN1_{quarter}$			L, O	$PIN3_{half}$	L, O	$PIN7_{half}$
	VASCULATURE	L	$PIN1_{half}$			L	$PIN3_{half}$	L	$PIN7_{half}$
MZ	LRC	-	-	U, I	$PIN2_{half}$	-	-	-	-
	EPIDERMIS			U, I	$PIN2_{max}$				
	CORTEX			L	$PIN2_{max}$				
				O, I	$PIN2_{half}$				
	ENDODERMIS	L	$PIN1_{fifth}$	-	-	L	$PIN3_{sixth}$	L	$PIN7_{sixth}$
	PERICYCLE	L, O	$PIN1_{half}$			L, O	$PIN3_{half}$	L, O	$PIN7_{half}$
	VASCULATURE	L	$PIN1_{max}$			L	$PIN3_{half}$	L	$PIN7_{max}$
SCN - ROOT CAP	QC	-	-	-	-	L	$PIN3_{half}$	-	-
	COL_I					L, U, I, O	$PIN3_{max}$		
	CORTEX_ENDO_I	L, O	$PIN1_{quarter}$	-	-	L, U, I, O	$PIN3_{third}$	L, O	$PIN7_{fifth}$
	LRC_EPI_I	-	-			-	-	L, O	$PIN7_{fifth}$
	COL					L, U, I, O	$PIN3_{max}$	L, U, I, O	$PIN7_{quarter}$

Table 2.5: **Differential PIN localization and strength for each tissue and zone of the root.** PIN localization on the cell membrane: U, upper; I, inner; L, lower; O, outer. PIN strength: a range of permeability values is assigned for each member of the PIN family, spanning from maximum of $2\,\mu m/s$ ($PIN1_{max}, PIN2_{max}, PIN3_{max}, PIN7_{max}$) to a minimum of $0.2\,\mu m/s$ ($PIN1_{min}$,$PIN2_{min}$,$PIN3_{min}$,$PIN7_{min}$). Zones and tissues of the root are set as defined in Figure 2.11.

$$\vec{J_a} = \begin{cases} -D_{aux} \cdot \vec{\nabla a}, & in\ the\ cell\ wall\ and\ cytoplasm \\ -(PIN_i(tissue, zone) \cdot \hat{n}) \cdot a_{in} + P_{IAAH} \cdot a_{out} & across\ cell\ membrane,\ PINs\ and\ passive\ influx \\ -(PIN_i(tissue, zone) \cdot \hat{n}) \cdot a_{in} + P_{AUX/LAX} \cdot a_{out} & across\ cell\ membrane,\ PINs\ and\ AUX/LAX \\ -(P_{ebg} \cdot \hat{n}) \cdot a_{in} + P_{AUX/LAX} \cdot a_{out} & across\ cell\ membrane,\ only\ background\ efflux\ and\ AUX/LAX \\ -(P_{ebg} \cdot \hat{n}) \cdot a_{in} + P_{IAAH} \cdot a_{out} & across\ cell\ membrane,\ only\ background\ efflux\ and\ passive\ influx \end{cases}$$

where $\hat{n}$ is the inward directed unit vector, perpendicular to the membrane; a_{in} stands for the auxin concentration in the cytosol at the grid point bordering the cell membrane and a_{out} stands for the auxin concentration in the cell wall grid point immediately adjacent to the cell membrane.

2.2.2.2.2 Reaction Terms Together with transport terms, auxin reaction terms include a basal production rate b and a decay rate δ_{IAA} (i.e. auxin half-life) in each cell, following the framework used in [50]. In addition, an auxin source in the SCN region (QC and columella initials) was introduced, as auxin produced in the QC both sustains stem cell niche activity and maintenance and acts as a long-range signal to fine tune the level of the early cytokinin response regulator ARR1 in the TZ.

Cytokinin Regulation on Auxin Degradation Besides the known input of cytokinin on PINs [31], new data from our lab (data not published) identified another ARR1-mediated input of cytokinin on auxin distribution. Specifically, through microarray analysis the gene GH3.17 (belonging to GRETCHEN HAGEN 3 (GH3) Group II family) was found as *ARR1* direct target genes. GH3 Group II family is involved in auxin homeostasis through the regulation of biologically active auxin levels mediating auxin conjugation with Asp and Glu amino acids, thus triggering auxin degradation.

To include this additional regulation, I measured the enzymatic degradation rate *in vivo* (details on the experimental procedure are provided in paragraph A.2.3 of the Appendix section), hypothesizing that at 5 days after germination auxin degradation rate was mediated by cytokinin effect. The derived degradation rate δ_{GH3} was inserted as a parameter in the model.

Therefore, the reaction terms were implemented as follows:

$$R(a) = \begin{cases} -\delta_{IAA} \cdot a + b_{IAA}, \ if\, auxin\, basal\, decay \\ -\delta_{GH3} \cdot a + b_{IAA}, \ if\, GH3 - mediated\, degradation \end{cases}, \quad (2.49)$$

$$b_{IAA} = \begin{cases} b & basal\, production\, everywhere \\ b_{QC} & source\, in\, QC\, and\, iCOL \end{cases} \quad (2.50)$$

2.2.3 Results

To assess to what extent our model correctly captures experimental evidence, the resulting steady-state heat map of auxin distribution in the root were compared to the expression of DR5 and DII-VENUS auxin reporters. According to these reporters, auxin distribution in the root is characterized by both an auxin maximum in the QC and high auxin levels in the DFZ (Figure 2.13 and 1.14 a, c).

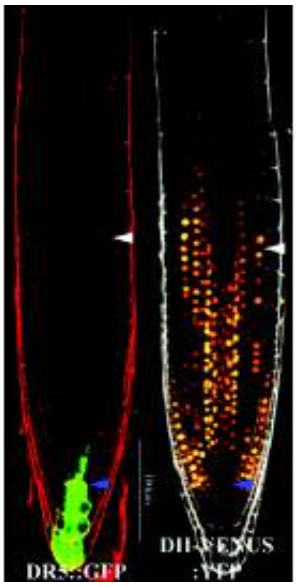


Figure 2.13: **Auxin levels inferred from auxin reporters**. Left, confocal image of a root meristem carrying DR5 reporter fused to GFP protein. Right, confocal image of a root meristem carrying DII-VENUS sensor fused to Yellow Fluorescent Protein (YFP). DII-VENUS expression is inversely proportional to auxin levels. Both reporters show an auxin maximum in the QC. DII-VENUS expression also shows a rise in auxin levels in the DFZ. Blue and white arrowheads indicate the QC and the cortex TB, respectively.

To first understand how the sole regulation of cytokinin on PINs could affect auxin distribution, I integrated in the model the measured PIN expression and strength, as reported in Methods and Table 2.5. The resulting heat map (Figure 2.14 a) does not reproduce the expected auxin distribution in the root tip as visualized by the auxin reporters (Figure 2.13 a). In particular the auxin maximum is not confined to the QC cells and the auxin levels do not rise in all tissues of the DFZ (see also Fig. 1.14 c). This result suggests that the sole control of cytokinin on PINs is not sufficient to determine the correct auxin graded distribution.

To test the effect of the coupled input of cytokinin on PINs and auxin degradation, I introduced the measured degradation rate δ_{GH3} (see paragraph 2.2.2.2.2 in the Methods section and A.2.3 in the Appendix section) in the auxin transport equation. By introducing this degradation rate, the resulting auxin graded distribution matches with

DR5:GFP and DII-VENUS expression, displaying an auxin maximum in the QC and a rise of auxin levels in all tissues of the DFZ (Fig. 2.14 b). Auxin degradation was introduced as a uniform parameter hypothesizing that this process is active in all root cells. However, confocal analysis of plants carrying the GH3.17 gene fused to the GFP protein revealed that the GH3.17 protein has a very specific expression pattern, being expressed only in in the outermost file of the root cap and lateral root cap (LRC) and in the differentiated epidermal cells (Fig. 2.14 c).

To investigate the role of spatially regulated auxin degradation, I thereby introduced in the model a local auxin degradation rate restricted to the GH3.17 expression domain. The steady-state simulation obtained including both cytokinin local control on auxin degradation (GH3.17 dependent effect) and cytokinin regulation of auxin transport still displays an auxin maximum in the QC and a rise of auxin level in the DFZ (Fig. 2.14 d).

Interestingly, as an emerging property, the simulation predicts the formation of an auxin "dip", i.e. a minimum specifically confined and aligned in the uppermost meristematic row of cells, as shown in the blow up of Figure 2.14 d). The longitudinal and transversal concentration profile for each tissue of the root at the steady state also show a maximum of auxin in the QC and a minimum in correspondence to the TB of all tissues (Figure 2.14 e, f).

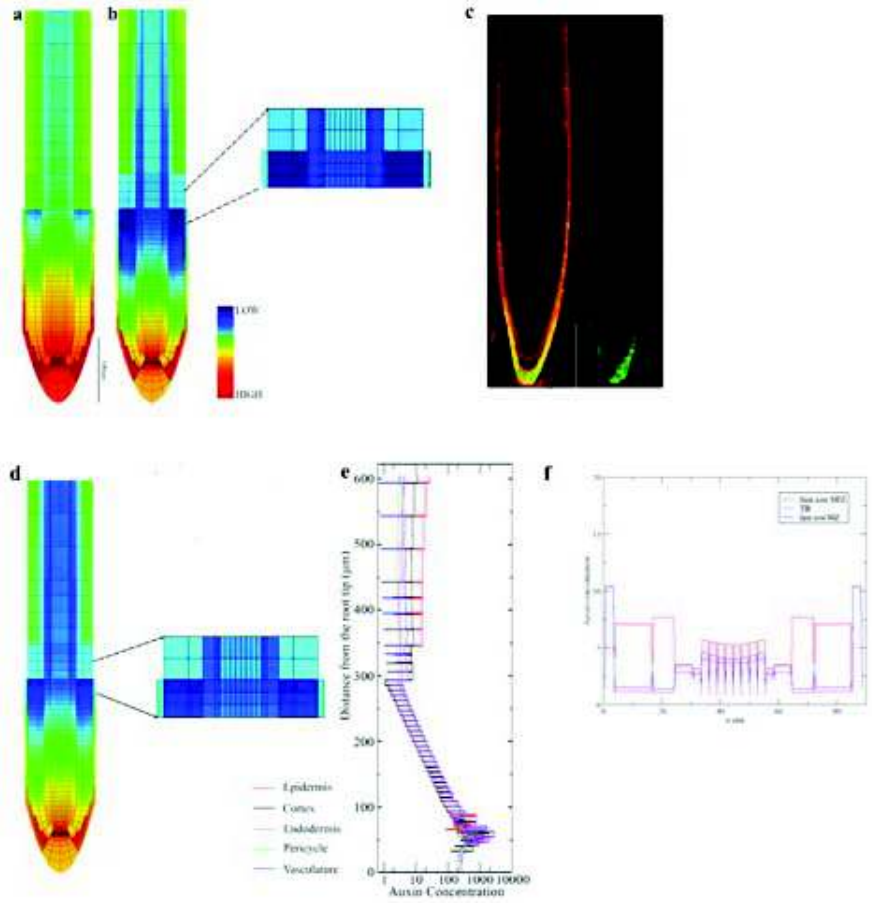


Figure 2.14: **Auxin heat map and concentration profile: wild-type roots.** Steady state simulations. The heat map resulting from the sole regulation on PINs (a) does not reproduce the experimentally observed auxin thresholds. Cytokinin input both on PINs and on auxin degradation (b) leads to the formation of a maximum in the QC and a rise of auxin levels as experimentally observed (Figure 2.13). c) Confocal image of a root meristem of plants carrying a GH3.17::GFP reporter. Note that the expression domain of GH3.17 (green green fluorescent protein) is in the outer tissues (LRC and epidermis DFZ) of the root. d) When localizing cytokinin regulation on auxin degradation in the GH3.17 domain, an auxin minimum emerges confined at the transition boundary (TB) (blow up), i.e. in correspondence to the last meristematic cell of each tissue. The longitudinal (e) and transversal (f) auxin concentration profiles relative to the heat map in (d) show the corresponding auxin gradient and highlight the formation of a minimum aligned at the TB (e, f).

Analyzing auxin flux for each tissue along the root (Figure 2.15), I noticed that the auxin minimum occurs upon a break in the reflux loop through external files (due to degradation), which results in minimum fluxes at the TB, as opposed to the high throughput converging in the QC that leads to the formation of an auxin maximum, which has been previously described [50, 69].

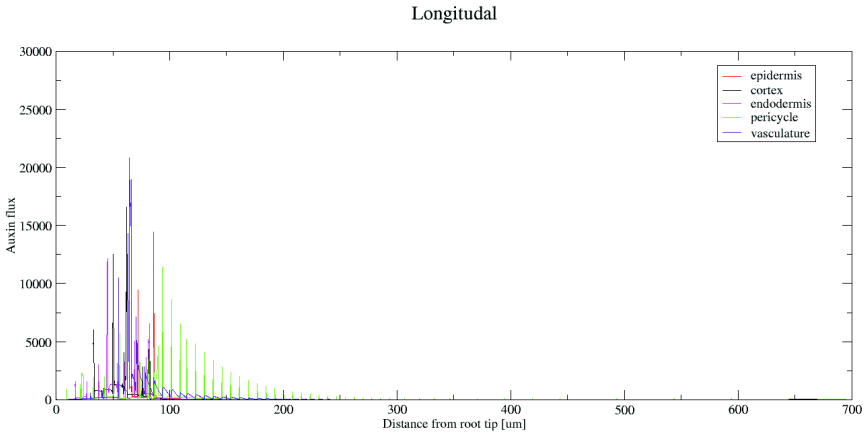


Figure 2.15: **Longitudinal profile of auxin flux across root cells**. An auxin maximum in the QC is accompanied by high fluxes, while an auxin minimum at the TB occurs in correspondence to minimum fluxes, mainly caused by a break in the reflux loop due to local auxin degradation.

The auxin "dip" matches with the position of the transition boundary, suggesting that it could act as a morphogenetic threshold necessary to induce a developmental switch towards differentiation. In fact specific morphogen thresholds are associated with distinct cell response [139, 142, 62] and specifically, during plant development, an auxin maximum in the QC is associated with root SCN specification [115] while low auxin levels are fundamental to drive cell towards differentiation [79, 69].

Measuring auxin levels at cellular resolution is still challenging, thus to associate auxin thresholds with cell specification and, more specifically, to assess whether this auxin minimum in the uppermost meristematic cells does exist in planta, would not feasible by only means of wet biology. Therefore, model predictions were experimentally validated at our lab, to indirectly provide evidence of the predicted auxin minimum *in*

vivo. Based on the model predictions, the absence of a localized GH3.17-mediated auxin degradation (i.e. *gh3-17* mutant, Fig. 2.14 a), or the introduction of a uniform auxin degradation parameter (i.e. ectopic GH3.17 expression, Fig. 2.14 b) result in changes of auxin graded distribution such that a minimum in the last meristematic cell is not formed. I then hypothesized that if the auxin minimum in the last meristematic cell acts as an auxin threshold necessary to position the transition boundary, its absence would indicate a mutant phenotype, where the transition boundary is shifted upward or downward altering meristem size. Indeed, the meristem of the *gh3.17-1* mutant root is longer than the wild-type (Fig. 2.16 a, c, g). Conversely plants where GH3.17 was ectopically expressed, display shorter meristems compared to wild-type plants (Fig.2.16 a, e, g). Experimental analyses on mutant plants were carried out at our lab.

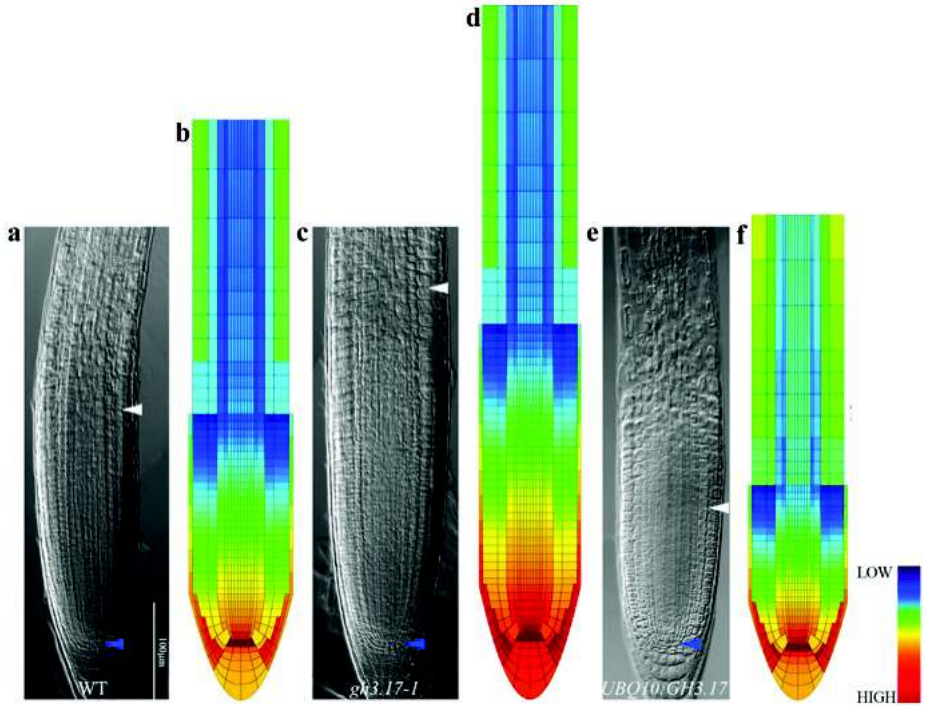


Figure 2.16: **Auxin heat map**: *gh3.17* **mutant and GH3.17 over-expressor plants.** Predicted steady state auxin pattern (b) simulated for wild type (WT) roots at 5 days after germination (a). A well-defined auxin minimum at the TB is restablished in *gh3.17-1* mutants (d) and UBQ10:GH3.17 over-expressors (f) when increasing or decreasing the number of MZ cells in the root layout, according to the size of their respective phenotype: *gh3.17* mutant has a longer meristem (c) while GH3.17 over-expressor (e) has a shorter meristem compared to WT (a). Blue and white arrowheads indicate the QC and the cortex TB, respectively.

Meristems of both *gh3.17-1* mutant roots and of roots ectopically expressing GH3.17 do have transition boundaries, however shifted as compared to wild type meristems. To assess whether the shift of the transition boundary is concomitant with (or results from) the repositioning of the auxin minimum, I simulated the *gh3.17-1* mutant and GH3.17 overexpression by modifying the auxin dynamics and the root layout. accordingly. To simulate *gh3.17-1* mutant meristematic cells were increased and local auxin degradation was neglected, while to simulate GH3.17 overexpressing plant meristematic cell were decreased and GH3.17 mediated auxin degradation was imposed in all cells. In both

simulations an auxin minimum still lays in the last meristematic cells (Fig. 2.16 b, d, f).

Interestingly, heat maps simulating the $gh3.17-1$ mutant predict higher auxin levels in the root tip (Fig. 2.16 d). Analysis of the root of *gh3.17-1* plants expressing either DII-VENUS or DR5:GFP, indicates high level of auxin indeed (experimental data not shown from our lab).

Dello Ioio et al. [32, 30] have previously shown that cytokinin dependent regulation of auxin polar transport mediated by SHY2 is sufficient to control the position of the transition zone. Plants mutated in this gene display in fact an enlarged root meristem [31]. In accordance, the model predicts that both local auxin degradation (GH3.17 dependent effect) and cytokinin dependent regulation of auxin polar transport (SHY2 dependent effect) are necessary to maintain the position of the auxin minimum and that it is sufficient to alter one of the two input to affect the position of the auxin minimum in turn (Figure 2.14a and 2.17 a, b). The additive effect of these two contributes is confirmed by the analysis of the double *shy2-31;gh3.17-1* mutant, whose root meristem is bigger than the parental (experimental data from our lab).

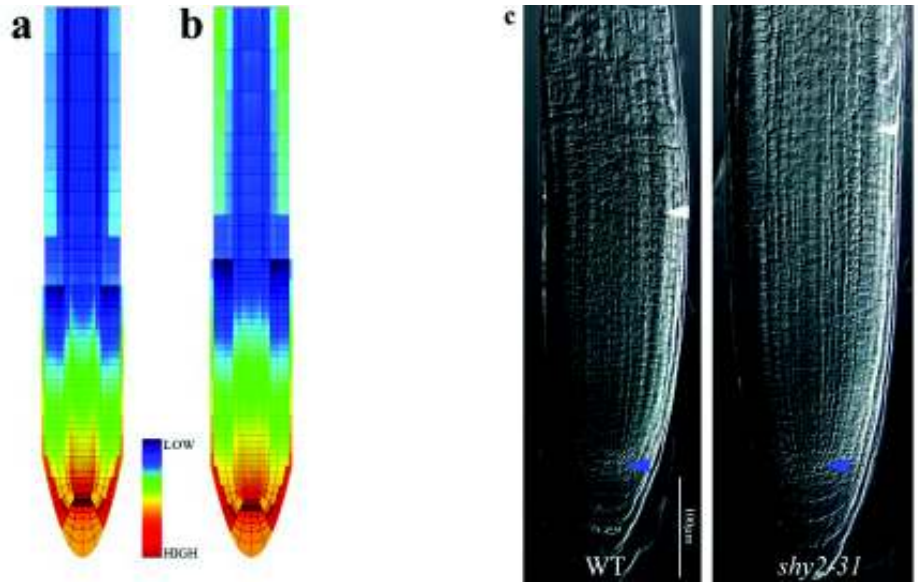


Figure 2.17: **Auxin heat map***: shy2-31* **mutant**. When no regulation of cytokinin on PINs is included in the model (a) the auxin pattern for a wild-type root does not reproduce the observed distribution (Figure 2.13) and a minimum at the TB. Auxin thresholds are re-established only by simulating a root with increased cell number in the root layout (b), according to the enlarged *shy2-31* mutant phenotype compared to the wild-type (WT) (c). Blue and white arrowheads indicate the QC and the cortex TB, respectively.

Hence a systems biology approach is able to predict that, by controlling auxin degradation and transport, cytokinin generates an auxin minimum ("dip") that is always associated with the position of the transition boundary: this resultstrongly suggests that this auxin dip a fundamental threshold driving cell towards differentiation.

Once elucidated the separate effect of cytokinin input on auxin degradation and PIN down-regulation, I finally investigated how coupling cytokinin inputs would influence the auxin minimum in order to control meristem size, explaining, for example, why depletion of cytokinin results in an enlarged meristem size while exogenous cytokinin application leads to meristem shrinking [32, 31](see 1.2.4.2 and Fig. 2.18 b, e, h). Cytokinin depletion was simulated neglecting local auxin degradation (GH3.17 effect) and increasing PIN permeability (SHY2 effect). As expected, auxin levels at the transition boundary do not reach a minimum (Fig. 2.18 c) suggesting that those cells would not

enter the differentiation program, thus causing the larger meristem, as experimentally observed [32, 31]. Accordingly, only by increasing the number of cells in the MZ, the minimum is re-established (Fig 2.18 d) as in the case of the *shy2-31* and *gh3.17-1* simulation (Fig. 2.17 b, c and Fig. 2.16 d). Conversely, simulations mimicking exogenous cytokinin treatment, where local auxin degradation (GH3.17 effect) was increased and PIN permeability decreased (SHY2 effect), show a broader auxin minimum (Fig. 2.18 a, f). This observation suggests that cells encounter a differentiation threshold within the MZ, leading to a premature cell differentiation responsible of the observed meristem shrinking [32, 31]. In accordance, the auxin minimum was re-established in the last meristematic cells only when cells in the MZ were decreased (Fig. 2.18 g).

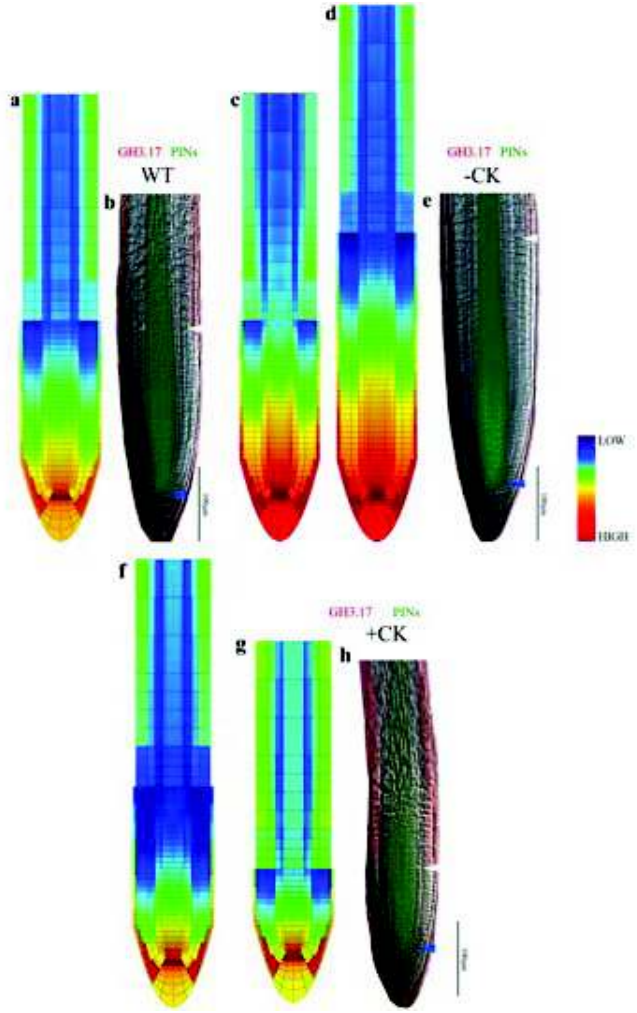


Figure 2.18: **Auxin heat map: cytokinin treatments.** a,b) Wild type (WT) simulation (a) and real root (b) shown as a control. Cytokinin depletion (no regulation both on PINs and auxin degradation) simulated on a wild type root layout (c) does not display the correct auxin thresholds, which are re-established increasing the number of cells according to *in vivo* observation (e). Conversely, mimicking cytokinin application (higher auxin local degradation and lower PIN strength) on a wild-type root (f) would result in minimum levels of auxin in most cells of the meristem. Also in this case, an auxin minimum is re-established decreasing the number of cells according to real root meristem size (h). Changes in the expression of GH3.17 (red) and of PINs (green) are sketched on the images of the cartoon roots (b,c,h). Blue and white arrowheads indicate the QC and the cortex TB, respectively.

I tested the robustness of the model both towards changes in auxin transport and spatial properties. To explore the influence of transporter permeability, I first tested the sensitivity of auxin distribution and concentration profiles to AUX/LAX transporter permeability, as main players of auxin transport together with PINs. In accordance with previous studies [6], these changes only affect auxin levels while the overall auxin graded distribution is maintained (see the longitudinal concentration profiles in Figure2.19 a). Variations in the auxin biosynthesis rate do not affect auxin pattern as well (Figure2.19 b). Auxin steady state distribution is also robust towards variation in the diffusion constant (Figure2.19 c). However, too high values of diffusion constant modify the slope of the gradient, which flatten out in the DFZ (profile relative to Dx2 in Figure 2.19 c).

To test the presence of bias in the choice of spatial parameters, and in particular of cell geometries, I modified the root a layout neglecting differences in cell length in the different zones. The resulting steady state auxin distribution does not differ from the scenario of the realistic root layout built for our simulations. Comparing this simulation to a wild type auxin heat map, I noticed that a dilution effect does not occur in the longer cells of the DFZ of the wild type. This is may due to the low strength of PINs in the DFZ, which impedes auxin to flow out of the cells. (Figure2.19 d).

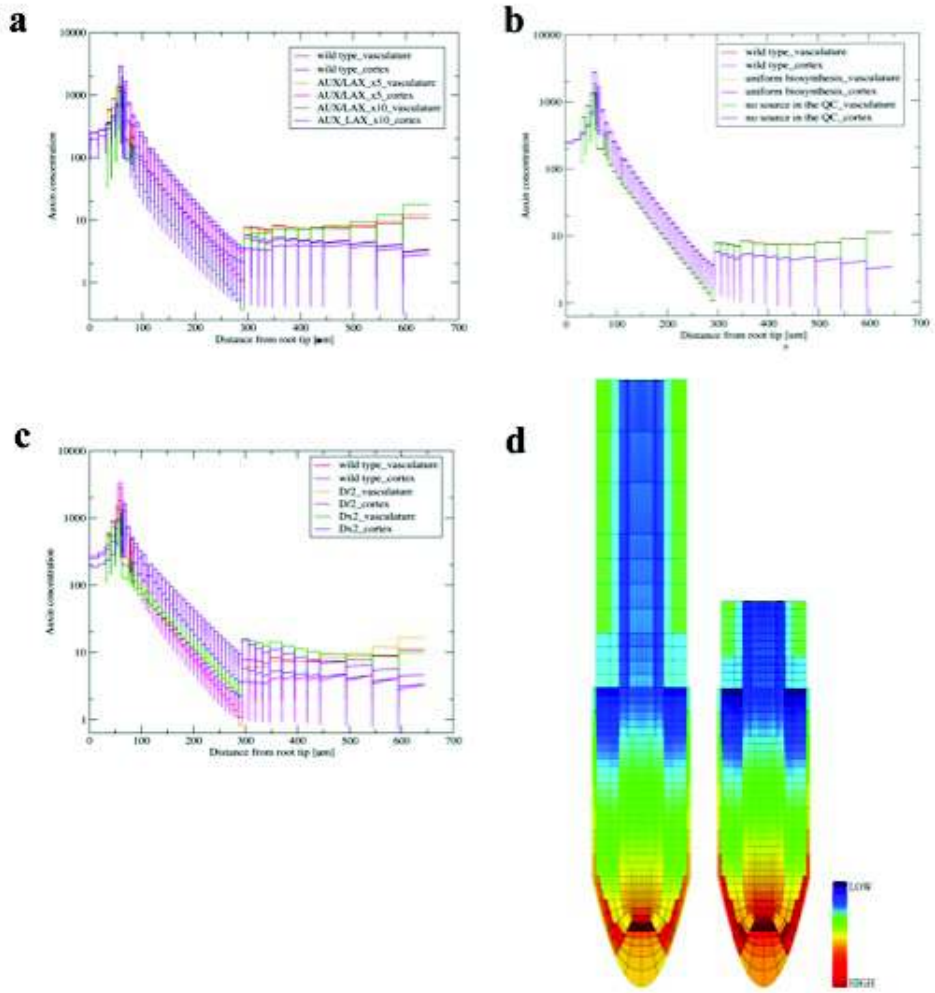


Figure 2.19: **Robustness analysis of the 2D model.** a) Longitudinal auxin concentration profile in the vasculature and cortex of wild type roots versus changes in AUX/LAX permeability. Despite changes in auxin levels, increasing AUX/LAX permeability does not affect the shape of the gradient b) Longitudinal concentration profile of wild type roots compared to changes in auxin biosynthesis: auxin gradients in the presence of a spatially uniform auxin biosynthesis or depleting auxin source in the QC overlap with the wild type one. c) Longitudinal concentration profile of wild type roots compared to changes in auxin diffusivity: while the maximum is preserved in the presence of a double or half value of the diffusion constant, the slope of the gradient in the DFZ changes, tending to flatten out. d) Auxin heat map of a root with no difference in cell length between the different zones: while small differences can be observed in the endodermis of the DFZ, auxin thresholds are not perturbed.

Chapter 3

Discussion and conclusion

Combining wet biology and mathematical and computational modelling results powerful in providing a quantitative understanding on the interplay between transport and reaction terms involved in auxin asymmetric distribution and on how cytokinin shapes auxin gradients.

The adopted systems biology approach allowed for investigating the system under study linking different scales (cellular, tissue, and organ) and techniques (experimental, mathematical, computational).

Through a one-dimensional theoretical approach, a recursive formula allows for linking auxin concentration between non adjacent cells, eventually leading to the derivation of a linear diffusion equation, where all transport components are embedded within the equivalent diffusivity parameter. As an ultimate goal, I envision that this formalism could be used as a tool for the estimation of parameters from auxin maps.

A two-dimensional computational model was further developed to integrate experimental evidence on the input of cytokinin on auxin gradients, occurring at cellular level. Importantly, the predictions of the model feedback on the genetics underlying the antagonistic interaction between auxin and cytokinin, elucidating a mechanistic picture of the circuit: by activating the ARR1 transcription factor, cytokinin regulates auxin polar transport via SHY2 [32] and its local degradation via GH3.17 (new data from our lab).

Both inputs are needed for the positioning of a local auxin minimum in the upper-

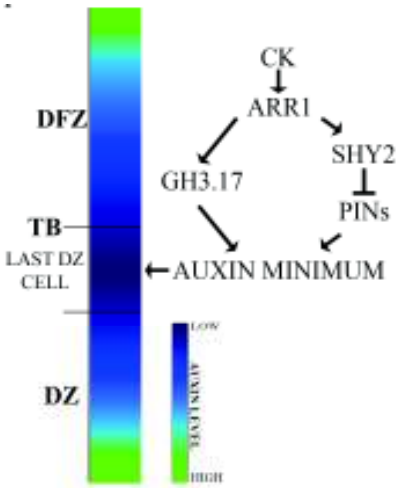


Figure 3.1: **Final scheme of the adopted systems biology approach**. According to the genetic model hypothesized in this work, cytokinin, through ARR1, down-regulates PINs (inducing SHY2 expression) and triggers local auxin degradation (inducing GH3.17 localized expression), affecting auxin distribution. Introducing this regulation in the 2D model the observed auxin thresholds can be obtained: a maximum in the QC and a rise of auxin levels in the DFZ (shades of green), while predicting the formation of a new threshold, which is an auxin minimum at the transition boundary (deep blue)

most meristematic cells. This auxin minimum acts as a morphogenetic threshold that instruct cells towards differentiation, thus positioning the transition boundary. In this way cytokinin sets meristem size and ensures a correct root growth. A final scheme of the proposed mechanism is provided in Figure 3.1.

Appendix A

Appendix

A.1 Analytical Solution of Diffusion-Advection-Reaction Equation

Given the full auxin transport equation with uniform transport and reaction terms:

$$\frac{\partial a}{\partial t} = D_{aux}\frac{\partial^2 a}{\partial x^2} - P_{PIN} \cdot \frac{\partial a}{\partial x} - \delta_a \cdot a + b_a$$

a steady state analysis on its analytical solution was worked out.

The equation at the steady state becomes:

$$D_{aux}\frac{\partial^2 a}{\partial x^2} - P_{PIN} \cdot \frac{\partial a}{\partial x} - \delta_a \cdot a + b_a = 0$$

Homogenous solution

$$D_{aux}\frac{\partial^2 a}{\partial x^2} - P_{PIN} \cdot \frac{\partial a}{\partial x} - \delta_a \cdot a = 0$$

Substituting the auxiliary equation:

$$D_{aux}\lambda(x)^2 - P_{PIN} \cdot \lambda(x) - \delta_a = 0$$

the roots of the polynomium are:

$$\lambda_{1/2}(x) = \frac{P_{PIN} \pm \sqrt{P_{PIN}^2 + 4D_{aux} \cdot \delta_a}}{2D_{aux}}$$

Considering that $\Delta = P_{PIN}^2 + 4D_{aux} \cdot \delta_a$ is always >0, the roots are real and distinct.

Thus the homogeneous solution will be of the form:

$$a_H(x) = A \cdot e^{\lambda_1 x} + B \cdot e^{\lambda_2 x}$$

Substituting $\lambda_{1/2}(x)$ in the homogeneous solution:

$$a_H(x) = A \cdot e^{\frac{P_{PIN}+\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}x} + B \cdot e^{\frac{P_{PIN}-\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}x}$$

To determine unknown coefficients A and B two equations are needed, which are given by boundary conditions at the left and right side of the domain (at the left we have no flux boundary condition and at the right a constant influx boundary condition):

$$\begin{cases} \frac{\partial a_H(x)}{\partial x} = 0 & x = 0 \\ \frac{\partial a_H(x)}{\partial x} = j_0 & x = L \end{cases}$$

The diffusive influx (from right to left) is given by: $J_{diff} = -D_{aux} \cdot \frac{\partial a_H(x)}{\partial x}$, $\frac{\partial a_H(x)}{\partial x} > 0$, and the advective influx given by: $J_{adv} = P_{PIN} \cdot a(x)$ (if $P_{PIN} < 0$, the wave will propagate from right to left).

Applying these conditions to $a_H(x)$:

$$\begin{cases} \lambda_1 A e^{\lambda_1 0} + B\lambda_2 e^{\lambda_2 0} =\!\!\rightarrow B = -A\frac{\lambda_1}{\lambda_2} & x = 0 \\ \frac{\partial a_H(L)}{\partial x} = j_0 \rightarrow \lambda_1 A \cdot e^{\lambda_1 L} + \lambda_2 B \cdot e^{\lambda_2 L} = j_0 \rightarrow A = \frac{j_0}{\lambda_1 (e^{\lambda_1 L} - e^{\lambda_2 L})} & x = L \end{cases}$$

Thus I obtained an analytical expression for the unknown coefficients:

$$\begin{cases} A = \frac{j_0}{\frac{P_{PIN}+\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}(e^{\frac{P_{PIN}+\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}L} - e^{\frac{P_{PIN}-\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}L})}; \\ B = -\frac{j_0}{\frac{P_{PIN}+\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}(e^{\frac{P_{PIN}+\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}L} - e^{\frac{P_{PIN}-\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}L})} \frac{\frac{P_{PIN}+\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}}{\frac{P_{PIN}-\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}}; \end{cases}$$

Particular solution

$$D_{aux}\frac{\partial^2 a}{\partial x^2} - P_{PIN} \cdot \frac{\partial a}{\partial x} - \delta_a \cdot a = -b_a$$

Since the non homogeneous term is a constant: $D'(-k_a) = D''(-k_a) = 0$, the particular solution will be a constant: $0 - 0 - \delta_a \cdot a_P = -b_a \rightarrow a_P = \frac{b_a}{\delta_a}$

General solution

$$a(x) = a_H(x) + a_P(x) = \frac{b_a}{\delta_a} + A \cdot e^{\frac{P_{PIN}+\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}x} + B \cdot e^{\frac{P_{PIN}-\sqrt{P_{PIN}^2+4D_{aux}\cdot\delta_a}}{2D_{aux}}x},$$

I can now obtain the expression for the characteristic length of the gradient ι, known as the distance reached by the morphogen when its concentrations falls as a_0/e:

$$a_H(\iota) = A \cdot e^{\lambda_1 \iota} + B \cdot e^{\lambda_2 \iota} = a_0/e$$

for the homogeneous solution, and

$$a(\iota) = \frac{k_a}{\delta_a} + A \cdot e^{\lambda_1 \iota} + B \cdot e^{\lambda_2 \iota} = a_0/e$$

for the general solution.

The numerical value of the characteristic length for the general solution has been calculated for known parameters values as $110\,\mu m$.

A.2 Experimental Methods

A.2.1 Transition zone measurement

To date, the measurement of the height of the transition zone in real roots, has been carried out "by eye", through the observation of root optical microscope images. To individuate differences between cells with meristematic and elongating/differentating property, difference in cell morphology and length is observed, taking as a reference cortex cells, as described in [101]. The number of dividing cells till the transition boundary (TB) of the cortex are counted from the initial cortex/endodermis ($iCOR/END$) cell.

For the purpose of my work, this procedure was integrated with a semi-automated method, using the tool $Cell-o-Tape$ [41]. This tool can be installed as a plugin in $Fiji$ ($ImageJ$), and applied to confocal images, setting a grey scale (as shown in Figure A.1). Once the user traces a line across cells (yellow line in Figure A.1), each cell is automatically detected (red dots) and numbered based on colour differences between cell walls (white/grey) and cytoplasms (black). In output, cell id-number and distance from the starting point at which the switch point between dividing and elongating cells

occurs is given. An example of the output is reported in the text box of Figure A.1.

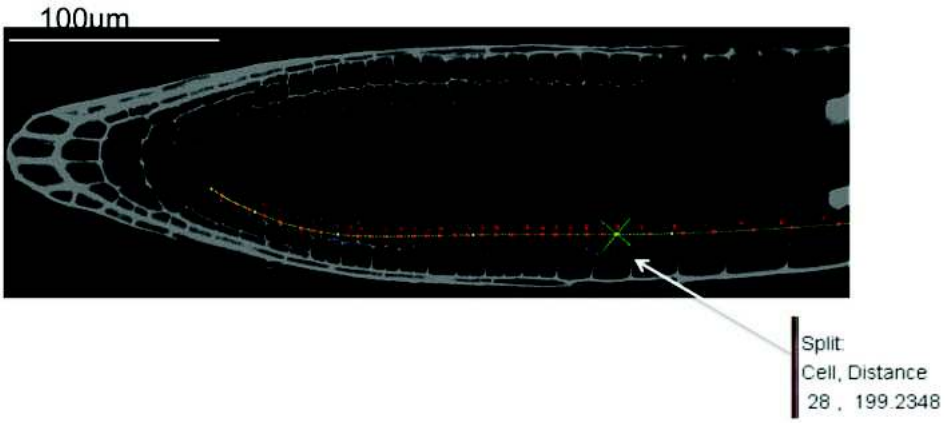


Figure A.1: **Measurement of transition zone position in *A.thaliana* roots using Cell-o-Tape.** The semi-automated tool allows for cell number counting and measurement of the distance at which a switch between dividing and elongating cells occurs, as indicated in the text box. Scale bar $100\mu m$.

I applied this procedure to confocal images of *Arabidopsis* roots. To be consistent with manual measurements, I traced a yellow line along the cortex cell file, choosing as starting point the initial cortex/endodermis ($iCOR/END$) cell and as ending point the fourth cell of the differentiation zone (useful for the software to better distinguish between different cell length). On average, the resulting "split" distance corresponded to what manually measured using the optical microscope. The results obtained for transition zone measurement applying $Cell-o-Tape$ on 12 roots are reported in Table 2.3.

A.2.2 PIN expression measurement

To compute PIN expression I used the software $ImageJ$. To quantify the GFP intensity for a given region, I chose the plugin $Measure\,RGB$, which gives a $RawIntDen$ (Raw Intensity Density) value referred to the sum of pixels intensity for the green channel. The obtained values were normalized for the chosen area. The selected areas were:

MZ (from the staminal cell excluded to the cortex TB); DFZ subdivided into region 1 of height y from TB, region 2 of height y from the end of the region 1, and region 3 approximated as region 2. PIN expression in each tissue of the chosen area was then quantified.

To compute relative PIN permeability values, for each member of the PIN family I first associated the maximum intensity of GFP (GFP_{max}) with a permeability value of $2\,\mu m/s$ (PIN_{max}), according to the average value estimated by [67]. The minimum value of permeability was assumed to be 10 times lower (PIN_{min}) (i.e., comparable to passive efflux P_{ebg}, in accordance with [50]). Differences in PIN expression in each zone and tissue were then captured by scaling permeability levels in the range, through the following proportion:

$$PIN_i(tissue,\, zone) = \frac{PIN_{i,max} \cdot GFP_{PIN_i(tissue,zone)}}{GFP_{max}},\ i = 1, 2, 3, 7 \tag{A.1}$$

where $GFP_{PIN_i(tissue,zone)}$ is the GFP intensity measured for each PIN in different tissues and zones, and $PIN_i(tissue,\, zone)$ is the resulting value falling in the range: $\{PIN_{i,half};\ PIN_{i,third};\ PIN_{i,quarter},\ PIN_{i,fifth};\ PIN_{i,sixth}\}$.

A.2.3 Degradation rate quantification

To quantify degradation rate, DII-VENUS plants grown on MS media were transferred on media containing $10\,\mu M$ of the transport inhibitor NPA and $5\,\mu M$ of auxin IAA (indole – 3 – acetic acid) in order to saturate the system with auxin and to block the auxin transport activity (i.e. to only allow to molecule to be degraded). After $2\,h$ of treatment plants were transferred on media containing $10\,\mu M$ NPA, as to observe only the auxin degradation phenomenon as a recovery of DII-VENUS expression. The recovery of DII-VENUS expression was analyzed in a time lapse confocal microscopy experiment for three time points: $30\,min$, $1\,h$, and $1.5\,h$, as shown in Figure A.2.

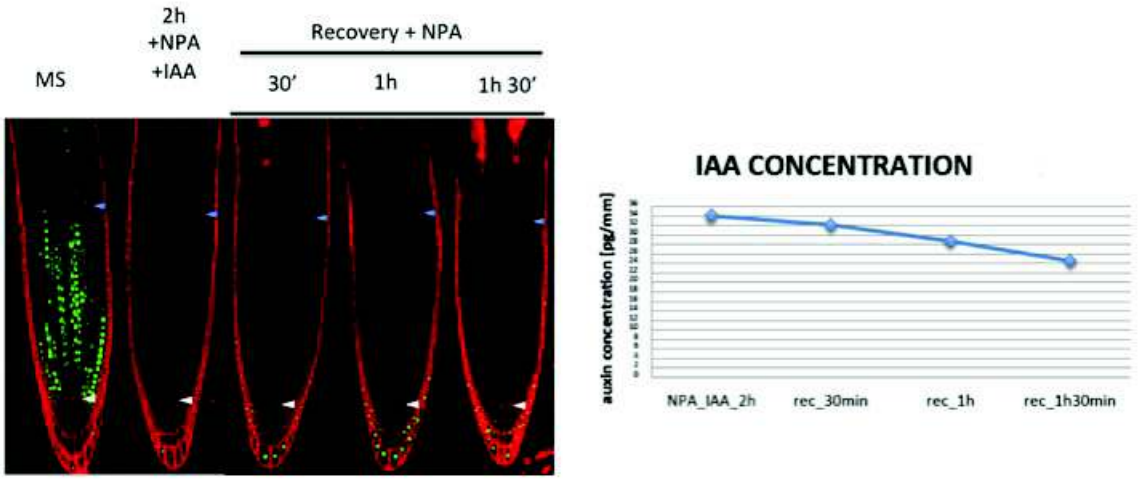


Figure A.2: **Measurement of auxin degradation rate in** *Arabidopsis* **roots.** Left, experimental setting: to measure auxin degradation rate, roots carrying DII-VENUS grown on MS media were transferred on media containing transport inhibitor NPA and auxin IAA, to saturate the system with auxin and block the auxin transport activity, allowing the molecule to be only degraded. After $2\,h$ of treatment plants were transferred on media containing NPA in order to observe only the auxin degradation phenomena as a recovery of DII-VENUS expression. The recovery of DII-VENUS expression was analyzed in a time lapse confocal microscopy experiment for three time points: $30\,min$, $1\,h$, and $1.5\,h$. Right. IAA concentration obtained at each time point, as an inversely proportional function of DII-VENUS expression.

To calculate the degradation rate, I assumed the process to occur with a first-order kinetics:

$$\frac{d[IAA](x,t)}{dt} = -\delta_{GH3} \cdot [IAA](x,t) \tag{A.2}$$

Thus, I quantified DII-VENUS expression on root confocal images at each time point using the software *ImageJ*. To translate the fluorescence intensity into a concentration value, I set DII-VENUS intensity measured for the DII-VENUS root (MS) equal to

the inverse of average auxin concentration value (DII-VENUS expression is inversely proportional to auxin values) in root seedlings (length $0 - 0.5\,mm$), as measured by Ljung et al. [75], equal to $1.8\,pg/mm$, then deriving auxin concentration values at each time point according to the equality:

$$[IAA](t) = \frac{GFP_{wt} \cdot [IAA]_{wt}}{GFP(t)} \tag{A.3}$$

Auxin concentration data at each time point were interpolated using Matlab curve fitting tool *CFtool* to the solution of the equation A.2, obtaining a numerical value for $\delta_{GH3} \sim 10^{-4}\,1/s$ (see Table 2.5).

A.2.4 Auxin maps

Confocal images of *Arabidopsis* roots expressing DII-VENUS were translated into auxin maps, which can be used for the estimation of parameters using the derived auxin transport equation. Auxin maps were obtained using the software *CellseT* [104], based on a cell segmentation method that extract information from confocal images: individual cells are identified and the expression of the fluorescent protein in the nucleus (as in the case of DII-VENUS) or on the cell membrane is combined with the geometry to produce a heat map representing the nuclei intensity of all cells in the image.

Interestingly, the experimentally-derived auxin heat map shows low levels of auxin at the TZ, as predicted by the 2D modelling results (see 2.2.3).

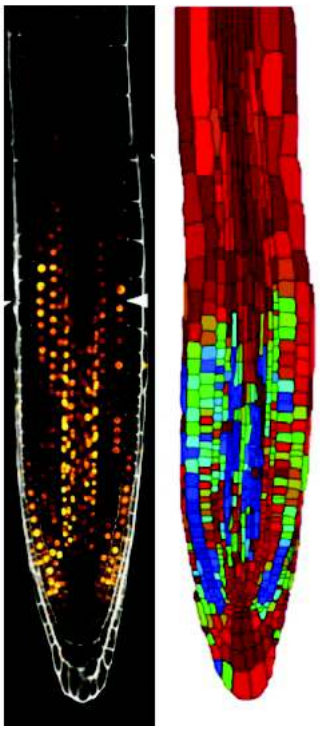


Figure A.3: **DII-VENUS expression translated into auxin heat map**. a) yellow, DII-VENUS nuclear marker. b) DII-VENUS derived auxin heat map after segmentation. Brighter colors represent higher levels of DII-VENUS detected in the nuclei of each cell.

A.3 Modelling Auxin and Cytokinin Interaction

The interplay between auxin and cytokinin has been treated in this thesis work neglecting cytokinin transport or gradient formation, as cytokinin transport mechanism in the root is still poorly understood and no morphogenetic role for cytokinin has been proved. However, I investigated whether root patterning could arise upon the interaction between auxin and cytokinin gradients or as a result of an activator-inhibitor mechanism.

A.3.1 Continuous model

Bishopp et al. demonstrated [13] that cytokinins are translocated from the shoot to the root tip via the phloem and that this transport pathway controls vascular patterning in the root apex. Thus, phloem-derived cytokinins are likely to form a gradient across vascular tissues [90], antagonizing auxin gradient. However, as no evidence of cytokinin polar transport are shown, I assumed the spreading of the hormone as a diffusion-decay mechanism [1].

Following the same procedure used to describe auxin gradient establishment (see 2.1.2.3), cytokinin transport equation can be written as:

$$\frac{\partial \mathrm{c}(\mathbf{x},\mathrm{t})}{\partial \mathrm{t}} = D_C \frac{\partial^2 c(\mathbf{x},t)}{\partial x^2} + b_C - \delta_c \cdot c(\mathbf{x},t) \tag{A.4}$$

, where c is cytokinin concentration and accounting for uniform biosynthesis b_C and degradation throughout the tissue, with degradation rate $\delta_c = 7.5 \cdot 10^{-5} s^{-1}$ [90] and diffusion coefficient in the cytosol $D_C = 600\,\mu m^2/s$, assuming cytokinin diffusivity equal to auxin diffusivity. Cytokinin source is taken in the phloem (in the plot corresponds to the end value of domain length).

Moreover, according to literature data, reaction terms in auxin and cytokinin transport equation can be coupled when studying their interaction: while cytokinin induces auxin degradation [experimental data from our lab and modelling], auxin inhibits cytokinin biosynthesis [59]. These regulations are implemented as saturation functions in the following system:

$$\begin{cases} \frac{\partial \mathrm{a}(\mathbf{x},\mathrm{t})}{\partial \mathrm{t}} = D \cdot \frac{\partial^2 \mathrm{a}(\mathbf{x},\mathrm{t})}{\partial \mathrm{x}^2} - P_{PIN} \cdot \frac{\partial(\mathbf{x},t)}{\partial x} - \frac{\delta_a \cdot a(x,t) \cdot c(x,t)}{1+c(x,t)} + b_a \\ \frac{\partial \mathrm{c}(\mathbf{x},\mathrm{t})}{\partial \mathrm{t}} = D_c \cdot \frac{\partial^2 \mathrm{c}(\mathbf{x},\mathrm{t})}{\partial \mathrm{x}^2} - \frac{b_c}{1+a(x,t)} - \delta_C \cdot c(x,t) \end{cases} \tag{A.5}$$

(A.6)

To evaluate how auxin and cytokinin gradients interact i, I implemented the system of equations in *Mathematica* environment, obtaining the graph reported in Figure A.4. The initial concentration for both auxin is ì $c_{0A} = 1\,a.u.$, whereas cytokinin con-

centration, it is assumed to be some orders of magnitude lower ($c_{0C} = 10^{-3}\, a.u.$), in accordance with literature data [127, 59].

According to the choice of parameter set, the two gradients intersect at $\sim 200\,\mu m$, at the same position at which the TZ is encountered (see section A.2.1 and Table 2.3). The opposite direction of the two gradients supports the hypothesis that the two hormones mediate cell differentiation (cytokinin) and division (auxin), respectively in the distal and apical meristem zone [32].

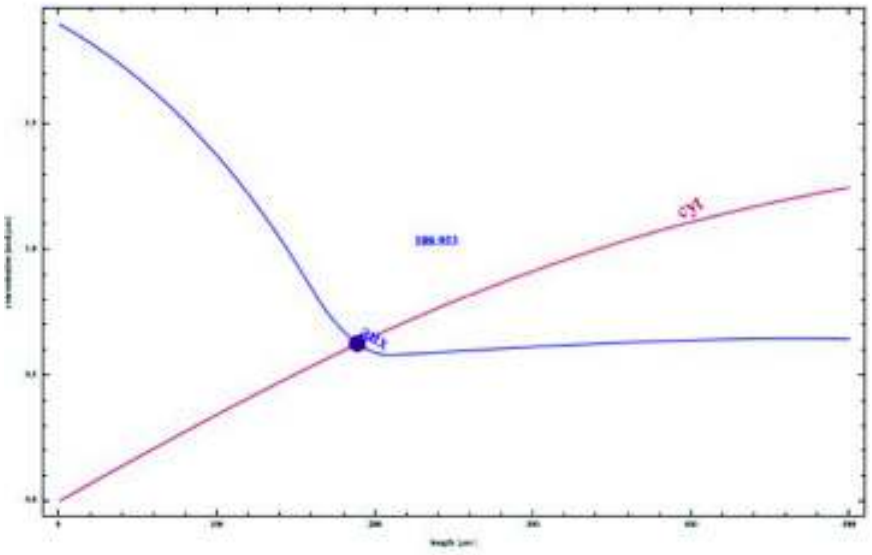


Figure A.4: **Auxin (red) and cytokinin (blue) gradients intersection in the root meristem**. Auxin concentration reaches a maximum at theQC (whose coordinate is positioned at $x = 0$); cytokinin gradient formed by diffusion - decay exponentially decreases from the shoot to the tip of the root. The two gradients intersect at $\sim 200\mu m$, where it is experimentally found the cross-talk between the two hormones [86].

A.3.2 Activator-Inhibitor Model

Given that auxin and cytokinin are both diffusible hormones and that they have specific domains of action, I attempted to study their interaction through the activator-inhibitor

model, according to the theorical background presented in the subsection 1.3.2. This model can be described through the system of equation in 1.6.

For the specific purpose of this work, I assumed auxin to be the activator as it stimulates cell division and its own transport [31, 32], while cytokinin is assumed to be the inhibitor, as it down-regulates auxin polar transporters (PINs) and induces auxin local degradation (see 1.2.4.2 and 2.2 for a full description).

Thus, the the partial differential equations (PDE) of the system 1.6 become:

$$\begin{cases} \frac{\partial A}{\partial t} = D_A \nabla^2 A + f(A, C) \\ \frac{\partial C}{\partial t} = D_C \nabla^2 C + g(A, C) \end{cases} \tag{A.7}$$

, A and D_A, C and D_C represent auxin and cytokinin concentration and diffusivity, respectively. The functions $f(A, C)$ and $g(A, C)$ represent reaction kinetics and vary with respect to the system under study.

Here, the two functions are set as:

$$\begin{cases} f(A, C) = & A - b \cdot A + A^2 C \\ g(A, C) = & A^2 - C \end{cases} \tag{A.8}$$

The dimensionless form of the equations can be written setting , $d = D_C/D_A$; $t = \gamma \tilde{t}$; $x = \gamma \tilde{x}$;:

$$\begin{cases} \partial_{\tilde{t}} A = \tilde{\nabla}^2 A + \gamma \cdot f(A, C) \\ \partial_{\tilde{t}} C = d\tilde{\nabla}^2 C + \gamma \cdot g(A, C) \end{cases} \tag{A.9}$$

This system was studied implementing in *Matlab* environment (the code used is provided at the end of the section). The domain was discretized using finite difference methods.

At the steady state (A_0, C_0), the following conditions are satisfied: $f(A_0, C_0) = 0$; $g(A_0, C_0) = 0$. In order for spatial patterns to arise, the system has to be perturbed, such as an instability emerges.

Introducing a perturbation w of the steady state:

$$w(t) = \begin{cases} A(t) - A_0 \\ C(t) - C_0 \end{cases} \tag{A.10}$$

The perturbation will be locally regulated by the equation:

$$w_t = \gamma \cdot J \cdot w(t) \tag{A.11}$$

J is the Jacobian matrix evaluated at (A_0, C_0): $J = \begin{pmatrix} \frac{\partial f}{\partial A} & \frac{\partial f}{\partial C} \\ \frac{\partial g}{\partial A} & \frac{\partial g}{\partial C} \end{pmatrix}$

This linear system is stable when all eigenvalues λ of A have Re (λ) <0. The characteristic polynomium of A is:

$$det(\gamma \cdot J - \lambda \cdot I) = \lambda^2 - \gamma \cdot (f_A + g_C) \cdot \lambda + \gamma^2 \cdot (f_A \cdot g_C - f_C \cdot g_A) = 0 \tag{A.12}$$

Thus, linear stability is ensured $(Re(\lambda) < 0)$ if:

$$\begin{cases} tr(J) = f_A + g_C < 0, \\ det(J) = f_A \cdot g_C - f_C \cdot g_A > 0 \end{cases} \tag{A.13}$$

```
function []=RDzeroflux(N,a,b,d,gamma)

N = 30; % number of cells
h = 1/N; % step
x = h*(0:N); % ascisse grid
y = h*(0:N); % ordinate grid
[xx,yy] = meshgrid(x,y); % mesh in 3D
dt = .01*h^2; % timestep
% parameters
a = 0.05;
b = 1;
```

```
d = 9;
gamma = 120;

% Inizialization
A0 = (a+b); %  A steady state
C0 = b/A0^2; % C steady state
A = (a+b)*ones(size(xx));
C = (b./A.^2);
A = A + .1*randn(size(xx)); % A steady state perturbation
C = C + .1*randn(size(xx)); % C steady state perturbation
t = 0;
tmax = 10;
nsteps = round(tmax/dt); % time steps

 % Finite difference method - A and C are discretized
    along NORTH,SOUTH,WEST,EAST direction

for n = 1:nsteps
t = t+dt;
AE = A(:,[2:N+1 N]);
AW = A(:,[2  1:N]);
AN = A([2  1:N],:);
AS = A([2:N+1 N],:);
CE = C(:,[2:N+1 N]);
CW = C(:,[2  1:N]);
CN = C([2  1:N],:);
CS = C([2:N+1 N],:);
u2v = (u.^2)./v;
A = C + gamma*dt*(a - b.*A + (A.^2)./C) + dt*(AE+AW+AN+AS-4*A)/h^2;
C = C + gamma*dt*(A.^2-C) + d*dt*(CE+CW+CN+CS-4*C)/h^2;
end
```

```
%Jacobian trace and determinant
tr = fA(A0,C0)+gC(A0,C0);
det = fA(A0,C0)*gC(A0,C0) - fC(A0,C0)*gA(A0,C0);
```

According to the chosen parameter set, the limit condition for instability to arise was found to be $d \geq 9$. The case for $d = 9$ is shown in Figure, where local peaks of the activator are formed in correspondence of local minimum of the activator . Considering $D_A = 600\,\mu m^2/s$ (see Table 2.4), this would imply that $D_C = 5400\,\mu m^2/s$, i.e. cytokinin diffusion coefficient would be much higher than auxin cytokinin diffusivity.

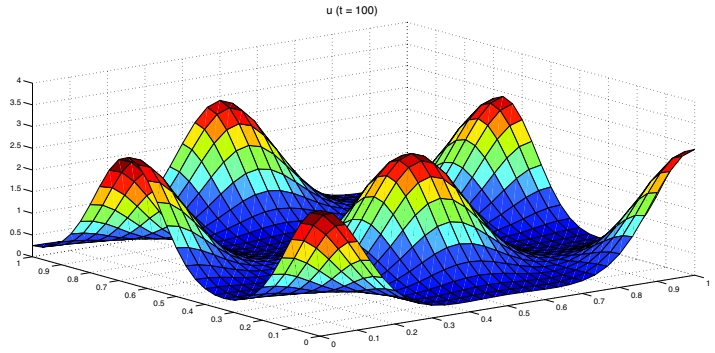


Figure A.5: **Pattern arising from an activator (auxin) -inhibitor (cytokinin) model**: if $D_C/D_A \geq 9$, a multipeak pattern would form. The heat map range spans from red, i.e. highest level or the activator, to blue, i.e. local minimum of the inhibitor.

A.4 Protein Contact Networks (PCN)

The modular (multiscale) feature of Systems Biology is conserved among the most diverse application. For intance, the idea that global behavior of the biological systems can be predicted by understanding its constituting modules can be traced in the identification of modules into proteins. Such a task can be solved through translation of

biological systems into graphs: a network-like structure whose elements are nodes and their mutual interactions are expressed as edges connecting them.

Indeed, during the first part of my PhD, besides the project presented in the previous chapters, I have been involved in a parallel systems biology project, which was specifically focused on protein contact network (PCN) analysis. Understanding the link between protein structure and function is a crucial question in biology. Describing a protein secondary structure as a contact network between its amino acid residues allows for understanding the truly link between structure and function in topological terms: the spatial distribution of residues in the crystal three-dimensional structure is translated into a network of inter-residue interactions, primarily responsible for the protein's three-dimensional structure and activity. his approach relies on the translation of biological systems into graphs.

The use of PCNs has revealed successful in describing processes such as protein folding, allosteric transition and mutation through mathematical descriptors. Through the research carried out on PCNs, I provided my scientific contribution to the field. Published papers are attached.

Biophysical Chemistry 165–166 (2012) 21–29

Contents lists available at SciVerse ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

Shedding light on protein–ligand binding by graph theory: The topological nature of allostery

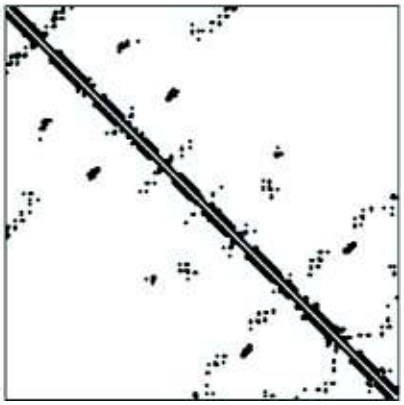
Micol De Ruvo [a], Alessandro Giuliani [b], Paola Paci [c], Daniele Santoni [d], Luisa Di Paola [a,*]

[a] Faculty of Engineering, Università CAMPUS BioMedico, Via A. del Portillo, 21, 00128 Roma, Italy
[b] Environment and Health Department, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161, Roma, Italy
[c] CNR-Institute of Systems Analysis and Computer Science (IASI), BioMathLab, viale Manzoni 30, 00185 Roma, Italy
[d] CNR-Institute of Systems Analysis and Computer Science (IASI), viale Manzoni 30, 00185 Roma, Italy

HIGHLIGHTS

- A graph theoretic method is proposed to analyze protein structure and function.
- The method is based on protein contact matrices.
- Application of the method shows the key role of topology in allosteric transitions.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 February 2012
Received in revised form 2 March 2012
Accepted 2 March 2012
Available online 12 March 2012

Keywords:
Biological thermodynamics
Complex networks
Protein contact network
Enzyme kinetics
Bioinformatics

ABSTRACT

Allostery is a very important feature of proteins; we propose a mesoscopic approach to allosteric mechanisms elucidation, based on protein contact matrices. The application of graph theory methods to the characterization of the allosteric process and, more broadly, to obtain the conformational changes upon binding, reveals key features of the protein function. The proposed method highlights the leading role played by topological over geometrical changes in allosteric transitions. Topological invariants were able to discriminate between true allosteric motions and generic protein motions upon binding.

1. Introduction

Proteins are effective "biological operators": if genetic material can be considered as an information storage device, the effective biological work (whether catalysis, structure formation, signalling, immune defence, energy production..) is carried out by protein molecules and the need of their correct structure is crucial for life.

* Corresponding author. Tel.: +39 06225419634; fax: +39 0622541456.
E-mail address: l.dipaola@unicampus.it (L. Di Paola).

0301-4622/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2012.03.001

In the introduction of [1], Tanford and Reynolds discuss the general attitude of proteins to fulfill biological tasks: "For every imaginable task in a living organism, for every little step in every imaginable task, there is a protein designed to carry it out. The ultimate objective of a task may be chemical or mechanical or to measure colour or fight off a foreign invader — there is no limit to what can be accomplished".

This versatility of functions is due to a proper combination of structure stability and flexibility [2], that comes from several intramolecular interactions, differentially sensitive to the environmental stimuli. The 'correctness' of a protein structure comes from a delicate interplay between molecule and the chemico-physical microenvironment it is embedded into.

In this work, we assume a mesoscopic view of protein–environment interactions: protein structures are considered as networks, i.e. as graphs having the residues as nodes and their mutual contacts as edges, where aminoacid residues are indexed by a non-ambiguous ordering. This view has been demonstrated to be predictive in applications ranging from domain predictions [3] to allosteric hot spots [4].

A protein network corresponds to a square matrix having as rows and columns the aminoacid residues, ordered accordingly to their location along the sequence.

In the case of structural modelling, in the $\{i, j\}$ location we will have a no zero entry if there is a contact between the i-th and j-th residues in the 3D structure (see Materials and methods).

In Fig. 1.1 we report a simple protein structure (recoverin) and the corresponding residue adjacency matrix; the aminoacids ordering along the sequence makes the structure–adjacency matrix relation unique.

The contact network can be considered as the highway on which any signal (e.g., a structural modification) can be transmitted throughout the protein. These signals are the basis of protein physiological action. This creates an immediate and natural link between allostery and complex network analysis: the usual graph theoretical descriptors are, thus, ideal candidates to analyze protein structural and functional changes. Then, for instance, the degree of a node (i.e., the number of contacts of a specific residue in the protein structure) provides a natural description of the aminoacid relevance for protein structural stability (microscale, single residue, and level of modelling).

On the other hand, the spectral graph analysis of the protein network singles out specific modules of the protein (intermediate scale, clusters of the network). Eventually, different proteins can be compared in terms of their global graph descriptors, to elucidate similar construction principles (macroscale, entire protein level).

We focus on the characterization of the changes a protein structure undergoes when it binds its ligand. This characterization, to be effective, should be able to discriminate allosteric effects from mere ligand binding. It is well known that conformational changes are intrinsic to the function of a variety of proteins, and that are somehow triggered by ligand binding [5]. We analyze five proteins as for the differences between free (apo) and bound (holo) conformations. This comparison has been carried out by both geometrical and topological views, demonstrating the relevance of local topological description for the allosteric character identification.

2. Materials and methods

2.1. Protein systems

Calcium binding proteins (CaBPs) are responsible for many regulatory physiological mechanisms [6], switched by calcium concentration transients.

CaBP's can be divided into two general classes: sensors and buffers. Proteins belonging to the former class, like calmodulin (CaM) and recoverin, translate the chemical signal of an increased Ca^{2+} concentration into diverse biochemical responses. Calcium binding to sensor proteins leads to a transition from a tense (T) to a relaxed (R) state, that results in the exposure of a large hydrophobic surface, allowing the protein to interact with other molecular targets to accomplish regulatory functions.

On the other hand, the capture of calcium ions by buffer (or carrier) proteins, such as parvalbumin and calbindin D_{9k}, is accompanied by minor conformational changes. Sensors are allosteric protein, while buffers are not.

Our analysis has been extended to human hemoglobin and serum albumin, both representing a paradigm in protein science. Hemoglobin is in charge for the transport of respiratory gases and its structure was firstly resolved by Perutz, which was awarded of the Nobel Prize for this discovery [7]. Hemoglobin quaternary structure accounts for four-chains (A, B, C, D), each containing a heme group, responsible for oxygen and carbon monoxide binding, with a strong allosteric attitude [8].

Human serum albumin (HSA) is the most abundant serum protein, accounting for different biological functions, most of them linked to its ability to catch and carry hydrophobic molecules [9]. It is one of the largest single chain protein, showing different domains (I, II and III), each divided into two subdomains, A and B [9]. It has different binding sites for endogenous and exogenous toxins, drugs and metabolites [10]. The interaction between sites is limited to steric hindrance (steric interactions between fatty acids and L-tryptophan [11]) and it is not correlated to structural variations, neither at global nor at local scale. The non-allosteric character of HSA binding sites promotes its attitude as a carrier, making it able to transport independently several different classes of hydrophobic molecules, with large binding affinities.

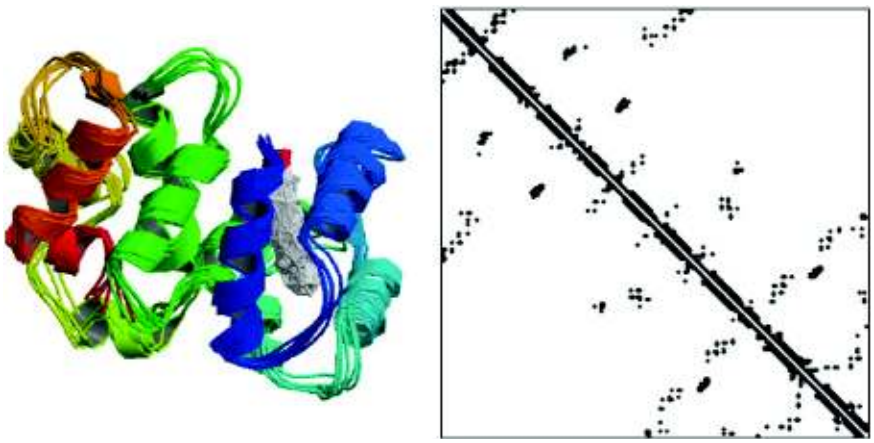


Fig. 1.1. Protein structure and adjacency matrix map for recoverin: contacts correspond to black dots.

2.2. Topological and geometrical comparison of protein structures

The protein structure was translated into a residue contact network, based on the Euclidean mutual distance matrix **d**, computed on the basis of the spatial positions of the α-carbons, extracted from the protein PDB file.

We established a cut-off for inter-residue distance ranging within $\mathcal{I} = [4-8]$ Å accounting for intramolecular noncovalent interactions [12]; thus, the corresponding unweighted protein structure graph was built up, whose adjacency matrix $A = \{a_{ij}\}$ is formally defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } d_{ij} \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

A key point is the identification of graph modules, namely groups of nodes showing a larger number of connections to each other with respect to those established with nodes outside the module itself. Such modules are strongly related to protein domains [13,14].

To single out meaningful modules, we applied the classical partitioning method based on the spectral analysis of the graph Laplacian matrix [15]. As a final result, the graph is divided into modules that are disjoint sets of nodes ("crispy" partition). Once the graph has been partitioned, for each node it is possible to evaluate two parameters that are a direct measure of its attitude to establish links with nodes belonging to the same module rather than with those lying into other modules [16]:

- the within-module z-score:

$$z_i = \frac{k_{is} - \bar{k}_{s_i}}{\sigma_{si}}$$

  k_{is} is the number of links the i-th node establishes with nodes belonging to its own cluster s_i; $\bar{k}_{s_i}$ is the average degree for nodes in cluster s_i, and LGRsv$_{si}$ the corresponding standard deviation;
- the participation coefficient, that describes the attitude of the node to connect to other nodes off its own cluster:

$$P = 1 - \left(\frac{k_{is}}{k_i}\right)^2$$

  k_i is the i-th node overall degree.

According to the values of P and z, it is possible to establish a cartography for the nodes [16], based on their role in terms of within- and between-modules connections.

High values of z combined with zero P are characteristic for nodes occupying a central position in their own module, while not even communicating with other modules; on the other way, nodes that are characterized by not-null P values make modules communicate with each other (signal transmission).

We make the hypothesis this latter process is the molecular engine behind the allosteric effect. The demonstration of a hypothetical percolation threshold of allosteric proteins in terms of such a description will constitute a further proof of our statement.

Table 1
List of proteins mentioned with reference to the specific name, PDB ID, family and class (the classification is based on the relative ligand binding function).

Protein name	PDB ID	
	Apo	Holo
EF-hand calcium binding proteins		
Calbindin D_{9k} (Cb-D_{9k}) (buffer)	1CLB	2BCA
Parvalbumin (PV) (buffer)	2NLN	1TTX
Recoverin (RC) (sensor)	1IKU	1JSA
Human hemoglobin (HbO2, HbCO)	2DN2	1GZX (O2), 1BBB (CO)
Human serum albumin (Ab)	1AO6	2VUE (bilirubin)

To investigate structural similarity between the apo and holo forms, we computed the Root Mean Square Distance (*RMSD*) [17] at a global and a local scale.

As we will discover in the next section, an active site residue going from a not-null to a zero value of P corresponds to a shift from a "prone to binding" to a "closed" state. Eventually, we introduced another topological parameter, the Hamming distance D between adjacency lists, that reports the overall number of matrix positions at which the corresponding values are different.

In details, two metrics have been computed: the first one (D_{av}) is the Hamming distance normalized to the corresponding connectivity of a completely connected graph ($\frac{N(N-1)}{2}$), while the other (D_{eff}) is normalized to the number of apo form effective contacts (unit values in the apo adjacency matrix) .

Global (entire protein) and local (single residue) metrics are summarized in Table 2.

Table 2
Global and local parameters employed for topological and structural analysis.

Global metrics		Local metrics	
Topological	Structural	Topological	Structural
D_{av}	RMSD	P,z (active sites)	*RMSD* per residue
D_{eff}			

3. Results

3.1. Global level analysis

At first, we computed the topological distances D_{av} and D_{eff} between the apo and holo forms for proteins in Table 1; results reported in Table 3 show that D_{av} scales with protein size: this is a logical consequence of the exponential decay of contacts with size [13,18].

On the other hand, D_{eff} does not appear related neither to the size nor to allostery, being strongly invariant across different protein systems, despite their size, kinetic properties and, as we will discover, huge differences in flexibility (measured by *RMSD*).

This suggests the contact matrix maintains the protein identity and mirrors the strong invariance we observed as for P–z global distributions of proteins [13].

On the other hand, results in Table 4 show a huge variation in *RMSD* values regardless of size or allosteric character. Thus, global *RMSD* value does not provide any useful information about allosteric effect.

3.2. Local level

Hemoglobin is assimilable to sensor proteins, it undergoes conformational changes due to its physiological activity (respiratory gases binding). The quaternary configuration of low affinity, deoxygenated hemoglobin is known as the tense (T) state, whereas the quaternary structure of the fully oxygenated high affinity form is known as the relaxed (R) state. The dissociation constant of the first oxygenation step is $K_T = 1 \cdot 10^{-4}$ M while that related to the last oxygenation step high affinity) is $K_R = 2 \cdot 10^{-6}$ M [19].

Table 3
D_{av} and D_{eff} for sample proteins.

	Cb-D_{9k}	PV	RC	HbO2	HbCO	Ab
N	75	108	188	574	574	1124
D_{av}	0.015	0.012	0.006	0.002	0.002	0.001
D_{eff}	0.58	0.62	0.53	0.56	0.60	0.56

Table 4
Global *RMSD* calculated for sample proteins.

	Cb-D_{9k}	PV	RC	HbO2	HbCO	Ab
N	75	108	188	574	574	1124
RMSD	1.84	3.59	11.45	0.30	1.21	0.94

When carbon monoxide is present, it competes strongly with oxygen at the heme binding sites (the dissociation constant is 200 times smaller than K_R), forming a very bright red form of hemoglobin called carboxyhemoglobin; the corresponding dissociation constants are $K_T = 1 \cdot 10^{-6}$ M and $K_R = 5 \cdot 7^{-8}$ M.

The results of *P*–*z* analysis of the deoxigenated (2DN2), oxygenated (1GZX) and carbon-monoxy (1BBB) forms, are shown in Table 5, where A–C and B–D chains are grouped, due to the symmetry of hemoglobin structure. It appears clear that in the holo states, *P* values in the active sites turn to 0 (this implies that the active site is decoupled from other modules). We interpreted this shift in terms of a T–R transition, coupled with an increase of ligand binding affinity. Also for the *CO*-bound complex all *P* values in the binding site turn to 0, confirming the higher affinity of hemoglobin for this ligand.

We have also reported the index ($\overline{P}_0$) that expresses the fraction of residues having $P = 0$, for each protein form. The net increase of null values in the *P* distribution in holo forms with respect to the apo one is very remarkable, given the basic invariance of *P*–*z* distribution, and it is consistent with the "closing of the gates" between modules, hindering the signal transmission throughout the protein.

HSA binds bilirubin, a toxic metabolite of heme (dissociation constant: $K_d = 10^{-7}$ M–10^{-8} M) at a high affinity site and acts as a buffer preventing the transfer of bilirubin from blood to tissues that otherwise would cause bilirubin encephalopathy [20]. Focusing our analysis on contacts belonging to this site the non-allosteric character of HSA domains corresponds to *P* values showing only minor changes upon binding, as reported in Table 6.

3.3. Calcium binding proteins

Recoverin is a calcium-binding protein involved in visual signal transduction; the protein molecule is composed of two domains, each bearing the calcium binding active sites [21]. Recoverin undergoes a typical T (tense)–R (relaxed) structural transition upon calcium binding [22]; the dissociation constants of the two sites in the holo state are $K_R = 1 \cdot 4^{-5}$ M for site 1 and $K_T = 9.1 \cdot 10^{-6}$ M for site 2, respectively. As we can observe in Table 7, our data confirm the higher affinity of site 2: all residues of this site have $P = 0$ in the holo state. Moreover, on the entire protein, the $P = 0$ nodes increase in number in the holo form, with respect to the apo one (Table 7).

Parvalbumins (PVs) are a family of small, acidic Ca^{2+} binding proteins found in muscle of vertebrates, in mammal brain and endocrine glands [23]. Binding and release of Ca^{2+} induce relatively small conformational changes [24], confined within the Ca^{2+} binding loop or in the close vicinity. Therefore, the main function of PV is presumably the control and modulation of intracellular Ca^{2+} signals [25].

Table 5
Hemoglobin *P* values for the active sites aminoacids relative to the deoxygenated (2DN2), complexed with oxygen (1GZX) and with carbon-monoxy (1BBB) forms.

	Chain A-C					Chain B-D				$\overline{P}_0$
	Leu	Phe	His	Val	Leu	Leu	Phe	His	Val	
2DN2 (apo)	0	0.30	0	0.43	0.39	0	0	0	0.26	0.13
1GZX (holo)	0	0.31	0	0	0	0	0	0	0	0.70
1BBB (holo)	0	0	0	0	0	0	0	0	0	0.75

Table 6
Apo (1AO6) and holo (2VUE) HSA *P* values for the active sites aminoacids.

Active site	1AO6 (apo)	2VUE (holo)
Arg	0.96	0.64
Leu	0.96	0.43
Val	0.98	0.75
Arg	0.84	0.67
Pro	0.75	0.94
Val	0.86	0.61
Met	0.86	0.67
Ala	0.75	0.51
Phe	0.82	0.89
Lys	0.75	0.89
Tyr	0.75	0.95
Glu	0.19	0.69
Ile	0.61	0.89
Arg	0.91	0.75
His	0.99	0.75
Phe	0.86	0.75
Leu	0.56	0.89
Phe	0.86	0.75
Tyr	0.69	0.75
Leu	0.75	0.56
Leu	0.67	0.86
Arg	0.75	0.84
Gly	0.51	0.91
Lys	0.89	0.75
$\overline{P}_0$	0.013	0.0125

Rat parvalbumin contains two equivalent Ca^{2+} binding sites, (dissociation constant $K_d = 11$nM) binding CA^{2+}in a noncooperative way [23].

In this study, apo and holo form of rat LGRb-parvalbumin are considered for the analysis: the two domains bind Ca^{2+} with similar affinities ($K_1 \approx K_2 = 4 \cdot 10^{-8}$ M) and, in this case too, alterations in *P* values reflect thermodynamic data; actually, although the number of $P = 0$ nodes does not increase in the holo state (there is no global conformational change), in the active site, either in site 1 and 2, the residues *P* value goes to 0 (Table 8).

We can conclude that, even if a local change occurs, binding sites affinities are not modified, as expected for a non-allosteric binding mechanism.

Calbindin D_{9k} (Cb-D_{9k}) is the smallest EF-hand protein, whose EF-hand pair consists of a "canonical" EF-hand domain with a Ca^{2+} binding loop (EF2) and a non-canonical (also termed "pseudo EF-hand") Ca^{2+} binding loop (EF1), typical for the S100 family proteins. The two EF-hand domains bind Ca^{2+} with similar affinities ($K_1 \approx K_2 = 4 \cdot 10^{-8}$ M) and positive cooperativity; it has been seen that there are long range effects that contribute to the positive cooperativity of calcium binding to calbindin. In Johnson et al.'s [26] study, the flexibility of the methyl groups of the apo versus the holo state has been examined: although the overall Ca^{2+} induced conformational change in Cb-D_{9k} is not pronounced, upon binding, the methyl groups (and thus the rest of the protein bound to them) at the end of the protein far from the binding sites becomes more flexible [27]. This behaviour is mirrored by the change in *P* values distribution depicted in Fig. 3.1.

Thus, we can again prove the consistency of our hypothesis: as it can be observed in Table 9, not only $\overline{P}_0$ values do not change, but even *P* values of the active sites remain substantially unchanged, pointing to the absence of evident modification in the topological structure of the whole protein.

Table 7
Apo (1IKU) and holo (1JSA) recoverin *P* values for the active sites aminoacids.

	Site 1					Site 2				$\overline{P}_0$
	Asp	Asn	Asp	Thr	Glu	Asp	Asn	Thr	Glu	
1IKU (apo)	0	0	0	0	0.67	0.19	0	0	0	0.54
1JSA (holo)	0.56	0	0.36	0.26	0	0	0	0	0	0.56

Table 8
Apo (2NLN) and holo (1TTX) parvalbumin *P* values for the active sites aminoacids.

	Site 1					Site 2				$\overline{P}_0$
	Asp	Ser	Tyr	Glu	Asp	Asp	Asp	Lys	Glu	
2NLN (apo)	0.61	0	0	0	0	0	0	0.16	0.75	0.41
1TTX (holo)	0	0	0	0	0	0	0	0	0	0.37

Table 9
Apo (1CLB) and holo (2BCA) calbindin D_{9k} *P* values for the active sites aminoacids.

	Site 1					Site 2				$\overline{P}_0$
	Ala	Glu	Asp	Gln	Glu	Asp	Gly	Glu	Glu	
1CLB (apo)	0	0	0	0.51	0.75	0.19	0.61	0.75	0.75	0.49
2BCA (holo)	0	0	0	0.36	0.19	0.17	0.36	0.69	0.21	0.39

Therefore, the allosteric mode (recoverin) amplifies the local change (in the active site) on a global scale rearrangement; on the contrary, the non-allosteric mode (parvalbumin) corresponds only to local scale topology variations (see Fig. 3.2(a)).

As depicted in Fig. 3.2(b), in the case of recoverin the distribution changes throughout the whole protein, while in the case of parvalbumin the modifications are confined to regions close to the active site.

Eventually, we computed *RMSD* for residues in the active site of proteins (Fig. 3.3): these values are quite smaller than global *RMSD*. This result, together with the lack of discrimination ability of RMSD as for allosteric/non-allosteric systems, is a further proof of the mainly topological nature of allosteric motion, while geometrical changes are common to any kind of protein–ligand binding.

4. Discussion

It is important to put our results into the general perspective of protein–ligand interactions, that have been largely examined in terms of ligand binding thermodynamics equilibrium [28,29]. Simple models refer to the very simple case of n active sites on a protein P, that are independent and identical (same microscopical dissociation constant k); in this case, the binding equilibrium is represented by the well-known Scatchard equation:

$$\frac{\nu}{c_L} = \frac{n}{k} - \frac{c_L}{k} \tag{4.1}$$

where c_L is the free ligand concentration and ν represents the equilibrium saturation degree, defined as:

$$\nu = \frac{\sum_{i=1}^{n} i \cdot c_{Pi}}{c_{P0}} \tag{4.2}$$

c_{Pi} is the ligand–protein complex (accounting for i molecules of L bound to the protein active sites, such that the overall protein concentration is $c_{P0} = c_P + \sum_{i=1}^{n} c_{Pi}$).

Applying the mass balance to all molecular species it is derived an expression in terms of directly observable variables (c_L, c_{P0} and c_{L0}):

$$\nu = \frac{c_{L0} - c_L}{c_{P0}} \tag{4.3}$$

where c_{L0} is the overall ligand concentration.

Once n is known, the corresponding fraction saturation degree y as function of c_L is described by the typical saturation hyperbolic curve:

$$y = \frac{\sum_{i=1}^{n} i \cdot c_{Pi}}{n \cdot c_{P0}} = \frac{\nu}{n} = \frac{c_L/k}{1 + c_L/k} \tag{4.4}$$

This representation is often oversimplified: frequently, the ligand interaction with the active site leads to a local and, more often, to a global conformational transition, thus changing the ligand affinity of the remaining free active sites. This phenomenon is addressed to as allostery and it results in a dissociation constant k varying with the saturation degree ν.

Allostery is a specific aspect of cooperativity, underlying many biochemical processes such as folding [30], biomacromolecules aggregation [31] and, indeed, biomacromolecules ligand binding.

Cooperativity derives from the complex chemico-physical nature of biomacromolecules, giving rise to a continuous exchange between local and global — over the all molecule — events [32]. Thus, the result of multiple local processes is almost always not additive with respect to the single process contribution; the difference is a measure of the ability of the local events to be propagated on a larger scale.

Let's set k_0 as the value of k at zero saturation degree, being $\Delta G_0^0 = -RTlog(k \cdot c_{rs})$ the corresponding standard specific Gibbs free energy (c_{rs} is the standard reference for the concentration, taken 1 M), the value of the dissociation Gibbs free energy at the saturation degree ν can be expressed as [28]:

$$\Delta G_\nu^0 = \Delta G_0^0 + RT\ \phi(\nu) \tag{4.5}$$

$\phi(\nu)$ is a measure of the site interaction energy, at a given saturation degree ν.

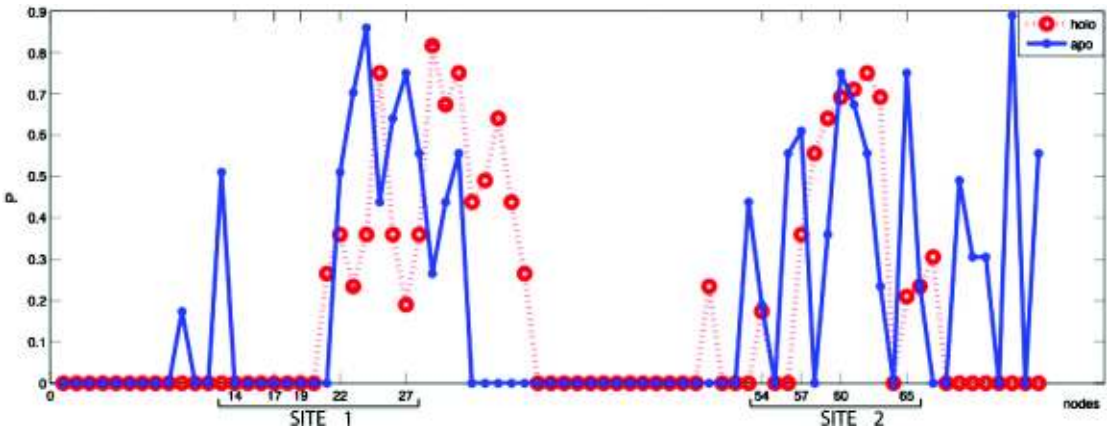


Fig. 3.1. *P* values distribution for apo (blue) and holo form (red) of calbindin; residues involved in the active sites are emphasized. Flexibility of the holo form is evident as *P* values of several residues out of the active sites turn into 0.

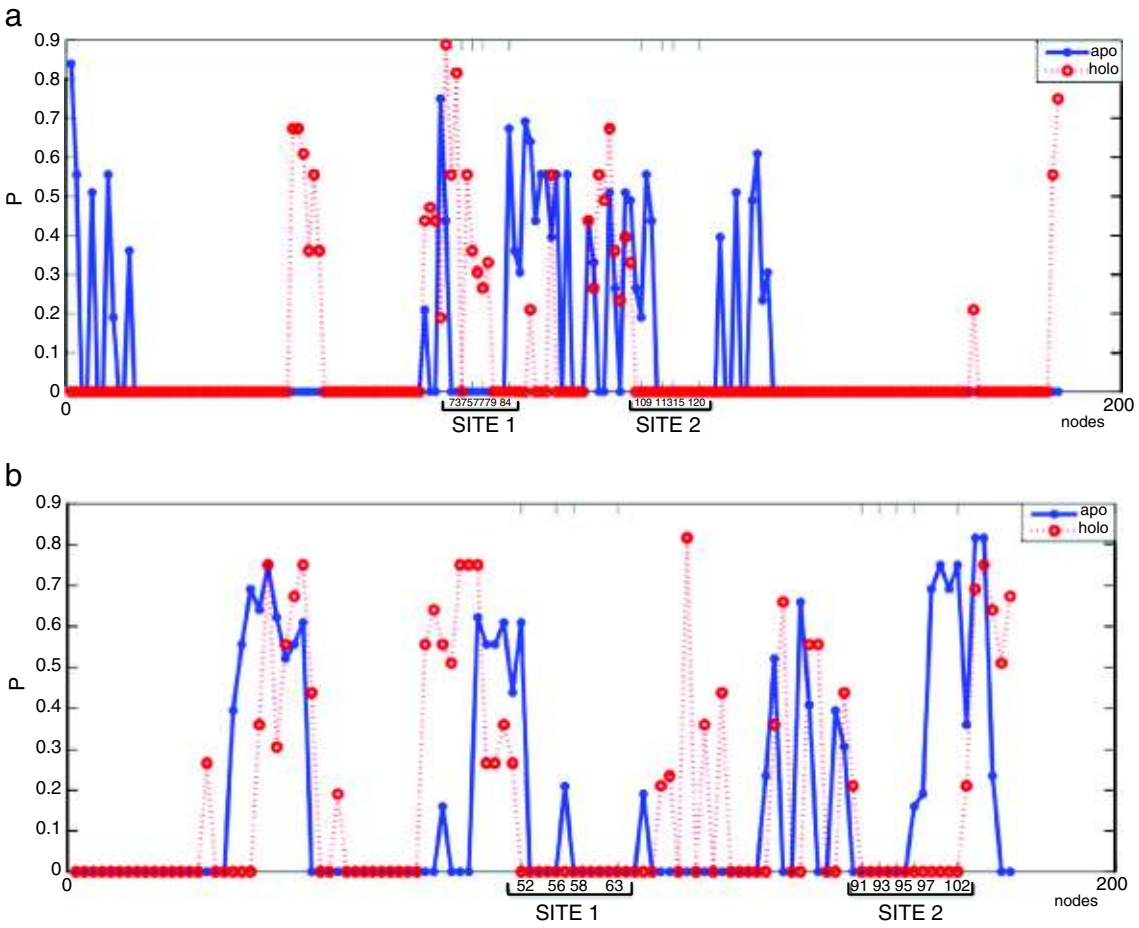


Fig. 3.2. *P* values distribution for apo (blue) and holo form (red) of recoverin a) and of parvalbumin b). Residues involved in the active sites are emphasized. Overall *P* values sharply change in the holo form of RC whereas, in the case of PV, holo form P distribution quite follows that of the apo one.

Since $\Delta G_{\nu}^{0} = -RT\, logk(\nu)$, it is obtained:

$$k(\nu) = k_0 \cdot exp(-\phi(\nu)) \tag{4.6}$$

Unfortunately, it is not straight and simple to derive a physically-based description of $\phi(\nu)$; in any case, when k depends on ν, if all the residue sites have initial identical affinity, an allosteric process is probably occurring.

A semiempirical approach [33] is based on an "all-or-none" scheme:

$$P + \alpha L \rightleftharpoons P \tag{4.7}$$

to which corresponds an apparent macroscopic dissociation constant

$$K_{\alpha} = \frac{c_P \cdot c_L^{\alpha}}{c_{P\alpha}} \tag{4.8}$$

the corresponding fractional saturation degree y is:

$$y = \frac{(c_L/K)^{\alpha}}{1 + (c_L/K)^{\alpha}} \tag{4.9}$$

where the parameter α, known as the Hill constant, is derived as:

$$\alpha = \frac{d\left\{log\left[\frac{y}{1-y}\right]\right\}}{d[log(L)]} \tag{4.10}$$

α lies in the range $[1, n]$ and it is a direct measure of the cooperativity of the binding process: when $\alpha = 1$, Eq. (4.10) equals to Eq. (4.4), describing a noncooperative system; on the other hand, when $\alpha = n$, the process is named as infinitively cooperative, indicating a simultaneous binding of ligand molecules to all protein active sites.

Classical molecular models for allosterism are based on the concept that interactions between sites is translated into conformational perturbations that locally modify binding site shape and, consequently, ligand affinity; this conformational change is thought to act also at a global scale, giving rise to different structural forms, whose transformation equilibria match to those of binding [28].

If the identical binding sites are located on highly similar chains that are collected together into a quaternary structure, the MWC model [34] describes the concerted motion of subunits due to ligand binding, that results into an allosteric behaviour; this model has been applied to describe the affinity of hemoglobin for respiratory gases (oxygen and carbon dioxide); if more complex conformational changes are involved, the KNF model [35] is able to describe the relationship between sites occupancy and affinities.

These idealized representations are based on a vision of the proteins, as a rigid objects that change mostly abruptly their shape from one conformational state to another. Recently, the linkage between conformational changes and allosterism has been thoroughly discussed [36]: the allosteric conformational change may be revised in terms of redistribution of conformation distribution that assets the native structure conformation ensemble. In other words, allosterism doesn't induce conformational changes to the native, apoprotein form, but it shifts the conformational state distribution for the apo form towards the conformations corresponding to the higher ligand affinities (R state, for the MWC model, for instance) for cooperative allosterism. In other words,

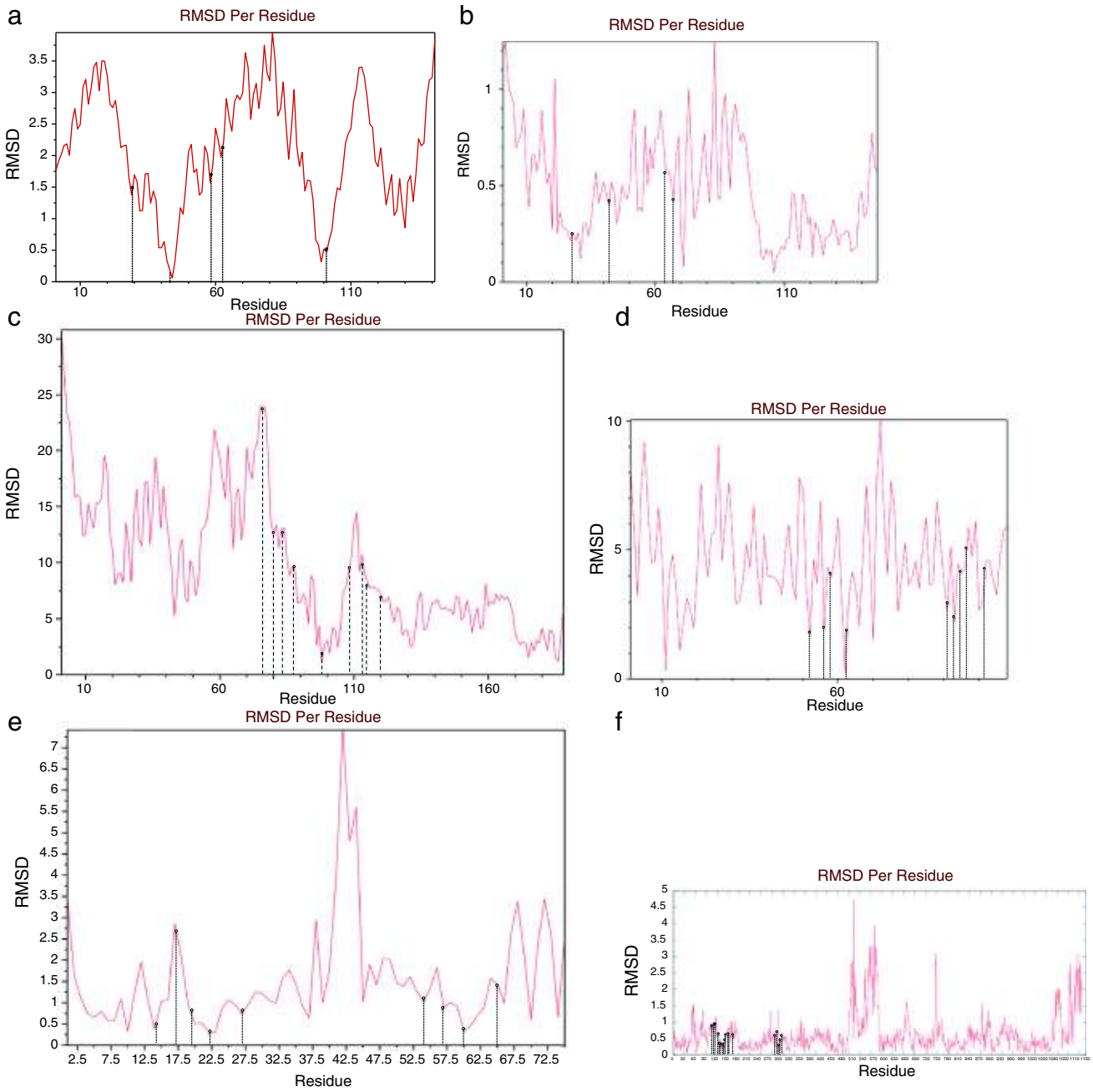


Fig. 3.3. *RMSD* per residue resulting from the superposition of the apo and holo forms of proteins: a) chain A and b) chain B of 2DN2-1GZX (Hb); c) 1IKU-1JSA (RC); d) 1TTX-2NLN (PV); e) 2BCA-1CLB (Cb-D_{9k}); f) 1AO6-2VUE (Ab). Dashed lines are traced to highlight active sites residues: for each protein, peaks correspondent to these residues are not prominent with respect to the global *RMSD* tendency.

the ligand binding doesn't create *de novo* conformations that are not energetically allowed in the apo form, but simply increase their probability of existence. Another point of view involves the dynamic fluctuations of protein atoms, that are strongly related to the intramolecular protein structure interactions, strongly influenced by the protein microenvironment [37–39]. Protein dynamics describe how proteins breathe in their environment, acting and reacting to external stimuli; to measure the entity of local and global fluctuations, the Root Mean Square Distance (*RMSD*) is successfully applied [40]. What is relevant for our approach is that, besides the different models of allosterism, a fundamental kinetic difference linked to the value of parameter α does exist between allosteric and not allosteric systems. Thus, it could be in principle highlighted by a topological/structural analysis.

The definition of what is allosteric and what is not, in our opinion, must be based on reaction kinetics; on the contrary, if we assume a purely structural definition of "a motion started at a given site and further generalized to other parts of the systems", each protein system can be considered as allosteric, as correctly stressed by Del Sol and Nussinov [4].

Applying our method, we were able to identify global motions discriminating apo and holo structures in all the studied systems (both allosteric and not allosteric).

The analysis presented demonstrates how topology is useful to discriminate proteins structure variation upon ligand binding, through the explanation of contact roles. Actually, our results indicate that the variation in *P* values is a peculiar hint to discriminate allosteric behaviour of sensor proteins despite to the rigid structure of the buffer ones. The negligible (but strongly invariant) topological motion of aminoacids does not provoke relevant global changes in contact network but, on a local scale, it leads to the variation in the role of contacts themselves;

meaning, the loss of boundary contacts within clusters being spreaded, let the entire structure being free to move. Thus, it can be easily understood that clusters individuated by spectral theory have a functional role.

5. Conclusion

We proposed a simple methodology, based on a network interpretation of protein structures, and applied it to the analysis of CaBP's, a well-known class of proteins involved in calcium metabolism by different molecular mechanisms, giving a proof-of-concept of the possibility to link chemico-physical features to protein graph invariants.

It can be therefore stated that the variation in P values is a peculiar hint to discriminate allosteric behaviour of sensor proteins despite to the rigid structure of the buffer ones.

We give a further proof of how graph theory provides a useful general framework for proteins science.

Future directions of our method could lead to the detection of altered values in the network modules, and/or in the *P–z* distribution of the aminoacids, thus identifying clusters that are pathologically altered in diseases, like Alzheimer's and Parkinson's. [25]

Furthermore, the identification of network invariants able to follow the variation of structure due to binding, in the physiological and pathological scenario, could be a starting point for drug design: the simulation *in silico* of synthetic CaBP's could be guided through the forecast of affinity for the active site through the network analysis of novel apo and holo forms.

The nature of global motions induced in a protein system by the binding of a ligand was investigated by means of molecular dynamics approaches by [41,42] , pointing to the very organized character of such motions. The conformational changes can be described as motion along one or few coherent degrees of freedom: only pairs of sites that couple to the same conformational degrees of freedom (e.g. sharing the same loading profile along the principal components of motions [43,44]) can be allosterically connected [41]. This 'coupling of sites' is the dynamic analogue to our static 'between modules' interaction as modeled by P index [43,44] . This analogy is made more cogent by the results of Park and Kim [45] which find a correlation between 'allosteric hotspots' (residues more deeply involved in allosteric signal transmission) and the 'betweeness' of the corresponding nodes in the protein contact network (a measure linked to the number of different pathways passing by a given node). The general consistency between dynamical (correlation between motions of residue pairs) and structural (distance in space between residue pairs) protein representations was given a proof-of-concept in [46], so in some sense 'closing the circle' between dynamical and topological allosteric behaviour pictures.

Our results confirm the hypothesis the allosteric signal transmission relies on non covalent contacts with only a minor (if any) role exerted by the peptide bonds backbone [47–49,4]. Indeed, topological metrics prevailed over geometrical one with reference to allosteric and non-allosteric discrimination. In this respect, it is worth noting that the 'global motion' is common to any protein system in the passage from apo to holo state. This is a consequence of the fact the protein molecule is a strongly connected entity.

On the other hand, geometrical and topological global motions have a very different status: while geometrical motions can vary a lot between different systems due to the relative flexibility of the structure, topological motions are strongly invariant across very different molecules. This allows us to hypothesize the 'contact network' as the invariant core responsible for maintaining protein identity, so allowing only relatively minor (and strongly concerted) motions. The fact topological changes are highly concerted implies the 'quality' of motions is more important than their mere entity. Allosteric 'effective' motions are of a very special kind in terms of their effects on protein networks: they 'close' the communication lines between active site and the rest of the molecule, affecting the inter-modules pathways. This behaviour was not observed in non-allosteric system and stresses the importance of protein network description in studying structure-function relations. Our result open a possible avenue so to rationalize the search for the so called 'allosteric-network' drugs devised by Csermely and Nussinov [50], in order to be really effective, our approach should be enlarged to natively-unfolded system (e.g. by the use of NMR data), given their relevance in cell signaling [51,52]. A last remark is linked to the need of assuming a thermodynamic based definition of allostery being the simple statement 'motion transmitted at distance from the ligand' being common to any substrate binding event.

References

[1] C. Tanford, J. Reynolds, Nature's Robots: a History of Proteins, Oxford University Press, 2004.
[2] K. Teilum, J.G. Olsen, B.B. Kragelund, Protein stability, flexibility and function, Biochimica et Biophysica Acta 1814 (8) (2011) 969–976.
[3] S. Vishveshwara, K. Brinda, N. Kannan, Protein structure: insights from graph theory, Journal of Theoretical and Computational Chemistry 1 (1) (2002) 187–212.
[4] A. del Sol, M. Araúzo-Bravo, D. Amoros, R. Nussinov, Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages, Genome Biology 8 (5) (2007) R92.
[5] F. Capozzi, F. Casadei, C. Luchinat, Ef-hand protein dynamics and evolution of calcium signal transduction: an NMR view, Journal of Biological Inorganic Chemistry 11 (8) (2006) 949–962.
[6] E. Permyakov, R. Kretsinger, Calcium Binding Proteins, Wiley Series in Protein and Peptide Science, John Wiley and Sons, New York-London, 2011.
[7] M. Perutz, M. Rossmann, A. Cullis, H. Muirhead, G. Will, A. North, Structure of hemoglobin, Nature 185 (4711) (1960) 416–422.
[8] J. Edsall, Hemoglobin and the origins of the concept of allosterism, Federation Proceedings 39 (2) (1980) 226–235.
[9] T.P. Jr, Serum albumin, Advances in Protein Chemistry 37 (1975) 161–245.
[10] X. He, D. Carter, Atomic structure and chemistry of human serum albumin, Nature 358 (1992) 209–215.
[11] V. Cunningham, L. Hay, H. Stoner, The binding of l-tryptophan to serum albumins in the presence of non-esterified fatty acids, Biochemical Journal 146 (1975) 653–658.
[12] I. Bahar, R. Jernigan, Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation, Journal of Molecular Biology 266 (1) (1997) 195–214.
[13] A. Krishnan, A. Giuliani, J. Zbilut, M. Tomita, Network scaling invariants help to elucidate basic topological principles of proteins, Journal of Proteome Research 6 (10) (2007) 3924–3934.
[14] A. Giuliani, L.D. Paola, R. Setola, Proteins as networks: a mesoscopic approach using haemoglobin molecule as case study, Current Proteomics 6 (4) (2009) 235–245.
[15] M. Newman, Finding community structure in networks using the eigenvectors of matrices, Physical Review E 74 (3) (2006) 036104.
[16] R. Guimerà, L. Amaral, Cartography of complex networks: modules and universal roles, Journal of Statistical Mechanics P02001 (2005) 1–13.
[17] O. Carugo, S. Pongor, A normalized root-mean-square distance for comparing protein three-dimensional structures, Protein Science 10 (7) (2001) 1470–1473.
[18] L. Di Paola, P. Paci, D. Santoni, M. De Ruvo, A. Giuliani, Proteins as sponges: a statistical journey along protein structure organization principles, Journal of Chemical Information and Modeling 52 (2) (2012) 474–482.
[19] K. Imai, Adair fitting to oxygen equilibrium curves of hemoglobin, Methods in Enzymology 232 (1994) 559–576.
[20] C. Petersen, C. Ha, K. Harohalli, J. Feix, N. Bhagavan, A dynamic model for bilirubin binding to human serum albumin, The Journal of Biological Chemistry 275 (28) (2000) 20985–20995.
[21] S. Permayakov, A. Cherskaya, I. Senin, A. Zargarov, S. Shulga-Morskoy, A. Alekseev, D. Zinchenko, V. Lipkin, P. Philippov, V. Uversky, E. Permyakov, Effects of mutations in the calcium-binding sites of recoverin on its calcium affinity: evidence for successive filling of calcium binding sites, Protein Engineering 13 (11) (2000) 783–790.
[22] J. Ames, T. Porumb, T. Tanaka, M. Ikura, L. Stryer, Amino-terminal myristoylation induces cooperative calcium binding to recoverin, The Journal of Biological Chemistry 270 (9) (1995) 4526–4533.
[23] M. Eberhard, P. Erne, Calcium and magnesium binding to rat parvalbumin, European Journal of Biochemistry 222 (1994) 21–26.
[24] W. Yang, H. Lee, H. Hellinga, J. Yang, Structural analysis, identification, and design of calcium-binding sites in proteins, Proteins 47 (3) (2002) 344–356.
[25] B. Schwaller, The continuing disappearance of "pure" Ca2+ buffers, Cellular and Molecular Life Sciences 66 (2) (2009) 275–300.
[26] S. Fischer, K. Olsen, K. Nam, M. Karplus, Unsuspected pathway of the allosteric transition in hemoglobin, PNAS 108 (14) (2011) 5608–5613.
[27] C. Bode, Network analysis of protein dynamics, FEBS Letters 581 (2007) 2776–2782.
[28] C. Cantor, P. Schimmel, The behaviour of biological macromolecules, WH Freeman and Company, New York, 1971.

[29] C. Tanford, Physical Chemistry of Macromolecules, John Wiley and Sons, New York-London, 1961.
[30] K. Dill, K. Fiebig, H. Chan, Cooperativity in protein-folding kinetics, PNAS 90 (1993) 1942–1946.
[31] D. Hamada, T. Tanaka, G. Tartaglia, A. Pawar, M. Vendruscolo, M. Kawamur, A. Tamura, N. Tanaka, C. Dobson, Competition between folding, native-state dimerisation and amyloid aggregation in beta-lactoglobulin, Journal of Molecular Biology 386 (2009) 878–890.
[32] E. Di Cera (Ed.), Thermodynamics in Biology, Oxford University Press, 2000.
[33] A. Hill, The combinations of haemoglobin with oxygen and with carbon monoxide, The Journal of Physiology 40 (1910) 4–7.
[34] J. Monod, J. Wyman, J. Changeux, On the nature of allosteric transitions: a plausible model, Journal of Molecular Biology 12 (1965) 88–118.
[35] D. Koshland, G. Nemethy, D. Filmer, Comparison of experimental binding data and theoretical models in proteins containing subunits, Biochemistry 5 (1966) 365–385.
[36] C. Tsai, A. del Sol, R. Nussinov, Allostery: absence of a change in shape does not imply that allostery is not at play, Journal of Molecular Biology 378 (1) (2008) 1–11.
[37] N. Goodey, S. Benkovic, Allosteric regulation and catalysis emerge via a common route, Nature Chemical Biology 4 (8) (2008) 474–482.
[38] M. Lee, J. Tsai, D. Baker, P. Kollman, Molecular dynamics in the endgame of protein structure prediction, Journal of Molecular Biology 313 (2001) 417–430.
[39] J. Wand, Dynamic activation of protein function: a view emerging from NMR spectroscopy, Nature Structural & Molecular Biology 8 (2001) 926–931.
[40] P. Steinbach, B. Brooks, Protein hydration elucidated by molecular dynamics simulation, PNAS 90 (1993) 9135–9139.
[41] S. Mitternacht, I.N. Berezovsky, Coherent conformational degrees of freedom as a structural basis for allosteric communication, PLoS Computational Biology 7 (12) (2011) e1002301.
[42] Z.N. Gerek, S.B. Ozkan, Change in allosteric network affects binding affinities of pdz domains: analysis through perturbation response scanning, PLoS Computational Biology 7 (10) (2011) e1002154.
[43] M. Bhattacharyya, A. Ghosh, P. Hansia, S. Vishveshwara, Allostery and conformational free energy changes in human tryptophanyl-trna synthetase from essential dynamics and structure networks, Proteins Structure & Function Bioinformatics 78 (3) (2010) 506–517.
[44] M. Bhattacharyya, S. Vishveshwara, Probing the allosteric mechanism in pyrrolysyl-trna synthetase using energy-weighted network formalism, Biochemistry 50 (28) (2011) 6225–6236.
[45] D.K. Keunwan Park, Modeling allosteric signal propagation using protein structure networks, BMC Bioinformatics 12 (Suppl. 1) (2011) 1471–2105.
[46] A. Kloczkowski, R. Jernigan, Z. Wu, G. Song, L. Yang, A. Kolinski, P. Pokarowski, Distance matrix-based approach to protein structure prediction, Journal of Structural and Functional Genomics 10 (1) (2009) 67–81.
[47] S. Lockless, R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families, Science 286 (5438) (1999) 295–299.
[48] G. Suel, S.W. Lockless, M. Wail, R. Ranganathan, Evolutionarily conserved networks of residues mediate allosteric communication in proteins, Nature Structural & Molecular Biology 10 (1) (2002) 59–69.
[49] A. del Sol, H. Fujihashi, D. Amoros, R. Nussinov, Residues crucial for maintaining short paths in network communication mediate signaling in proteins, Molecular Systems Biology 2 (2006) 0019.
[50] R. Nussinov, C. Tsai, P. Csermely, Allo-network drugs: harnessing allostery in cellular networks, Trends in Pharmacological Sciences 32 (2011) 686–693.
[51] A. Dunker, I. Silman, V. Uversky, J. Sussman, Function and structure of inherently disordered proteins, Current Opinion in Structural Biology 18 (2008) 756–764.
[52] V. Uversky, Protein folding revisited. a polypeptide chain at the folding–misfolding–nonfolding cross-roads: which way to go? Cellular and Molecular Life Sciences 60 (9) (2003) 1852–1871.

Send Orders of Reprints at reprints@benthamscience.net
Advances in Protein and Peptide Sciences, Vol. 1, **2013, 000-000**

CHAPTER 6

Proteins as Netwoks: Usefulness of Graph Theory in Protein Science

Alessandro Giuliani[1,*], Luisa Di Paola[2], Paola Paci[3], Micol De Ruvo[2], Caterina Arcangeli[4], Daniele Santoni[5] and Massimo Celino[4]

[1]Department of Environment and Health, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy; [2]University Campus Biomedico, via Alvaro del Portillo 21, 00128 Rome, Italy; [3]CNR-Institute of Systems Analysis and Computer Science (IASI), BioMathLab, viale Manzoni 30, 00185 Rome, Italy; [4]Ente per le Nuove Tecnologie, l'Energia e l'Ambiente, ENEA, C.R. Casaccia, CP 2400, I-00100 Roma, Italy and [5]CNR-Institute of Systems Analysis and Computer Science (IASI), viale Manzoni 30, 00185 Rome, Italy.

Abstract: The consideration of protein molecules as graphs whose nodes and edges are respectively the aminoacid residues and the non covalent contacts between them permits to develop very effective models of protein behavior. The network paradigm not only allows for a drastic reduction of the amount of information needed to represent a protein 3D structure, but provides a set of quantitative descriptors derived from mathematical graph theory endowed with crucial information as for protein biological role. Here we will briefly comment about the relevance of graph descriptors in registering the relevant features of the molecular dynamics of solvation process and sketch the possibility to consider contact map as the 'macromolecular' analogue of small organic molecules structural formula.

Keywords: Protein Contact Maps, Systems Biology, Chemo-informatics, Computational Biology, Molecular Dynamics.

1. INTRODUCTION

In our 2008 *CPPS* paper entitled '*Proteins as Networks: Usefulness of Graph theory in protein science*' [1], we recognized, as the primary cause of the appropriateness of graph theory for protein science, the presence of a natural

***Address correspondence to Alessandro Giuliani:** Department of Environment and Health, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy; Tel: ++39 0649902579; E-mail: alessandro.giuliani@iss.it

Ben M. Dunn (Ed)

'basic discretization level' constituted by the different amino acid residues and the fact these discrete elements were unequivocally ordered along the primary structure. These two facts have as a consequence that any possible view of a protein molecule made by N residues, can be traced back to an NxN adjacency (or equivalency correlation) matrix A in which the generic $a_{i,j}$ element corresponds to a value assigned to the interaction between i and j residues.

This interaction can be the covariance of the i and j trajectories in time (molecular dynamics view), the similarity as for a given chemico-physical property (Hydrophobicity patterning as measured by RQA), or the Euclidean distance between the two residues in the X-ray crystal structure (Tertiary structure view). As we remarked in the paper '. this allows us to approach the studied system in an unbiased way avoiding unjustified or arbitrary assumptions.'.

Since 2008 we witnessed a huge flourishing of scientific works adopting the 'protein as network' view that in some way fulfilled our predictions (to be honest a very easy to accomplish prediction, given four years ago many 'proteins as networks' papers were already published) [5-,7, 9]. In this short introductory review, we would like to go a step further with respect to the (already well established) convergence between graph theoretical descriptions and structural (dynamical, chemico-physical) features of protein molecules. Our focus will be mainly directed toward the 'hybridization' of network-based descriptions at different levels of definition of the protein systems so as to sketch a possible way to generate mesoscopic principles of protein folding and physiology in a mainly data driven way.

In the meanwhile, the concept of folding, that is a central pillar in protein science, has been largely revised, also thanks to the introduction of topology as a key determinant in the protein folding dynamics and biomolecular structure stability [10-14].

Here we report a brief survey of these updating notes, with a special regard to the application of the network paradigm [15-17] to the protein folding process. To complete the picture, we are reporting some other applications of the protein contact network formalism to identify key features of the protein molecules,

spanning from the identification of functional domains to the structural properties of the protein molecules linked to their stability and function in the biological environment. This new field has been conquering a growing interest in the protein science community, due also to the simple formalism able to catch the key features of the protein function, out of a large set of data.

2. PROTEIN FOLDING, SOLVATION AND TOPOLOGY

The theory of folding is an outstanding case of application of general and simple energetic rules to describe the self-organization of complex structures able to interact in an ordered regulatory fashion with the environment, that being a main feature of the biological world [18] In other words, the biological sphere distinguishes from the non-biological reality since the general laws pointing to random motions and uniform properties distribution are not complied by the biological systems that act - apparently - against the natural tendency of entropy (disorder) increase described by the thermodynamics laws.

In this sense, in the last years, strong evidence supporting the topology determinant role has been enlightened [2- 8, 19, 21–23]; specifically, the definition of a contact order [20] defines the major role played by with respect to short range inter-residue interactions in the folding process. Long Range Order (LRO = the number of spatial contacts intervening between residues located more than approximately 12 residues apart along the sequence) has been demonstrated to be strongly correlated to the folding rates [11–14].

Thus, the folding theory has been enriched by this strong "topological flavour", able to catch some key features, such as the folding rates, predicted on the basis of the final topology of the intra-molecular interaction networks. It is worth noting this strict dependence of the folding process from the end-product topology, while completely consistent with engineering based control theory, implies the acquisition of a top-down causation style deeply in contrast with the more usual pure bottom-up approach of molecular biology. This promises to represent a general paradigm shift promoted by the development of network-based approaches in systems biology: the perhaps most evident proof is the dependence of the essential character of gene mutations in yeast by the location of the affected

enzyme in the whole metabolic network. By considering the yeast metabolic network as a graph having different metabolites as nodes and the enzymes catalyzing the transformation of metabolite i into metabolite j as arcs, it is immediate to note [15] that only the inactivation of enzymes interrupting the communication between i and j metabolites in the absence of alternative pathway are lethal for the entire organism. This is a straightforward demonstration of how the 'essentiality' feature of a given enzyme (local element) descends from the general wiring of the entire network so pointing to a clear top-down causation.

In our opinion, the conciliation between the protein folding dynamics (and kinetics) and the identification of structural and functional modules in protein structures needs to pass through the analysis of the dynamic networks [3], that are able to detect the concerted motions of protein regions, that are involved in minor conformational changes (allosteric transitions and solvation).

The concept has been further extended by the use of LRO [13] that, as introduced above, can be considered as an average contact order, computed including only the spatial contacts between residues whose distance along the sequence exceeds a given threshold. It has been found that when this threshold is set at 12 units in sequence, the corresponding LRO shows a strong negative correlation with protein folding rates. This result complies with the identification of structural and functional modules of 12 and 25 residues as key short range clusters responsible of folding [10, 18]); at last, this corresponds surprisingly to the same picture of our previous paper, of a critical cluster dimension in folding is 12 (6+6) residues.

Protein folding dynamics and energetics are strictly linked to the solvation effects, due to the hydrophobic interactions driving the three-dimensional structure formation through funnel pathways, characterizing folding kinetics [4].

Preliminary data from our group point to a strong correlation between topological descriptors fluctuation and typical structural parameters, employed in the analysis of molecular dynamics of protein solvation; we applied the method to the protein of hemoglobin, in the early transient phase of protein solvation as soon as the protein structure is put into the aqueous environment.

We report results that are based on the application of descriptive multidimensional statistics to the dynamics of structural and topological descriptors.

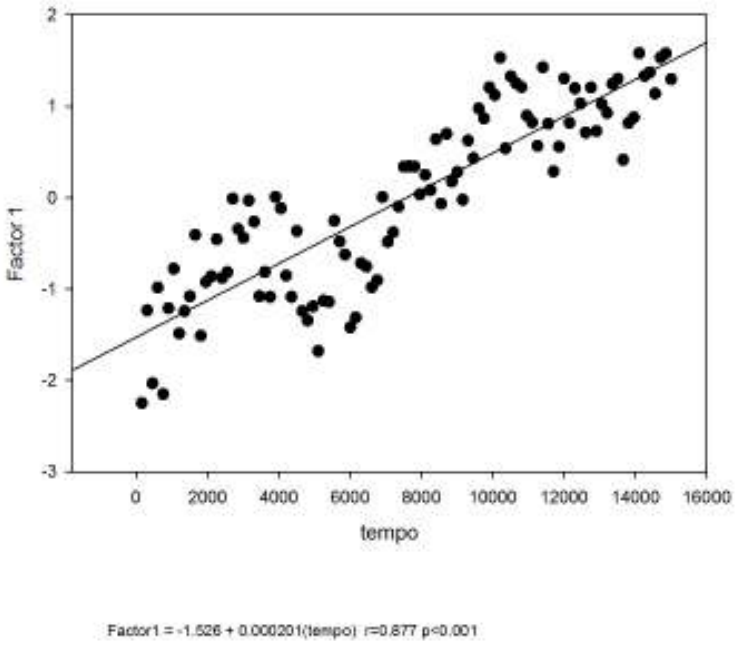


Figure 1: The main order parameter of topological/structural variation (Factor 1) shows a linear correlation with simulation time, topology strictly follows the progression of solvation process.

The dynamics has been extended over a time range 16000 ps wide that catch the first phase of the protein solvation, starting from the crystal structure, fed to the MD simulator (in this case, GROMACS) as PDB file, directly obtained from the Brookhaven depository.

We applied a method discussed thoroughly in [7], computing simultaneously topological and structural parameters of the different structures (hemoglobin conformations in this case) and then catching their mutual correlation structures in terms of emergent correlated variation.

The fluctuations of almost all the considered topological and structural parameters during solvation is constrained into very small coefficients of variation (CV

around 1%) with the only exception of purely structural descriptors as void fraction and shape index showing a CV around 20% as expected by the simple fact we are dealing with relatively small modifications of the same protein molecule (in [7] the CVs of the same parameters computed over a random set of different proteins had CVs around 1000%). From purely statistical considerations, we should expect that for the action of range restriction effect [16], the mutual correlation between the different descriptors in such low variance data set should be negligible. In contrast, when the data were analyzed by means of Principal Component Analysis (PCA), the emergence of a strongly between descriptors correlation structure was observed with a first principal component (Factor 1) explaining the 30% of the information carried by the original 14 descriptors. Still more important, Factor 1 was demonstrated to strictly follow the simulation time (see Fig. **1** where the vector points represent 100 consecutive snapshot configurations sampling the simulation period), being a strong proof of the sensitivity of topological descriptors to relatively small configuration changes.

These are very preliminary data but it is worth noting how average shortest path (*asp*) is the most relevant topological descriptor to link PCNs to global protein behavior. The same native relation of *asp* with protein global properties we observed at this very tiny scale of variation was observed at the very large scale of an heterogeneous set made of 911 proteins widely differing in both size and shape [7] so pointing to asp as a finely tuned (evolutionarily optimized?) parameter of protein structures.

3. FUNCTIONALITY OF PROTEINS IS BASED ON THE TOPOLOGY OF RESIDUE CONTACT NETWORKS

The fact that the adjacency matrix of a protein contact network (PCN) conveys all the information needed to reconstruct the entire 3D information of the protein molecule can hardly be overestimated. On a purely information theory point of view, the PCN has a dramatically reduced burden of information: each residue is considered as concentrated on the location of its alpha-carbon, the actual distances between residues are totally ignored maintaining only the binary code attaching a 1 to distances inside the [4-8 angstrom] interval and 0 elsewhere. If, despite this reduction, we can recover back the entire structure, this means PCNs can be

considered as a sort of 'structural formulas' for protein macromolecules. This proposal appears relevant in the analysis of allostery character of ligand-binding proteins; in a recent paper of our group [6], we analysed the topological properties at both global and local (active site) level; in the same work, we also applied a clusterization methodology giving rise to the so called 'dentist's chair' (*i.e.*, the peculiar disposition PCNs nodes in terms of Intra-Cluster and Between-Cluster Connectivity, see Fig. **7** of the previous paper), with the specific purpose of highlighting the specific topological role played by the residues of the active site, so to verify whether their topological role may experience some kind of variation upon binding (and consequent general conformational change, in allosteric structures). This approach revealed a peculiar "topological signature" for residues in the active site for hemoglobin, consistent with the different functionality of structures when comparing physiological and pathological forms [9].

When observed at the whole protein scale, the wiring properties of the apo- and holo- molecular forms (corresponding to the molecule free and linked to the ligand, respectively) are surprisingly similar, and this applies for both allosteric and non-allosteric proteins (such as the serum albumin); the same result is reflected into the reduced RMSD variations that we observed for the apo- and holo- forms for each analysed protein structure.

On the other hand, the clusterization results point to a clear discrimination between allosteric and non-allosteric proteins when only active site residues are considered: the allosteric character emerged as the change of availability of active site residues to establish inter-cluster links upon binding, passing from an high inter-cluster connectivity proneness to zero. Thus, the 'dentist chair' invariant wiring distribution of residues in the space spanned by the Intra-Cluster (k) and Between-Cluster (P) axes that in the 2008 was little more than a statistical curiosity, here finds a biochemical counterpart in terms of allosteric effective motions of proteins. The interruption of Between-Cluster communication has an immediate counterpart in terms of *asp* (at least for those paths originating from active site), so again suggesting its central role in topology-structure-function relations.

The strict link between *asp* and the ability of the protein molecules to transmit signals throughout the whole structure is probably at the very heart of the

importance of this topological parameter constituting a natural 'bridge' between function and structure. In this sense, the very peculiar properties of the protein structures are reflected in the singularity of the corresponding contact networks, which can be easily interpreted in terms of smart and optimized interaction graphs that keep the necessary stability and versatility to perform their own biological activity.

In this sense, the concept of allostery has been recently revised [17, 24–26], coupling the ligand-binding and conformational transition into a brand new scenario, based on the conformation distribution shift towards more favorable states upon environment stimuli. This smart adaptation to the requirements of the biological activity is easily understood through the network logic in terms of "creative elements", able to react and transmit signals and, thus, modulate the entire structural response of the protein [5]. The parallelism with social networks, often invoked, thus traces back to a common building role for life networks that may inspire even a unified theory of living dynamics, based on the complex networks formalism. Protein contact networks are small bricks at the molecular scale, but we're still fascinated by the idea that understanding the molecular interactions may help to trace a route to unravel as well larger scale networks.

ACKNOWLEDGMENTS AND CONFLICT OF INTEREST

The authors were partially funded by project BioGlue of Lazio Region, PI prof. Roberto Setola of University Campus-Biomedico. The authors declare they have no conflict of interests related to this chapter.

REFERENCES

[1] Krishnan, A; Zbilut, J.P.; Tomita M.; Giuliani, A. Proteins as networks: usefulness of graph theory in protein science. Current Protein and Peptide Sci. **2008**, *9* (1): 28-38.

[2] Alm, E; Baker, D. Matching theory and experiment in protein folding. *Curr*. *Opin*. *Struct*. *Biol*., **1999**, 189-196.

[3] Bode, C.; Kovács, I.A.; Szalay M.S.; Palotai R.; Korcsmáros T.; Csermely P. Network analysis of protein dynamics. *FEBS*. *Lett*., **2007**,*581*(15): 2776–2782.

[4] Cheung, M.S.; Garcia, A.E.; Onuchic, J.N. Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc*.*Natl*. *Acad*.*Sci*. *USA*., **2002**, 99(2): 685–690.

[5] Csermely P. Creative elements: network-based predictions of active centres in proteins, cellular and social networks. *Trends*. *Biochem*. *Sci*;**2008**, 33*:* 569–576.

[6] De Ruvo, M.; Giuliani, A.; Paci, P.; Santoni D.; Di Paola, L. Shedding light on protein-ligand binding by graph theory: The topological nature of allostery. *Biophys. Chem.*,**2012**,(165-166): 21–29.

[7] Di Paola, L.; Paci, P.; Santoni, D.; De Ruvo,M.; Giuliani, A. Proteins as sponges: A statistical journey along protein structure organization principles. *J. Chem. Inf. Model.*, **2012**, 52(2):474–482.

[8] Dokholyan N.V.; Li, L.; Ding, F.; Shakhnovich, E.I. Topological determinants of protein folding. *Proc.Natl.Acad.Sci. USA.,* **2002**, 99(13):8637–8641.

[9] Giuliani, A.; Di Paola, L.; Setola, R. Proteins as networks: a mesoscopic approach using haemoglobin molecule as case study. *Curr. Proteomics*., **2009**, 6(4): 235–245.

[10] Grantcharova, V.P.; Riddle, D.S.; Santiago, J.V.; Baker D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src sh3 domain. *Nat. Struct. Biol.,* **1998**, 5: 714–720.

[11] Gromiha, M.M. Importance of native-state topology for determining the folding rate of two-state proteins. *J Chem. Inf. Comput. Sci*., **2003**,43:1481–1485.

[12] Gromiha, M.M. Multiple contact network is a key determinant to protein folding rates. *J. Chem. Inf. Model*., **2009**, *49*,1130–1135.

[13] Gromiha, M.M.; Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *J. Mol. Biol*., **2001**, 310:27–32.

[14] Gromiha, M.M.; Thangakani A.M.; Selvaraj S. Fold-rate: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res., **2006**, 34:W70–W74.

[15] Palumbo, M.C.; Colosimo, A.; Giuliani, A.; Farina, L. Functional Essentiality from Topology features in metabolic networks: a case study in yeast. *FEBS Letter.,* **2005**, 579:4642-4646.

[16] Bobko, P.; Roth, P.L.; Bobko, C. Correcting the Effect Size of d for Range Restriction and Unreliability. *Organizational Research Methods*., **2001**, 4:46-61.

[17] Gunasekaran, K.; Ma, B.; Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins*., **2004**, 57: 433–443.

[18] Itzhaki, L.S.; Otzen, D.E.; Fersht A.R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol*., **1995**, 254: 260-288.

[19] Jung, J.; Buglass, A.J.; Lee, E. Topological quantities determining the folding/unfolding rate of two-state folding proteins. *J. Solution. Chem*., **2010**, 39: 943–958.

[20] Plaxco, K.W.; Simons, K.T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol*., **1998**, 277:985–994.

[21] Portman, J.J. Cooperativity and protein folding rates. *Curr. Opin. Struct. Biol.,* **2010**, 20: 11–15.

[22] Sosnick, T.R. Kinetic barriers and the role of topology in protein and RNA folding. *Protein Sci.,* **2008**,17: 1308–1318.

[23] Sosnick, T.R. The folding of single domain proteins – have we reached a consensus? *Curr. Opin. Struct. Biol*., **2011**, 21(1):12–24.

[24] Tsai, C.J.; del Sol, A.; Nussinov, R. Allostery: Absence of a change in shape does not imply that allostery is not at play. *J Mol. Biol*., **2008**, 378(1): 1–11.

[25] Tsai, C.J.; del Sol, A.; Nussinov, R. Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Mol. Biosyst*., **2009**, *5*:207–216.

[26] Vendruscolo, M. The statistical theory of allostery. *Nature. Chem. Biol*., **2011**, 7: 411–412.

CHEMICAL REVIEWS

Review

pubs.acs.org/CR

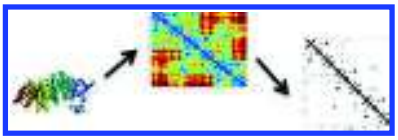
Protein Contact Networks: An Emerging Paradigm in Chemistry

L. Di Paola,[†] M. De Ruvo,[‡] P. Paci,[‡] D. Santoni,[§] and A. Giuliani*[,∥]

[†]Faculty of Engineering, Università CAMPUS BioMedico, Via A. del Portillo, 21, 00128 Roma, Italy

[‡]BioMathLab, [§]CNR-Institute of Systems Analysis and Computer Science (IASI), viale Manzoni 30, 00185 Roma, Italy

[∥]Environment and Health Department, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161, Roma, Italy



CONTENTS

1. INTRODUCTION

Topology is at the very heart of chemistry. This stems from the fact that chemical thought, since its prescientific alchemic origins, focused on the mutual relations between different entities expressed in terms of natural numbers instead of continuous quantities. This is the case in the concept of valence (e.g., atomic species A combines with atomic species B in the ratio 1:2 or 2:3) as well as of the periodic table, in which the discrete character of the atoms is implicit in the very same structure of a two-entry (period and group) matrix.[1] Chemical "primitives" are thus very often relational concepts that are naturally translated into the most widespread topological object of the whole science: the structural formula having atomic species as nodes and covalent bonds as edges connecting them. Structural formulas constitute an extremely efficient symbolic language carrying a very peculiar idea of what a structure is. While in physics structures are generally considered as consequences of a force field shaping a continuous space, so that the emerging structures are simply "energetically allowed" configurations in this mainly continuous space, chemistry assigns to a given structure an autonomous meaning by itself and not only as a consequence of an external force field.

The molecular graph (structural formula) relative to a given organic molecule is a condensate of the knowledge relative to that molecule: no other "scientific language" has an information storage and retrieval efficiency comparable to structural formulas. As a matter of fact, they can be used as the sole input for the computation of thousands of chemicophysical descriptors ranging from quantum chemistry to "bulk" properties, like melting point or partition coefficients,[2] and the knowledge of structural formula alone is, in many cases, sufficient to predict the interaction of the molecule with biological systems.[3] Descriptors based on bidimensional molecular graphs were demonstrated to outperform on many occasions, as in the prediction of receptor binding, sophisticated three-dimensional models, thus giving another proof of the unique role played by pure topology in chemistry.[4] Thus, chemical scholars could safely (and proudly) consider the recent surge of interest in graph-theoretical and, in general, network-based approaches in both physics and biology as nothing particularly novel for them.

Chemistry has already exploited graph theory methods: on the molecular scale, the chemical graph theory[5,6] has been harnessing the topological sketch of molecules into nodes (atoms) and links (chemical bonds) to derive mathematical descriptors of molecular structures, trying to delineate an ontology of molecules and predict their properties, on the sole basis of the molecular graph wiring. This method has been applied to derive the chemicophysical properties of alkanes, similarly to other methods that rely on the properties prevision from a group contribution application (UNIFAC[7] and UNIQUAC[8]).

Biological chemistry, additionally, poses intriguing issues regarding the analysis of complex kinetic schemes, made up of several chemical reactions with nonlinear kinetic expression for the corresponding reaction rate (Michaelis–Menten kinetic rate for enzymatic reactions). In this framework, the classical analytical approach to derive the dynamics of reactive systems[9] is unsatisfactory, due to high computational and modelistic

Received: June 11, 2012

   dx.doi.org/10.1021/cr3002356 | Chem. Rev. XXXX, XXX, XXX–XXX

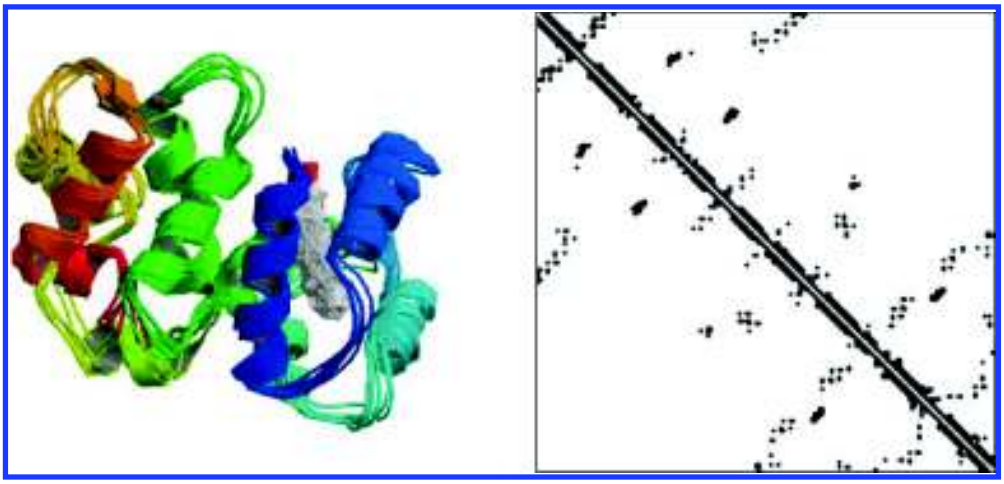


Figure 1. Recoverin 3D structure (left) and correspondent adjacency matrix.

burden required for a complete kinetic representation. Mathematics and chemistry meet on the common ground of the chemical reaction network theory (CNRT) that is explicitly aimed at analyzing complex biochemical reaction networks in terms of their topological emerging features.[10−13]

Nowadays, many different fields of investigation ranging from systems biology to electrical engineering, sociology, and statistical mechanics converge into the shared operational paradigm of complex network analysis.[14] A massive advancement in the elucidation of general behavior of network systems made possible the generation of brand new graph theoretical descriptors, at both single node and entire graph level, that could be useful in many fields of chemistry.

More specifically, in this review we will deal with the protein 3D structures in terms of contact networks between amino acid residues. This case allows for a straightforward formalization in topological terms: the role of nodes (residues) and edges (contacts) is devoid of any ambiguity and the introduction of van der Waals radii of amino acids allows us to assign a motivated threshold for assigning contacts and building the network.[15−17]

On the other hand, the Protein Data Bank (PDB) collects thousands of very reliable X-ray-resolved molecular structures, allowing scientists to perform sufficiently populated statistical enquiries to highlight relevant shared properties of protein structures or to go in-depth into specific themes (e.g., topological signatures of allostery), as well as to identify residues potentially crucial for activity and stability of proteins.

From a purely theoretical point of view, the reduction of a protein structure (that in its full rank corresponds to the three-dimensional coordinates relative to all the atoms of the molecule) to a binary contact matrix between the α-carbons of the residues represents a dramatic collapse. How many relevant properties of protein 3D structures (and consequently of possible consequences in terms of protein physiological role) are kept alive (and, hopefully, exalted by the filtering out of not relevant information) by the consideration of a protein as a contact network? How firmly based is the guess that adjacency matrices having as rows and columns amino acid residues (see Figure 1) could in the future play the same role the structural formula plays for organic chemistry? Relying on a single nonambiguous and physically motivated ordering of nodes (the primary structure) dramatically enlarges the realism of contact networks with respect to other kinds of networks (e.g., gene expression correlation networks) for which no such support is possible.

The inter-residue contact network has been yet largely explored in terms of inter-residue contacts frequencies under the quasichemical approximation;[18−20] as a matter of fact, in the seminal work of Miyazawa and Jernigan,[18] the amino acid hydrophobicity is assessed on the basis of the frequency of contacts of the corresponding residues as emerging from the analysis of a large number of structures.

In this way, residues involved more frequently in noncovalent interactions (mainly of hydrophobic nature, for hypothesis) are addressed to be of similar hydrophobic character. Application and confirmation of this view emerge from more recent works,[19,21] where the thermal stability of proteins belonging to thermophiles or psychrophiles has been inspected through the inter-residue interaction potential. The main result is that a characteristic distribution of inter-residues is able to provide the protein structure with the required flexibility to adapt to the environment.

The two above referenced works[19,21] are, in any case, only a statistical application over a huge number of proteins; what we really want to know is the character of information about a single and specific molecule that we can derive from its residue contact graph.

A very immediate example of this single molecule information is the fact that protein secondary structure can be reproduced with no errors on the sole basis of an adjacency matrix.[22] Similar considerations hold true for protein folding rate,[23−26] while normal-mode analysis confirmed that mean square displacement of highly contacted residues is substantially limited (nearly 20% of maximal movement range[27]). From another perspective, the presence of highly invariant patterns of graph descriptors shared by all the proteins, irrespective of their general shape and size, points to still unknown mesoscopic invariants (formally an analogue to valence considerations) on the very basis of protein-like behavior, irrespective for both fibrous and globular structures.[28,29] The scope of this review is, by briefly discussing some applications in this rapidly emerging field, to sketch an at least initial answer to the quest for a new "structural formula" language for proteins. This quest will be pursued in the following chapters by presenting side-by-side the different complex network invariants developed by graph theory and their protein counterparts.

dx.doi.org/10.1021/cr3002356 | *Chem. Rev.* XXXX, XXX, XXX−XXX

2. GRAPH THEORY AND PROTEIN CONTACT NETWORKS

2.1. Elements of Graph Theory

The classic Könisberg bridge problem introduced graph theory in 18th century. The problem had the following formulation: does there exist a walk crossing each of the seven bridges of Könisberg exactly once? The solution to this problem appeared in "Solutio Problematis ad geometriam situs pertinentis" in 1736 by Euler.[30] This was the first time a problem was codified in terms of nodes and edges linking nodes. This structure was called a graph.

A graph G is a mathematical object used to model complex structures and it is made of a finite set of vertices (or nodes) V and a collection of edges E connecting two vertices.

A graph $G = (V, E)$ can be represented as a plane figure by drawing a line between two nodes u and v and an edge $e = (u, v) \in E$ (Figure 2).

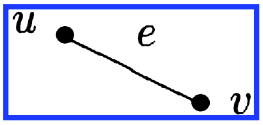


Figure 2. Example of an undirect graph comprising two nodes and an edge.

A graph $G = (V, E)$ can be represented by its adjacency matrix **A**; given an order of $V = \{v_1, v_2, \ldots v_n\}$, we define the generic element of the matrix $A_{i,j}$ as follows:

$$A_{ij} = \begin{cases} A_{i,j} = 1 & \text{if } (v_i, v_j) \in E \\ A_{i,j} = 0 & \text{otherwise} \end{cases}$$

The adjacency matrix of a graph is unique with respect to the chosen ordering of nodes. In the case of proteins, where the ordering of nodes (residues) corresponds to the residue sequence (primary structure), we can state that its corresponding network is unique. This is one extremely strong consequence that establishes a 1 to 1 correspondence between the molecule and its corresponding graph.

Let $v \in V$ be a vertex of a graph G; the neighborhood of v is the set $N(v) = \{u \in G \mid e(u,v) \in E\}$. Two vertices u and v are adjacent or neighbors, when $e = (u,v) \in E$ ($u \in N(v)$ or $v \in N(u)$). The degree k_i of the ith node is the number of its neighbors, defined on the basis of the adjacency matrix as

$$k_i = \sum_{j=1}^{N} A_{ij}$$

When $k_i = 0$, the ith node is said to be isolated in G, whereas if $k_i = 1$, it is said to be a leaf of the graph.

Information may be attached to edges, in this case we call the graph weighted and we refer to the weights as "costs". A weighted graph is defined as $G = (V, E, W)$, where W is a function assigning to each edge of the graph a weight:

$$W: E \rightarrow \mathbb{R}$$

The adjacency matrix **A** of a weighted graph is defined as follows:

$$A_{ij} = \begin{cases} A_{i,j} = w(v_i, v_j) & \text{if } (v_i, v_j) \in E \\ A_{i,j} = 0 & \text{otherwise} \end{cases}$$

The degree of a node in a weighted graph is defined as

$$k(v) = \sum_{i=1}^{N} w(u_i, v)$$

where $u_i \in N(v)$.

2.2. Protein Contact Networks (PCNs)

A protein structure is a complex three-dimensional object, formally defined by the coordinates in 3D space of its atoms.[31,32] Since the first works on the subject in the early 1960s,[33] a large number of protein molecular structures has been resolved, now accessible on devoted web databases.[34] The large availability of protein molecular structures has not solved yet many of the issues regarding the strict relationship between structure and function in the protein universe.

Thus, an emerging need in protein science is to define simple descriptors, able to describe each protein structure with few numerical variables, hopefully representative of the functionally relevant properties of the analyzed structure.

Protein structure and function rely on the complex network of inter-residue interactions that intervene in forming and keeping the molecular structure and in the protein biological activity.

Thus, the residues interactions are a good starting point to define the protein interaction network;[20,27,35] in this framework, the molecular structure needs to be translated into a simpler picture, cutting out the redundant information embedded in the complete spatial position of all atoms.

The most immediate choice is collapsing it into its α-carbon location (thereinafter indicated as C_α): correspondingly, the position of the entire amino acid in the sequence is collapsed into the corresponding C_α.

The spatial position of C_α is still reminiscent of the protein backbone; thus residues that are immediately close in sequence are separated by a length of 3−4 Å, corresponding to the peptide bond length[36] (see Figure 3); other α-carbons have a position that recalls the secondary domains and still reproduce, even in a very bare representation, the key features of the three-dimensional structure.

As soon as the complex protein structure architecture has been reduced to a simpler picture in terms of C_α position, the spatial topology can be further reduced to a contact topology that represents the network of inter-residue interactions, primarily responsible for the protein's three-dimensional structure and activity. Thus, the interaction topology is derived by the spatial distribution of residues in the crystal three-dimensional structure and represents the overall intramolecular potential.

Specifically, starting from the C_α spatial distribution, the distance matrix $\mathbf{d} = \{d_{ij}\}$ is computed, the generic element d_{ij} being the Euclidean distance in the 3D space between the ith and jth residues (holding the sequence order). The interaction topology is then computed on the basis of **d**: if the distance d_{ij} falls into a given spatial interval $\mathcal{I}$ (said cutoff), a link exists between the ith and the jth residues. The definition of the type of the graph (unweighted or weighted) is made in order to describe a given kind of interaction, in a more or less detailed fashion.

dx.doi.org/10.1021/cr3002356 | Chem. Rev. XXXX, XXX, XXX−XXX

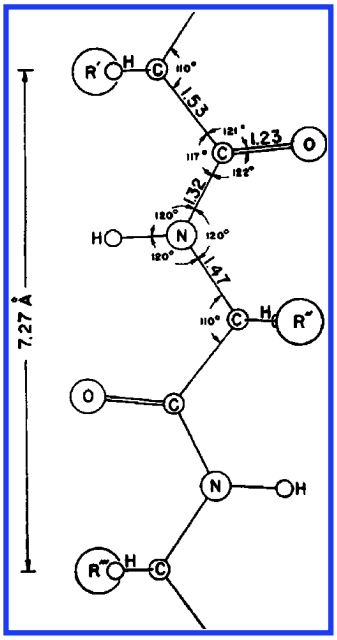


Figure 3. Geometry of the peptide bond: the upper threshold of 8 Å, commonly introduced in the analysis of PCNs, roughly corresponds to two peptide bond lengths.[36]

The choice of $\mathcal{I}$ determines the kind of interactions included in the analysis.[17,37] Most authors[15,16,38,39] consider only an upper threshold (around 8 Å) to cut off negligible interactions; some others, conversely, introduce also a lower limit, around 4 Å, that corresponds to the average value of the peptide bond length, so to eliminate the "noise" due to the "obliged" contacts coming from sequence proximity. In this way, only significant noncovalent interactions are included in the analysis, with the purpose of including only those interactions that may be modified upon slight environment changes, such in the case of biological response to environment stimuli.

Many authors use unweighted graphs to represent PCNs,[16,35,40−46] in order to infer several properties while keeping minimal information. On the other hand, some other groups propose a description for the PCNs using weighted graphs that is based on a side chain level reduction of the whole protein structure. In this case, all the information regarding the spatial position of atoms is kept and the single residue is represented by all of its atoms. Then, the distance matrix is computed over all protein atoms, which are labeled according to the residue they belong to. The strength of the interaction between two residues is measured as the number of their atoms whose distance lies within $\mathcal{I}$.[15,39,46−49]

Eventually, a straightforward way to establish a weighted protein contact network is to take the inverse of the distance among two residues as a direct measurement of their mutual interaction: the closer they lie, the stronger their mutual interaction.[50,51]

Another kind of representation is based on the same criterion but adopts as nodes the 20 different amino acids, which are combined through the peptide bond backbone in the protein primary structure. The link between two residues is represented by the number of links the residues of those types establish in the three-dimensional structure, according to the distance matrix **d** and the cutoff interval $\mathcal{I}$, as a rule. This method can be applied on an ensemble of protein structures,[43,52] in order to find a common rule of protein structure construction, in terms of more probable contacts between residues.

This representation, while keeping track of the nature of the interacting residues, destroys the one-to-one correspondence with the original 3D structure, given that different structures can give rise to the same representation in a way analogously to the structure isomerism in organic chemistry. Figure 4 reports the two kind of formulas.

The first emerging property of the PCNs is the degree of the corresponding graph, i.e., the average number of links each node (residue) establishes with neighbors. It is a direct measure of connectivity attitude of residues within the interaction network and it is strictly linked to the attitude of residue to establish noncovalent interactions with other residues.[28,38,40,46,49,54−56] The average degree, on the other hand, is a measure of the overall protein connectivity that is a rough index of the protein stability.

The contact density of a protein decreases exponentially with the number of residues; thus, bigger proteins are much less compact than smaller ones, giving rise to bigger cavities and a more fuzzy distinction between internal and external milieu.[57,58]

The degree distribution defines the graph model, allowing us to classify the network into already established network classes endowed with specific features (e.g., random graphs, scale-free

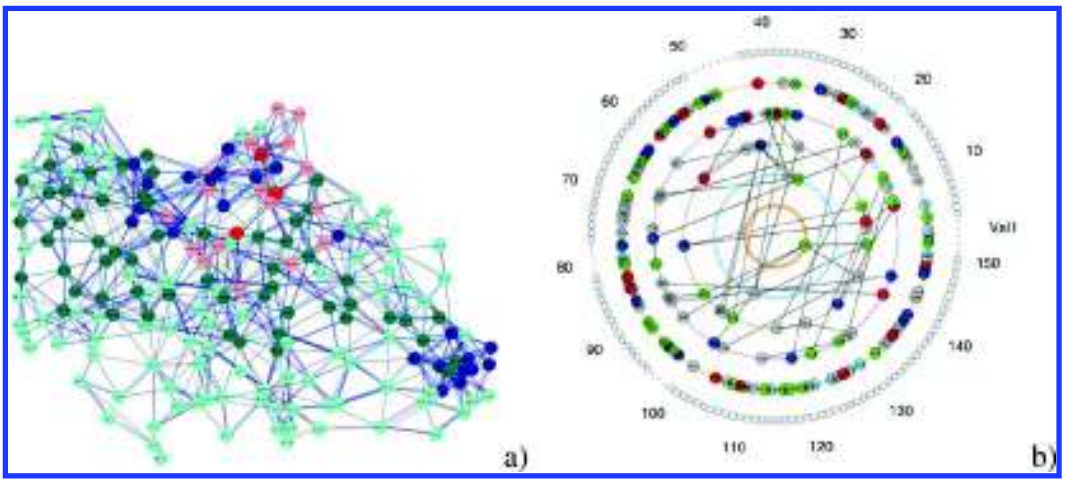


Figure 4. Graph protein formulas: (a) contact map[53] and (b) wheel diagram.[27]

dx.doi.org/10.1021/cr3002356 | Chem. Rev. XXXX, XXX, XXX−XXX

networks, regular lattices) as we will show in the next paragraphs.

2.3. Shortest Paths, Average Path Length, and Diameter

In a graph G, the distance $sp_{v,u}$ between any two vertices $v,u \in V$ is given by the length of the shortest path between the vertices, that is, the minimal number of edges that need to be crossed to travel from vertex v to vertex u. The shortest path between two vertices is not necessarily unique, since different paths may exist with identical length. In a graph, if no path exists connecting two nodes v, $u \in V$, we say that those nodes belong to different connected components; in such a case, we call the graph disconnected.

All PCNs are connected graphs at first glance. The so-called "percolation threshold" of a PCN can be estimated as the number of edges to be destroyed in order for the PCN to lose its connectivity.

This concept becomes relevant when we focus on long-range contacts (i.e., contacts between residues far away on the sequence,[23,59] which were demonstrated to be of crucial importance in protein folding rates,[23,25] as we will see in more detail below.

The diameter $\mathrm{diam}(G) = \max\{sp_{v,u} | v, u \in V_G\}$ of a graph is defined as the maximal distance of any pair of vertices. The average or characteristic length $l(G) = \langle sp_{v,u} \rangle$ is defined as the average distance between all pairs of vertices; the average inverse path length (efficiency) is defined as $\mathrm{eff}(G) = \langle 1/sp_{v,u} \rangle$; this descriptor is particularly suitable when components are disconnected (in this case, the contribution of infinite distances corresponds to zero efficiency).

The shortest path $sp_{v,u}$ between two residues of a PCNs represents a molecular shortcut that connect the residues through a mutual interaction pathway. In this sense, the smaller the $sp_{u,v}$, the tighter the relationship between the two nodes, which are strictly correlated, regardless of their distance in a sequence. These tight relations are thought to be responsible for the allosteric response in protein ligand binding[42,60−65] and in the concerted motions of distinct protein regions in protein dynamics.[64,66−71]

In general, still preliminary evidence from our group (work in progress) points to the average shortest path as the most crucial network invariant to link topology to both molecules' dynamics and the general thermodynamical properties of the protein molecules.

2.4. Clustering on Graphs

Identifying clusters on a network is a more complicated task than computing the average shortest path. The clustering coefficient measures the cliquishness of a typical neighborhood (a local property). One possible definition is the following:[29,30,40,72] let us define the clustering coefficient of the ith node C_i as

$$C_i = \frac{\text{the number of connected neighbor pairs}}{\frac{1}{2}k_i(k_i - 1)} \tag{2.1}$$

where k_i is the degree of the ith vertex; the average clustering coefficient C of the graph is the average of C_i values over all nodes.

For social networks, C_i and C have intuitive meanings: C_i reflects the extent to which friends of i are also friends of each other; thus, C measures the cliquishness of a typical friendship circle.

Another definition for C is[73,74]

$$C = \frac{3 \times (\text{the number of triangles on a graph})}{\text{the number of connected triples of vertices}} \tag{2.2}$$

where a "triangle" corresponds to three vertices that are each connected to each other and a "connected triple" means a vertex that is connected to an (unordered) pair of other vertices. The factor of 3 in the numerator accounts for the fact that each triangle contributes to three connected triples of vertices, one for each of its three vertices; thus, the value of C lies strictly in the range from zero to one.

With regard to PCNs, the clustering coefficient referred to the ith residue measures the triangles number insisting on it;[15,28,29,38,40,45−47,49,56,75] thus, high clustering coefficient nodes are central in communities with a large number of interconnecting links, corresponding to high local stability. In other words, we can infer that mutation producing depletion of such nodes may cause dramatic changes in the protein structure.[29]

2.4.1. Spectral Clustering. The spectral analysis of a graph allows one to identify clusters in the network by minimizing the value of parameter Z defined as[45]

$$Z = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 A_{ij}$$

where x_i and x_j represent the position of nodes i and j in the network and A_{ij} is the adjacency matrix. The minimum of Z corresponds to the second smallest eigenvalue of the Laplacian matrix $\mathbf{L}$ of $\{A_{ij}\}$, also known as the Kirchoff matrix, defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

where $\mathbf{D}$ is the degree matrix, which is a diagonal matrix in which $\{D_{ii}\} = k_i$. Once $\mathbf{L}$ eigenvalues λ are computed, the second smallest eigenvalue λ_2 corresponds to the minimum value of Z (the first one provides a trivial solution[45]). The components of the corresponding eigenvector v_2, known as the Fiedler eigenvector, refer to single nodes and define two clusters depending on the sign of each component. Nodes are parted into two clusters according to the sign of the corresponding component in v_2. This process can be iterated on both subnetworks until all the components of v_2 show the same sign.

Identification of clusters in PCNs has a strong impact on detecting structural and functional domains in proteins.[50,54,76,77] The presence of folding clusters is a key point in the molecular development of the funnel folding pathways theory, which provides the most reasonable molecular mechanism for protein folding, out of the random approaches of residues, in order to form the favorable inter-residue interaction network, providing stability to the tertiary structure.[78,79]

The reliable identification of clusters in the PCNs allows for the definition of descriptors at the single residue level, relying on the PCNs partition structure.

2.4.2. Intracluster and Extracluster Parameters. Once the clustering process is performed, two parameters, z_i and P_i, representing the modularity rate for each node,[80] can be computed. These two parameters are defined as

$$z_i = \frac{k_{is} - \bar{k}_{is}}{\sigma_{is}} \tag{2.3}$$

dx.doi.org/10.1021/cr3002356 | Chem. Rev. XXXX, XXX, XXX−XXX

$$P_i = 1 - \sum_{s=1}^{n_M} \left(\frac{k_{is}}{k_i} \right)^2 \quad (2.4)$$

where k_{is} is the number of links the ith node establishes with nodes belonging to its own cluster s_i; $\overline{k}_{si}$ is the average degree for nodes in cluster s_i, σ_{si} is the corresponding standard deviation, and n_M is the number of clusters to which the ith node belongs to. The spectral clustering performs a "crispy" partition, namely, clusters are disjoint sets of nodes, thus eq 2.4 becomes

$$P_i = 1 - \left(\frac{k_{is}}{k_i} \right)^2 \quad (2.5)$$

These parameters have been introduced to discriminate nodes according to their topological role in the so-called Guimerà and Amaral's cartography,[80] the aim of which is the classification of nodes in a modular network, relying on intra- and intermodule connectivities.[81]

In their seminal work,[80] Guimerà and Amaral demonstrated that the relative importance of each node in maintaining the global graph connectivity can be traced back to its location in the P, z plane.

Once the network is partitioned into a set of meaningful communities, it is possible to compute statistics for how connected each hub (a hub is a node having an extremely high degree of connectivity) is both within its own community and to other communities: hubs endowed with strong connections within functional modules were assumed to be interacting with their partners at once (party hubs); conversely, those with a low correlation were assumed to link together multiple modules (date hubs), playing a global role in the network. It is worth stressing that although both hub types have similar essentiality in the network, as the characteristic path length increases, deleting given hubs, the network begins to disintegrate, since hubs provide the coordination between functional modules. To make a comparison, party hubs should correspond to Guimerà "provincial hubs", which have many links within their module but few outside, whereas date hubs could be "nonhub connectors" or "connector hubs", both of which have links to several different modules; they could also fall into the "kinless" roles, since very few nodes are actually found in these categories.[82] Considering network motifs, it was observed that party hub network motifs control a local topological structure and stay together inside protein complexes, at a lower level of the network. On the other hand, date hub network motifs control the global topological structure and act as the connectors among signal pathways, at a high level of the network. Network motifs should not be merely considered as a connection pattern derived from topological structures but also as functional elements organizing the modules for biological processes.[67]

Spectral clustering of PCNs produces characteristic P–z diagrams, referred to as "dentist's chair", due to their shape.[28,29,58] This shape is strongly invariant with respect to the protein molecule, as shown in Figure 5; panel a refers to a typical diagram derived from the analysis of a single protein structure, while panel b shows the superposition of a structure analyses of 1420 proteins. The fact the general shape of the graph remains substantially invariant on going from one to 1420 proteins is an impressive proof of the robustness of the P, z organization of PCNs.

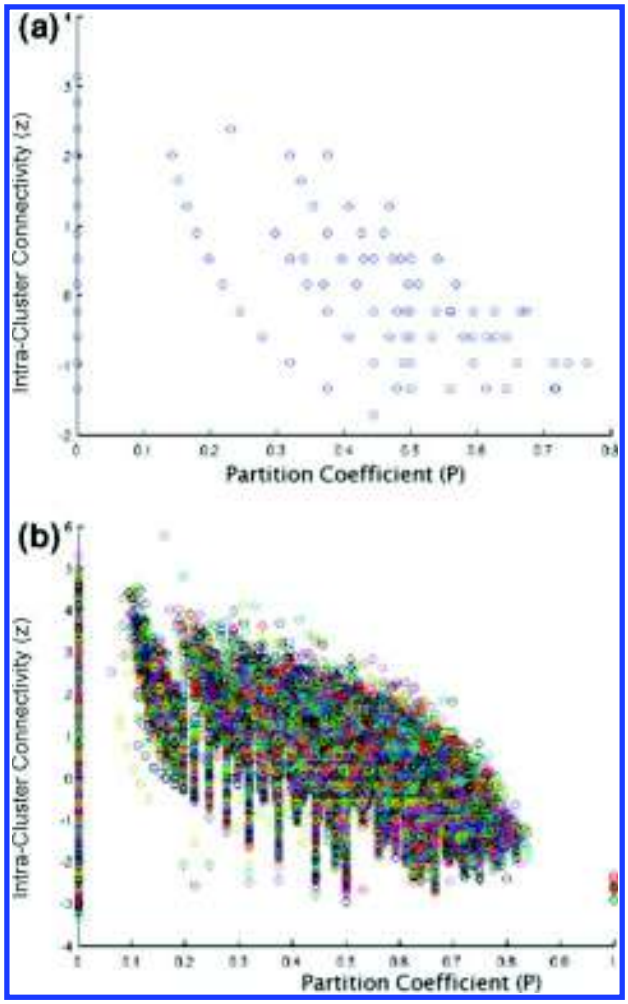


Figure 5. P vs z plot (dentist's chair) (a) for a single protein, where each point identifies a residue, and (b) the superposition of structures analysis of 1420 proteins[58].

The strong invariance of the P, z portraits of PCNs irrespective of both protein general shape and size is extremely intriguing, given it suggests the existence of still hidden mesoscopic principles of protein structures analogous to valence rules in general chemistry.

2.5. Network Centralities

The centrality of a node deals with its topological features in network wiring. The term "central" stems from the origin of this concept in the definition of key, central indeed, nodes of social networks: people, in other words, that are responsible for the stability and activity of the network. This "social science" origin of the concept of centrality was found to have a correlation in PCNs by Csermely.[83]

Centrality can be computed in different ways, using different weights to evaluate and compare the importance of a node (degree, clustering coefficient, for instance). They are almost equivalent definitions that point to the same attitude of central nodes, to establish strong local interactions, in their own local community, able to stabilize the whole network structure.

The role of central nodes in modifying the network structure according to their centrality values is the starting point to define the property of centrality–lethality,[81,84] which emerges as a key element in the analysis of biological networks, where central nodes represent a prerequisite for the organism survival: for

instance, a shortage or a depletion of a central protein in a protein−protein interaction network does lead to the death of the organism.[85]

Central nodes in PCNs correspond to residues crucial for both the protein structure folding and stability. Thus, the centrality of a node can be a measure of the biological consequences of its mutation; for instance, the highly detrimental mutation of hemoglobin that causes sickle cell anemia is due to just a substitution of one residue (glutamic acid is replaced in position 6 by valine) that produces dramatic changes in the protein structure and function. This effect is widely reflected in the high centrality value of this specific residue.[29]

The easiest and the most natural way to define centrality is provided by the so-called degree centrality that simply counts the number of connections for each node, its degree, i.e., the number of nodes it is directly connected with; in this case, the degree centrality of a node v_i corresponds to its degree. Hubs are, thus, the central nodes of a network, according to this paradigm.

2.5.1. Path-Based Centralities: Closeness and Between-ness. Closeness centrality, as well as between-ness, belong to the class of shortest path-based centrality measures. The closeness centrality provides information about how close a node is to all other nodes. The closeness of a node v_i is defined as

$$c(v_i) = \frac{1}{\sum_{j=1}^{n} \mathrm{sp}_{i,j}}$$

The closeness centrality is connected to the aptitude of a node to participate in the signal transmission throughout the protein structure. High closeness centrality nodes were demonstrated to correspond to residues located in the active site of ligand-binding proteins or to evolutionary conserved residues.[41,52,72,86,87]

It is worth noting that closeness centrality, at odds with degree centrality, it is not solely based on local features of the network but takes into account the location of the node in the global context of the network it is embedded into. In this respect, closeness, as well as between-ness, are genuine systemic properties that are computed at the single node level, thus establishing a "top-down" causative process. This is probably the reason for the efficiency of this kind of network invariants to single out relevant general properties of the proteins.

This is formally analogous to what happens in basic chemistry, where the properties (i.e., acidity, electronegativity) of the hydrogen atom in the CH_4 molecule are different from those of the hydrogen atoms in H_2O or H_2 molecules, because of the general molecular context they are embedded into.

This is the same philosophy of single node (residue) descriptors, implicitly taking into account the whole context and so overcoming a purely reductionist view.

Between-ness measures the ability of a vertex to monitor communication between other vertices; every vertex that is part of a shortest path between two other vertices can monitor and influence communication between them. In this view, a vertex is central if lots of shortest paths connecting any two other nodes cross it. Let $\sigma_{v,u}$ denote the number of shortest paths between two vertices v, $u \in V$ and let $\sigma_{v,u}(s)$, where $s \in V$, be the number of shortest paths between v and u crossing s; trivially $\sigma_{v,u} \geq \sigma_{v,u}(s)$.

$$\mathrm{betw}(s) = \sum_{v \in V, v \neq s} \sum_{u \in V, u \neq s} \frac{\sigma_{v,u}(s)}{\sigma_{v,u}}$$

In biological networks (e.g., protein−protein interaction network), the nodes with higher between-ness were demonstrated to be the main regulators.[84,88,89]

In PCNs, since the between-ness centrality is based on shortest paths, it comes immediately clear that this index is strongly linked to the centrality of nodes (residues) in terms of their capability to transfer signals throughout the protein molecule.[43,56,72,86] Thus, the depletion of residues having high between-ness centrality values is supposed to interrupt the allosteric communication among regions of the proteins that lie far apart.

2.6. Network Assortativity and Nodes Property Distribution

Newman suggested[90] that an important driving factor in the formation of communities was the preference of nodes to connect to other nodes that possess similar characteristics; he defined this behavior as assortativity. The concept of assortativity is a very general one, so in the case of protein structures, we could identify the "behavior" of different residues in terms of their hydrophobic/hydrophilic character so that an assortative structure will correspond to a network in which similar hydrophobicity residues will be preferentially in contact with each other compared to what is expected by pure chance. In this example, the "behavior" of the nodes corresponds to a feature (hydrophobicity) independent of the pure network wiring and can be equated to a "coloring" of the nodes, whose relations with the underlying topological support constituted by network wiring is investigated. Along similar lines, we could think of assortative social networks in which friends (nodes in direct contact) tends to share the same political ideas, income classes, or professional activities. On a different heading, we can think of assortativity as an "internal" description of network wiring in which nodes are defined in terms of their connection patterns. Actually, in some networks high-degree nodes preferentially connect to other high-degree nodes (assortative networks), whereas in other types of networks high-degree nodes connect to low-degree nodes (disassortative networks); in particular, numerical evidence from experimental data have shown that many biological networks exhibit a negative assortativity coefficient and are therefore claimed to be examples of disassortative mixing.[40,91]

Assortativity r is defined as the Pearson correlation coefficient of degrees at either ends of an edge, and it varies as $-1 \leq r \leq 1$;[92] r is a very simple measure of the probability of a high-degree node to form edges with other high-degree nodes. When the r value is close to 1, the network is addressed to as assortative, whereas values of r close to -1 are characteristic of disassortative networks. Random graphs are purely nonassortative networks, since by definition, links between nodes, in this case,

$$k(v) = \sum_{i=1}^{N} w(u_i, v)$$

are placed at random.

In the case of external "coloring" assortativity, the index r, instead of being computed on the nodes degree, can be computed over the feature of interest, the one used to "color" the nodes.

dx.doi.org/10.1021/cr3002356 | Chem. Rev. XXXX, XXX, XXX−XXX

Thus, in a recent work,[57] Di Paola et al. demonstrated the lack of any clearly defined "hydrophobic core" in proteins, for which the arrangement of fractal structures was demonstrated not to have a clear-cut separation between internal and external milieu by means of network assortativity measures based on the hydrophobicity of nodes. Moreover, the presence of both assortative and disassortative structuring (hydrophobic−hydrophobic and hydrophilic−hydrophobic) in proteins highlighted the presence of different "folding logic" contemporarily present in the protein world, probably as a consequence of the varying relevance of hydrophobic and electronic forces in the folding process.

Generally speaking, the distribution of a given feature of the nodes can be explored through the combined definition of diadicity and heterophilicity,[93] measuring the tendency of nodes with similar properties to form links. Given a key physical property, if nodes show an attitude to establish preferentially links with similar nodes, the network is named as dyadic, otherwise it is said to be antidyadic or heterophilic.[93]

Let n_1 and n_0 respectively denote the number of node possessing or not a specific property; e_{10} and e_{11} are the number of edges connecting homologous and heterelogous nodes, respectively. The heterophilicity score H is then defined as

$$H = \frac{e_{10}}{e_{10,r}} \tag{2.6}$$

where $e_{10,r}$ is the random value in case of uniform distribution of the property among nodes that depends on the number of possible edges $E = N(N-1)/2$, $N = n_1 + n_0$ beingthe number of nodes:

$$e_{10,r} = En_1(N - n_1) \tag{2.7}$$

Analogously, as for the homologous contacts, it is defined the dyadicity D as

$$D = \frac{e_{11}}{e_{11,r}} \tag{2.8}$$

and the corresponding value for random homologous nodes is

$$e_{11,r} = E\frac{n_1(n_1 - 1)}{2} \tag{2.9}$$

Thus, dyadic networks have D values larger than 1 and, on the other hand, H values lower than unity.

The above-described network invariants provide a description that can be traced back to the single node of a network, but the effective values of the descriptors strongly depend on the general wiring architecture of the whole graph, again a systemic top-down causation metric. The dyadic character of PCNs was exploited by Alves and colleagues[94] to define simple hydrophobicity scores to profile protein structure. Single residue hydrophobicity was demonstrated to be strongly correlated with the corresponding network invariants:[56] these systemic properties strictly depend upon the "general class" the specific graph pertains to. Below we will briefly present the main classes of wiring architectures.

2.7. Models of Graphs

2.7.1. Random Graphs. One of the simplest and oldest network models is the random graph model,[95] which was introduced by Solomonoff and Rapoport[96] and studied extensively by Erdös and Rènyi;[97−99] according to their works, there are two different random graph models.

One is called $G_{n,m}$ and is the set of all graphs consisting of n vertices and m edges, and it is built by throwing down m edges between vertex pairs chosen at random from n initially unconnected vertices.

The other is called $G_{n,p}$ and it is the set of all graphs consisting of n vertices, where each pair is connected together with independent probability p. In order to generate a graph sampled uniformly at random from the set $G_{n,p}$, initially unconnected vertices are taken and each pair of them is joined with an edge with probability p ($1 - p$ being the probability of being unconnected). Thus, the presence or absence of an edge between two vertices is independent of the presence or absence of any other edge, so that each edge may be considered to be present with independent probability p. The two models are essentially equivalent in the limit of a large number of nodes n. Since $G_{n,p}$ is somewhat simpler to work with than $G_{n,m}$, it is usual to refer to it as a random graph $G_{n,p}$.

A vertex in a random graph is connected with equal probability p to each of the $N - 1$ other vertices in the graph, and hence, the probability p_k that it has degree k is given by the binomial distribution

$$p_k = \binom{N}{k} p^k (1 - p)^{N-k} \tag{2.10}$$

Noting that the average degree of a vertex in the network is $z = (N - 1)p$, we can also write this as

$$p_k = \frac{(N-1)!}{k!(N-1-k)!}\frac{z^k}{(N-1)^k}\left(1 - \frac{z}{N-1}\right)^{N-k} \simeq \frac{z^k e^{-z}}{k!} \tag{2.11}$$

where the second equality gets exact as $N \to \infty$; in this case, p_k corresponds to the bell-shaped curve that peaks on the average value (Figure 6b).

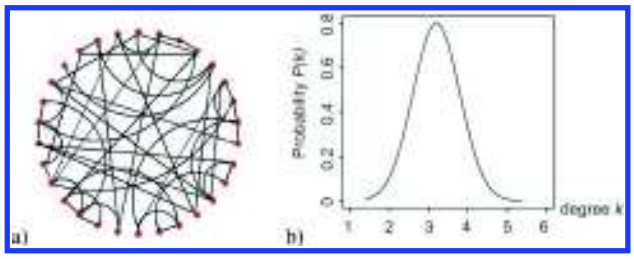


Figure 6. Random graph: (a) a sample picture, where most nodes have three or four links, and (b) the bell-shaped degree distribution.

Random graphs have been employed extensively as models of real-world networks of various types, particularly in epidemiology,[74] where the spreading of a disease through a community strongly depends on the pattern of contacts between infected subjects and those susceptible to it.

However, as a model of a real-world network, a random graph has some serious shortcomings. Perhaps the most serious one is its degree distribution, which is quite unlike those seen in most real-world networks.[92] On the other hand, the random graph has many desirable properties; specifically, many of its properties can be calculated exactly.[92]

The random graph model has been applied to PCNs to test their connectivity (degree) distribution.[48,75] Specifically, the protein dynamic properties have been explored in terms of

random graphs, since the unbiased corresponding network dynamics can be put into the perspective of the random evolution of the protein structure, due to random, Brownian motion of protein segments to get up to the final, stable conformation.[100]

Further, the generic random graph model is introduced as a reference to test the property of the network as a specialized random graph (small-world network).[15,35,38,40,41,43,47,49,72,101−103] This comparison has a tight link with the common assumption of the random coil structure as being a reference state in folding thermodynamics: the random coil has the corresponding translation in terms of a "graph formula" into the random graph model that represents a random network of residue interactions, corresponding to a random distribution of the inter-residue distance.

In their work,[104] Bartoli and colleagues demonstrated PCNs are very far from random graph behavior, this was particularly evident when they projected simulated networks together with real PCNs in the bidimensional space, spanned by the clustering coefficient and characteristic path length (see Figure 7).

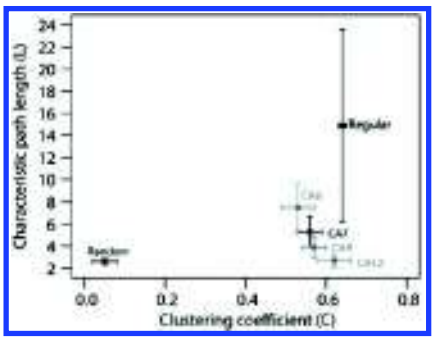


Figure 7. Characteristic path length vs Clustering Coefficient (Figure 3 in ref 104): sample protein classes are labeled as CA#, the label "random" refers to collection of random graphs, whereas "regular" points to periodic lattices.

The authors demonstrated the difference between random graph and contact maps derive from the existence of the covalent backbone, that imposes very strict constraints to the contact that can be established between residues. This feature makes PCNs to more similar to the so-called scale-free graphs.

2.7.2. Scale-Free Graphs. Since many years from the seminal work of Erdös and Rènyi,[97] all complex networks are treated commonly as random graphs. This paradigm was outdated by the pioneristic work of Barabàsi,[105] in which the topology of the World Wide Web was studied, formerly thought to show a bell-shaped degree distribution, as in the case of random graphs.

Instead, by counting how many Web pages have exactly k links the authors showed that the distribution followed a so-called power law, namely, the probability that any node is connected to k other nodes is

$$p_k = \alpha k^{-\gamma}$$

where γ is the degree exponent and α is the proportionality constant. The value of γ determines many properties of the system. The smaller the value of γ, the more important the role of the hubs is in the network. Whereas for $\gamma > 3$ the hubs are not relevant, for $2 < \gamma < 3$ there is a hierarchy of hubs, with the most connected hubs being in contact with a small fraction of all nodes, and for $\gamma = 2$, a hub network emerges, with the largest hubs being in contact with a large fraction of all nodes.[107] In general, the unusual properties of scale-free networks are valid only for $\gamma < 3$, such as a high degree of robustness against accidental node failures.[85] For $\gamma > 3$, however, most unusual features are absent, and in many respects, the scale-free network behaves like a random one.[85] As for the World Wide Web, Barabàsi[105] found that the value of γ for incoming links was approximately 2; this means that any node has roughly a probability 4 times bigger to have half the number of incoming links than another node.

Different from a Poisson degree distribution of random networks, a power law distribution does not have a peak, but it is described by a continuously decreasing function (Figure 8):

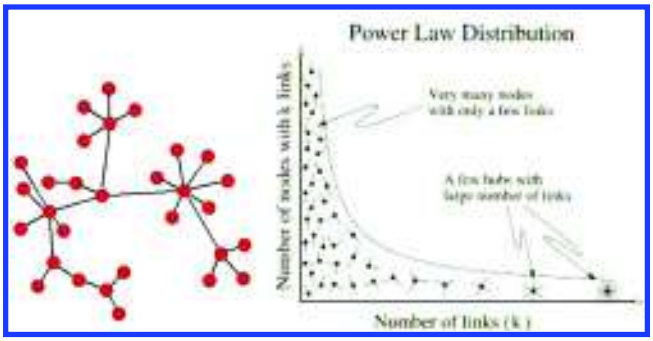


Figure 8. Scale-free networks: (a) a sample scale-free networks, in which few nodes have many links, and (b) the degree distribution of the scale-free graph power law.[106]

in this case, it is evident that a specific characteristic average degree does not exist; in other words, these networks do not converge toward a characteristic degree, at increasing number of nodes. On the contrary, in scale-free networks, the average degree progressively increases with sampling dimension, because the (very rare) high-degree nodes are sampled with a higher probability. The lack of a characteristic degree is on the basis of the denomination "scale free" for this kind of architecture.

This is in strong contrast to random networks, for which the degree of all nodes is in the vicinity of the average degree, which could be considered typical. However, as Barabàsi and colleagues wrote in,[107] scale-free networks could easily be called scale-rich as well, as their main feature is the coexistence of nodes of widely different degrees (scales), ranging from nodes with one or two links to major hubs.

In contrast to the democratic distribution of links typical of random networks, power laws describe systems in which few hubs dominate:[105] networks that are characterized by a power-law degree distribution are highly nonuniform, most of the nodes having only a few links. Only few nodes with a very large number of links, which are often called hubs, hold these nodes together.

A key feature of many complex systems is their robustness, which refers to the system's ability to respond to changes in the external conditions or internal organization while maintaining relatively normal behavior.[107] In a random network, disabling a substantial number of nodes will result in an inevitable functional disintegration of a network, breaking the network into isolated node clusters.[107]

Scale-free networks do not have a critical threshold for disintegration (percolation threshold[108]): they are amazingly

robust against accidental failures: even if 80% of randomly selected nodes fail, the remaining 20% still form a compact cluster with a path connecting any two nodes.[107] This is because random failure is likely to affect mainly the several small degree nodes, whose removal does not disrupt the networks integrity.[85] This reliance on hubs, on the other hand, induces a so-called attack vulnerability: the removal of a few key hubs splinters the system into small isolated node clusters.[85]

Scale-free architecture can exhibit the so-called "small word property".[38,104] The small word model has its roots in the observation that many real-world networks show the following two properties: (i) the small-world effect (i.e., small average shortest path length) and (ii) high clustering or transitivity, meaning that there is a heightened probability that two vertices will be connected directly to one another if they have another neighboring vertex in common.

The former property is quantified by the characteristic path length (or average shortest path) l of the graph, while the second property is computed as the clustering coefficient C. Thus, small-world effect means that the average shortest path in the network scales logarithmically with graph size[73,109,110]

$$l \propto \log(N)$$

where N is the number of nodes.

PCNs were analyzed as for their scale-free properties, in order to identify crucial binding sites.[43,59] The small-world behavior of protein structure networks was shown for the first time by Vendruscolo et al.[43] and later confirmed in several works.[38,75] As we stretched before, it was shown that small-world behavior of an inter-residue contact graph is conditioned by the backbone connectivity.[104]

According to both,[59,104] PCNs are not "pure small-world" networks, given that no explicit hub is present, so they must be considered as "a class of network in its own", generated by the very peculiar constraint to mantain a continuous (covalent) backbone joining the nodes in a fixed sequence.[59,104]

Nevertheless, the most important feature of small-world architecture, i.e., the presence of shortcuts allowing for an efficient signal transmission at long distance, is present in PCNs and it is the very basis of their physiological role (allostery, dynamical properties, folding rate, etc.) (Figure 9).

In this respect, it is relevant to go more in-depth into the link existing between a given topology and the dynamical behavior it can host. As a matter of fact, according to a pattern-based computational approach,[111] modular dynamic organization follows modular topological organization; this assumption has been applied to biological neural networks, showing that the dynamic behavior of neural networks might be coordinated through different topological features,[111] such as network modularity and the presence of central hub nodes. A similar topology/dynamics relation seems to hold for contact networks, too. As a matter of fact, allosteric "hot spots",[65] where the motion is generalized from a local excitation to the entire protein structure, correspond to central residue contacts, which were demonstrated to be crucial for efficient allosteric communications.[41,59,66,72,86]

The field of relations between molecular dynamics trajectories and topological contact network description is a very important avenue of research in protein science.[62−64,66,70]

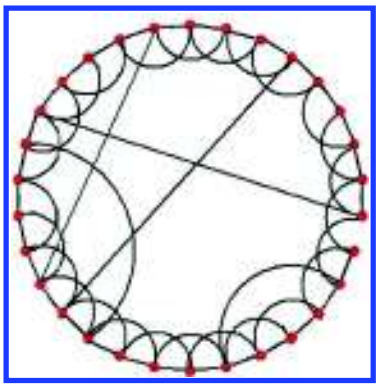


Figure 9. An example of a small-world network: most nodes are linked only to their immediate neighbors, while few edges generate shortcuts between distant regions of the network.

3. APPLICATIONS

3.1. Networks and Interactions

It is well-known that proteins interact among themselves and with other molecules to perform their biological functions;[69] crucial factors in all interactions are the shape and chemical properties of the pockets located on protein surfaces, which show high affinity to binding sites. In a recent work,[112] the analysis of topological properties of the pocket similarity network demonstrated that highly connected pockets (hubs) generate similar concavity patterns on different protein surfaces. These similarities go hand-in-hand with similar biological functions that imply similar pockets.[112] In addition, they found that maximum connected components in the pocket similarity networks have a small-world and scale-free scaling. The analysis of the physicochemical features of hub pockets leads to the investigation of more functional implications from the similarity network model, which provided new insights into structural genomics and have great potential for applications in functional genomics.[113] The future purpose is to develop a classification method to divide similar pockets into small groups and afterward to compile this evolutionary information into a library of functional templates.

This work delineates a possible link between network wiring and common function of utmost interest for the development of contact-based meaningful formulas. By briefly describing direct translation of graph theoretical descriptors into meaningful protein functional properties, we gave a proof-of-concept of the general relevance of the proposed formalism. Now we will go more in-depth into some of these topology−function relations, but a leading leitmotiv can be already stated: the structure−function link passes through a topological bottleneck, the contact network, that allows for a consistent and very efficient formalism to be applied to the study of macromolecules.

3.2. Protein Structure Classification

Proteins can be considered as modular geometric objects composed of blocks, so allowing for a peptide-fragment-based partition.[114] For instance, it is well-known that globular proteins are made up of regular secondary structures (α-helices and β-strands) and nonregular secondary regions, called loops, that join regular secondary structures and lack the regularity of torsion angles for consecutive residues; actually, many families of proteins evolved to perform multiple functions, with variations in loop regions on a relatively conserved secondary structure framework. Considering this, Tendulkar et al.[114] developed an unconventional scheme of loops and secondary structure classification: the clustering of the peptide fragments

(three to six amino acid residues) was only based on the backbone structural similarity computed on first-order (length related) and second-order (area related) geometric invariants.

These invariants were obtained from the tetrahedra generated from the α-carbons' relative position in the fragment; in this way, the authors overcame the difficulties coming from the need of finding out the superimposing transformation for all the possible pairs of fragments and the need of considering hydrogen-bonding patterns. This allowed a faster and more reliable protein structure classification. In addition, clusters were differentiated as "functional" (mainly made up of loop regions) if more than 70% of the peptides of the cluster belong to a common superfamily or as "structural" (otherwise made up of regular secondary structures). This partition resulted in that 90% of the clusters belonged to the latter group, indicating that the conformation types are not merely a result of sequence homology. Moreover, this result is in line with the incredibly small number of "different folds" present in the protein world,[115] and it points to a strong invariant modularity of protein structure, whose basic brick seems to be a fundamental "protein word" six to eight residues long.[86,90] The explicit consideration of a protein as a contact network allowed one to base the "search for modularity principles" on a more rigorous and simple procedure.

Visheveshwara et al.[45] demonstrated that signal transmission through protein structure (e.g., during allosteric transitions) happens through noncovalent contacts; according to the authors, proteins can be considered as modular structures optimal for signal transmission, this optimal character being related to peculiar features of the corresponding contact networks.

Actually, in signaling proteins, modular domains can act as switches mediating activation, repression, and integration of different input functions; the finding of physical connectivity in coevolving networks suggested that there might be mechanically coupled elements in the atomic molecular structure, allowing for efficient long-range propagation of local perturbations.[62,63,66,68,70,86,109] Therefore, it is assumed[86] that the interactions of the most central residue contacts, responsible for the intermodular interactions, are mainly involved in information transfer between functional domains, as they maintain the shortest paths between all amino acid residues, whereas intramodular regions (which include most of the ligand binding sites) form a flexible pocket; this implies that the modular architecture of the active site relates to its subfunctions. Thus, functional specificity and regulation would depend on the communication between modules: due to the plausible functional independence of modules, changes in boundary residues may lead to new functions or to functional alterations eventually occurring in a changing environment. Likewise, as reported in ref 87, site-specific correlated motions in proteins are key determinants of function: most sites seem to act in a largely independent manner, robust to perturbations relative to other sites, and a few positions form coevolving linked networks through the structure.

It is worth noting that the most evolutionarily conserved sites largely correspond to the elements of the contact networks that guarantee the inter module communication.[66,86] This is a key point with respect to our quest for a paradigm of "contact graph as protein structural formula".

Evolution dynamics seems to keep the structure of the contact graph responsible for protein internal signal transduction largely invariant. Hence, it is evident that modularity can be identified as a focus component in evolvability, because it can both supply mutational robustness through the isolation of components and provide fast adaptation through the recombination of parts or by altering the connections between the modules.[116] In this framework, Wuchty[103] analyzed the topology of protein domain networks (the large-scale version of the PCNs, where functional and structural domains in a protein are the nodes of the network) and found a strong scale-free character, with small-world property. This suggests a hierarchical organization of domains that favors high short-range connectivity, along with few long-range interactions.

3.2.1. Modularity in Allosteric Proteins. Several investigations were pursued on the role of modularity in allosteric proteins: a detailed study involving 13 allosteric proteins[86] showed that functional sites are often contained within one module, and in a few cases they are located in two or more modules. Moreover, modules containing functional sites exhibited high modularity, suggesting that modularity can be useful to identify functional domains.

Another study[117] focused on the modular nature of biological ligands, composed of chemical fragments, and of the nucleotide-binding pockets, composed of fragment-specific protein structural motifs, that exhibit a modular organization and are responsible for binding different regions of their ligands. To investigate whether ligand chemical modularity could influence modularity of binding sites, authors chose nucleotides as paradigmatic ligands, since these molecules can be described as composed of well-defined fragments (nucleobase, ribose and phosphates) and are quite abundant either in nature or in protein ligands within protein structure databases.

Modularity seems to have both evolutionary and functional implications; actually, although binding motifs could not necessarily have the same evolutionary origin, the authors[117] demonstrated that they can be functionally interchangeable, showing that binding pockets can be decomposed into small modules instead of being treated as whole functional units. The identification of these protein motifs and their modular location suggests new hints to identify undetected binding sites based on the spatial proximity of the motifs on protein surfaces, which can lead to the assignment of ligands to protein structures in light of the design of biologically active chemicals.[117]

In recent research,[118] an allosteric protein was considered as both a modular and a dynamical computational device, in which each allosteric component has an input and an output: the input is a "modifier", a molecule that binds to and locally perturbs the structure of the component, and the output is the fraction of time that the component spends in each conformation when the allosteric transition is at equilibrium. In this way, a strict relationship among allostery, modularity, and complexity in networks is stressed: according to the authors, the structural domains in allosteric proteins exist in distinct conformational states with different biological activity. The transition between the conformations of these components is induced in either a concerted or a sequential fashion. Thus, they presented a modular and scalable modeling methodology, consisting in a set of modular structures and interaction rules implemented in the allosteric network compiler (ANC), which alleviates the combinatorial complexity of network computing: in detail, a thermodynamically grounded treatment of allostery is described in which ligands and other molecules interact with each conformation of the protein noncooperatively, distinguishing only its conformational state and not its state of covalent

dx.doi.org/10.1021/cr3002356 | Chem. Rev. XXXX, XXX, XXX–XXX

modification at distant sites. The reduction of the parameters required for both simulation and model construction allows for a very efficient modeling of the interaction of a protein with individual ligands, as well as for the prediction of the response to mixtures of ligands. Therefore, it can be easily understood that ANC has potential advantages in the creation of predictive models, since the modularity of ANC structures can afford the combination of different synthetic subsystems to generate more complex behavior.

De Ruvo et al.[119] highlighted the contemporary presence of a strong invariance of the general contact network and very specific concerted motions of local elements of the network changing global modularity features as the hallmark of the "allosteric computation" sketched by ref 64. This result is in line with the ones by del Sol et al.[41,72,86] indicating preferred "allosteric paths" along the protein structure. The allosteric property, in terms of interaction network, is a direct consequence of the small-world features of the PCNs.

Nussinov and colleagues[60] extended the concept of allostery to the coordinated behavior of ensembles of proteins in close contact between them. According to the proposed model, allosteric signal propagation does not stop at the "end" of a protein but may be dynamically transmitted across the cell by a protein−protein interaction network.

This hypothesis (perfectly plausible from a chemicophysical point of view), if confirmed, could have very important consequences on both basic biology (scientists still do not have a reasonable hypothesis for the generalization of the molecular stimuli) and pharmacology ("allo-network drugs" able to give rise to such a generalized allosteric transition could be the ideal candidates in a lot of therapeutic interventions). In this scenario, the concept of allostery has been recently revised:[44,61,62,120] the coupling of the ligand-binding and conformational transition, at the very early elucidation of the molecular mechanism of allostery, has been recently enriched, according to the statistical thermodynamic methods. In other words, all the protein conformations, formerly related to different ligand states of the protein (relaxed and tense in the MWC model[121]) are always present in the protein ensemble. The observed conformational transitions, from a tense to relaxed state in the hemoglobin binding of oxygen, for instance, are viewed as statistical shifts of the conformational states distribution toward the more probable conformation. This "configuration selection" model asks for a link between network invariants relative to the different configurations and their abundance.

This picture is consistent with the superposition of allosteric and folding properties: the same residues are responsible of both features.[122] In other words, the "allo-network drugs" paradigm is made plausible by the observed allosteric modification upon ligand binding, passing through the same residue patches (folding elements) and transmitting the folding information throughout the forming structure.

3.2.2. Protein Folding. The relationship between protein folding and protein sequence-based features has been a central issue in protein science since the 1960s;[123−125] as a matter of fact, a direct link between the protein sequence and folding thermodynamics represents a kind of "philosophic stone" for protein science scholars.

In this framework, a promising approach is based upon the representation of protein structures in terms of PCNs, and contact network analysis is likely to catch some folding relevant features.

The nowadays prevailing paradigm states that folding depends more on native state topology rather than on interatomic interactions.[100,126] Specifically, the locality of the residue contacts has been demonstrated to have a key role in folding rate, through the definition of contact order, CO, expressed as[127]

$$\mathrm{CO} = \frac{\sum \Delta S_{ij}}{LN}$$

where ΔS_{ij} is the distance in sequence among the ith and jth residues experiencing a contact, L is the overall number of contacts over N nodes, and the sum is extended over the whole set of contacts. CO demonstrated a strong correlation with folding rates, even coming to be a predictor for them.[23−26]

The concept has been further extended into a long-range order parameter, LRO, defined as[23]

$$\mathrm{LRO} = \frac{\sum n_{ij}}{N}$$

$$n_{ij} = \begin{cases} 1 & |i - j| > 12 \\ 0 & \text{otherwise} \end{cases}$$

n_{ij} represents active links among residues distant by more than 12 unities in sequence. This parameter has shown a strong negative correlation with protein folding rates, showing that the establishment of long-range interactions slows down the folding process, nonetheless yielding more stable protein structures.

It is noteworthy that the cutoff for the link definition (8 Å) and the distance in sequence for LRO have been found on the basis of the best correlation with the folding rates values, but they correspond to well-known structural and functional thresholds (hydrophobic interaction length[20] and modules of 12 and 25 residues as key short-range clusters responsible of folding[128,129]). Following a similar approach, set on a protein structural network, matching with an analysis of the hydrophobicity-based recurrence plots, Krishnan et al.[28] found that a critical cluster dimension in folding is 12 (6 + 6) residues, in strong compliance with results emerging from the CO and LRO method.

Vendruscolo and co-workers[42,43] carried out a folding analysis by PCNs, focusing on its small-world property, which can be exploited to derive "key residues" in the protein folding process. Indeed, they found very small clusters (made up of three residues) that represent the folding nuclei. This result matches with the assumption of cooperativity of the protein folding mechanism: cooperativity, as matter of fact, is a feature of many biological process and it is invoked to describe the nonlinear, nonadditional nature of many complex processes, such as folding and ligand binding. From a network perspective, the small-world paradigm is a good theoretical framework to introduce the holistic nature of biological processes involving biomacromolecules that interact with the environment by experiencing conformational changes to adapt to their function. A reductionistic style would not be capable of describing such a scenario given there could be no distinct definition for subprocesses, since the different regions in biomacromolecules are strongly interrelated through the interaction network that defines the protein system as a whole.

4. CONCLUSIONS

The problem of protein sequence−structure−function relation was classically formalized in two distinct ways:

(1) From a global prediction perspective, the problem had the form of a function mapping the specific location of a given element (residue) on a monodimensional array (the sequence) onto a three-dimensional coordinate vector in Euclidean space. This mapping was approached by attaching to the residues some chemicophysical attributes (hydrophobicity, molecular weight, electronic properties, etc.), by imposing on the proposed solutions some chemicophysical motivated constraints (statistical potentials, simulated energy landscapes, etc.) or by adopting a purely statistical phenomenological approach, like the computation of sequence similarity with other proteins whose 3D structure was known.

(2) On a local perspective, the discrimination of the most crucial residues for biological function was accomplished by site-directed mutagenesis, residue conservation estimation based on phylogenetic analysis, and structural considerations.

Different combinations of the two above approaches were present in the literature so that, as often happens in many fields of science, the most common (and in many cases successful) approaches were the most eclectic ones, whose general philosophy was the maximization of the consistency among the different perspectives.[113] The most relevant point to be stressed, at least in our opinion, is that in the last two decades the sequence−structure−function problem underwent a profound change in its formalization. The discovery of natively unfolded proteins as well as of natively unfolded patches inside otherwise structured systems[29] started the growing recognition that protein disorder plays a crucial role in both protein−protein interaction and protein folding.[130,131] This awareness casts doubts on the linear pathway going from sequence to function across the mediation of a rigid crystal structure. Similar suggestions came from the research focusing on prion-like and thermophilic systems where completely different behaviors characterized systems with practically identical 3D structures and/or sequences.[132] The deterministic task of mapping a given linear arrangement of symbols into a three-dimensional spatial coordinate system turned into a mainly probabilistic affair of delineating the "mesoscopic features"[133] at both entire protein and specific residues or domain levels. Some of these mesoscopic features (e.g., flexibility, allosteric properties) are related to the dynamical behavior of the system at hand and some other (e.g., folding rate, thermal stability, aggregation propensity) to kinetic and thermodynamic characteristics. What is generally recognized is that by only adding an extra, mainly dynamic, mesoscopic dimension to the sequence−structure static perspective we can get a satisfactory picture of the physiological role exerted by protein systems. The consideration of proteins as contact graphs holds the promise to represent a general and consistent formalization to obtain mesoscopic views of proteins, while the above-mentioned "classical" requirements of attaining both "global" and "residue centered" consistent views are maintained. Since the seminal work by Maxwell on the conditions of the rigidity of systems of elements that hold together by edges (the so-called Maxwell−Cremona contact rule[134]), graphlike formalizations were recognized as deeply intertwined with mesoscopic properties of systems. The same relation between contact graphs and structure stability was at the base of symmetry groups in crystallography[135] and in rotational properties of organic molecules considered as chemical graphs.[136] Here we show how biologically crucial properties like allostery and signal transmission have their immediate graph-theoretical counterpart in terms of average shortest path or clustering coefficient. Last but not least, the possibility of "coloring" both nodes and edges of the graph by means of meaningful descriptors (e.g., hydropobic/hydrophilic character of residues, effective distance, or energetic characterization of edges) dramatically enlarges the reach of network-inspired approaches. The main theme of the present review is to put these theoretical suggestions into practice, showing that in many instances graph-theoretical descriptors can be profitably used to predict physiologically relevant properties of proteins. We are conscious that we are still in a mainly phenomenological phase of the work and the described topics are still tackled on a case-by-case approach, but nevertheless, some interesting invariants are emerging. Probably the most intriguing one is the constancy of the between modules/within module contacts distribution as depicted in the P, z graphs, which are remarkably invariant across protein systems, largely varying in both size and shape. This invariance (probably linked to the peculiar degree distribution of protein contact networks) is reminiscent of "valencelike" constraints of structural formulas. The presence of such invariants, as well as the demonstrated contact graph formalization, allows for the conservation of all the relevant information linked to protein 3D configuration. Moreover, the filtering out of the noise that obscures the functionally relevant information, as in the case of allosteric behavior prediction,[119] makes us confident that contact graphs will become, in the near future, the "structural formulas" of proteins.

AUTHOR INFORMATION

Corresponding Author

*E-mail: alessandro.giuliani@iss.it.

Notes

The authors declare no competing financial interest.

Biographies

Luisa Di Paola was trained as a Chemical Engineer at the University "La Sapienza"; she got her Ph.D. in Industrial Chemical Processing at the same university. During her doctorate, she was a visiting scholar at the University of California—Berkeley, under the supervision of Prof. J. M. Prausnitz, where she acquired the methods of molecular thermodynamics, with special regard to biomedical and biotechnological applications. Currently, she is Assistant Professor in Chemical and Biochemical Engineering Fundamentals at the University "Campus Biomedico" in Rome. Her reasearch focuses on integration of biophysical chemistry tools with novel methodologies to analyze complex biological systems (protein structures, biological signaling),

biotechnological methods for biofuel production, modeling of transport phenomena in artificial organs (oxygenators, artificial kidney and liver), and physiological compartmental models for pharmacokinetics.

Micol De Ruvo was born in 1986 and received her M.S. degree in Biomedical Engineering in 2010 from "Campus Bio-Medico" University of Rome. Currently, she is a Ph.D. student in Biomedical Engineering at "Campus Bio-Medico" University of Rome and is carrying out research regarding the modeling of hormone diffusion and cell growth in plants, in collaboration with Prof. S. Sabatini of the Laboratory of Functional Genomics and Proteomics of Model Systems at the Biotechnology Department "Charles Darwin" of the University "La Sapienza" in Rome. During her graduate training, she applied the molecular thermodynamics methods to the analysis of protein structure and function. Her present research interests cover the application of mathematical and computational modeling to biology, with a major focus on systems biology approaches applied to both protein contact network and hormones interaction and cellular growth during plant development.

Paola Paci was trained as a theoretical physicist and has worked over 10 years in the field of computer modeling. She holds a degree in Physics from the University of Rome "La Sapienza", a Ph.D in Physics from the University of Pavia "A. Volta", and a Master in Bioinformatics from the University of Rome "La Sapienza". She spent two years at the International School for Advanced Studies (SISSA) in Trieste for her theoretical studies on the structure of condensed matter. Nowadays, her research interests include physics applied to medicine and biology, bioinformatics, and systems biology as a researcher at the Institute for System Analysis and Computer Science (IASI) of CNR in Rome.

Daniele Santoni was born in Rome, Italy, and graduated in Mathematics at the University of Rome "La Sapienza". After obtaining a Master in Bioinformatics, he obtained a Ph.D. in Genetics and Molecular Biology. Since 2011 he has served as a research scientist at the Institute for System Analysis and Computer Science of the National Research Council of Italy in Rome. His research is mainly on the area of bioinformatics and system biology and focuses on the application of information theory and graph theory to biological systems.

Alessandro Giuliani was born in 1959 in Rome, Italy, and graduated in Biological Sciences at the Rome University 'La Sapienza' in 1982. He served as research scientist at Sigma-Tau pharmaceutical industries from 1985 to 1997, and from 1997 to now he is Senior Scientist at Istituto Superiore di Sanità. His research is devoted to the mathematical and statistical formalization and analysis of biological and chemical problems with a particular emphasis on multidimensional statistics and nonlinear dynamics inspired methods. He is the author of around 200 scientific papers ranging from structure−activity relationships in medicinal chemistry and toxicology to protein science, molecular biology, animal behavior, and epidemiology.

REFERENCES

(1) Bensaude-Vincent, B.; Stengers, I. *A History of Chemistry*; Harvard University Press: Boston, 1996.
(2) Todeschini, R.; Consonni, V. In *Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry*; Manhold, R., Kubinyi, H., Timmermann, H., Eds.; Wiley VCH: New York, 2000; Vol. *11*.
(3) Schultz, T.; Cronin, M. T. D.; Walker, J.; Aptula, A. *J. Mol. Struct. THEOCHEM* **2003**, *622*, 1.
(4) Bender, A.; Glen, R. *J. Chem. Inf. Model.* **2005**, *45*, 1369.
(5) Bonchev, D.; Rouvray, D. *Chemical Graph Theory: Introduction and Fundamentals*; Gordon & Breach Science Publishers: New York, 1990.

(6) Mekenyan, O.; Bonchev, D.; Trinajstic, N. *Int. J. Quantum Chem.* **1980**, *18*, 369.
(7) Fredenslund, A.; Gmehling, J.; Rasmussen, P. *Vapor−Liquid Equilibria Using UNIFAC: a Group Contribution Method*; Elsevier Scientific: New York, 1979.
(8) Abrams, D.; Prausnitz, J. *AIChE J.* **1975**, *21*, 116.
(9) Levenspiel, O. *Chemical Reaction Engineering*, 3rd ed.; Wiley: New York, 1999.
(10) Papin, J.; Price, N.; Wiback, S.; Fell, D.; Palsson, B. *Trends Biochem. Sci.* **2003**, *28*, 250.
(11) Schuster, S.; Felli, D.; Dandekar, T. *Nat. Biotechnol.* **2000**, *18*, 326.
(12) Feinberg, M. *Chem. Eng. Sci.* **1987**, *42*, 2229.
(13) Feinberg, M. *Chem. Eng. Sci.* **1988**, *43*, 1.
(14) Palumbo, M.; Farina, L.; Colosimo, A.; Tun, K.; Dhar, P. K.; Giuliani, A. *Curr. Bioinf.* **2006**, *1*, 219.
(15) Aftabuddin, M.; Kundu, S. *Physica A* **2006**, *369*, 895.
(16) Barah, P.; Sinha, S. *Pramana* **2008**, *71*, 369.
(17) da Silveira, C.; Pires, D.; Minardi, R.; Ribeiro, C.; Veloso, C.; Lopes, J.; Meira, W., Jr; Neshich, G.; Ramos, C.; Habesch, R.; Santoro, M. *Proteins* **2009**, *74*, 727.
(18) Miyazawa, S.; Jernigan, R. *Macromolecules* **1985**, *18*, 534.
(19) Goldstein, R. *Protein Sci.* **2007**, *16*, 1887.
(20) Bahar, I.; Jernigan, R. *J. Mol. Biol.* **1997**, *266*, 195.
(21) Metpally, R.; Reddy, B. *BMC Genomics* **2009**, *10*, 11.
(22) Webber, C. L. J.; Giuliani, A.; Zbilut, J.; Colosimo, A. *Proteins* **2001**, *44*, 292.
(23) Gromiha, M.; Selvara, J. S. *J. Mol. Biol.* **2001**, *310*, 27.
(24) Gromiha, M. *J. Chem. Inf. Model.* **2003**, *43*, 1481.
(25) Gromiha, M.; Thangakani, A.; SelvaraJ, S. *Nucleic Acids Res.* **2006**, *34*, W70.
(26) Gromiha, M. *J. Chem. Inf. Model.* **2009**, *49*, 1130.
(27) Sun, W.; He, J. *Biopolymers* **2010**, *93*, 904.
(28) Krishnan, A.; Zbilut, J. P.; Tomita, M.; Giuliani, A. *Curr. Protein. Pept. Sc.* **2008**, *9*, 28.
(29) Giuliani, A.; Di Paola, L.; Setola, R. *Curr. Proteomics* **2009**, *6*, 235.
(30) Cohen, R.; Havlin, S. *Complex Networks: Structure, Robustness and Function*; Cambridge University Press: Cambridge, UK, 2010.
(31) Ilari, A.; Savino, C. *Methods Mol. Biol.* **2008**, *452*, 63.
(32) Wutrich, K. *Science* **1989**, *243*, 45.
(33) Perutz, M.; Rossmann, M.; Cullis, A.; Muirhead, H.; Will, G.; North, A. C. T. *Nature* **1960**, *185*, 416.
(34) Berman, H. *Acta Crystallogr. A* **2008**, *64*, 88.
(35) Vijayabaskar, M.; Vishveshwara, S. *Biophys. J.* **2010**, *99*, 3704.
(36) Pauling, L.; Corey, R.; Branson, H. *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37*, 205.
(37) Afonnikov, D.; Morozov, A.; Kolchanov, N. *Biophysics* **2006**, *51*, 56.
(38) Bagler, G.; Sinha, S. *Physica A* **2005**, *346*, 27.
(39) Brinda, K.; Surolia, A.; Vishveshwara, S. *Biochem. J.* **2005**, *391*, 1.
(40) Bagler, G.; Sinha, S. *Bioinformatics* **2007**, *23*, 1760.
(41) del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R. *Mol. Syst. Biol.* **2006**, *2*, 2006.0019.
(42) Vendruscolo, M.; Paci, E.; Dobson, C.; Karplus, M. *Nature* **2001**, *409*, 641.
(43) Vendruscolo, M.; Dokholyan, N.; Paci, E.; Karplus, M. *Phys. Rev. E* **2002**, *65*, 061910.
(44) Vendruscolo, M. *Nat. Chem. Biol.* **2011**, *7*, 411.
(45) Vishveshwara, S.; Brinda, K.; Kannan, N. *J. Theor. Comput. Chem.* **2002**, *1*, 1.
(46) Vishveshwara, S.; Ghosh, A.; Hansia, P. *Curr. Protein Pept. Sci.* **2009**, *10*, 146.
(47) Brinda, K.; Vishveshwara, S. *Biophys. J.* **2005**, *89*, 4159.
(48) Brinda, K.; Vishveshwara, S.; Vishveshwara, S. *Mol. Biosyst.* **2010**, *6*, 391.
(49) Kundu, S. *Physica A* **2005**, *346*, 104.
(50) Kannan, N.; Vishveshwara, S. J. *Mol. Biol.* **1999**, *292*, 441.
(51) Yan, Y.; Zhang, S.; Wu, F. *Proteome Sci.* **2011**, *9*, S17.
(52) Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. *J. Mol. Biol.* **2004**, *344*, 1135.
(53) Doncheva, N.; Klein, K.; Domingues, F.; Albrecht, M. *Trends Biochem. Sci.* **2011**, *36*, 179.
(54) Krishnadev, O.; Brinda, K.; Vishveshwara, S. *Proteins* **2005**, *61*, 152.
(55) Sathyapriya, R.; Vishveshwara, S. *Proteins* **2007**, *68*, 541.
(56) Sengupta, D.; Kundu, S. *Physica A* **2012**, *391*, 4266.
(57) Di Paola, L.; Paci, P.; Santoni, D.; De Ruvo, M.; Giuliani, A. *J. Chem. Inf. Model.* **2012**, *52*, 474.
(58) Krishnan, A.; Giuliani, A.; Zbilut, J.; Tomita, M. *J. Proteome Res.* **2007**, *6*, 3924.
(59) Estrada, E. *Biophys. J.* **2010**, *98*, 890.
(60) Nussinov, R.; Tsai, C.; Csermely, P. *Trends Pharmacol. Sci.* **2011**, *32*, 686.
(61) Tsai, C.; del Sol, A.; Nussinov, R. *J. Mol. Biol.* **2008**, *378*, 1.
(62) Tsai, C.; del Sol, A.; Nussinov, R. *Mol. Biosyst.* **2009**, *5*, 207.
(63) Clarkson, M.; Gilmore, S.; Edgell, M.; Lee, A. *Biochemistry* **2006**, *45*, 7693.
(64) Daily, M.; Gray, J. *PLoS Comput. Biol.* **2009**, *5*, e1000293.
(65) Kim, D.; Park, K. *BMC Bioinf.* **2011**, *12*, 1471.
(66) Bode, C.; Kovács, I.; Szalay, M.; Palotai, R.; Korcsmáros, T.; Csermely, P. *FEBS Lett.* **2007**, *581*, 2776.
(67) Guangxu, J.; Zhang, S.; Zhang, X.; Chen, L. *PLoS ONE* **2007**, *2*, e1207.
(68) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128.
(69) Teilum, K.; Olsen, J. G.; Kragelund, B. B. *Biochim. Biophys. Acta* **2011**, *1814*, 969.
(70) Yang, L.; Song, G.; Jernigan, R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12347.
(71) Emerson, I.; Gothandam, K. *Physica A* **2012**, *391*, 905.
(72) del Sol, A.; O'Meara, P. *Proteins* **2005**, *58*, 672.
(73) Newman, M. E. *J. Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 404.
(74) Newman, M. E. J.; Strogatz, S. H.; Watts, D. *J. Phys. Rev. E* **2001**, *64*, 026118.
(75) Greene, L.; Highman, V. *J. Mol. Biol.* **2003**, *334*, 781.
(76) Lang, S. Protein domain decomposition using spectral graph partitioning. Ph.D. thesis, Studienarbeit am ITI Wagner Fakultat fur Informatik Universitat Karlsruhe (TH), 2007.
(77) van Mieghem, P.; P, X. G.; Schumm; Trajanovski, S.; Wang, H. *Phys. Rev. E* **2010**, *82*, 1.
(78) Tsai, C.; Kumar, S.; Ma, B.; Nussinov, R. *Protein Sci.* **1999**, *8*, 1181.
(79) Ma, B.; Kumar, S.; Tsai, C.; Nussinov, R. *Protein Eng.* **1999**, *12*, 713.
(80) Guimerà, R.; Sales-Pardo, M.; Amaral, L. A. N. *Nat. Phys.* **2006**, *3*, 63.
(81) Gursoy, A.; Keskin, O.; Nussinov, R. *Biochem. Soc. Trans.* **2008**, *36*, 1398.
(82) Agarwaal, S.; Deane, C.; Porter, M.; Jones, N. *PLoS Comput. Biol.* **2010**, *6*, e1000817.
(83) Csermely, P. *Trends Biochem. Sci.* **2008**, *33*, 569.
(84) Jeong, H.; Mason, S.; Barabàsi, A.; Oltvai, Z. *Nature* **2001**, *411*, 41.
(85) Albert, R.; Jeong, H.; Barabàsi, A. *Nature* **2000**, *406*, 378.
(86) del Sol, A.; Araúzo-Bravo, M.; Amoros, D.; Nussinov, R. *Genome Biol.* **2007**, *8*, R92.
(87) Süel, G.; Lockless, S.; Wall, M.; Ranganathan, R. *Nat. Struct. Biol.* **2002**, *10*, 59.
(88) Girvan, M.; Newman, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 7821.
(89) Yu, H.; Kim, P.; Sprecher, E.; Trifonov, V.; Gerstein, M. *PLoS Comput. Biol.* **2007**, *3*, e59.
(90) Newman, M. E. *J. Phys. Rev. E* **2006**, *74*, 1.
(91) Xu, X.; Zhang, J.; Sun, J.; Small, M. *Phys. Rev. E* **2009**, *80*, 1.
(92) Newman, M. E. J.; Watts, D. J.; Strogatz, S. H. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 2566.

(93) Park, J.; Barabàsi, A. L. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 17916.
(94) Alves, N.; Aleksenko, V.; Hansmann, U. *J. Phys: Condens. Matter* **2005**, *17*, S1595.
(95) Bollobas, B. *Random Graphs*; Academic Press: New York, 1985.
(96) Rapoport, A.; Solomonoff, R. *Math. Biophys.* **1951**, *18*, 107.
(97) Erdos, P.; Renyi, A. *Publ. Math.* **1959**, *6*, 290.
(98) Erdos, P.; Renyi, A. *Publ. Math. Inst. Hung. Acad. Sci.* **1960**, *5*, 17.
(99) Erdos, P.; Renyi, A. *Acta Math. Hung.* **1961**, *12*, 261.
(100) Dokholyan, N.; Li, L.; Ding, F.; Shakhnovich, E. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 8637.
(101) Milencovic, T.; Filippis, I.; Lappe, M.; Przulj, N. *PLoS ONE* **2009**, *4*, e5967.
(102) Sharan, R.; Ulitsky, I.; Shamir, R. *Mol. Syst. Biol.* **2007**, *3*, 1.
(103) Wuchty, S. *Mol. Biol. Evol.* **2001**, *18*, 1694.
(104) Bartoli, L.; Fariselli, P.; Casadio, R. *Phys. Biol.* **2008**, *4*, 1.
(105) Albert, R.; Jeong, H.; Barabàsi, A. *Nature* **1999**, *401*, 130.
(106) Barabàsi, A. *Linked: How Everything Is Connected to Everything Else and What It Means*; Penguin Group: New York, 2003.
(107) Barabási, A.; Oltvai, Z. *Nat. Rev. Gen.* **2004**, *5*, 101.
(108) Vàsquez, A.; Moreno, Y. *Phys. Rev. E* **2003**, *67*, 015101.
(109) Atilgan, A.; Akan, P.; Baysal, C. *Biophys. J.* **2004**, *86*, 85.
(110) Albert, R.; Barabàsi, A. *Rev. Mod. Phys.* **2002**, *74*, 47.
(111) Müller-Linow, M.; Hilgetag, C. C.; Hütt, M. T. *PLoS Comput. Biol.* **2008**, *4*, 1.
(112) Liu, Z.; Wu, L.; Wang, Y.; Zhang, X.; Chen, L. *Protein Pept. Lett.* **2008**, *15*, 448.
(113) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. *PLoS ONE* **2011**, *6*, e28766.
(114) Tendulkar, A. V.; Joshi, A.; Sohoni, M.; Wangikar, P. *J. Mol. Biol.* **2004**, *338*, 611.
(115) Govindarajan, S.; Recabarren, R.; Goldstein, R. *Proteins* **1999**, *35*, 408.
(116) Hintze, A.; Adami, C. *Biol. Direct.* **2010**, *5*, 1.
(117) Gherardini, P.; Ausiello, G.; Russell, R.; Helmer-Citterich, M. *Nucleic Acids Res.* **2010**, *38*, 3809.
(118) Ollivier, J.; Shahrezaei, V.; Swain, P. *PLoS Comput. Biol.* **2010**, *6*, e1000975.
(119) De Ruvo, M.; Giuliani, A.; Paci, P.; Santoni, D.; Di Paola, L. *Biophys. Chem.* **2012**, *165*−*166*, 21.
(120) Gunasekaran, K.; Ma, B.; Nussinov, R. *Proteins* **2004**, *57*, 433.
(121) Monod, J.; Wyman, J.; Changeux, J. *J. Mol. Biol.* **1965**, *12*, 88.
(122) Hu, Z.; Bowen, D.; Southerland, W.; del Sol, A.; Pan, Y.; Nussinov, R.; Ma, B. *PLoS Comput. Biol.* **2007**, *3*, 1097.
(123) Anfinsen, C. *Science* **1973**, *181*, 223.
(124) Karplus, M. *Fold. Des.* **1997**, *2*, S69.
(125) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 20.
(126) Alm, E.; Baker, D. *Curr. Opin. Struct. Biol.* **1999**, *2*, 189.
(127) Plaxco, K.; Simons, K.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985.
(128) Grantcharova, V.; Riddle, D.; Santiago, J.; Baker, D. *Nat. Struct. Biol.* **1998**, *5*, 714.
(129) Itzhaki, L.; Otzen, D.; Fersht, A. *J. Mol. Biol.* **1995**, *254*, 260.
(130) Csermely, P.; Sandhu, K.; Hazai, E.; Hoksza, Z.; Kiss, H.; Miozzo, F.; Veres, D.; Piazza, F.; Nussinov, R. *Curr. Protein Pept. Sci.* **2012**, *13*, 19.
(131) Uversky, V. *Cell. Mol. Life Sci.* **2003**, *60*, 1852.
(132) Giuliani, A.; Benigni, R.; Sirabella, P.; Zbilut, J.; Colosimo, A. *Biophys. J.* **2000**, *78*, 136.
(133) Laughlin, R.; Pines, D.; Schmalian, J.; Stojkovic, B.; Wolynes, P. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 32.
(134) Maxwell, J. *J. Philos.* **1864**, *4*, 250.
(135) Spek, A. *Acta Crystallogr. D* **2009**, *65*, 148.
(136) Gunawardena, J. *Chemical Reaction Network Theory for In-Silico Biologists*; Harvard University: Cambridge, MA, 2003.

Structural and Functional Analysis of Hemoglobin and Serum Albumin Through Protein Long-Range Interaction Networks

Paola Paci[1], Luisa Di Paola[2], Daniele Santoni[3], Micol De Ruvo[2] and Alessandro Giuliani*[,4]

[1]*CNR-Institute of Systems Analysis and Computer Science (IASI),BioMathLab, viale Manzoni 30, 00185 Roma, Italy;* [2]*Faculty of Engineering ,Università CAMPUS BioMedico, Via A. delPortillo,, 21, 00128 Roma, Italy;* [3]*CNR-Institute of Systems Analysis and Computer Science (IASI), Viale Manzoni, 30, 00185 Roma, Italy;* [4]*Environment and Health Department, Istituto Superiore di Sanità, Viale Regina, Elena 299, 00161, Roma, Italy*

Abstract: Long-range contacts in protein structures were demonstrated to be predictive of different physiological properties of hemoglobin and albumin proteins. Complex networks based approach was demonstrated to highlight basic principles of protein folding and activity. The presence of a natural scaling region ending at an approximate threshold of 120-150 residues shared by proteins of different size and quaternary structure was highlighted. This threshold is reminiscent of the typical size for a macromolecule to have a binding site sensible to environmental regulation.

Keywords: Allosteric effect, Bioinformatics, Graph theory, Structural biology, Computational biochemistry, Topological invariants.

1. INTRODUCTION

One of the most challenging task in structural biology is to understand and predict protein folding on the basis of the primary structure[1–3]. Thus, the folding mechanism has been largely explored in order to unravel the relationship between aminoacid properties and their attitude to be involved in intramolecular, non-covalent bonds, largely responsible of the folding kinetics and protein stability [2–10].

Protein folding is known to be the result of cooperative mechanisms, structural changes and chemical interactions, that occur in parallel and allow the molecule to reach the native tertiary structure. Long-range interactions between far-away residues in sequence play a significant role in determining the three-dimensional structure of the proteins, since they are essential for highly-cooperative stabilization of the native conformation, whereas the short-range interactions accelerate the folding and unfolding transitions[10–16].

The interactions between the amino acid residues within a protein can be intended in terms of a protein contact network (PCN) in which aminoacid residues represent nodes and the interactions (mainly non-bonded, non-covalent) among them correspond to undirected edges [17].

In order to isolate the peculiar contribution of long-range contacts, it is possible to focus on sub-networks made only of residues faraway along the sequence.

*Address correspondence to this author at the Environment and Health Department, Istituto Superiore di Sanità, Viale Regina Elena299, 00161, Roma, Italy;
Tel:+39 0649902579; E-mail: alessandro.giuliani@iss.it

Long-range interaction networks (LINs) are used to investigate the correlation of the long-range interactions with topological (such assortativity) and biophysical properties (such as folding rate [18]). LINs display peculiar properties in terms of network invariants: assortativity is a very relevant descriptor in this respect.

The coefficient of assortativity *r* [19] is a global quantitative measure of degree correlations in a network, and takes values ranging from -1 to 1. In [18], *r* values were found markedly positive for both PCNs and LINs with respect to other networks of different origins [19]. Keeping in mind the degree of a node corresponds to the edges it is involved into, an high positive *r* indicates the tendency of connections rich residues to be in contact. On the contrary, a negative *r* value points to an opposite behaviour.

The coefficient of assortativity shows a positive correlation with protein folding rate by speeding up the formation of both short- and long-range contacts.

More in general, LINs topological parameters can effectively represent the structural and functional properties required for fast information transfer among the residues, facilitating biochemical/kinetic functions (allostery, stability and folding rate)[18].

Several studies [3, 14,20–23] emphasized the dominance of hydrophobic residues in protein folding. Poupon and Mornon [20] showed a striking correspondence between the conserved hydrophobic positions of a protein and the intermediates formed during the folding initial stages. Aftabbudin and Kundu [23] performed a comparative topological study of the hydrophobic, hydrophilic and charged residues contact networks showing hydrophobic residues are mostly responsible for the overall topological features of a protein. Selvaraj and Gromiha [14] identified the role of hy-

1570-1646/12 $58.00+.00

drophobic clusters in folding of β-barrel proteins and stressed the key role of medium-and long-range interactions in the formation and stability of hydrophobic clusters [24].

In order to quantify the overall effect of local and non-local contacts on the folding kinetics, contact order (*CO*) was introduced [12,16]. This parameter identifies the average distance in sequence of effective contacts between residues. The higher *CO*, the stronger is the effect of long-range interactions on protein folding and stability. *CO* has been demonstrated to linearly correlate with the logarithm of the folding rate *k*, allowing to predict with a good precision the folding rate of proteins [25].

The concept of contact order can be further extended to the Long Range Order (*LRO)* [12],that measures the average number of contacts that occur between residues whose distance in sequence is larger than a given threshold (long-range contacts).

LRO too has been computed setting the threshold of sequence distance at 12 [26]: this threshold was found to be an optimal value for folding rate prediction [12]. As *CO, LRO* shows a correlation close to -0.8 with the folding rate, pointing to a strong influence of long-range contacts in slowing-down folding.

In this work, we try and approach long-range interactions by the concurrent perspectives of the degree and hydrophobic assortativity, in order to shed light on general mesoscopic principles governing protein folding process. The choice of the two model proteins, albumin and hemoglobin, was dictated by the need of having two extremes in the space of the protein behaviour: a very efficient allosteric system (hemoglobin) and a relatively "dull" system (albumin) with a pure storage role. The opposite dynamic properties of the two selected systems could have a counterpart in the LIN architecture, given long-range contacts are known to mediate allosteric behaviour [27].

We extended the analysis on another set of binding proteins, accounting for enzymes implied in biological mechanism (catalase and acetate kinase) or exploited in biotechnological processes (cellobiohydrolase and lipase). This extension was made in order to give a proof-of-concept to the general trends highlighted by the hemoglobin- albumin comparison. Details for this protein test set are provided in Table 1.

The choice of these enzymes was dictated by the hypothesis the need of having a fine tunable active site could be an important driver of LIN organization.

2. METHODS

2.1. Protein Contact Graph and General Topological Indexes

PDB files provide complete information about the atom position in 3D crystal structure of a protein; this information can be exploited to derive the between residues interaction map.

The corresponding Protein Contact Network (PCN) is a network whose nodes are the residues (spatially identified by their α-carbons) and edges exist between two residues if their mutual distance lies in a given length range (I= [4−8]Å); the network adjacency matrix $\mathbf{A}=\{a_{ij}\}$ is therefore defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } d_{ij} \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

This approach results useful to analyze chemico-physical and functional protein properties [18,24,26,31–37].

LINs are a special class of PCNs, including only links between residues whose distance in sequence is larger than a given threshold L. Thus, the corresponding adjacency matrix $\mathbf{A}^{(L)}=\{a^{(L)}\}$ is modified with respect to the 3.1 as:

$$a_{ij}^{(L)} = \begin{cases} 1 & \text{if } d_{ij} \in \mathcal{I} \wedge |i-j| \geq L \\ 0 & \text{otherwise} \end{cases} \qquad L \in \mathbb{N} \tag{3.2}$$

In this paper, we deal with some topological descriptors that can be extracted from $\mathbf{A}^{(L)}$ [38]:

- *density*: ratio between the actual number of edges E and the maximum value of possible links $E_{MAX}(L)$;
- *avdegree*: the average value of node degree computed over all the residues;
- *avshortpath*: the shortest path is the minimum number of links connecting two residues; this value, averaged over all the residue pairs, is the average shortest path;
- *DBA*: Degree-Based Assortativity, computed as the Pearson correlation coefficient between the two vectors containing the degree values for incident nodes in LINs;
- *HBAKD*: Hydrophobic-Based Assortativity, corresponding to the Pearson correlation coefficient between incident residue hydrophobicity lists; hydrophobicity scores are based on the Kyte-Doolittle scale.
- *LRO*: long range order is defined as [12]:

$$LRO(L) = \frac{1}{N \cdot E} \cdot \sum_{i,j}^{N} \Delta S_{ij}^{(L)}$$

Table 1. Protein test set.

PDBcode	Description	Reference	Residues	Chains
1GPI	Cellobiohydrolase	[28]	431	1
1TUU	Acetatekinase	[29]	798	2
8CAT	Liver catalase	[30]	996	2
3GUU	Lipase A	(to be published)	863	2

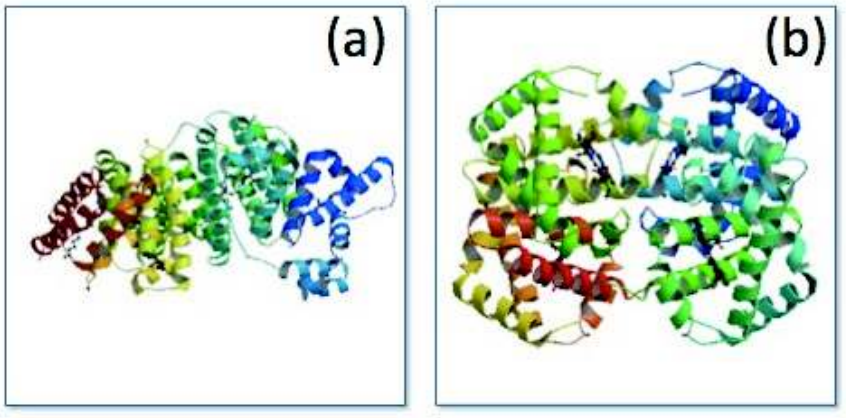


Fig. (3.1). Albumin (a) and hemoglobin (b) 3D protein structures.

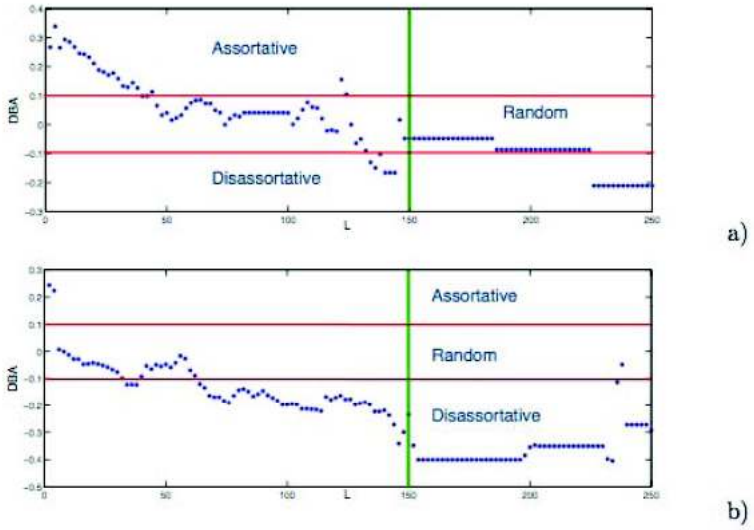


Fig. (4.1). *DBA*: a) serum albumin; b) hemoglobin.

where $\Delta S^{(L)}$ is the distance in sequence between incident residues (nodes) for a long-range threshold L (i.e., including only contacts between residues distant L at least in sequence); E is the number of links in the LIN and N is the number of residues.

Proteins of main interest are human hemoglobin and serum albumin, whose structures are available in PDB repository, corresponding to 1HBB (574aa) and 1E7I (585aa) codes, respectively (Fig. **3.1**).

Test set was made of 1GPI, 1TUU, 8CAT, 3GUU whose specifics are reported in Table 1.

Graphical representation of networks have been realized by YED (http://www.yworks.com/en/products_yed_about.html).

3. RESULTS AND DISCUSSION

LINs were analyzed in terms of the above mentioned descriptors as function of L.

3.1. Assortativity

DBA values were partitioned into three different intervals [19]: greater than 0.1; in the range (-0.1÷0.1); smaller than -0.1, identifying assortative, random and disassortative behaviour, respectively. This classification derives from the comparison inside a very diverse set of networks [19].

The plots in 4.1show*DBA*values of albumin and hemoglobin networks as a function of L. Albumin LINs are assortative up to a very long sequence distance, while hemoglobin LINs loose assortativity at very short sequence distance. This is in line with the relative rigidity of albumin structure with respect to the hemoglobin flexibility, as mirrored in topological descriptors[27].

As can be observed, there is an overall decreasing slope in both protein networks till reaching a stability around L=150, meaning that for values of L greater than this threshold, the edges of the network are conserved, so the corresponding nodes are in contact even if they are far (at least 150 aminoacids) in the sequence. It is worth noting that 150 aminoacids is near to the maximal single domain length as computed in[34]. With regards to *HBAKD*

(Fig. **4.2**), we cannot refer to any *a-priori* classification but it is remarkable the different behaviour of the two systems. Albumin displays assortative behaviour (r>0.1) only for very-long-range interactions (L>150). This is reminiscent of the creation of inter-domains contacts that close the structure into a compact whole at the end of the folding process. This accounts for the above mentioned rigidity of the protein.

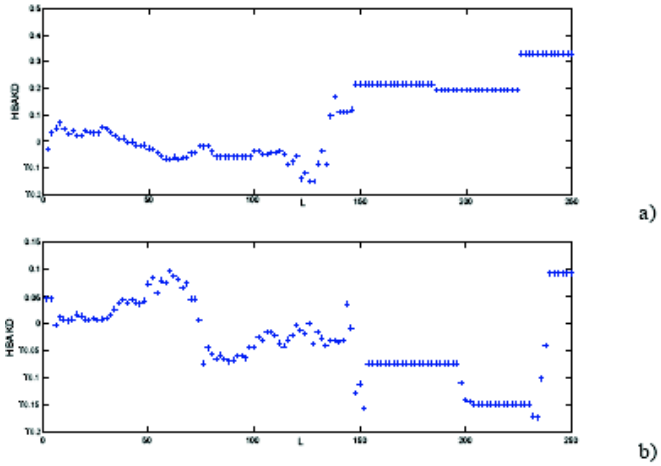


Fig. (4.2). *H BAK D* : a) serum albumin; b) hemoglobin.

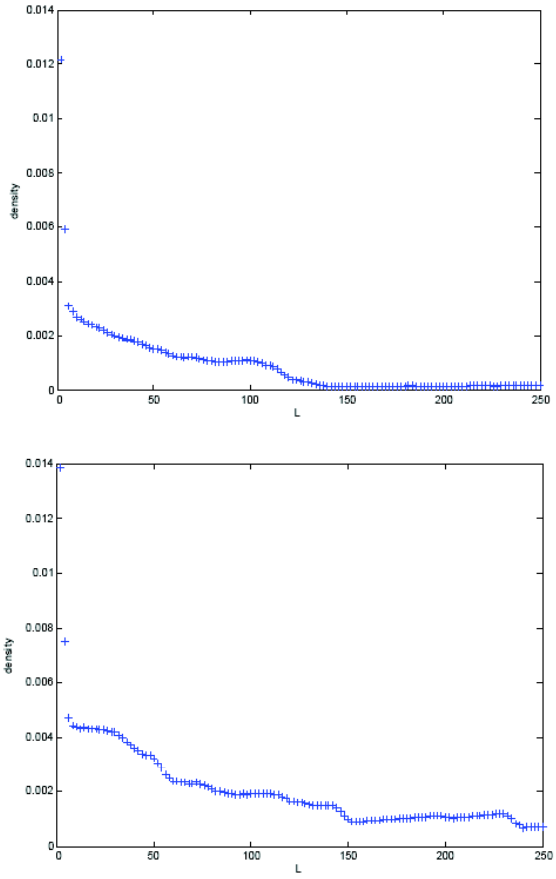


Fig. (4.3). Contact density as function of L: serum albumin (upper panel); hemoglobin (lower panel).

On the contrary, hemoglobin never reaches an assortative asset, consistently with the need of maintaining a large scale flexibility of the whole molecule. In any case, the discontinuity at L=150 is justified by the length of the chains constituting the quaternary structure.

The lack of hydrophobic assortativity for both structures in the case of low L values is consistent with a novel vision of protein molecules as open and sponge-like structures, very different from the old concept of compact objects, stabilized by the segregation of hydrophobic aminoacids in the inmost core[39].

The test protein set showed a very similar behavior that will be described in details in a following(LRO) paragraph.

3.2. Topological Indexes

Density endeavor along with L is strikingly similar for both proteins (Fig. **4.3**): there is a steep decrease in density at very low L values, while keeping a smooth variation up to L =150, where a smaller step can be observed, again reminiscent of the domain size. The very low density of long-range contacts witnesses the open and flexible attitude of protein molecules. Accordingly, a similar behaviour can be observed as for the average degree for both systems (Fig. **4.4**) and test protein set (data not shown).

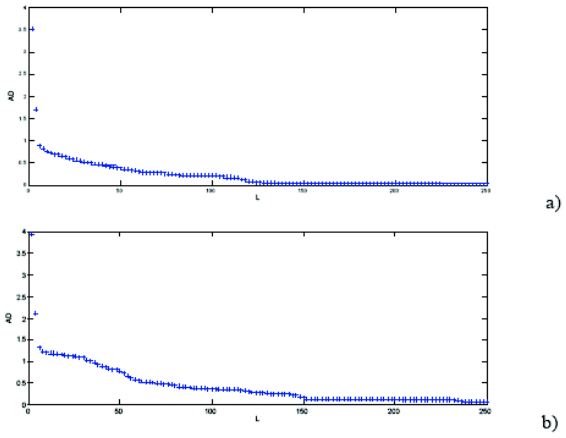


Fig. (4.4). Average degree as function of L: a) serum albumin; b) hemoglobin.

3.3. Long Range Order (LRO)

In Fig. (**4.5**), the *LRO* scaling with L is reported for both albumin and hemoglobin: the threshold at about 150 residues is remarkably clear and much more evident with respect to other indexes.

The same qualitative behaviour with L varying between 120 and 200 is evident as well in the test protein set (Fig. **4.6**). Keeping in mind we are dealing with proteins endowed with different size and quaternary structure (see Table 1), the presence of a largely invariant scaling of LINs wiring points to a common feature of protein structure organization.

3.4. Pictorial Sketch of Graphs

To complete the analysis, we report a graphical sketch of the protein contact networks for the two protein systems

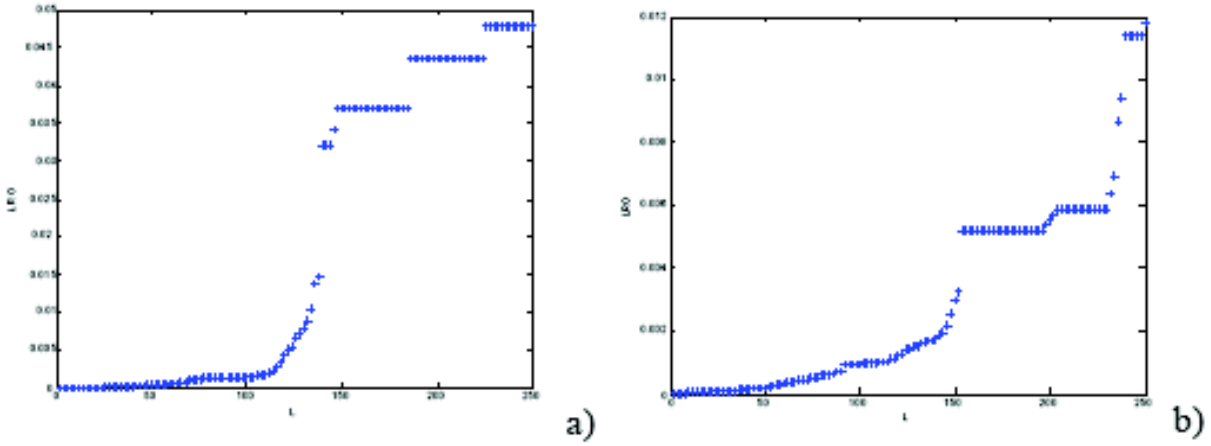


Fig. (4.5). LRO along with L: a) albumin; b) hemoglobin.

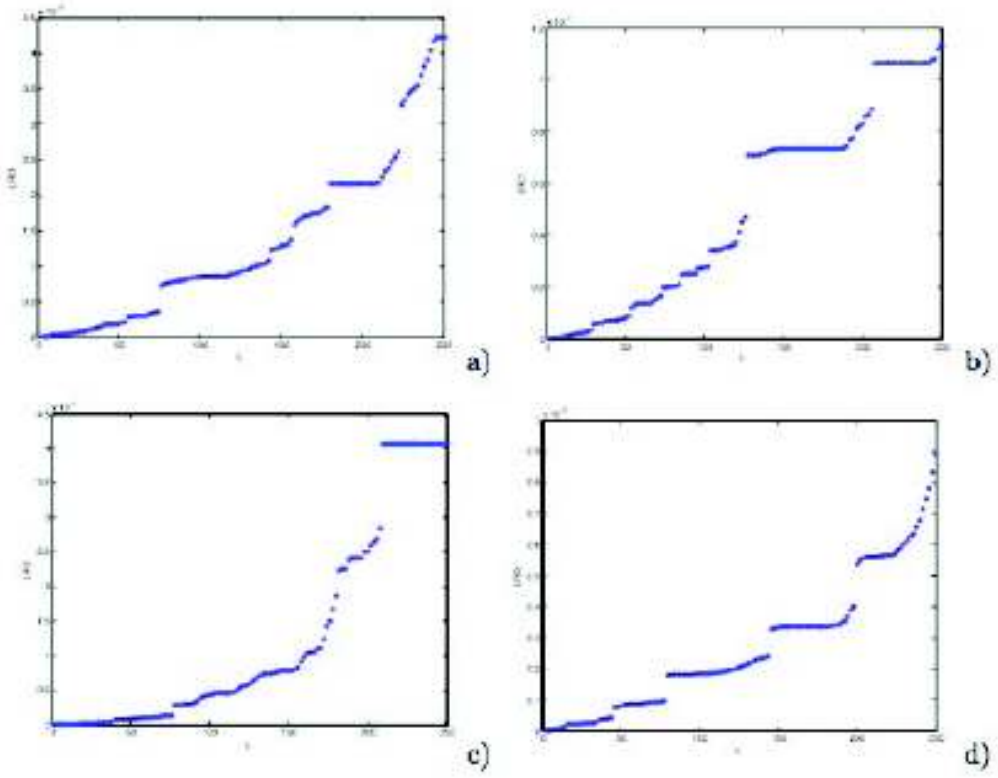


Fig. (4.6). LRO along with L: a) 1GPI Cellobiohydrolase, b) 1TUU Acetate Kinase, c) 8CAT Liver Catalase and d) 3GUU Lipase.

at different L (Figs. **4.7** and **4.8**); it is worth noting how both albumin and hemoglobin loose the total connectivity at very low L values, consistently with their average degree (Figs. **4.4**, **4.7** and **4.8**). This is in line with the results of [40], showing the main topological features of PCNs derive from the presence of a continuous backbone; as a matter of fact, the whole connectivity is lost as soon as L gets larger than 4,excluding the adjacent residues contacts, due to the peptide backbone. Relying on covalent, peptide bonds for network connectivity, allows for large motions of the proteins keeping alive the molecule integrity.

4. CONCLUSIONS

The novelty of our work relies on the link between functionality and topological descriptors. To our knowledge, this is the first computation of hydrophobicity protein based contact networks assortativity. Even if these are preliminary data, we expect this computation could be of use for protein folding studies. The difference in allosteric character of the two main systems was suggested to be the major determinant of the topological differences between them. The analysis, extended to four more proteins, showed the presence of a general 120-150 residues domain size, even

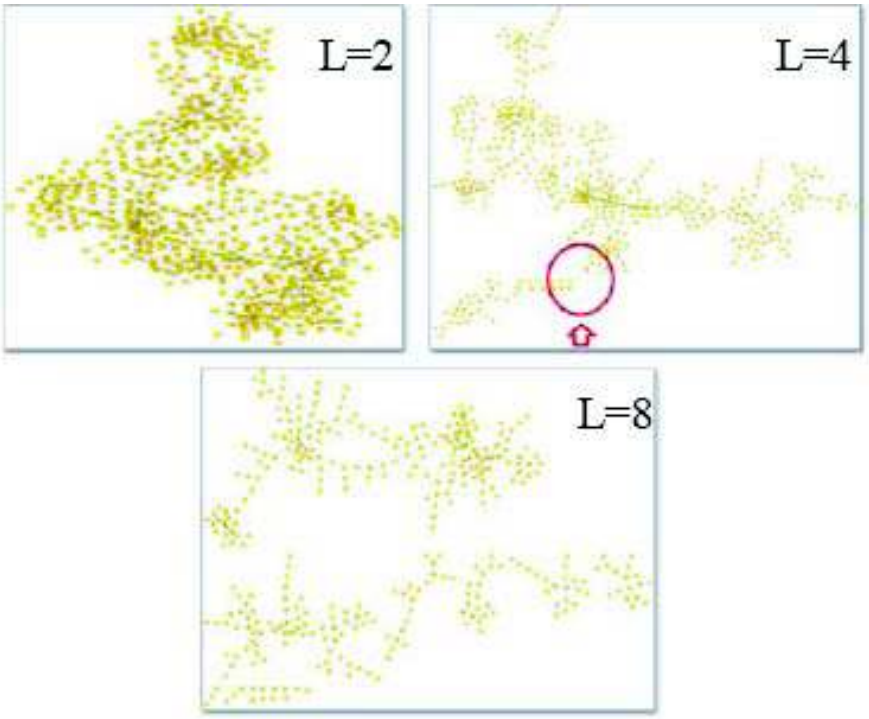


Fig. (4.7). LINs: albumin.

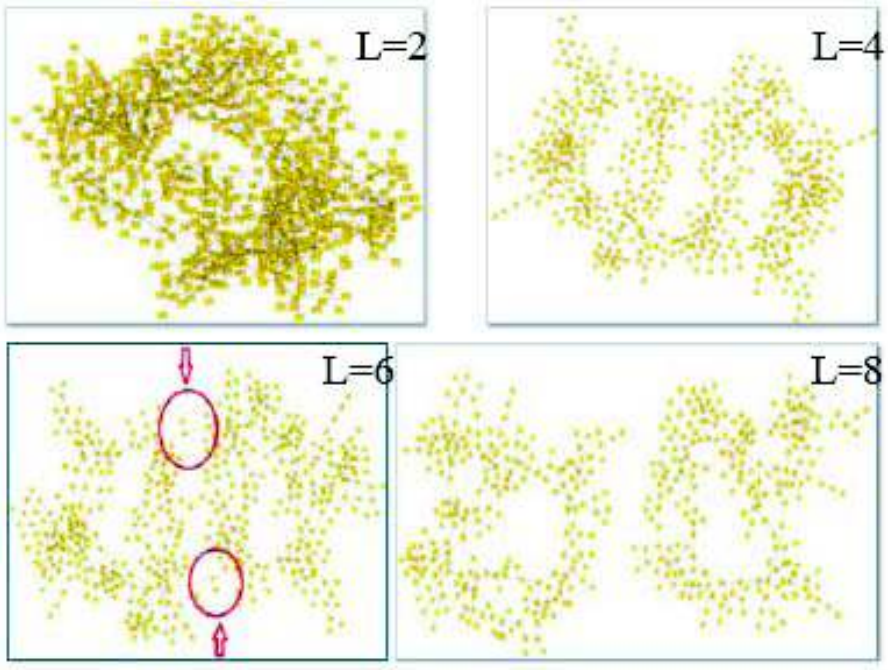


Fig. (4.8). LINs: hemoglobin.

if this threshold is accomplished in a totally different manner in the molecules. In some cases, the domain refers to the same chain, whereas in some other cases, different chains correspond to different domains. All in all, proteins under analysis share the same functionality class of binding proteins. It can be argued that the domain threshold corresponds to a "critical size" required to host a tunable binding site. Networks based approaches promise to allow us to discover still unknown universal principles of protein structure organization.

ABBREVIATIONS

CO = Contact order

DBA = Degree based assortativity

LIN = Long-range interaction networks

LRO = Long-range order

PCN = Protein contact network

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENT

We thank the "Consorzio inter-universitario per le Applicazioni di Supercalcolo Per Università e Ricerca " (CASPUR) for computing resources and support. This work is partially supported by the FLAGSHIP "InterOmics" project (PB.P05) funded and supported by the Italian MIUR and CNR organizations. We wish to thank the two anonymous reviewers for helping us to ameliorate our work.

REFERENCES

[1] Anfinsen, C. Principles that govern the folding of protein chains. *Science,* **973**, *181*, 223-230.

[2] Dobson, C. Protein folding and misfolding. *Nature,* **2003**, *426*, 884-890.

[3] Dobson, C. and Karplus, M. The fundamental of protein folding: bringing together theory and experiment. *Curr. Opin. Struct. Biol.,* **1999**, *9*(1), 92-101.

[4] Dill, K. *Biochemistry,* **1985**, *24*(6), 1501-1509.

[5] Dinner, A.; Šali,A.; Smith, L.; Dobson, C. and Karplus, M. Understanding folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* **2000**, *25*(7), 331-339.

[6] Fersht, A. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.,* **1997**, *7*(1), 3-9.

[7] Gianni, S.; Ivarsson, Y.; Jemth, P.; Brunori, M.; Travaglini A. and Allocatelli, C. Identification and Characterization of protein folding intermediates. *Biophys. Chem.,* **2007**, *128*(2-3), 105-113.

[8] Privalov, P. Intermediate states in protein folding. *J. Mol. Biol.,* **1996**, *258*(5),707-725.

[9] Shakhnovich, E. Theoretical studies of protein folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.,* **1997**, *7*(1), 29-40.

[10] Baker, D. A surprising simplicity to protein folding. *Nature,* **2000**, *405*, 39-42.

[11] Taketomi, H.; Ueda, Y. and Go, N. Studies on protein folding, unfolding and fluctuations by computer simulations. *Intl. J. Pept. Protein Res.,* **1975**, *7*(6), 445-459.

[12] Gromiha, M. and Selvaraj, S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate. *J. Mol. Biol.,* **2001**, *310*, 27-32.

[13] Gromiha, M. Importance of native state topology for determining the folding rate of two-state proteins. *J. Chem. Inf. Comp. Sci.,* **2003**, *43*, 1481-1485.

[14] Gromiha, M. and Selvaraj, S. Inter-residue interactions in protein folding and stability. *Progr. Biophys. Mol. Biol.,* **2004**, *86*(2), 235-277.

[15] Gromiha, M. Multiple contact network is a key determinant to protein folding rate. *J. Chem. Inf. Model,* **2009**, *49*, 1130-1135.

[16] Plaxco, K.; Simons, K. and Baker, D. Contact order, transitions state placement and the refolding rates of single domain proteins. *J. Mol. Biol.,* **1998**, *277*, 985-994.

[17] Sengupta, D. and Kundu, S. Protein contact networks at different length scales and role of hydrophobic, hydrophilic and charged residues in protein's structural organization *Arvix Preprint*, **2010**.

[18] Bagler, G. and Sinha, S. Assortative mixing in protein contact networks and protein folding kinetics. *Bioinformatics,* **2007**, *23*(14), 1760-1767.

[19] Newman, M. Assortative mixing in networks. *Phys. Rev. Lett.,* **2002**, *89*, 208701.

[20] Poupon, A. and Mornon, J. Poulation of hydrophobic aminoacids within protein globular domains: identification of conserved 'topo-hydrophobic' positions. *Proteins Struct. Funct. Bioinf.,* **1998**, *33*(3), 329-342.

[21] Dobson, C. Protein folding and misfolding. *Nature,* **2003**, *426*(6968), 884-890.

[22] Dobson, C. Principles of protein folding, misfolding and aggregation. *Semin. Cell Dev. Biol.,* **2004**, *15*(1), 3-6.

[23] Aftabuddin, M. and Kundu, S. AMINONET a tool to construct and to visualize amino acid networks, and to calculate topological parameters. *J. Appl. Cryst.,* **2010**, *43*, 367-369.

[24] Sengupta, D. and Kundu, S. Distributer computing and Networking. *Cornell University Library* **2010**.

[25] Gromiha, M.; Thangakani, A. and Selvaraj, S. FOLD-RATE: prediction of protein folding rates from aminoacid sequences. *Nucleic Acids Res.,* **2006**, *34*, W70-W74.

[26] Vendruscolo, M.; Paci, E.; Dobson, C. and Karplus, M. Three key-residues form a critical contactr network in a protein folding transition state. *Nature,* **2001**, *409*, 641-645.

[27] De Ruvo, M.; Giuliani, A.; Paci, P.; Santoni, D. and Di Paola, L. Shedding light on protein-ligand binding by graph theory: the topological nature of allostery. *Biophys. Chem.,* **2012,** *165-166,* 21-29.

[28] Munoz, I.; Ubahyasekera, W.; Henriksson, H.; Szabo, I.; Pettersson, G.; Johansson, G.; Wowbray, S. and Stahleberg, J. Family 7 cellobiohydrolase from Phanerochaete chrysosporium : crystal structure of the catalytic module Cel7D at 1.32 A resolution and homology models of the isozyme. *J. Mol. Biol*., **2001**, *14*(314), 1097-1111.

[29] Gorrell, A.; Lawrence, S. and Ferry, J. Structural and kinetic analyses of arginine residues in the active site of acetyate kinase from Methanosarcina thermophyla. *J. Biol. Chem.,* **2005**, *280*(11), 10731-10742.

[30] Fita, I. and Rossmann, M. The NADPH binding site on beef liver catalase. *Proc. Natl. Acad. Sci. USA,* **1985**, *82*(6), 1604-1608.

[31] Brinda, B. and Vishveshwara, S. A network representation of protein structures: implications for protein stability *Biophys. J.,* **2005**, *89*(6), 4159-4170.

[32] Greene, L.H. and Highman, V.A. Uncovering network systems within protein structures. *J. Mol. Biol.,* **2003**, *334*, 781-791.

[33] Gursoy, A.; Keskin, O. and Nussinov, R. Topological properties of protein interaction networks from a topological perspective. *Biochem. Soc. Trans.,* **2008**, *36*(6), 1398-1403.

[34] Zbilut, J.; Chua, G.; Krishnan, A.; Bossa, C.; Rother, K.; Webber, C. and Giuliani, A. A topologically related singularity suggests a maximum preferred size in protein domains. *Proteins Struct. Funct. Bioinf.,* **2007**, *66*, 621-629.

[35] Süel, M.; Lockless, S.; Wall, M. and Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins *Nature Struct. Biol.,* **2002**, *10*(1), 59-69.

[36] Vendruscolo, M.; Dokholyan, N.; Paci, E. and Karplus, M. Small-world view of amino acids that play a key role in protein folding *Phys. Rev. E.,* **2002**, *65*, 061910.

[37] Giuliani, A.; Di Paola, L. and Setola, R. Proteins as networks: a mesoscopic approach using hemoglobin molecule as case study *Curr. Proteomics,* **2009**, *6*(4), 235-245.

[38] Newman, M. *Networks: An introduction;* OxfordUniversity Press, USA, **2010**.

[39] Di Paola, L.; Paci, P.; Santoni, D.; De Ruvo, M. and Giuliani, A. Proteins as sponges: a statistical journey along protein structure organization principles. *J. Chem. Inf.,* **2012**, *52*(2), 474-482.

[40] Bartoli, L.; Fariselli, P. and Casadio, R. The effect of backbone on the small-world properties of protein contact maps *Phys. Biol.,* **2008**, *4*(4), 1-5.

Received: 00 00, 2012 Revised: 00 00, 2012 Accepted: 00 00, 2012

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Proteins as Sponges: A Statistical Journey along Protein Structure Organization Principles

Luisa Di Paola,[†] Paola Paci,[‡] Daniele Santoni,[¶] Micol De Ruvo,[§] and Alessandro Giuliani*[,∥]

[†]Università Campus Biomedico, via Alvaro del Portillo 21, 00128 Rome, Italy

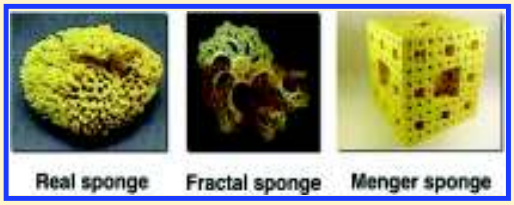
[‡]CNR-Institute of Systems Analysis and Computer Science (IASI), BioMathLab, viale Manzoni 30, 00185 Rome, Italy

[¶]CNR-Institute of Systems Analysis and Computer Science (IASI), viale Manzoni 30, 00185 Rome, Italy

[§]Università Campus Biomedico, via Alvaro del Portillo 21, 00128 Rome, Italy

[∥]Department of Environment and Health, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy

ABSTRACT: The analysis of a large database of protein structures by means of topological and shape indexes inspired by complex network and fractal analysis shed light on some organizational principles of proteins. Proteins appear much more similar to "fractal" sponges than to closely packed spheres, casting doubts on the tenability of the hydrophobic core concept. Principal component analysis highlighted three main order parameters shaping the protein universe: (1) "size", with the consequent generation of progressively less dense and more empty structures at an increasing number of residues, (2) "microscopic structuring", linked to the existence of a spectrum going from the prevalence of heterologous (different hydrophobicity) to the prevalence of homologous (similar hydrophobicity) contacts, and (3) "fractal shape", an organizing protein data set along a continuum going from approximately linear to very intermingled structures. Perhaps the time has come for seriously taking into consideration the real relevance of time-honored principles like the hydrophobic core and hydrophobic effect.



INTRODUCTION

In their extremely innovative and provocative paper, Banerij and Ghosh[1] explicitly state "A student of protein structure is constantly reminded of several myths prevalent in this paradigm. He (she, at any rate) studies that the globular proteins are so compactly packed that their interior mimics that of solids, but finds it a bit irreconcilable with reports of inhomogeneus packing in protein interior and the presence of cavities therein". The authors continue, claiming that the elusive character of the presence of the time honored "hydrophobic core" is the main driver of folding and suggest a much more realistic fractal folding of proteins, giving rise to objects more similar to sponges than to densely packed spheres.

Building upon the above statements and our previous experience demonstrating an exponential scaling of density of contacts rapidly fading away with increasing size,[2] in this work we approach a large scale statistical study to substantiate the correlations between different "shape" viewpoints of actual protein structures.

The concept of shape is one of the most fundamental (and consequently most elusive) concepts in science. The intuitive (while geometrically rigorous) definition of shape deals with the fulfillment of certain constraints linking the different dimensions of a given entity. Thus, a circle shape is defined by the fulfillment, by a set of points, of the constraint of an invariant distance from a special point called the center, while a triangular shape corresponds to a 180° sum of the internal angles formed by a set of three incident segments. When we move away from the world of regular geometrical entities, defining shapes becomes much less simple, and a wide spectrum of possible quantitative descriptors of the shape of natural objects comes into life. The case of proteins, with their diverse beautiful and intermingled three-dimensional architectures is paradigmatic of this multiplicity. Several methods have been proposed to characterize proteins shape, almost all of these methods focused on the proteins' surface representations, this was motivated by the fact that surface geometry plays the most relevant biological role because it delineates the interface between the molecule and its environment, i.e., the region where physiologically meaningful interactions take place.[3] As a consequence, protein shape has been often defined with reference to a finite set of points, a space curve, or a surface.[4] Among the wide variety of approaches proposed to describe a molecular surface, van der Waals surface (VdW) refers to the union of the atoms (modeled as balls) according to their van der Waals radii. The solvent accessibile surface (SAS) is a measure for quantitatively determining the interaction tendencies of a protein, delineated by the center of a probe sphere (typically a water molecule) rolling on top of the VDW surface. By removing a layer of solvent radius depth from the SAS model, the molecular surface (MS) can be obtained (Figure 1).

In these models, the surface of the molecule is depicted as a polyhedron, triangular facets link a triplet of surface atoms

Received: October 26, 2011
Published: January 11, 2012

   dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474–482

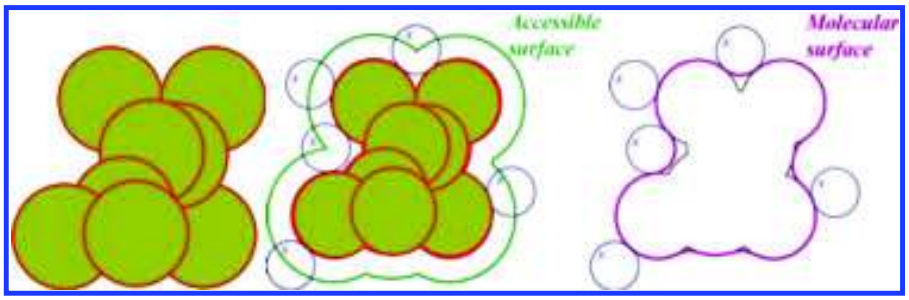


Figure 1. VDW, SAS, and MS surfaces.

blocking a probe sphere, whereas an edge links two atoms that allow the same probe sphere to roll from one blocking position to another. The vertices of the shape are exactly the atoms with a strictly positive accessible surface area.[4]

Anthony Hopfinger developed a "molecular shape analysis" on the basis of the comparison of electrostatic fields of small organic molecules[5] extended to proteins by Arteca.[6] The Hopfinger–Arteca paradigm gives a physical flavor (electrostatic forces) to the purely topological character of shape. In particular, shape descriptors have been exploited to individuate protrusions and cavities of a known input structure. The consideration of shape we adopted in this work, instead of being surface-based, concentrates on the idea of shape as a "three-dimensional distribution of mass", neglecting the explicit hidden/exposed separation of surface-based approaches. This separation, as we will explain, is implicitly taken into consideration in our work by two indexes (corrHBMJ and corrHBKD) corresponding to the quantification of internal/external asymmetry of hydrophobicity distribution in terms of the Pearson correlation coefficient between distance from the center of mass and relative hydrophobicity of residues. In her fundamental work on cell shapes,[7] Carole Heckman explicitly states "Earlier work suggested that certain restrictions apply to the geometric configuration that can be assumed by cells. Such restrictions were indicated by the finding that the values of certain shape descriptors were highly correlated with one another. This was surprising because these descriptors appeared to measure dissimilar geometrical properties of the cell. The present research confirms that the high levels of correlation are due to geometrical constraints on cell shape". This is a striking "back-to-the-root" simplification of the concept of shape: a given form is nothing more and nothing less than a set of constraints imposed on different shape descriptors (analogously to the circle and triangle examples quoted before), consequently the shape of complex objects like proteins (or cells in the case of Heckman) can be interpreted as a correlation structure imposed to a set of diverse shape descriptors. This is the classical way of spectroscopy. After all, a particular molecule (or mixture) is identified as a set of peaks, where the correlation is imposed to an initially independent set of descriptors (different wavelengths intensities) by the particular composition of the analyzed sample. Consistent to this "form-as-correlation" idea, the statistical paradigm we assumed was the principal component analysis in which the major fluxes of correlation of the studied descriptors correspond to the organization principles of protein structures.[8−10] This approach allows for a bottom-up, data-driven description of the main structural and topological determinants of the protein structures not imposing any particular theoretical frame on the data.[11−13] In this work, we demonstrate that the hydrophobic core is a pertinent concept only for very small proteins being linked to the microscopic (few residues) but not macroscopic protein structural organization, giving a proof-of-concept to both the Banerji–Ghosh hypothesis and to our previous contact density scaling results.[1,2] Local structure (single contacts) was demonstrated to be the place for the :hydrophobicity-coupling" of classical folding models.[14] This local organization was captured by a component of its own independent from global structure organization. Fractal dimension was independent from the above-sketched global and local organizations, while being connected with another component of shape description, namely, the relative fibrous/globular shape of proteins. This tripartite description of protein topology and structure into "size", "local contacts", and "shape complexity" was demonstrated to be strictly linked to the formation of cavities and promises to be of use for the elucidation of interesting aspects of protein physiology and dynamics.

RESULTS AND DISCUSSION

Descriptive Statistics. Before entering the "correlation business" to model the emerging relations linking the variables, it is worth considering the univariate distribution of each descriptor (Table 1) to get a dimensional idea of the variance/

Table 1. Simple Statistics

variable	*N*	mean	std dev	minimum	maximum
corrHBKD	911	−0.085	0.083	−0.548	0.196
corrHBMJ	911	−0.1408	0.1041	−0.526	0.089
N	911	613.698	656.130	50	6894
E	911	403023	1317913	1225	23760171
AS	911	0.306	0.163	0.003	0.938
R_G	911	12.775	5.138	4.76	48.197
R_{Gh}	911	12.233	5.287	3.973	47.878
R_{Gp}	911	13.062	5.076	4.915	48.343
MFD	911	2.356	0.403	0.930	4.148
H	911	0.982	0.044	0.805	1.209
D	911	1.361	0.134	1.008	2.306
HBAKD	911	0.005	0.038	−0.152	0.141
HBAMJ	911	0.045	0.035	−0.14	0.173
DBA	911	0.156	0.07	−0.118	0.373

invariance of each index in the considered data set. Obviously, we expect a relevant range of variation candidates for a given descriptor to become a major order parameter of the data set (i.e., a measure potentially effective to discriminate different protein architectures), while a relatively invariant descriptor indicates a measure that is almost identical across protein structures. Thus, while potentially important for its common

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

role in all proteins, it is not endowed with classification ability. Our PCA-based approach clearly favors the "wide range variation" descriptors. As for hydrophobic core, we can safely say, according to ref 1, that is a myth or in any case a very "low power" concept to look at protein structures at large. Both Kyte−Doolitle and Myiazawa−Jernigan[15] based the Pearson correlation between hydrophobicity and the center of the protein (coorHBKD and corrHBMJ) very low on average (−0.08 and −0.14, respectively), accounting for 0.64% and 0.20% of residue localization in protein structures, respectively. It is worth noting the big spread (especially for corrHBKD) of this index going from frankly negative, consistent with the existence of hydrophobic core values (minimum = −0.548), to paradoxically positive (maximum = 0.196) values.

Size variables were designed in order to get the largest possible variation range to allow a significant scaling of the descriptors with protein dimensions. Thus, the smallest protein of our collection is 50 residues long, while the larger one has N = 6894. This range of variation, for algebraic reasons, is still greater for E. The fact these two size descriptors vary over three (four in the case of E) orders of magnitude makes them less manageable as synthetic size measures with respect to radius of gyration variables (R_G, R_{Gh}, R_{Gp}). These variables are constrained into much more manageable variation ranges (approximately from 4 to 50). This mathematical well conditioning makes R_G variables the most correlated with the emerging "size component" (PC1) with which they will be shown to have almost unitary correlation coefficients (loadings, Table 2).

Table 2. Loading Matrix of Components[a]

variables	PC1	PC2	PC3	PC4
corrHBKD	**0.604**	−0.278	0.323	**-0.441**
corrHBMJ	**0.767**	−0.258	0.271	−0.171
N	**0.870**	0.172	−0.357	0.039
E	**0.702**	0.180	**-0.411**	0.101
AS	0.202	−0.203	**0.654**	**0.34**
R_G	**0.970**	0.060	−0.041	0.071
R_{Gh}	**0.974**	0.040	−0.037	0.068
R_{Gp}	**0.966**	0.070	−0.043	0.071
MFD	−0.296	0.258	**-0.729**	−0.152
H	−0.135	**-0.702**	−0.204	0.400
D	−0.224	**0.419**	0.337	−0.236
HBAKD	0.026	**0.483**	0.137	**0.649**
HBAMJ	0.030	**0.865**	0.314	−0.022
DBA	0.112	0.030	0.007	**-0.356**

[a]Bolded values represent the loadings of the variables most relevant for interpretation of each component.

As for the general shape variables, AS goes from perfect spherical symmetry (minimum = 0.003) to a straight linear structure (maximum = 0.94), spanning the entire theoretical asymmetry space with a small bias toward more globular shapes (mean = 0.30). This wide range of variation candidates AS as a possible order parameter of the data set. Fractal dimension (MFD) has an average value of 2.36 that is in strict concordance with previous results.[1] We inserted in our data set (despite the general invariance of this descriptor, getting a standard deviation of 0.40, that confirms the existence of a typical "fractal signature" of protein molecules) some extreme variants going from a totally linear object (minimum = 0.93) to an extremely intermingled structure (maximum = 4.15). It is worth noting that the lowest outliers are in some way artificial constructs (1mow, MFD = 0.93, artificial endonuclease; 1gk4, MFD = 1.10, Vimentin coil), while the paradoxical four dimensions object at the MFD maximum (2pn8, MFD = 4.15, thiredoxin peroxidase) gives rise to an extremely complex quaternary structure that can be hardly recognized as a single unitary object, being more similar to a "magic circle" of relatively loosely connected elements. In Figure 2, these extreme

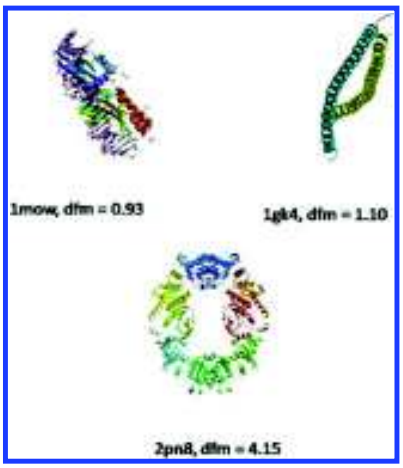


Figure 2. Example of extreme protein structures.

structures are reported. The "normal" range of variation (95% confidence limits over the data set) of MFD goes from 1.71 to 3.01, consistent with the general meaning of fractal dimension.

Topology descriptors were based on the proteins considered as contact networks, where α-carbons of the residues represent the network nodes and their thresholded (see Material and Methods) mutual distances the edges.[16] Besides DBA, that is based solely on graph invariants without any reference to the chemico−physical characterization of nodes (residues), these descriptors all redound to the influence of hydrophobicity on the probability of contacts. In the cases of HBAKD and HBAMJ, the departure from complete independence between hydrophobicity and contact probability is expressed by correlation coefficient metrics. Zero value correlation points to the complete independence of hydrophobicity and contact probabilities, while a negative value implies a tendency toward dissimilar hydrophobicity contacts, and a positive value implies a tendency toward similar hydrophobicity contacts. It is worth noting a substantial independence between hydrophobicity and contact probability with a very minor tendency of Myiazawa−Jernigan (HBAMJ) of preferring similar contacts over HBAKD, reminiscent of the character of statistical potential of the MJ index.[15] Notwithstanding the almost perfect "average" independence of hydrophobicity and contact probability on the entire data set, both HBAKD and HBAMJ display a nontrivial range of variation going approximately from −0.15 to 0.15 that will give rise to the second principal component (PC2), thus representing the second most relevant order parameter shaping protein structural organization. Similar considerations hold for H and D descriptors, in which the perfect independence corresponds to unitary value. In the case of DBA, we are dealing with a pure "architectural" principle related to the tendency of high degree (high number of contacts) nodes to be connected to each other (negative values of DBA,

again over a correlation coefficient metrics having −1 and +1 as theoretical extremes) or to be actively kept apart (positive values). Again, we have a substantial average symmetry of the two choices (mean DBA = 0.16) but with a multiplicity of architectural solutions going from DBA = −0.11 to DBA = 0.37. In the presence of a strong common compact core we expect a positive DBA, while a negative DBA is a marker of a very distributed architecture with almost independent "local cores" for different parts of the structure.

Correlation Structure and the Emerging of General Organizational Principles. Having described in the previous paragraph the univariate structure of the different descriptors of the data set focusing attention on both the location (mean values, "typical protein" view) and variability (standard deviation, range, "ordination of the protein set" view), we now shift to the most relevant part of the work: the emerging of consistent correlation fluxes (principal components) shaping the entire data set collecting the consistent variation across multiple descriptors. The fact that principal components are each other independent by construction[17] and globally provide the maximum parsimonious representation of the data set reassures us these components represent the basic independent factors describing protein molecules configurations. The interpretation of the loading matrix (being the loadings of the correlation coefficients of the original variables with the components) allows us to sketch an interpretation of the meaning of the extracted components. This interpretation will in turn be verified by its ability to predict "external variables" that did not enter into the component construction and that in this case were the two eps and eps_1 scores. The original 14 dimensions variable space, when analyzed by means of PCA, collapsed to a four component bona fide signal solution,[13,18] globally explaining the 71% of total variability, with a by far most relevant order parameter (PC1) responsible for the 37.4% of variance. The scaling of eigenvalues with component number is reported in Figure 3, where the threshold between

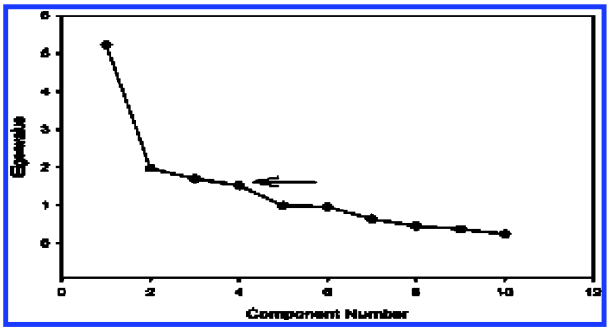


Figure 3. Saling of eigenvalues with component number. The arrow marks the threshold between "estimated signal" and "noise floor", substantiating our choice of four components to be analyzed.

"estimated signal" and "noise floor" is marked by an arrow. The loading matrix is reported in Table 2; the loadings of the variables most relevant for the interpretation of each component are bolded.

In Table 3, the percentage of variation explained by each component is reported. The near to maximal correlation of PC1 with size variables allows us to identify this component as "protein size". The fact that size is the main order parameter shaping the data set is a consequence of the presence of similar

Table 3. Percentage of Explained Variance by Each Component

PC1	PC2	PC3	PC4
37.4	14.1	12.2	8.2

construction principles common to all protein molecules. It is worth noting physical size (R'_GS) is much more relevant (near unit loadings) for the component meaning than indirect indexes, as N or E (edges). This is an emerging property of the PCA solution reassuring us of the "well conditioning" of the component space that is not driven by the presence of outliers

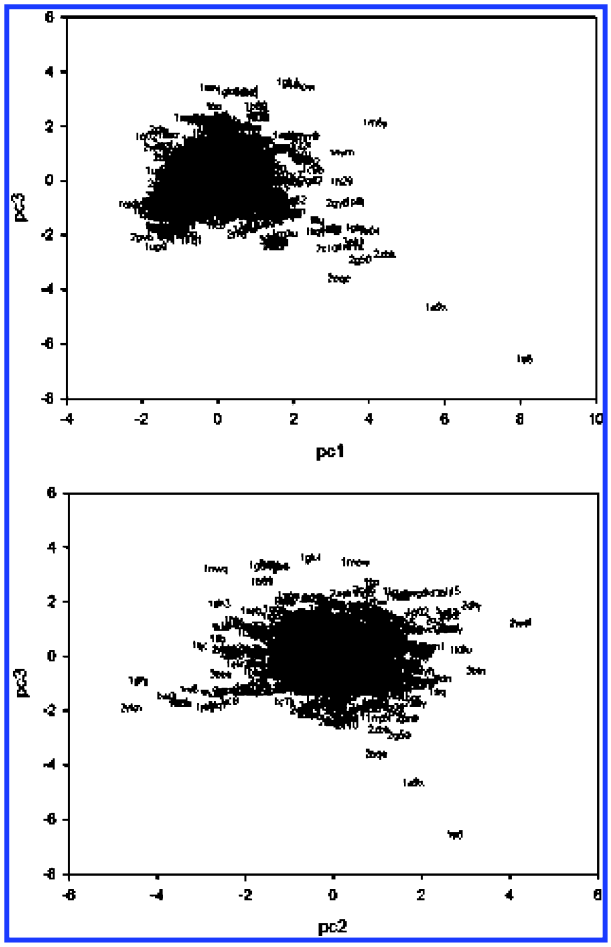


Figure 4. Two views of the component space are reported. PC1-PC3 in the upper panel and PC2-PC3 in the lower panel.

but by a consistent scattering of the entire data set in the component space. This is confirmed by a simple look at Figure 4, where two views (PC1-PC3 and PC2-PC3 planes) of the component space are reported. The (relatively few) outliers go together with a globally continuous distribution of the component scores of the entire protein data set. As for the size component (PC1), it is worth noting both corrHBKD and corrHBMJ have a relevant positive loading on the component. This comes from the fact the "hydrophobic core" hypothesis is only tenable for very small proteins, while it loses any relevance (with the consequent converging of corrHBKD and corrHBMJ to zero) at increasing size. This interpretation is

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

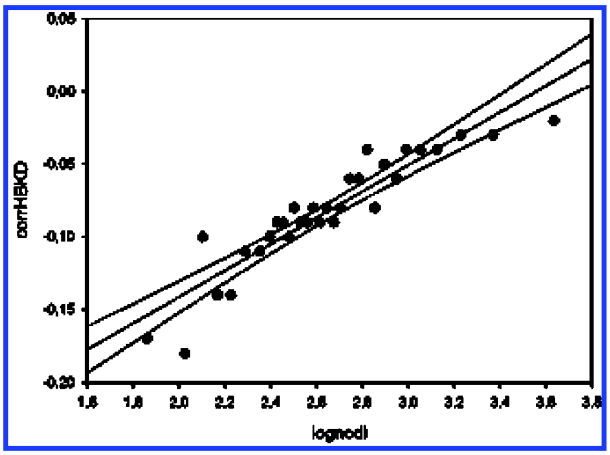


Figure 5. Average corrHBKD value is plotted vs average log(N). Here, corrHBKD = −0.32 + 0.091 × log(N), and the correlation between corrHBKD and the size is r = 0.92.

given a proof-of-concept in Figure 5, where the relation between size (expressed as logarithm of the number of residues) and corrHBKD is reported. In order to normalize the scattering of corrHBKD present in small molecules, we applied to the data a smoothing procedure. The protein data set was ordered along increasing N and partitioned into nonoverlapping sets of 30 molecules each. For each set, the average corrHBKD value is plotted versus the average log(N), obtaining a striking correlation between corrHBKD and size (r = 0.92) confirming our hypothesis. This result allows us to hypothesize that the origin of the "myth" of the "hydrophobic core" dates back to an age where the protein molecules whose structure was known was largely biased toward the "small proteins" end. The fading away of a closely packed hydrophobic core with size is consistent with the demonstrated progressive decrease in the density of contacts with the generation of less dense (and consequently with an increasing number of voids) architectures for larger proteins.[2]

The second component (PC2) explains the 14.1% of total variance and is clearly linked to topology hydrophobicity-based variables (H, D, HBAKD, HBAMJ). Basically, PC2 orders protein structures along an axis going from the prevalence of "opposite polarity" (low component scores) to a prevalence of "same polarity" (high component scores) contacts. As we stressed in the previous paragraph, the extremes of positive/negative assortativity go from −0.15 to 0.15 (with a theoretical range going from −1 to +1); nevertheless; this difference is sufficient to give rise to a relevant component of protein organization, pointing to a different relative balance of microscopic forces (hydrophobic effect, hydrogen bonds, electrostatic forces, etc.) intervening in different proteins folding mechanisms. This different relevance of physical forces, in our opinion, deserves a more detailed investigation given the implicit unitary character of force fields approximation used in simulation and theoretical studies. In any case, this local topology component is independent from both size and general shape (PC3) that in turn explains 12.2% of total variance. AS and MFD are the two variables leading PC3 with linear structures (high asymmetry), being less complex than globular structures (AS and MFD enter PC3 with opposite loadings). All in all, PC3 is a "general shape" or "relative complexity" index with low scores (high MFD) pointing to very complex structures and high scores to less complex architectures.

We were not able to attach a clear meaning to PC4; the loading profile points to very specific architectural constraints possibly linked to class-specific features. It is worth noting that DBA has a non-null loading only on this component.

Modeling with Components: The Prediction of Cavities and Empty Spaces. The protein structure picture emerging from the above results can be metaphorically imagined as a sponge with increasing space for cavities at increasing size and complexity of the molecule. This is actually what happens with real sponges that in turn can be considered as fractal, large-scale invariant, objects. Figure 6 reports a real

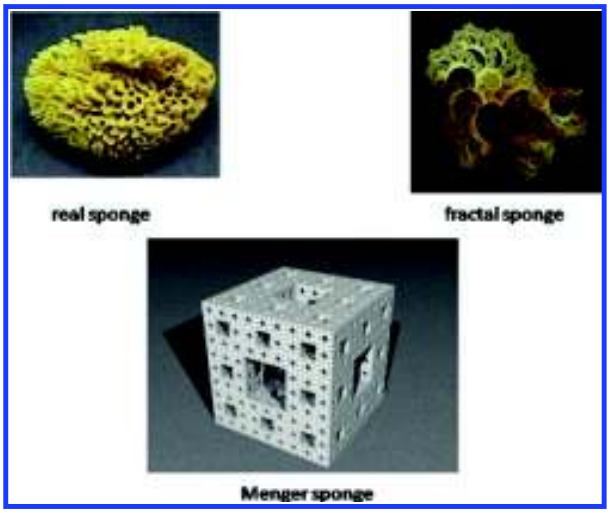


Figure 6. Natural and mathematical sponges.

sponge from the Mediterranean Sea, a so-called "fractal sponge" computationally generated, and the so-called "Menger sponge",[19] a mathematical 3-D object, which is a 3-D generalization of the unidimensional Cantor set, obtained by application of the same iterative void/dense rule at different scales. According to the "proteins as sponges" model, we expect PC1 and PC3 (i.e., size and fractal dimension) must be able to efficiently model the eps (or eps_1) variables quantifying the relative amount of voids (cavities) in the molecular volume. This comes from the consideration that the number of voids increases with both size (PC1) and fractal complexity (PC3). On the same heading, the microscopic topology−hydrophobic organization principles collected in PC2 must not have any role in modeling voids. This was actually the case; eps and eps_1 were maximally correlated between them (Pearson r = 0.96), giving a proof-of-concept of the substantial robustness of void estimation with different measurement paradigms. PC1 and PC3 were both significantly correlated to eps_1 (Pearson r = 0.69 and 0.41 for PC1-eps_1 and PC3-eps_1, respectively), while PC2 scored a very low −0.14 Pearson r value with eps_1. When we modeled eps_1 by both PC1 and PC3 (whose effect in the estimation of eps_1 can be supposed as purely additive given the components are each other linearly independent by construction) using a linear multiple regression paradigm, we obtain a striking Pearson correlation r = 0.80 between

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

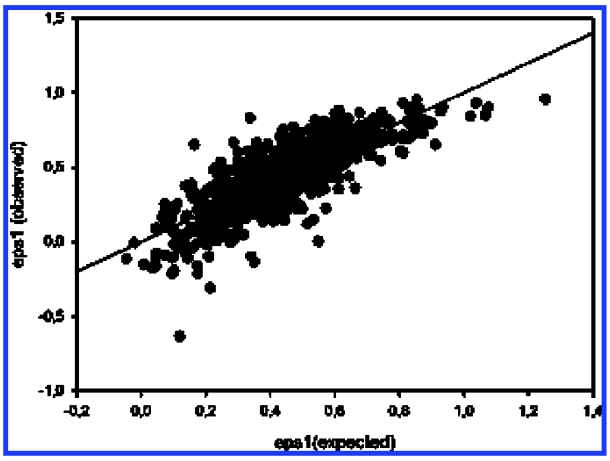


Figure 7. Each point represents a protein with the expected eps_1 on the horizontal axis (given by $eps_1 = 0.44 + 0.15 \times PC1 + 0.09 \times PC3$) and the observed eps_1 on the vertical axis. The correlation between observed eps_1 and expected eps_1 is $r = 0.804$ ($p < 0.0001$).

estimated and observed eps_1. Figure 7 reports model fitting and the correspondent linear equation linking voids to PC1 and PC3 scores. PC1 has a larger role than PC3 in void estimation, as expected by the prevailing role of size over other order parameters in shaping our data set. On the other hand, we need both the PC1 and PC3 "organizing principles" to efficiently model the portion of empty volume of the different proteins.

Table 4. Variables Describing Each Protein Structure

size	hydrophobic core	general shape	topology
N	corrHBKD	AS	H
E	corrHBMJ	MFD	D
R_G			HBKD
R_{Gh}			HBMJ

CONCLUSION

The effective prediction of empty space volume of 911 proteins by means of the extracted components gives a proof-of-concept of the relevance of the organizing principles we derived from PCA. Our results allow for a refinement of the general concepts used for describing proteins and cast doubts on firmly accepted concepts as the existence of hydrophobic core and the presence of common energetic principles for protein folding. A puzzling result of our analysis is the emerging of a component (PC2) ordering proteins from a preference for heterophilyc (low values of PC2) to homophilic (high values of PC2) contacts between adjacent residues. If only a single alternative was the preferred one, shared by all protein systems, we should not observe any component related to this mechanism (principal components follow the directions of maximal variance of the data set) on the contrary, the second most relevant organizing principle after size is the nature of local contacts. Using an architectural metaphor we can equate PC1 (size) and PC3 (shape) to global features of a building (its dimensions and form respectively) while PC2 has to do with the nature of the forces keeping it together at the microscopic/mesoscopic level (bricks disposition, junctions, arches...). These components, exactly like in architecture, are each other independent and while we have a clear appreciation of what size and general shape are, we are still in trouble to have a decent rationalization of the emerging fact the relative importance of physical forces driving folding (at the basis of the local contacts) can vary among different proteins. On the other hand the general statements set forth by Banerij and Ghosh[1] are largely confirmed by our analysis that in turn was able to confirm and give an explicit quantification to the fractal scaling hypothesis of the authors.[1] On another heading, the fractal sponge-like structure of proteins with water filled cavities asks for different paradigm of protein dynamics and physiology.

MATERIALS AND METHODS

We started from a repository of 1000 sequences, provided by the most widely used protein structure classification system

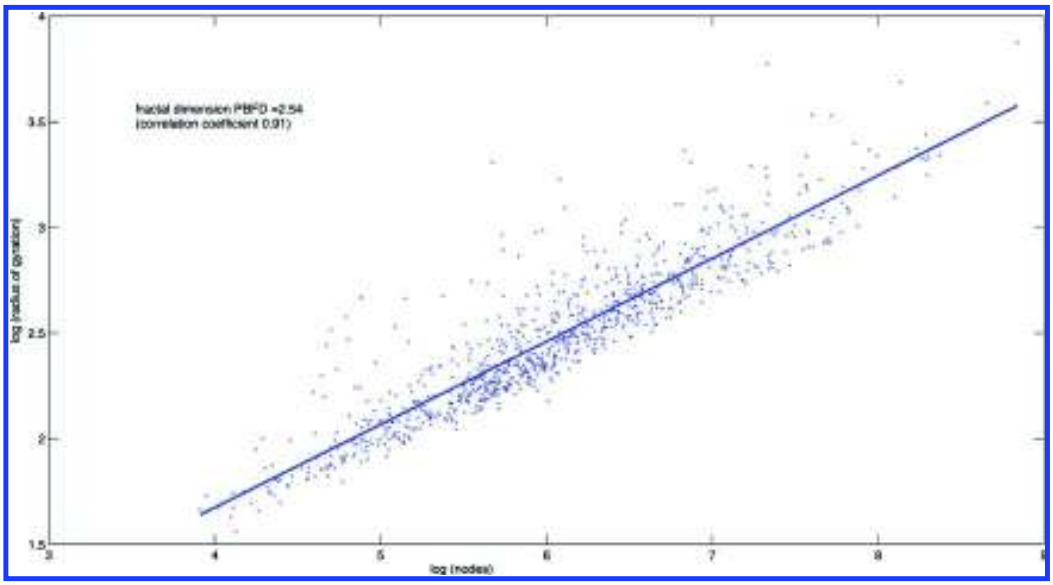


Figure 8. Scaling of protein radius of gyration R_g with the number of residue N (protein backbone fractal dimension).

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

CATH (SHREC'10, http://www.loria.fr/mavridis/SHREC_10); the superfamilies are randomly selected from CATH v3.3.

For each sequence, we extracted the corresponding PDB code, on which the analysis has been performed. The presence of a wide range of CATH classes guaranteed the consistency of analysis. The overall data set comprises 911 protein structures because several sequences belong to the same quaternary protein structure.

Each protein structure was described by means of 14 variables that were in turn submitted to a principal component analysis (PCA) and by two external variables consisting in two indexes (eps, eps_1) proportional to the "amount of cavitation" of the structures that were modeled by the extracted components to give a "proof-of-concept" of the components interpretation. The variables can be roughly classified into a-priori "groups" as shown in Table 4.

Topological parameters refer to the protein contact graphs that have been attained starting from the structural information embedded into PDB files, i.e., the spatial positions for all atoms in protein molecules.[16] We extracted the position of α-carbons representing the whole residue location; the distance matrix $\mathbf{d} = \{d_{ij}\}$ has been computed, the generic element d_{ij} being the Euclidean distance in the 3-D space between the i-th and j-th residues (holding the sequence order). Once the spatial distances are known, the corresponding contact graph has been obtained. Residues are the graph nodes, and links represent between residue contacts. A contact is established when a between residue distance is in the range 4−8 Å. Thus, the corresponding adjacency matrix $\mathbf{A} = \{a_{ij}\}$ is defined as

$$a_{ij} = \begin{cases} 1 & \text{if } d_{ij} \in [4-8] \\ 0 & \text{if } d_{ij} \notin [4-8] \end{cases}$$

The consideration of the protein structure as such a simplified graph was demonstrated to be very effective allowing for reconstructing the basic features of protein configuration in space as secondary structure,[20] proteins subunits and structural domains,[16,21,22] active sites,[23] residues relevant for protein folding kinetics and general stability of proteins,[24,25] allosteric signal pathways,[26] and many others.

As follows, a detailed description for each variable is provided.

- N: number of residues
- E: number of possible edges in the case of fully connected network, $E = N(N-1)/2$
- R_G: radius of gyration of the whole protein structure, defined as

$$R_G = \sqrt{\frac{1}{2}\frac{\sum_{i=1}^{N} m_i r_i^2}{\sum_{i=1}^{N} m_i}} \quad (1)$$

where m_i is the molecular mass of the i-th residue, and r_i is the corresponding distance (referred to the α-carbon position) from the center of mass, whose coordinates are the mass-weighted average of the α-carbons values.

- R_{Gh} and R_{Gp}: radius of gyration of hydrophobic and polar residues, respectively, defined similarly to eq 1, but it includes only the hydrophobic and polar residues contribution[27]
- corrHBKD and corrHBMJ: the hydrophobic core notion relies on a nonuniform distribution of hydrophobic residues that should stand closer to the center of mass of the whole protein structure. To test this hypothesis, we computed the Pearson correlation coefficients between two vectors, one reporting the distance from the center of mass for each residue and the other the corresponding hydrophobicity value (we applied this method to the Kyte−Doolittle hydrophobicity, corrHBKD, and Miyazawa−Jernigan scores, corrHBMJ); a strong negative correlation between hydrophobicity and distance from the center of mass would be proof of the core presence.
- AS: asymmetry. This is the departure from equality of the major displacements on the X,Y,Z of protein structures; high values of AS point to the existence of a major "elongation" axis and thus to linear (fibrous) protein; low values point to symmetric and thus very globular structures. AS is defined such that it is comprised between 0 and 1, with the values close to 0.5 corresponding to discoidal molecules.
- MFD: protein fractal nature has been recognized as a useful tool to interpret structural data and molecule function. On the other hand, many scaling laws apply to different protein properties (for a detailed review of protein structure fractal nature see ref 28). Mass fractal dimension (MFD), also known as Haussdorff dimension, refers to the inner structure of proteins. It has been computed by evaluating the mass comprised within spheres of varying radii into the protein 3-D structure; if the structure is not a compact, yet a fractal object, the following dependence shows up

$$M \sim R^{MFD} \quad (2)$$

The value of MFD = 3 corresponds to compact objects. MFD has been computed measuring the mass of residues contained within spheres of different radii, R, centered on the protein center-of-mass. Specifically, we chose 20 different radiuses equally spaced in the range of $[\overline{d}_{CM} - 10 \times \delta, \overline{d}_{CM} + 10 \times \delta]$, $\overline{d}_{CM}$ being the mean value of the distances of residues from the center-of-mass of the whole protein structure, with $\delta = \overline{d}_{CM}/400$ corresponding to the steps of different values of R .

In our work, we state proteins are more similar to sponges than to compact spheres; this is something more than a metaphor. A well-known mathematical model for fractal sponges is the Menger or the Menger−Sierpinsky sponge.[29] It is built up by applying an infinite number of operations of volume subtraction. Starting from a cube whose faces are divided into nine squares, the primitive cube is parted into 27 smaller cubes. Those placed at the center of every face (6) plus that located at the very center of the cube are eliminated. Then, the initial volume has been reducted of a fraction 20/27. By repeating this operation for each resulting cube, a sponge-like object is obtained. The mass fractal dimension and the void fraction, or porosity, corresponding to the n-th stage are therefore

$$d_f^{(n)} = \frac{\log 20^n}{\log 3^n} = \log_3 20 \simeq 2.73 \quad (3)$$

$$\text{eps}^{(n)} = 1 - \left(\frac{20}{27}\right)^n \quad (4)$$

It is worth noting that sponge fractal dimension is in the same order of magnitude as proteins. Moreover, this mathematical object allows for a rigorous definition of porosity that

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

is exactly what we did in our study. We computed the MFD for each single protein, but a very similar scaling behavior could emerge also analyzing the whole protein population. In this case, the radius of gyration R_G of the protein scales with the number of residues N as[30]

$$R_G \sim N^{\nu} \tag{5}$$

ν is a fractional scaling exponent that depends on the residue−solvent interactions. In a "good solvent", protein residues interact preferentially with the solvent molecules rather than with each other. The protein structures are then stretched into the solvent environment, and the corresponding value for ν is 3/5.

The fractional exponent ν is found to be the inverse of the protein backbone fractal dimension (PBFD) describing the scaling relationship between polymer length and number of residues, interpreted as rulers of fixed unitary length.

Analyzing protein backbone scaling (eq 5) on the whole set of 911 proteins, in order to determine the protein backbone fractal dimension, we obtained PBFD = 2.54 (Figure 8). The corresponding protein fractal dimension PBFD = 2.54 is in good agreement with both literature data[30] and our MFD computations. This gives further strength to our sponge-like model demonstrating that a large part of the residue is in contact with solvent.

- H and D: protein graphs describe the connectivity of residues (nodes) notwithstanding of the specific chemico−physical nature of each residue. In order to correlate topological descriptor to chemico−physical properties, combined descriptors are required, accounting for both classes of properties (topological and chemico−physical ones). Given a key physical property (for instance, hydrophobicity), if nodes show an attitude to preferentially established links with other similar nodes, the network is named dyadic; otherwise, if nodes preferentially link to dissimilar ones, the network is said antidyadic.[31]

  Let n_1 and n_0 denote, respectively, the number of node possessing or not a specific property (discretized hydrophobicity, in the case of point); e_{10} and e_{11} represent the number of edges connecting homologous or heterelogous nodes, respectively. The heterophilicity score H is then defined as

$$H = \frac{e_{10}}{e_{10,r}} \tag{6}$$

where $e_{10,r}$ is the expected value in the case of uniform distribution of the property among nodes, that depends on E. It is finally found

$$e_{10,r} = E \times n_1 \times (N - n_1) \tag{7}$$

Analogously, as for the homologous contacts, it is defined the dyadicity D as

$$D = \frac{e_{11}}{e_{11,r}} \tag{8}$$

and the corresponding value for uniform distribution is

$$e_{11,r} = E \times \frac{n_1 \times (n_1 - 1)}{2} \tag{9}$$

- Assortativity: this is another index for the proneness of nodes to connect to other nodes possessing similar features.[32] It is computed as the Pearson correlation coefficient between the two vectors containing a selected property for pairs of incident nodes in a network. For instance, in some networks, high-degree nodes preferentially connect to other high-degree nodes (assortative networks), whereas in other types of networks high-degree nodes preferentially connect to low-degree ones (disassortative networks). For the first class of networks, the correspondent assortativity index is close to 1, while it is close to −1 for the second class of networks.

We extended the analysis also to hydrophobicity related assortativity. Specifically, we adopted two kinds of scores defined as follows

- HBAKD, HBAMJ: hydrophobic-based assortativity correspondent to the Pearson correlation coefficient between incident residue hydrophobicity lists; hydrophobicity scores are based on two scales: Kyte−Doolittle and Miyazawa−Jernigan.
- DBA: degree-based assortativity that is Pearson correlation coefficient between incident residues computed over their respective degree (namely, the number of contacts each residue engages in the 3-D structure)
- eps and eps_1: porosity or void fraction is a parameter featuring the structure of porous, fractal media.[33] On the molecular scale, this definition applies to porous biomacromolecular structures;[28] hence, we defined two slightly different void fractions for the protein volume

$$\begin{aligned} \text{eps} &= \frac{V_f}{V} \\ &= 1 - \frac{V_{\text{residue}}}{V};\ \text{eps}_1 \\ &= 1 - \frac{V_{\text{residue}}}{V_1} \end{aligned} \tag{10}$$

where V_f is the free volume within the protein structure, being the complementary part of V_{residue} (volume occupied by residue atoms with respect to the overall protein volume). The definition of the two parameters slightly differs only for the overall protein volume definition: V represents the average volume between the three spherical volumes $[V_X, V_Y, V_z]$, evaluated at three diameters corresponding to the maximum distance between residues in the three coordinates; V_1 corresponds to spherical volumes $[V_X', V_Y', V_z']$ related to the maximum distance along the three spatial directions of residues from the center of mass.

Principal component analysis: PCA was computed by SAS software in terms of the eigenvectors of the pairwise correlation matrix of the descriptors. The use of a correlation matrix instead of the covariance matrix implies a normalization of the data that in this case is mandatory given the variables have heterogeneous dimensions and range of variation. Noise floor and the consequent selection of meaningful components comes from a visual scree test. As we aptly demonstrated in ref 34, even a very minor component in terms of explained variance can have a signal-like character. The only way to decide about the signal character of a component can be only "semantic" and not "syntactic". Shortly, if a very small component is found to correlate with an external variable not explicitly put into analysis, this implies it carries relevant information and cannot be considered as pure noise. This feature is of crucial

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

importance (and routinely used) for the analysis of large scale high-throughput microarray data, where the presence of an overwhelming first principal "size" component corresponds to the characteristic transcriptome of the tissue of origin and relegates the usable part of information for sample discrimination to very minor components explaining a few percent of the total variability (for a very accurate and thorough explanation of this approach see ref 35). The demonstrated correlation between extracted components and external variables (eps and eps_1) reassure us about the signal-like character of extracted components, while the substantial identity (correlation coefficients around 0.9) between the components of the general data set and the components relative to reduced sets (data not shown) is proof of the robustness of the obtained solution.

AUTHOR INFORMATION

Corresponding Author
*E-mail: alessandro.giuliani@iss.it.

ACKNOWLEDGMENTS

We thank the "Consorzio interuniversitario per le Applicazioni di Supercalcolo Per Università e Ricerca" (CASPUR) for computing resources and support. Work partially granted by Italy–U.S.A. Cooperation Project "Systems Effects of Chemicals: Evaluation of High-Throughput Profiling".

REFERENCES

(1) Banerji, A.; Ghosh, I. *PloS One* **2009**, *4*, e7361.
(2) Zbilut, J.; Chua, G.; Krishnan, A.; Bossa, C.; Rother, K.; Webber, C.; Giuliani, A. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 621–629.
(3) Natarajan, V.; Koehl, P.; Wang, Y.; Hamann, B. Visual Analysis of Biomolecular Surfaces. In *Mathematical Methods for Visualization in Medicine and Life Sciences*; Linsen, L., Hagen, H., Hamann, B., Eds.; Mathematics and Visualization Series; Springer Verlag: Berlin, 2007.
(4) Wang, Y. Ph.D. Thesis, Department of Computer Science, Duke University, 2004.
(5) Hopfinger, A. *J. Med. Chem.* **1983**, *26*, 990–996.
(6) Arteca, G. *Biopolymers* **1993**, *33*, 1829–1841.
(7) Heckman, C. *Cytometry* **1990**, *11*, 771–783.
(8) Colafranceschi, M.; Colosimo, A.; Zbilut, J.; Uversky, V.; Giuliani, A. *J. Chem. Inf. Model.* **2005**, *45*, 183–189.
(9) Emberly, E.; Mukhopadhyay, R.; Tang, C.; Wingreen, N. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 91–98.
(10) Bloom, J.; Drummond, D.; Arnold, F.; Wilke, C. *Mol. Biol. Evol.* **2006**, *23*, 1751–1761.
(11) Benigni, R.; Giuliani, A. *Am. J. Physiol.* **1994**, *266*, R1697–R1704.
(12) Christie, O. *Chemom. Intell. Lab. Syst.* **1995**, *29*, 177–188.
(13) Preisendorfer, R. *Principal Component Analysis in Meteorology and Oceanography*, Elsevier; Amsterdam, 1988.
(14) Dill, K. *Biochemistry* **1990**, *29*, 7133–7155.
(15) Vajda, S.; Sippl, M.; Novotny, J. *Curr. Opin. Struct. Biol.* **1997**, *7*, 222–228.
(16) Giuliani, A.; Di Paola, L.; Setola, R. *Curr. Proteomics* **2009**, *6*, 235–245.
(17) Cotta-Ramusino, M.; Benigni, R.; Passerini, L.; Giuliani, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 248–254.
(18) Jackson, D. *Ecology* **1993**, *74*, 2204–2214.
(19) Takeda, M.; Kirihara, S.; Miyamoto, Y.; Kazuaki, S.; Honda, K. *Phys. Rev. Lett.* **2004**, *92*, 093902.
(20) Webber, C.; Giuliani, A.; Zbilut, J.; Colosimo, A. *Protein: Struct., Funct., Bioinf.* **2001**, *44*, 292–303.
(21) Giuliani, A; Zbilut, J.; Tomita, M. *J. Proteome Res.* **2007**, *6*, 3924–3934.
(22) Kannan, N.; Vishveshwara, S. *J. Mol. Biol.* **1999**, *292*, 441–464.
(23) Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. *J. Mol. Biol.* **2004**, *344*, 1135–1146.
(24) Gromiha, M. M.; Selvaraj, S. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235–277.
(25) Krishnan, A.; A; Giuliani, A. J. Z.; Tomita, M. *PLoS ONE* **2008**, *3*, e2149.
(26) Sol, A. D.; Fujihashi, H.; Amoros, D.; Nussinov, R. *Mol. Syst. Biol.* **2006**, *2*, 0019.
(27) Alves, N.; Aleksenko, V.; Hansmann, U. *J. Phys: Condens. Matter* **2005**, *17*, S1595.
(28) Enright, M. B.; Leitner, D. M. *Phys. Rev. E* **2005**, *71*, 011912.
(29) Russ, J. *Fractal Surfaces*; Plenum Press: New York, 1994.
(30) Dewey, G. *Fractals in Molecular Biophysics*; Oxford University Press, New York, 1998.
(31) Park, J.; Barabasi, A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17916–17920.
(32) Newman, M. *Phys. Rev. E* **2006**, *74*, 056108.
(33) Elias-Kohav, T.; Moshe, S.; Avnir, D. *Chem. Eng. Sci.* **1991**, *46*, 2787–2798.
(34) Giuliani, A.; Colosimo, A.; Benigni, R.; Zbilut, J. *Phys. Lett. A* **1998**, *247*, 47–52.
(35) Roden, J.; King, B.; Trout, D.; Mortazavi, A.; Wold, B.; Hart, C. *BMC Bioinf.* **2006**, *7*, 194.

Characterizing Protein Shape by a Volume Distribution Asymmetry Index

Nicola Arrigo[1], Paola Paci[2], Luisa Di Paola[1*], Daniele Santoni[3], Micol De Ruvo[1], Alessandro Giuliani[5] and Filippo Castiglione[4]

[1]*Università Campus Biomedico, 00128 Rome, Italy*

[2]*CNR-Institute of Systems Analysis and Computer Science "Antonio Ruberti", BioMathLab, 00185 Rome, Italy*

[3]*CNR-Institute of Systems Analysis and Computer Science "Antonio Ruberti", 00185 Rome, Italy*

[4]*CNR-Institute for Computing Applications "Mauro Picone", National Research Council, 00185 Rome, Italy*

[5]*Department of Environment and Health, Istituto Superiore di Sanità, 00161 - Rome, Italy*

Abstract: A fully quantitative shape index relying upon the asymmetry of mass distribution of protein molecules along the three space dimensions is proposed. Multidimensional statistical analysis, based on principal component extraction and subsequent linear discriminant analysis, showed the presence of three major 'attractor forms' roughly correspondent to rod-like, discoidal and spherical shapes. This classification of protein shapes was in turn demonstrated to be strictly connected with topological features of proteins, as emerging from complex network invariants of their contact maps.

Keywords: Protein shape, protein contact network, topological indices, principal component analysis.

1. INTRODUCTION

It is commonly stated that the activity of a protein is somewhat encoded into its shape [1]. A rough classification of proteins on the basis of their shape, identifies two distinct classes: globins (near spherical molecules) and scleroproteins (rod-like or fibrous). Fibrous proteins are for the major part mainly structural elements (for instance, collagen in the connective tissue); on the other hand, globins are apt to many different tasks, often subdued to the presence of specific interaction sites located on the protein surface [2].

The concept of molecular shape is somewhat elusive: the identification of quantitative descriptors for the molecular structure is, thus, a potentially very interesting avenue of research[3].

Several methods have been proposed to characterize proteins shape[5, 4]: so far, shape analysis has been limited to protein surface representation, assuming believing that surface as the privileged view given is a key factor, because is the region where it is where biologically meaningful interactions take place [6]. Actually, geomitetric shape has been often defined with reference to a finite set of points, a space curve, or a surface [7], instead of considering the overall volume of a molecule, that is specifically the purpose of this work, building upon previous results in which we demonstrated both the lack of any marked separation between protein internal and external milieu and the basic fractal structure of protein fold (Di Paola *et al*. JCIM).

*Address correspondence to this author at the Università Campus Biomedico, via Alvaro del Portillo 21, 00128 Rome, Italy;
Tel: ++39 06 225419634; Fax: ++39 0622541456;
E-mail: l.dipaola@unicampus.it#

Literature provides several different approaches to describe the molecular surfaces: among these, Van der Waals surface(VdW) refers to the union of atoms (modeled as balls) according to their van der Waals radiuses [8]; the Solvent AccessibleSurface (SAS), originally proposed by Lee and Richard [9], is the surface traced out by the center of a probe sphere (typically a water molecule) rolling on top of the VdW surface: in this way, the overall protein molecule hindrance comprises also the hydration shell. The Solvent Excluded Surface (SES) is the result of the SAS erosion by the same probe [10]. For a graphic representation, see Fig. (**1**).

Hopfinger developed a useful method for small molecules named 'molecular shape analysis' [12], based on the comparison of electrostatic fields, later adapted by Arteca and Mezey to define shape descriptors of macromolecules [13].

Some authors have focused on the detection of protrusions and cavities of known input structures, provided by shape descriptors.

The Connelly function[10], the most used and known, is derived as follows: in any point on the surface, a sphere is centered, having a diameter as large as a water molecule. If the fraction of the sphere volume within the SES volume (see dashed sphere in Fig. **1**) is smaller than 0.5, the surface is considered as locally convex, otherwise concave.

Formally, for any point $x \in M$, let us consider the ball B (r, x)centered at x with radius r : if S (r, x) = δB (r, x) is the boundary of B (r, x) and SI the portion of S (r, x) contained within the surface, the Connolly function fr : M → R is defined as:

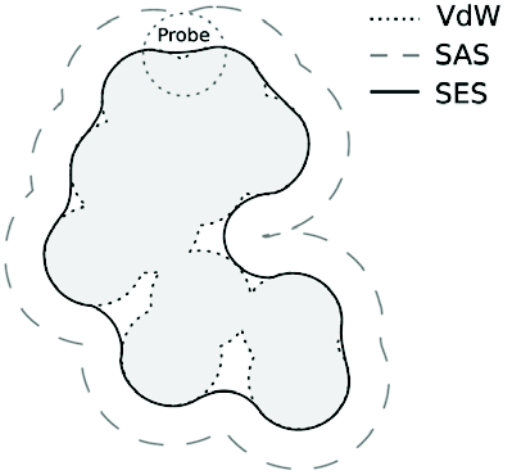


Fig. (1). VdW, SAS and SES molecular surfaces [9].

$$f_r(x) = \frac{Area(S_I)}{r^2} \quad \#$$

High values of f_r (x) indicate that the surface around x is largely concave, while low values point to a prevalent convexity around x. Røgen and Sinclair introduced protein shape descriptors based on backbone [14].

Although curvature-based methods (as Connelly's) well identify points located at local protrusions and cavities, they all depend on a pre-fixed value r (the neighborhood size); in many cases, it is desirable that the function value can also give some clues about the length scale of the conformational feature the function refers to.

All these models have a strong 'theoretical flavor' and are concentrated on the molecule surface shape. On the contrary, we adopted a mainly statistical bottom-up approach in order to derive a coarse-grain, but easily computable and free from *a priori* constraints shape index. At odds with surface-based approach, the proposed index is based on the volume distribution of the atoms along the three axes of the space.

The starting point of this work is that the most interesting geometrical templates in structural biochemistry are the sphere, the disc and the cylinder; thus, we decided to rely upon the amount of symmetry of the volume distribution on the three dimensional space so to develop a global index catching the relative 'spherical', 'discoidal' or 'cylinder' character of the studied structure.

A data set spanning the entire range of variation of protein shapes, from perfect sphere to almost perfect cylinders, was developed in order to check by means of a correlative approach based on Principal Component Analysis (PCA) [15], the consistency of the proposed index with relevant size, geometry and topology related properties of protein structures.

The demonstrated ability of the proposed method not only to discriminate different shapes but to discover the shape variability typical of a functional protein class (membrane proteins) confirms the relevance of volume based shape representation.

2. METHODS

In this work, we perform an analysis of the three-dimensional protein structures along the canonical axes, as reported in PDB files, containing the relevant information about biomolecular structures.

As a first step, we identify the Center of Mass (CM) of the molecule, reducing each amino acid residue to the correspondent α -carbon. In the case of the sphere, the center of mass coincides with its geometrical center and the distance from the CM to the surface of the molecule is identical along each of the three dimensions. In the case of the disc, the CM represents its center, two dimensions have an almost identical elongation and the third is not relevant; in the case of the cylinder, there is just one relevant dimension and the CM is located approximately at the middle point of the cylinder main axis.

Once identified the CM, the maximal distance of α–carbons from the CM is computed along the three axes; the three values R_x ,R_y , R_z represent the radius of hypothetical spheres (Fig. **2**), whose volume provides indication concerning the length of the object along a specific direction:

$$R_x = max(x_{CDM} - x_i)$$

$$R_y = max(y_{CDM} - y_i)$$

$$R_z = max(z_{CDM} - z_i)$$

The corresponding equivalent spherical volumes are then computed:

$$V_x = \frac{4}{3}\pi R_x^3$$

$$V_y = \frac{4}{3}\pi R_y^3$$

$$V_z = \frac{4}{3}\pi R_z^3$$

In the case of the sphere, $V_x = V_y = V_z$, whereas for a generic non-spherical molecule, these three volumes differ from each other.

Let's now introduce a shape space, in which each protein is identified by a vector ρ defined as follows:

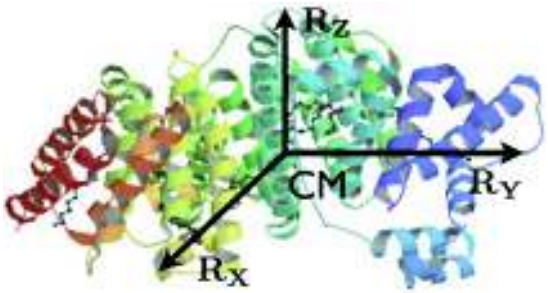


Fig. (2). Radiuses R_x, $R_y$and $R_z$in the case of HSA (PDB code 1E7I).

$$\rho = \left[\frac{V_x}{V_x + V_y + V_z}, \frac{V_y}{V_x + V_y + V_z}, \frac{V_z}{V_x + V_y + V_z} \right]$$

The reference shapes correspond to the following points in the shape space (Fig. **3**):

- Sphere: $S = \left[\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3} \right]$;
- Disc: $D_1 = \left[\frac{1}{2} \quad \frac{1}{2} \quad 0 \right]$; $D_2 = \left[\frac{1}{2} \quad 0 \quad \frac{1}{2} \right]$; $D_3 = \left[0 \quad \frac{1}{2} \quad \frac{1}{2} \right]$;
- Rod-like: $R_1 = [1 \quad 0 \quad 0]$; $R_2 = [0 \quad 1 \quad 0]$; $R_3 = [0 \quad 0 \quad 1]$.

The tetrahedron represented in Fig. (**3**) is the space of possible protein molecular shapes. Clearly, a nearly spherical molecule is represented by a point closely located to S; on the contrary, the largest distance from S accounts for rod-like proteins. Thus we use the ratio between the actual distance of the protein from S with the maximum distance correspondent to perfect rod-like shape; this maximum distance is:

$$d_1 \equiv \overline{SR_1} = \sqrt{\left(\frac{1}{3} - 1\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2} = \frac{\sqrt{6}}{3} \quad \#$$

On the other hand, the distance related to disc template is:

$$d_2 \equiv \overline{SD_1} = \sqrt{\left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{1}{3}\right)^2} = \frac{\sqrt{6}}{6} = \frac{d_1}{2} \quad \#$$

Thus, let us define a normalized distance $\xi = \frac{d}{d_1}$, being the distance of a generic point in the shape space from S; according to ξ values, molecules can be classified as follows:

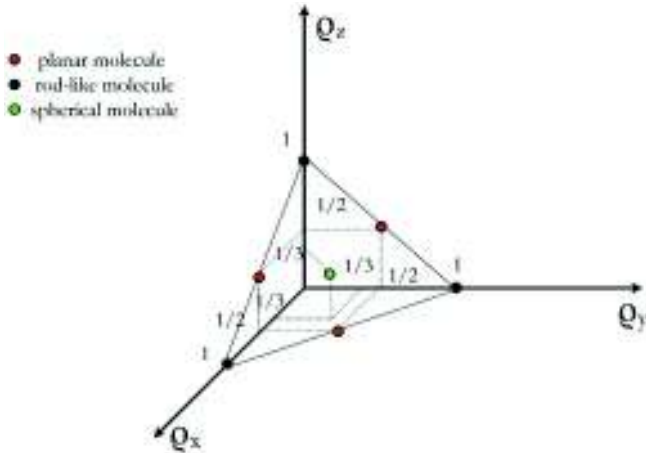


Fig. (3). Shape space: each point of the space represents a molecule in terms of its own shape; green, red and black dots refer to spheres, discs and rods, respectively.

$$\begin{cases} 0 \le \xi < 0.4 & spherical \\ 0.4 \le \xi < 0.6 & discoidal \\ 0.6 \le \xi \le 1 & rod-like \end{cases} \quad \#$$

where ξ is an asymmetry index, given its character of departurefrom perfect symmetry in space.

To prove the effectiveness of our index, we tested it on a 40 proteins data set, half of which chosen amongglobular shapes and half among fibrous.

In Table **1** the values of ξ are reported along with the classification consistent with our proposal and the number of residues(nodes). 3D structures of three proteins belonging to spherical, discoidal and rod-like groups are shown in Fig. **4 a**), **b**), **c**) respectively.

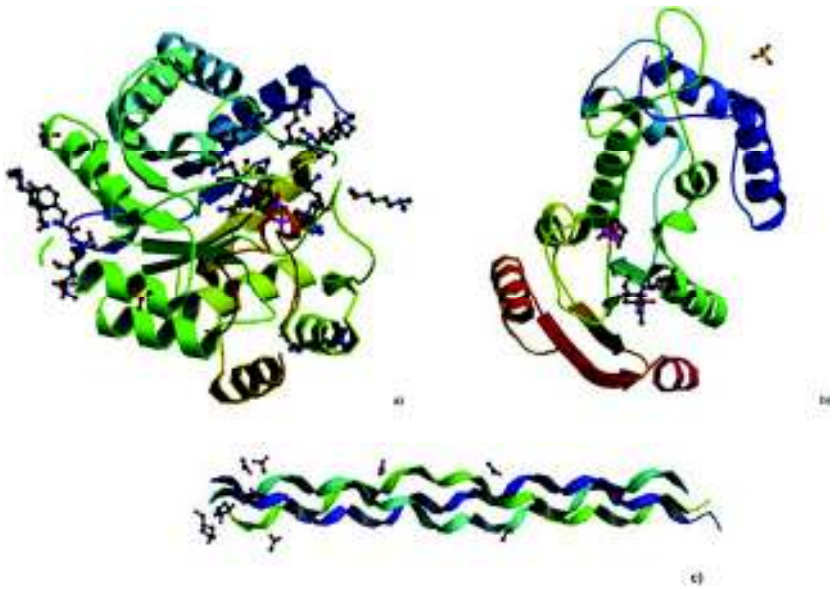


Fig. (4). Protein reference sample structures : **a**) 3ln3 (globular): putative reductase; **b**) 3kou (planar): cyclic ADP ribose hydrolase; **c**) 1cgd (rod-like): collagen peptide.

Table 1. Protein Data Set: PDB Codes are Reported in the First Column; Nodes = Number of Residues; ξ is the Asymmetry Index; Shape = Class

PDB ID	Nodes	ξ	Shape
1cgd#	60 #	1#	Rodlike#
1cag#	58#	1#	Rodlike#
3qc7#	172#	0,99#	Rodlike#
1h6w#	161#	0,94#	Rodlike#
2kt9#	116 #	0,76#	Rodlike#
1qqk#	254#	0,73#	Rodlike#
1emw#	88 #	0,71#	Rodlike#
1gr3#	132#	0,65#	Rodlike#
3dmw#	87#	0,65#	Rodlike#
3hon#	55 #	0,63#	Rodlike#
3p46#	39#	0,57#	Discoidal#
3js7#	174 #	0,55 #	Discoidal#
3hal#	262#	0,54#	Discoidal#
3jqu#	173#	0,52#	Discoidal#
1azz#	727#	0,51#	Discoidal#
1v1h#	610#	0,51#	Discoidal#
2dkm#	104#	0,50#	Discoidal#
2v53#	292#	0,47#	Discoidal#
2vl5#	869#	0,42 #	Discoidal#
3kou #	482 #	0,41 #	Discoidal#
1bkv #	68 #	0,41 #	discoidal#
2otp #	387 #	0,37 #	Spherical#
3qae #	371 #	0,36 #	Spherical#
3lqb #	207 #	0,32 #	Spherical#
1ao6 #	1556 #	0,30 #	Spherical#
2xvq #	1135 #	0,21 #	Spherical#
2vuf #	1105 #	0,21 #	Spherical#
2xw0 #	1135 #	0,20 #	Spherical#
3gbl #	97 #	0,18 #	Spherical#
3jxp #	307 #	0,16 #	Spherical#
2y3z #	351 #	0,13 #	Spherical#
3po6 #	263 #	0,12 #	Spherical#
3ln3 #	331 #	0,11 #	Spherical#
3p43 #	126 #	0,11 #	Spherical#
2q2m #	152 #	0,10 #	Spherical#
2jtd #	112 #	0,10 #	Spherical#
3p4h #	142 #	0,08 #	Spherical#
1uz2 #	158 #	0,05 #	Spherical#
3q1q #	112 #	0,05 #	Spherical#
3npo #	169 #	0,02 #	Spherical#

#

In order to put into perspective the proposed asymmetry index, we introduce some topological descriptors, based on a protein structure representation in terms of inter-residue contact graphs [16].

As a matter of fact, the 3D crystal structure of a protein canbe translated into a contact matrix among α -carbons that in turn can be considered as a network with α-carbons as nodes and the contacts between them as edges. This kind of formalization isextremely useful to study protein properties at all [17, 18, 19].

Starting from the spatial position of α -carbons, in the PDB files, the mutual residue distance matrix **d** = $\{d_{ij}\}$ is computed : the generic element d_{ij} is the Euclidean distance in the 3D space between the i-th and j-th residue, holding the primary structure ordering. A link is established between two residues if their mutual distance lies in the range 4 − 8Å; the contact graph adjacency matrix A =$\{a_{ij}\}$ is therefore defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } d_{ij} \in [4-8] \text{ Angstrom} \\ 0 & \text{otherwise} \end{cases} \quad \#$$

Some topological descriptors can be extracted from **A** [20]:

- *N* : number of nodes (residues) in the graph;
- *E* : number of edges connecting the graph nodes;
- *density* : ratio between the actual number of edges *E* and the maximum value
- *N* (*N*− 1)/ 2, corresponding to the complete graph;
- *avdegree* : the average of node degrees k_i , where $k_i = \sum_{j=1}^{N} a_{ij}$ is the number of links involving the i-th node;
- *avshortpath* : the shortest path is the minimum number oflinks connecting two residues; this value, averaged over allthe residue pairs, is the average shortest path;
- *diameter* : the longest shortest path;
- *avcluscoeff*: the clustering coefficient $C_i = \sum_{j,m\in N,\, j\neq m} \frac{2}{k_i \cdot (k_i - 1)} a_{ij} a_{jm} a_{mi}$ is a measure of connectivity on a local scale, for the i-th node: it measures the connectivity of the sub-graph made of nodes con-

nected to thei-th node. C_i averaged over the whole set of nodes is the *avcluscoeff*.

3. RESULTS

We computed the asymmetry ξ and the seven above mentioned topological properties for each protein of the data set. In order to evaluate the correlation of ξ with the other parameters, a multivariate data analysis is required. To this aim, we applied PCA to the data matrix, containing all the computed properties for each protein in the data set.

The presence of a specific component highly correlated withξ (PC2) is a consequence of the selection of a data set spanning the entire range from spherical to rod-like structures. On the other hand, protein size as measured by N is the main order parameter shaping the data set (PC1).

Results are reported in Table **2** in terms of component loadings, i. e., of correlation coefficients between principal components(PCs) and original variables. A high absolute value of the correlation coefficient (loading) between a variable and a component is used as guide for the structural interpretation of the extracted components.

PCA highlighted a three component solution as explaining the by far most important (and reasonably signal-like) part of information correspondent to the 86% of total variance, with

PC1 explaining the 47% of variability, while PC2 and PC3 accounting for 25% and 14% respectively.

Not surprisingly, the first component (PC1) corresponds toprotein size: the number of nodes *N*, as well as the number of links *E*, are strongly related to this component.

Both contact density and clustering coefficient negatively scale with size, confirming previous results,[19]. As shown in Fig. (**5**), density neatly scales with size (here, number of nodes).

The second component (PC2) identifies the 'general shape',since asymmetry has the highest correlation. It has to be stressed that *diameter* and *avcluscoeff* bring considerable contributions to PC2, suggesting that topology influences general shape.

In the case of PC3, the only relevant descriptor is *avdegree*. Therefore, this component is a topologic invariant, since it is neither influenced by size nor by general shape.

Table 2. Principal Component (PC) Pattern

	PC1	PC2	PC3
ξ(AS)	-0.37	0.80	-0.20
nodes (*N*)	0.92	-0.02	-0.07
avdegree	0.13	0.27	0.94
edges (*E*)	0.92	-0.02	-0.004
density	-0.7	0.24	-0.30
avshortpath	0.82	0.52	-0.13
diameter	0.69	0.64	-0.17
avcluscoeff	-0.55	0.76	0.20

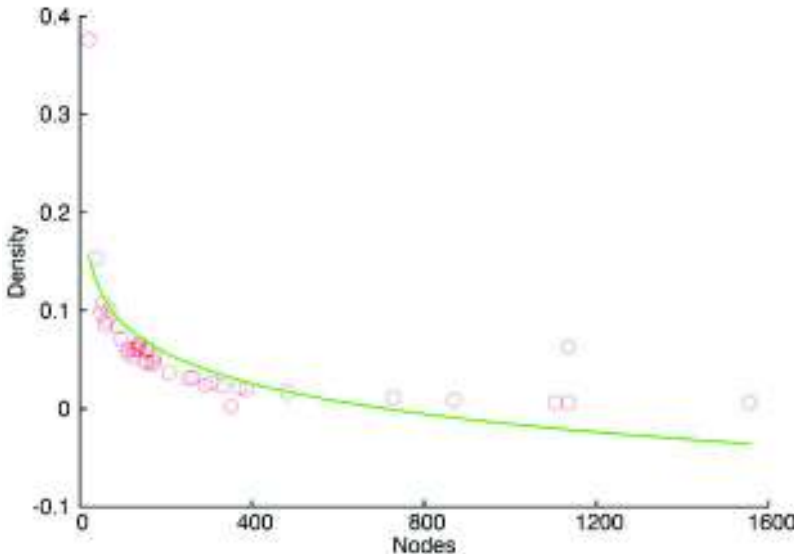


Fig. (5). Effect of protein size on density (open red circles). The data follows a power low behavior (green line).

Afterwards, we repeated the analysis taking out asymmetry, thus evaluating the ability of sole topologic features to predict protein shape.

The 'reduced' principal components (PC1r - PC3r, i.e., those components generated without the explicit contribution of ξ)are indeed able to perform a very significant classification of the three groups of rod-like, discoidal, and spherical proteins, as reported in Table **3**, where the classification matrix based on linear discriminant analysis based PC1r-PC3r is reported.

As evident from Table **3**, the discriminant analysis allows for an 84.4% of correct classification.

The efficacy of this discrimination can be appreciated in Fig. (**6**), reporting the space spanned by the first two reduced space components, where the space is approximately subdivided into three regions, correspondent to the three classes.

The ability of the components to separate the three classes is a proof-of-concept of the fact ξ is consistent with other features of protein organization (as those used for generating PC1r-

PC3r). As it can be observed in Table **4**, PC1r still represents size, but *avshortpath* is now concentrated on PC1r.

Furthermore, *avcluscoeff* is now the most sensitive descriptor to PC2r, being clusterization linked to shape; on the contrary, *avdegree* does not change appreciably, confirming it isa general invariant property. The 'reduced' component space explains the 88% of total variability as follows: PC1r = 52%,PC2r = 21% and PC3r = 15%.

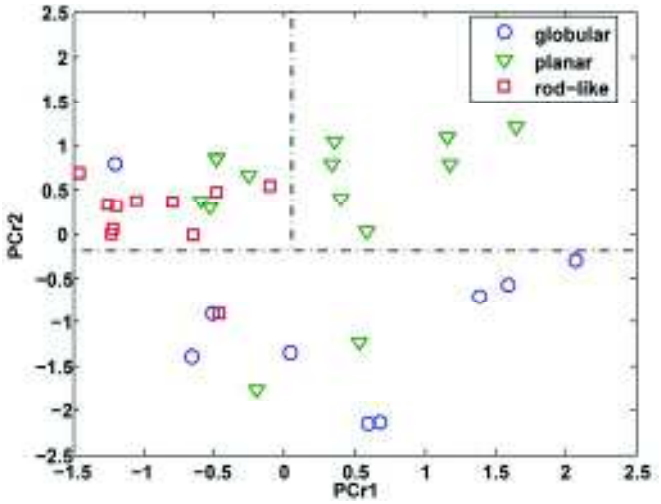


Fig. (6). Cartographic representation of PC2 vs PC1 on the basis of protein shape index ξ.

In order to check the relevance of the proposed index with an independent data set, asymmetry index was computed on a sample made of three different classes of proteins: globins, membrane and fibrous proteins. While 'globin' and 'fibrous' classifications refer directly to the protein shape, the denomination 'membrane' has to do only with the location of the molecule in the cell. According to the presence in membrane proteins of a part of the structure in the form of an elongated (mainly alpha helix) patch inside the membrane, we expect that membrane proteins must lie in between 'glo-

Table 3. Linear Discriminant Analysis Results. Topology is Able to Predict General Shape

	Estimated Shape			
Original shape	rod	disc	sphere	Total
rod	11	0	3	14
%	78.57	0.00	21.43	100
disc	0	8	1	9
%	0.00	88.89	11.11	100
sphere	1	1	9	11
%	9.09	9.09	81.82	100

Table 4. PCr Component Loadings

	PC1r	**PC2r**	**PC3r**
nodes (N)	0.91	-0.12	0.04
avdegree	0.14	0.53	-0.82
edges (E)	0.91	-0.11	-0.03
density	-0.67	0.26	0.45
avshortpath	0.87	0.38	0.23
diameter	0.76	0.49	0.32
avcluscoeff	-0.46	0.86	0.08

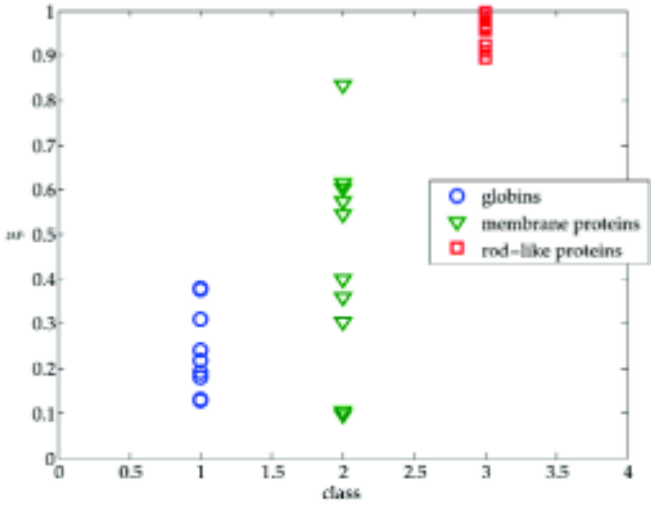


Fig. (7). Structural class of proteins discriminated on the basis of the asymmetry index.

bin' and 'fibrous' shapes as for their asymmetry index values. In the meantime, we do expect membrane proteins do have an higher variability with respect to the two other classes as for their asymmetry values. Here, we report result for the shape index for the above three classes of proteins; blue dots are proteins sharing the same globin-fold pattern, resulting in a spheroidal structure; green triangles are membrane proteins known to have widely different shapes with a slight prevalence for elongated forms (at least for the membrane embedded part of the structure), red squares correspond to fibrous proteins, having an elongated, rod-likemolecular shape. As shown in figure, fibrous protein segregate in the upper part of the figure, with asymmetry index close to maximum (Mean = 0.96, Std. Dev. = 0.03) ; globins, that are approximately spherical, locate, as expected, on the bottom, in a wider area with respect to rod-like structures (Mean= 0.24,

Std.Dev. =0.09). Membrane proteins, finally, spread out in the wide central part of the figure (Mean = 0.44, Std.Dev. = 0.24),consistently with their morphological variability going hand-in-hand with a tendency toward elongated shape as for their membrane-embedded part. The above results were highly statistically significant for both mean (Students t-test) and variance (F-test) pairwise comparisons. Fibrous vs. membrane comparison scored a t value =6.8 (p ¡ 0.0001) and an F value = 44.35 ($p \leq 0.0001$); globin vs membrane comparison scored a t value= 2.53 ($p \leq 0.03$) and an F value = 6.05 ($p \leq 0.02$), eventually globin vs. fibrous comparison scored a t-value = 21.2 ($p \leq 0.0001$) and an F-value = 7.34 ($p \leq 0.008$). The ability of the index not only to discriminate between different classes but to account for the internal variance of the membrane proteins is afurther proof of their possible use as a simple quantitative shapeindex to study diff erent protein folds.

4. CONCLUSION

As previously suggested by Holm and Sander [1], the generation of a principal component space based on the mutual correlation of different shape features allows for the identification of 'attractor shapes' acting as ideal templates rationalizing the apparently wild variety of protein forms. In this work, the same strategy was adopted in order to validate a global shape index allowing for a quantitative appreciation of the position of a given structure in the continuum spanning from very asymmetric fibrous structured to approximately globular shapes.

The possibility to discriminate the pertaining of a given molecule to the 'rod-like', 'discoidal' and 'spherical' attractors by the components of a feature space, not explicitly taking into account the proposed index, was a proof-of-concept of both the existence of such attractors and the consistency of the asymmetry descriptor here defined. Focusing on the quantification of symmetry, in order to build a general shape descriptor is notonly one of the many possible choices. In contrast, symmetry, as aptly explained by Goodsell and Olson[2], is a crucial property for rationalizing structure, function and even evolution history of protein molecules. Here it is sufficient to remind the role played by protein internal structural symmetries in allosteric effects, folding and cell localization [2] and the importance of detecting sequence-based symmetries, for both the modeling of sequence-structure relations and the protein evolution by gene duplication [21].

Our results point to the possibility to sketch a quantitative formalization of a so far largely qualitative concept as protein form, that could have very relevant outcomes in protein science.

This hope is substantiated by the strong, and still largely unexploited, link between general shape information and graph theoretical properties of protein contact networks.

ACKNOWLEDGMENTS

We thank the "Consorzio interuniversitario per le Applicazioni di Supercalcolo Per Universita e Ricerca" (CASPUR) for computing resources and support.

REFERENCES

[1] L. Holm, C. Sander, Mapping the protein universe, Science 273 (5275)(1996) 595–602.
[2] D. Goodsell, A. Olson, Structural symmetry and protein function, AnnuRev Biophys Biomol Struct 29 (105-53).
[3] E. Callaway, The shape of protein structures to come, Nature 449 (2007)765.
[4] W. Taylor, A. May, N. Brown, A. Asz´odi, Protein structure: geometry,topology and classification, Rep Prog Phys 64 (2001) 517.
[5] J. Ponomarenko, H. Bui, W. Li, N. Fusseder, P. Bourne, A. Sette, B. Peters, Ellipro: a new structure-based tool for the prediction of antibodyepitopes, BMC Bioinformatics 9 (2008) 514.
[6] V. Natarajan, P. Koehl, Y. Wang, B. Hamann, Visual Analysis of-Biomolecular Surfaces.
[7] Y. Wang, Geometric and topological methods in protein structure analysis,Ph.D. thesis, Department of Computer Science, Duke University(2004).
[8] A. Bondi, van der waals volumes and radii, The Journal of PhysicalChemistry 68 (3) (1964) 441–451.
[9] B. Lee, F. Richards, he interpretation of protein structures: Estimation ofstatic accessibility, J Mol Biol 55 (3) (1971) 379.
[10] M. Connolly, Solvent-accessible surfaces of proteins and nucleic acids,Science 221 (4612) (1983) 709–713.
[11] J. Giard, J. Ambroise, J. Gala, B. Macq, Regression applied to proteinbinding site prediction and comparison with classification, BMC Bioinformatics10 (2009) 276. doi:10.1186/1471-2105-10-276.

[12] A. Hopfinger, Theory and application of molecular potential energy fieldsin molecular shape analysis: a quantitative structure-activity relationshipstudy of 2,4-diamino-5-benzylpyrimidines as dihydrofolate reductase inhibitors,J Med Chem 26 (7) (1983) 990–996.

[13] M. Connelly, Shapes of small molecules and proteins.URL http://www.netsci.org/Science/Compchem/feature14.html.

[14] P. Røgen, R. Sinclair, Computing a new family of shape descriptors forprotein structures, J. Chem. Inf. Comput. Sci. (43) (2003) 1740–1747.

[15] R. Preisendorfer, Principal Component Analysis in Meteorology andOceanography, Elsevier, 1988.

[16] A Giuliani, L Di Paola, R Setola, Proteins as networks: A mesoscopicapproach using haemoglobin molecule as case study, Curr Proteomics6 (235-245).

[17] A Giuliani, J. Zbilut, M. Tomita, Network scaling invariants help to elucidatebasic topological principles of proteins, Journal of Proteome Research6 (10) (2007) 3924–3934.

[18] G. Bagler, S. Sinha, Assortative mixing in protein contact networks andprotein folding kinetics, Bioinformatics 23 (14) (2007) 1760–1767.

[19] J. Zbilut, G. Chua, A. Krishnan, C. Bossa, K. Rother, C. Webber, A. Giuliani,A topologically related singularity suggests a maximum preferredsize for protein domains, Prot Struct Funct Bioinf 66 (3) (2007) 621–629.

[20] M. Newman, Networks: An introduction, Oxford University Press, USA,2010.

[21] X. Ji, H. C. Y. Xiao, Hidden symmetries in the primary sequences ofbeta-barrel family, Computational Biology and Chemistry 31 (1) (2007)61–63.

[#]
Received: November 21, 2011 Revised: December 20, 2011 Accepted: December 23, 2011

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Modules Identification in Protein Structures: The Topological and Geometrical Solutions

Setareh Tasdighian,[†] Luisa Di Paola,*[,‡] Micol De Ruvo,[§] Paola Paci,[§] Daniele Santoni,[⊥] Pasquale Palumbo,[⊥,§] Giampiero Mei,[⊥] Almerinda Di Venere,[∥] and Alessandro Giuliani[∥]

[†]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium

[‡]Faculty of Engineering, Università CAMPUS BioMedico, Via A. del Portillo, 21, 00128 Roma, Italy

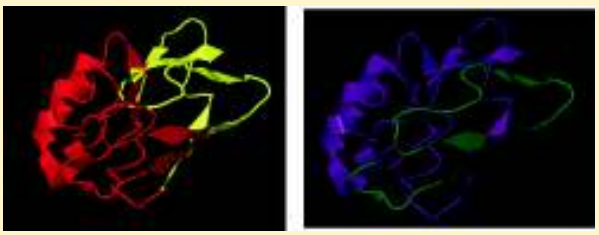
[§]CNR-Institute of Systems Analysis and Computer Science (IASI), viale Manzoni 30, 00185 Roma, Italy

[⊥]Department of Experimental Medicine and Surgery, University of Rome "Tor Vergata", via Montpellier 1, 00133 Rome, Italy

[∥]Environment and Health Department, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161, Roma, Italy

Supporting Information

ABSTRACT: The identification of modules in protein structures has major relevance in structural biology, with consequences in protein stability and functional classification, adding new perspectives in drug design. In this work, we present the comparison between a topological (spectral clustering) and a geometrical (*k*-means) approach to module identification, in the frame of a multiscale analysis of the protein architecture principles. The global consistency of an adjacency matrix based technique (spectral clustering) and a method based on full rank geometrical information (*k*-means) give a proof-of-concept of the relevance of protein contact networks in structure determination. The peculiar "small-world" character of protein contact graphs is established as well, pointing to average shortest path as a mesoscopic crucial variable to maximize the efficiency of within-molecule signal transmission. The specific nature of protein architecture indicates topological approach as the most proper one to highlight protein functional domains, and two new representations linking sequence and topological role of aminoacids are demonstrated to be of use for protein structural analysis. Here we present a case study regarding azurin, a small copper protein implied in the *Pseudomonas aeruginosa* respiratory chain. Its pocket molecular shape and its electron transfer function have challenged the method, highlighting its potentiality to catch jointly the structure and function features of protein structures through their decomposition into modules.



INTRODUCTION

Modularity is a central concept in Biology; as a matter of fact, the greater part of biological investigation for centuries has been devoted to looking for the most meaningful clusterization of structures and systems in tissues, organs, and physiological regulation circuits. The main objective of Systems Biology is the exploitation of the modular architecture of biological regulation at different organization scales, as explicitly stated by Denis Noble[1] that equates Systems Biology research to time-honored Physiology studies. The intuition that global behavior of the biological systems can be predicted by a clear picture of the constituting modules is present, for instance, in the task of identification of modules in a complex network of interacting proteins (PPI).[2] This approach relies on the translation of biological systems into graphs: a networklike structure whose elements are nodes and their mutual interactions are expressed as edges connecting them. In protein science, the search for the optimal decomposition of 3D structure into functionally meaningful modules is of utmost importance.[3]

Much work has been devoted to the definition and analysis of the amino-acid residue contact networks. Since the Holm and Sander work on protein comparison by interresidue distance matrices,[4] the field had a huge flourishing with interest in contact networks, describing intramolecular bonds between residues in a protein structure.[5−34] The recent indication of a possible reconstruction of the global 3D information of a protein molecule on the sole basis of its contact map[35] is only one of the myriad of proofs (e.g., refs 12, 14, 19, and 36−39) of the importance of the protein-as-contact-graph paradigm that promises to play the role structural formula played in organic chemistry as the most synthetic and immediate representation able to recover an information-rich picture of the biomolecular system under analysis.

From our viewpoint, the translation of a protein three-dimensional structure into a contact network is a rigorous and

Received: April 10, 2013
Published: December 1, 2013

simple starting point to search for modularity principles embedded into protein structures.

Modularity detection has evolutionary and physiological implications: as a matter of fact, protein domains are highly conserved throughout a whole proteome, keeping the same function in different protein structures.[40] Thus, the detection of functional regions (domains) has a relevant outcome in the phylogenetic analysis of the protein families. The automatic identification of these domains in protein structures relies on methods (CATH,[40] SCOP,[41] and FSSP[42]) based on secondary and supersecondary structural elements recognition.[43]

Modules identification in protein networks is strictly related to the topological role played by different nodes;[20,44] this purely topological characterization (a residue is defined on the sole basis of its contact pattern in the network) allowed for the identification of key residues in folding process,[45] holding a strong correlation between the strongly connected residue (hub) depletion and the lethality of the corresponding mutation.[17]

The identification of hubs in a network is generally based on the estimation of node "centrality" by different measurement paradigms.[46,47] The higher the betweenness, the stronger the role played in network robustness, so a removal of a high-centrality node is supposed to result in an abrupt network structure modification.

Roughly speaking, a node (amino-acid) endowed with high betweenness centrality is a node by which a lot of "shortest paths" (i.e., the most efficient paths linking one residue to another along the contact network) pass by. The basic analogy in this case is with a transmission network: if a node crucial for maintaining the most efficient (shortest) paths between different regions of the system is perturbed, we can expect a general detrimental effect on the entire system. One possible way to implement this general philosophy of network transmission efficiency (that has its biochemical counterpart in phenomena like allostery[14] and folding[48]) is by shifting from the perspective of paths linking different nodes to the separate consideration of between- and within-module communication. A simple analogy with the partition between highways (between modules) and local streets (within modules) could be of use to introduce this point.

The so-called cartography of Guimerà−Amaral,[49] a method based on complex network clustering, provides at the same time node classification according to their topological role in terms of between- and within-module connections (cartography) and the module (cluster) identification. Previous observations by our group[17,22] highlighted the fact any protein molecule, independently of its general shape and size, gave rise to very similar node role distribution in the Guimerà−Amaral cartography, provided the clustering was "optimal" in a global statistical sense (well-separated clusters), suggesting a sort of "optimal wiring" for a protein system. This feature allows us to check the reliability of different clustering procedures to analyze protein structures by exploring such maps (we call these "dentist's chairs" for their shape).

In this work we focused on the comparison between a geometrical clustering, taking into consideration the physical Euclidean distance between different residues coming from their actual three-dimensional coordinates in space (*k*-means on residues described by their *X*, *Y*, *Z* coordinates[50]) and a topological clustering, based on the discrete contact matrix in which each residue is described only in terms of its contacts (Shi−Malik spectral clustering[51]). We will show how the two clustering procedures, as applied on a data set of allosteric and nonallosteric proteins in their bound (holo) and free (apo) forms, gave rise to very consistent partitions. This gives a proof-of-concept of the isomorphism between contact networks and three-dimensional configuration. The application of both spectral and *k*-means clustering methods and consequent comparison is novel and first applied in this work.

The statistical analysis of different modularity descriptors allowed us to recognize some general principles of protein architecture. Once stated the statistical superposition of the two methods, we demonstrated, in a specific study, how the topological method overcomes the geometrical one with respect to functional modules identification.

Moreover, we tested both methods on a case study, regarding the azurin, a small copper protein with a remarkable pocket shape. We demonstrated that the nontrivial shape, able to decouple pure geometry and topology, provides a discrimination case for the two methods, putting into light the potentiality of the topological approach (spectral clustering) in catching jointly structural and functional features of the protein molecules.

Reminding the contact matrix represents a dramatic collapse of information with respect to *X*, *Y*, *Z* space, this result points to the peculiar role played by contacts topological metrics in protein function.

MATERIALS AND METHODS

Protein Data Set. Statistical analysis was performed on a protein data set comprising 5 different protein molecules, for a total of 11 protein structures, with different allosteric properties; we analyzed both the apo and the holo forms (listed in Table 1),

Table 1. Protein Data Set

protein name	PDB ID	
	apo	holo
calcium binding proteins		
calbindin D_{9k}	1CLB	2BCA
parvalbumin (PV)	2NLN	1TTX
recoverin (RC)	1IKU	1JSA
human hemoglobin (Hb_{O2}, Hb_{CO})	2DN2	1GZX (O_2), 1BBB (CO)
human serum albumin (Ab)	1AO6	1E7I (stearic acid)

to detect major descriptors, able to catch the allosteric nature of their function. Their biological role and chemicophysical properties are described thoroughly in ref 14. This choice was dictated by the need to have a consistent, albeit relatively small, set of proteins with common shared properties. The crucial role that between modules communication exerts in allosteric systems and its variation between holo and apo forms[14] are relevant reasons that prompted us to focus on the present data set.

The specific case study taken into account to investigate the peculiar properties of geometrical and topological approaches refers to the azurin structure (PDB code 1AZU, from *Pseudomonas aeruginosa*). Azurin is a small (14.6 kDa) redox protein that works as a mediator in the electron transfer system of denitrifying bacteria. Its ability to induce apoptosis in mammalian cells[52] has raised a growing interest in studying its binding capability to the mammalian tumor suppressor p53,[53] since it could act as a new therapeutic agent against cancer cells.[52] Azurin tertiary structure[54] is characterized by an eight strand, b-barrel motif,[55] which surrounds an hydrophobic cavity containing a buried tryptophan residue, whose peculiar fluorescence dynamics allowed a number of unique structural features of the molecule to be revealed. For instance, high pressure measure-

ments revealed that even at 3000 bar the hydrophobic core is not accessible to water molecules, thanks to the rather though and rigid scaffolding, which guarantees the protein a relevant conformational stability.[54] Indeed, enlarging the volume within the tryptophan cavity by site-directed mutagenesis[56] enhanced the protein flexibility, leading to a less stable structure.[54]

The presence of a remarkable pocket shape involving residues in the middle of the sequence decouples topological and geometrical views of the molecular structure so providing a very important test to distinguish topological and geometrical description of protein structure.

Definition of Protein Contact Network. The protein contact network is an undirected, unweighted graph; it is built on the basis of the distance matrix **d**, whose generic element d_{ij} records the Euclidean distance between the ith and the jth residue (measured between the corresponding α carbons). Then, we established a cutoff for inter-residue distance ranging within $\mathcal{I} = [4-8]$ Å accounting for intramolecular noncovalent interactions;[57] thus, the corresponding unweighted protein structure graph was built up, whose adjacency matrix $\mathbf{A} = \{A_{ij}\}$ is formally defined as follows:

$$A_{ij} = \begin{cases} 1 & \text{if } d_{ij} \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

Once **A** has been defined, different topological descriptors can be directly computed:

1. adeg (average degree): the node degree k_i represents the number of contacts the aminoacid is involved into, and it is defined as

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{1}$$

where N is the overall number of nodes; the average degree is the mean of k_i over the whole node set. The node degree is a direct measure of a single node relevance in the global wiring, while the average value is an indicator of the overall intramolecular interaction strength that, last, corresponds to protein stability and stiffness.

2. acc (average clustering coefficient): the clustering coefficient C_i is defined as the measure of how much the neighbors of a single node are close to each other:

$$C_i = \sum_{j,m\in N, j\neq m} \frac{2A_{ij}A_{jm}A_{mi}}{k_i(k_i-1)} \tag{2}$$

Residues with high clustering coefficients are likely to be crucial for the protein structure stability, since they play a key role in keeping the community on by the establishment of mutual relationships.

The average clustering coefficient is the mean of C_i over all nodes and represents a mean value of the local wiring strength in the protein structure.

3. asp (average shortest path): the shortest path sp_{ij} indicates the minimum number of links that connect the ith and jth residues; the average value over the whole set of residues pairs is called asp, that is a measure of the interresidue communication across the whole network. The minimization of asp is crucial for proteins, especially in the case of concerted motions, like those occurring in allosteric transitions.[14,37]

Clustering Algoritms. Clustering procedures are aimed at partitioning the whole network into communities (modules) that present specific intracommunity and intercommunity connectivity patterns; in this respect, the intramodule and intermodule connectivities are represented respectively by the two following parameters:[49]

- Within-module z-score

$$z_i = \frac{k_i - \overline{k_{si}}}{\sigma_{si}} \tag{3}$$

k_i is the overall node ith degree, k_{si} is the average value of the within-module degree and σ_{si} is its corresponding standard deviation of k_{si} distribution.

- Participation coefficient that describes the attitude of the node to connect to other nodes outside of its own module

$$P_i = 1 - \left(\frac{k_{si}}{k_i}\right)^2 \tag{4}$$

According to the P and z values, Guimerà et al.[49] established a cartography for the nodes, based on their role in terms of intra- and intermodule connections, on the basis of P/z values of residues (Table 2).

Table 2. Guimerà–Amaral Role Cartography

role	z	P
R1: ultraperipheral node	<2.5	<0.05
R2: peripheral node	<2.5	0.05 < P < 0.625
R3: nonhub connectors	<2.5	0.625 < P < 0.8
R4: nonhub kinless nodes	<2.5	>0.8
R5: provincial hubs	>2.5	<0.3
R6: connector hubs	>2.5	0.3 < P < 0.75
R7: kinless hubs	>2.5	>0.75

P/z curves have a special shape for protein structures,[22] called the dentist's chair, that is absent in other biological networks.[2] We recognized that a noticeable threshold for P is 0.75: indeed, nodes with $P = 0.75$ play a really special role in their community, since they share exactly half of their links with nodes belonging to their own cluster, while the remaining half is spent to establish connections with nodes belonging to other communities (from eq 4, for $P = 0.75$, $(k_{si}k_i)^2 = 0.25$, thus $k_{si} = 0.5k_i$). In this respect, they represent a frontier between the inner part of the community, with a strong homeland structure, and the outside world. On the other hand, nodes with $P > 0.75$ are mostly devoted to establish connections with other communities than to participate to their own community stability and structure: for this reason, we report the percentage of both classes of nodes (see Figure 1). Hereinafter, we are indicating with $P_{0.75}$ the percentage of nodes having a value of $P = 0.75$ (enclosed in the area sketched in Figure 1) and denoting with $P_{>0.75}$ the percentage of those whose P value exceeds 0.75.

On a general perspective, the distribution of residues in the P/z plane gives a mesoscopic view of the protein from the viewpoint of clusters (modules): the higher the percentage of nodes with $P > 0.75$, the wider the module's concertated motions; the higher $P = 0.75$ residues number, the larger the between-module contact surface.

Another very informative clustering representation is what we call the clustering color map that projects the results of topological network analysis on the sequence space (Figure 2).

The modularity M is a global parameter, able to catch the presence of well identifiable clusters, i.e. how well the separation

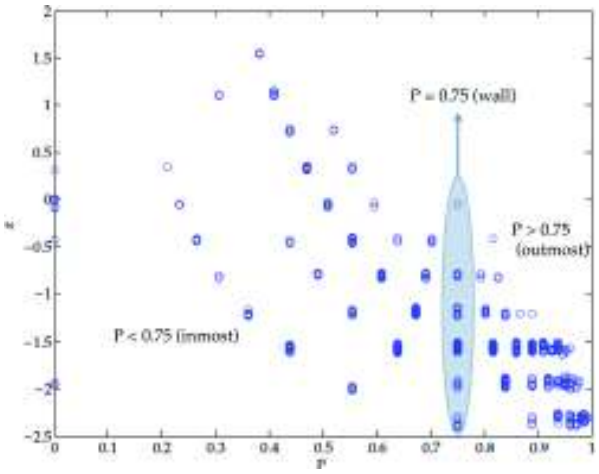


Figure 1. Dentist's chair map. The different roles of nodes on the basis of P values are shown in Table 2. The map refers to the hemoglobin structure (PDB code 1HBB) partitioned into 4 clusters.

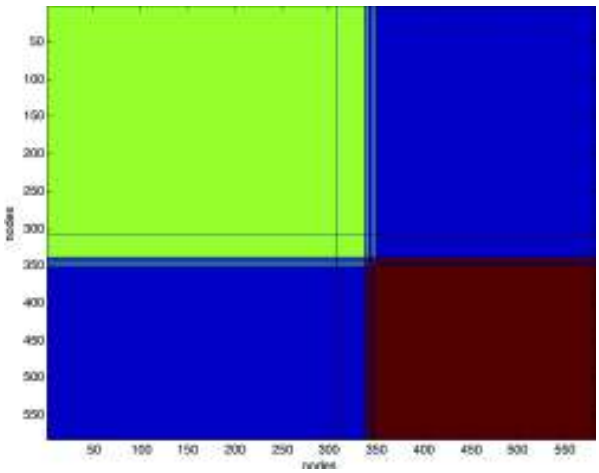


Figure 2. Clustering color map. The reports the map coming from human serum albumin structure (PDB code 1E7I) decomposition into two modules. The blue background represents (i, j) residue couples belonging to different clusters, whereas the couples pertaining to the same cluster are coded with the same color (other than blue). Coordinates of the map correspond to the protein sequence, thus the residues belonging to a given cluster can be easily identified accordingly. The interruptions of the continuity between sequence and cluster location correspond to long-range contacts in the molecule, putting into contact far in sequence residues; the corresponding residues show high value of P, pointing to their character of intermodule communication.

of amino-acids in modules is supported by their actual locations; it is defined as:[58]

$$M = \frac{\text{between-cluster variance}}{\text{total variance}} = \frac{\sum_{i=1}^{N} r_{i,S_i}{}^2}{\sum_{i=1}^{N} r_i^2} \tag{5}$$

where $r_{i,Si}$ represents the distance of each residue from the protein center of mass, when the residues coordinates correspond to those of its own cluster center of mass; r_i is the distance of a node from the center of mass of the whole structure: so, the higher M, the higher the system's modularity. M corresponds to an R^2 ordinary statistics, its numerator being the variance explained by the model, whereas its denominator is the total variance. With a number of clusters N we get a trivial unitary value for M (no intracluster variance). When approaching an optimal partition number, M approaches to a plateau.[58] We compute M with respect to spectral (Shi−Malik $M_{(sc)}$) and k-means ($M_{(m)}$) clustering algorithms.

Finally, to estimate the superposition of the two algorithms, we introduce the Rand index R;[59] it is generally defined in terms of number of pairs that change mutually the clusters they belong, so it is a direct measurement of the percentage of elements that are (or are not) in the same cluster for a method A and are not (or are) any more in the same cluster as for the method B. Specifically, it is defined as

$$\text{Rand} = 1 - [\text{number of pairs that are (are not) in the same cluster(A/B)}]/\left(\frac{N}{2}\right) \tag{6}$$

The closer Rand is to unity, the more the two methods match.

Shi−Malik (Spectral Clustering). Spectral Clustering is one of the most well-known algorithms for clustering.[51] The algorithm benefits from some concepts in the linear algebra related to graph Laplacians and similarity matrixes.

We have performed the algorithm application on the adjacency matrix **A**, by defining the corresponding Laplacian matrix **L** defined as:

$$\mathbf{L} = \mathbf{A} - \mathbf{D} \tag{7}$$

where **D** is the degree matrix, defined as the diagonal matrix whose generic non-null element corresponds to the ith residue degree ($D_{ii} = k_i$).

The next step has been to calculate the eigenvalues of **L** and put them in an ascending order. For mathematical reasons, the first eigenvalue is always equal to zero, then, $\lambda_1 = 0 < \lambda_2 < ... < \lambda_N$. In this way, the corresponding eigenvectors are ordered accordingly: $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_N$. The second eigenvector $\boldsymbol{v}_2$, named Fiedler's vector is used to split the set into two clusters (and, further, each resulting subset into two smaller subsets iteratively) according to the sign of its corresponding components.

The method presents advantages and drawbacks: the spectral clustering algorithm is easy to implement and stable under perturbation of the data set. It is an exact method, so it is not sensible to the initial conditions choice; however, it requires the computation of the Laplacian eigenvector for each step, that is a burdening computational step. Moreover, the method is not recommended for large graphs, since spectral graph partitioning falls in the class of NP-hard problems, having a high-order complexity with respect to the number of set elements.

k-Means. k-Means is an unsupervised learning and partitioning clustering algorithm, targetted at partitioning a data set into k clusters.[60] In the first step, the algorithm chooses k data points as centroids and then assigns each data point to its nearest centroid. In the next step k-means calculates a value called sum square error (SSE), for each cluster, defined as

$$\text{SSE}_k = \sum_{i=1}^{n_k} (x_i - c_k)^2 \tag{8}$$

where n_k is the number of data points in the kth cluster; x_i represents the coordinates of the generic ith node in the kth cluster, whose centroid (the center of mass of the cluster nodes) position is c_k. Next, k-means tries to vary the choice of centroid to find out the minimum possible value for the SSE_k and finally

dx.doi.org/10.1021/ci400218v | J. Chem. Inf. Model. 2014, 54, 159−168

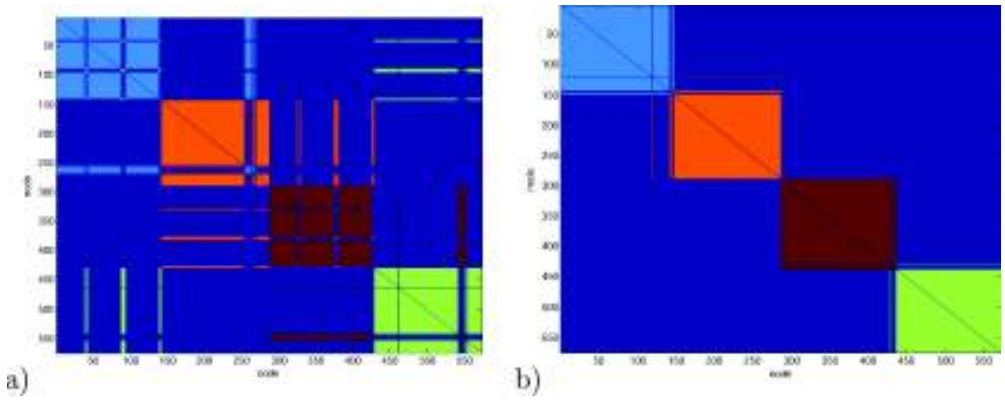


Figure 3. Clustering color map of hemoglobin residues: (a) spectral clustering; (b) *k*-means. The general similarity of the two partitions is noteworthy. Spectral clustering (a) is endowed with horizontal (vertical) lines starting from the cluster core and going forward; they represent the residues with high *P* values, establishing between-cluster links.[61] The geometrical character of *k*-means, on the other hand, results into a sharp transition partition along the sequence, tightly linked to residue spatial position.

Table 3. Topological Descriptors for the Protein Data Set[a]

PDB code	K_{OPT}	$Rand_{(K_{OPT})}$	adeg	acc	asp	N/K_{OPT}	$P^{(sc)}_{0.75}$	$P^{(sc)}_{>0.75}$	$M_{(sc)}$	$M_{(km)}$	A/H
					hemoglobin						
1HBB	4	0.9047	3.9390	0.2744	6.3922	143.5	10.10	57.14	0.6083	0.6083	A
1BBB	4	0.9278	3.9059	0.2688	6.4042	143.5	10.98	58.89	0.6047	0.6050	H
1GZX	4	0.9282	3.9321	0.2734	6.3970	143.5	10.63	56.62	0.6161	0.6159	H
					albumin						
1AO6	2	0.9794	3.6419	0.2402	10.7505	291	19.46	61.42	0.4407	0.4407	A
1E7I	2	0.8960	3.5069	0.2425	8.0870	291	24.05	63.40	0.3424	0.3395	H
					calbindin						
1CLB	4	0.8782	3.2667	0.3083	3.1168	18.75	16	68	0.5216	0.5031	A
2BCA	4	0.9052	3.48	0.3087	3.0897	18.75	9.33	56	0.5705	0.5925	H
					parvalbumin						
2NLN	8	0.8787	4.2037	0.3435	3.1722	13.5	11.11	66.67	0.7756	0.8036	A
1TTX	8	0.8714	3.5688	0.2845	3.4373	13.5	7.34	64.22	0.7470	0.7537	H
					recoverin						
1IKU	8	0.8588	3.5638	0.2624	4.3864	23.5	14.36	68.62	0.7454	0.7819	A
1JSA	8	0.8693	3.2819	0.2555	4.7849	23.5	17.55	69.68	0.7651	0.8882	H

[a] *P* and *M* values refer to the Shi−Malik algorithm application, K_{OPT} is estimated by the *k*-means application, and A/H refers to the apo/holo form.

updates the centroid of the *k*th cluster to the data point which gives the minimum SSE_k.

When all the centroids are updated (no unit changes its centroid at the next iteration), *k*-means calculates a more general SSE value for all the data set, defined as

$$\mathrm{SSE} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_i - c_k)^2$$

If SSE is less than a given threshold, the algorithm stops; otherwise, it goes back to the first step and starts over by choosing different initial centroids.

To choose the correct number of clusters, one has to run *k*-means over and over again for different values of *k* and calculate the SSE in each case. The optimal number of clusters would be the one corresponding to the minimum SSE. One of the central issue of the *k*-means method deals is the choice of the initial centroids and with how this choice affects the final results of the algorithm.

The method presents some pros and cons: *k*-means is fast and easy to implement. On the other hand, the final clustering results depend on the initial choice of the centroids. Moreover, it has to be run it several times to determine the right number of clusters. Finally, if the number of clusters *k* is defined, it takes a time that increases largely with the number of clusters but depends less strongly on the number of elements. So, it represents a good choice in the case of few clusters partition among a large number of elements, as in the case of modules identification in protein contact network.

RESULTS AND DISCUSSION

Statistical Analysis. The results of the topological analysis and clustering are sketched out in Figure 3, where all descriptors refer to each protein of the data set.

To get a direct picture of the cluster partition, in Figure 3 a color map representation of clusters for spectral clustering and *k*-means is shown, relative to the partition into four clusters of the hemoglobin structure (PDB code 1HBB).

Looking at results reported in Table 4 (coming from the analysis of the raw data of Table 3) it is worth noting the concordance between the spectral and the *k*-means clustering as measured by the Rand index Rand[59] is very high. Rand has an average value of 0.90; it never goes below 0.86 and has a maximal value of 0.98. Keeping in mind that the Rand maximum value

dx.doi.org/10.1021/ci400218v | J. Chem. Inf. Model. 2014, 54, 159−168

Table 4. Statistical Analysis of Topological Descriptors

	mean	std dev	minimum	maximum
K_{OPT}	5.09091	2.42712	2	8
N	329.81818	239.70015	75	282
$Rand_{(K_{OPT})}$	0.89979	0.03511	0.85880	0.97940
adeg	3.66279	0.29652	3.26670	4.20370
acc	0.27838	0.03122	0.24020	0.34350
asp	5.45618	2.44072	3.08970	10.75050
N/K_{OPT}	102.18182	108.83865	13.5	291
$P^{(sc)}_{0.75}$	13.71909	5.06835	7.34	24.05
$P^{(sc)}_{>0.75}$	62.78727	5.09175	56	69.68
$M_{(sc)}$	0.61249	0.14079	0.34240	0.34240
$M_{(km)}$	0.63022	0.16559	0.33950	0.88820

(complete concordance of the two classifications) is equal to 1, we can safely state that the apparently dramatic collapse of information going from actual geometric coordinates of residues to the corresponding adjacency matrix maintains the relevant mesoscopic features (domains) of macromolecule structure.

Having safely established the basic coherence between the spectral and *k*-means clustering procedures (image in light of the preservation of the relevant spatial information by the adjacency matrix), we can go more in depth into the different modular character of the studied proteins as estimated by different descriptors. For this reason, we compute the Pearson correlation coefficient r between different descriptors so to get a consistent picture of protein modularity. $M_{(sc)}$ and $M_{(km)}$ measure the proportion of variance explained by clusterization (spectral and *k*-means respectively, see Materials and Methods), so they keep track of the amount of "modularity" of the studied protein correspondent to the optimal (K_{OPT}) number of clustering. For simple statistical reasons, the value of both $M_{(sc)}$ and $M_{(km)}$ grows with K_{OPT} ($r(K_{OPT}, M_{(sc)}) = 0.94$ and $r(K_{OPT}, M_{(km)}) = 0.94$ in our data set), but the important point is that spectral and *k*-means clusterizations give the same estimation of the relative modularity of the analyzed proteins, with a mutual correlation near to unity ($r(M_{(km)}, M_{(sc)}) = 0.98$). This correlation between the estimation of modularity relative to the two methods does not depend on number of clusters (K_{OPT}); as a matter of fact, when separated out of the common effect of K_{OPT}, the correlation coefficient between $M_{(km)}$ and $M_{(sc)}$ is still very high ($r(M_{(km)}, M_{(sc)})_{K_{OPT}} = 0.84$).

The partial correlation technique allows us to discover some other interesting features about the definition of a modularity metrics. In this work we obtained three different viewpoints on "degree of modularity":

1. a global one ($M_{(sc)}$ and $M_{(km)}$) computed over the entire molecule;
2. a local one measured at the level of single residue (acc);
3. a mesoscopic, indirect one, correspondent to asp.

The relation between asp and modularity asks for further clarification. We can imagine two opposite scenarios for the relation between the characteristic length of between residues shortest paths (asp) and the degree of clusterization:

Scenario 1: If a clear and well discriminated clusterization does exist (let us keep in mind the original statistical definition of a well formed cluster is "a set of elements for which the intracluster distances are much lower than the between-cluster distances"), the between-cluster separation (as a mountain range between two cities) is a hurdle to the establishment of short-cuts between residues pertaining to different clusters, so implying a POSITIVE correlation between asp characteristic length and global modularity ($M_{(sc)}$).

Scenario 2: On the other hand, if a high number of "local paths" is present (this means high values of acc, local clustering measure) due to a strong clusterization (cluster compactness comes from the richness of edges linking residues pertaining to the same cluster), it is sufficient to establish relatively few "short cuts" between clusters to serve the entire cluster community and optimize between-cluster communication. In this case, the clusterization degree has an opposite effect with respect to scenario 1, and asp should have a NEGATIVE correlation with global modularity ($M_{(sc)}$). In other words, the relative importance of the two drivers of modularity (within-cluster compactness and between-cluster separation) is at the basis on the two possible effects of modularity on the between-residues communication efficiency, as expressed by asp. Scenario 2 has to do with the "small-world" character of networks:[62] small-world networks are those graphs occupying a middle ground between regular and random networks, showing high local clustering of elements, like regular networks, but also short path lengths between elements, like random networks. Small-worldness implies emergent features very different from both random and regular networks, such as a strong resilience to damage and a collective dynamics with few coherent modes.[63] Which of the two above architectural principles are typical of protein graphs wiring can be immediately decided by the mutual relations holding between acc, asp, and $M_{(sc)}$, i.e., between-modularity degree at local, mesoscopic, and global scale. The direct Pearson pairwise correlation coefficients between asp, acc, and $M_{(sc)}$ are reported in Table 5.

Table 5. Correlation Matrix between Different Modularity Descriptors

	asp	acc	$M_{(sc)}$
asp	1	−0.77	−0.64
acc	−0.77	1	0.41
$M_{(sc)}$	−0.64	0.41	1

It is worth noting the prevalence of scenario 2 implying an optimized choice of between clusters paths so to minimize asp: the higher the global modularity (as measured by $M_{(sc)}$), the shorter the characteristic path length ($r(\text{asp}, M_{(sc)}) = -0.64$, $p = 0.03$), thus it can be argued local compactness (acc) is used by protein network wiring as a communication efficiency enhancer ($r(\text{asp}, \text{acc}) = -0.77$, $p = 0.006$). There is no strong and significant direct relation between local clustering and global clustering ($r(M_{(sc)}, \text{acc}) = 0.41$, $p = 0.2067$); this interpretation is strengthened by the computed values of partial correlation between $M_{(sc)}$ and acc, demonstrated by the common relation with asp. The correlation drops from 0.41 to a counterintuitive (but practically null) $r(\text{acc}, M_{(sc)})_{\text{asp}} = -0.16$; on the other hand, the correlation coefficient between asp and acc, separated out of $M_{(sc)}$ remains practically invariant $r(\text{acc}, \text{asp})_{M(sc)} = 0.72$, $p = 0.019$) and the same happens as for the relation between asp and $M_{(sc)}$ separated out of the common relation with acc ($r(\text{asp}, M_{(sc)})_{\text{acc}} = 0.55$). Thus, it is confirmed that the mesoscopic character of asp mediating the local residue scale (acc) and global ($M_{(sc)}$) network modularity, pointing to the crucial role played by characteristic path length (asp) in protein functionality. Only through of the mediation of modules organization, the local (acc) and global ($M_{(sc)}$) structures are put into relation.

dx.doi.org/10.1021/ci400218v | J. Chem. Inf. Model. 2014, 54, 159−168

The point of between-module efficient communication will be further demonstrated to be the main determinant of the specific differences between topological and geometrical approaches when analyzing in depth azurin molecules. Moreover the link between most efficient paths and actual inside molecule signal transmission will be demonstrated in the case of electron transfer.

We get some more clues about the residues maintaining the global network communication efficiency through the analysis of dentist-chair graphs: the variable $P_{0.75}$ identifies residues having exactly the same number of connections within its own cluster and with neighboring cluster residues (see Materials and Methods). There is an excess of these "communicating residues" that represent on average the 13% of aminoacids. The $P = 0.75$ threshold parts the dentist-chair into two zones: on the left we have the proper "well-clustered" population of residues (intracluster connections outnumber intercluster ones) responsible for "keeping alive" the modular properties of the molecule, whereas on the right side we find the residues whose intercluster relations outnumber intracluster ones. As expected, Rand is negatively correlated with $P_{>0.75}$ ($r(\text{Rand}, P_{>0.75}) = -0.64$, $p = 0.03$) given these residues fuzzify the modular structure decreasing the spectral-geometrical correlation. We hypothesize "frontier residues" are the most crucial ones for cluster communication, while over-the-frontier residues roughly correspond to the residues Csermely and Nussinov indicate as "creative elements", endowed with a crucial role in protein dynamics by allowing the system to attain different configurations.[39,64] The statistical analysis of our data set allowed to sketch some peculiar properties of protein modular organization consistently with the small-world character of contact graphs. Well-formed domains are not maximally separated and distinct parts of the molecule, but modules whose identity must go hand-in-hand with richness of between-clusters contacts so to optimize the efficiency of signal transmission through the entire molecule. The coexistence of cluster identity and between-modules connections is a conundrum for *k*-means clustering based on geometry, but not for topological spectral clustering; for this reason in the next section, we will abandon the general perspective to go in depth into the analysis of a single protein structure.

Case Study. Figure 3 shows a different character of the *k*-means and spectral clustering solutions: while *k*-means strictly follows sequence order, spectral clustering highlights the peculiar role of residues in charge of intermodule communication. These residues appear in the as the disarranged lines in the spectral clustering profile, interrupting the strict superposition between structural cluster and sequence position. Figure 3 refers to hemoglobin, whose residues (nodes) mismatching the cluster location in the two methods are lying at the interlocking interfaces between the folded chains. The hemoglobin case, anyway, is too straightforward and in some sense trivial: the symmetric tetrameric structure is well characterized by both methods, even though some details are different. Thus we moved on a specific case study, regarding the azurin (arsenate reductase), a small, monomeric copper protein, whose structure presents a pocket enclosing the prosthetic group (copper; Figure 4).

We tested the two clustering methods as for detection of the pocket structure, and results are shown in Figures 5 and 6 (partitioned into two clusters). In this case, the two methods mismatch largely, showing two different pictures: while *k*-means simply splits the sequence into the two terminal regions (Figure 5), the result for spectral clustering tells of a strongly entangled structure, where the sequence position and the cluster location

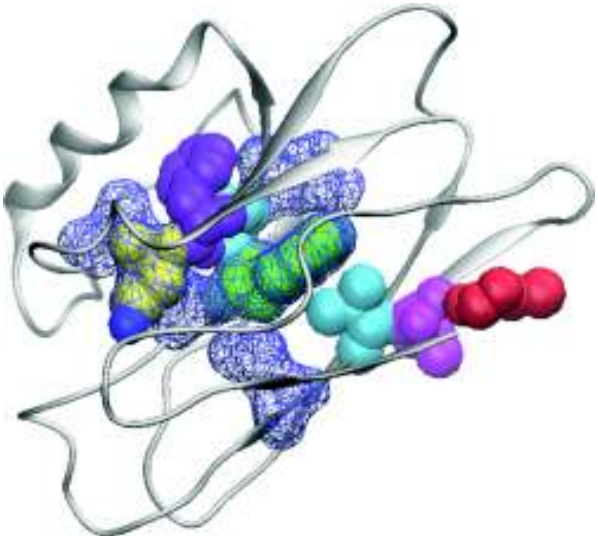


Figure 4. Ribbon structure of azurin. The model shows from right to left (filled van der Waals) the preferential pathway obtained in pulse radiolysis measurements from Cys 3 to Cu.[65] The residues are in blue, and empty wireframes are the aminoacids that form the hydrophobic cavity of azurin.[66]

are loosely related. However, still a partition in two gross sequence emerges, the first (green zone in color map of Figure 5a) comprising residues located in both terminals, and a central cluster, including the copper pocket (red zone in the color map of Figure 5a). In Figure 6a, the same result is reported in the usual ribbonlike representation where the clusters correspond to different colors on the sequence. It is evident that in Figure 6a that the copper pocket comprises only one cluster (red), while both terminals fall in the other cluster (yellow region). On the contrary, *k*-means clustering results in a pocket splitting in two different clusters that, in turn, correspond to the two terminals (purple and green regions in Figure 6b); differences in clustering are mirrored in a low Rand index (0.52) as well.

Additionally, we report the SSE and $M_{(\text{km})}$ profiles for the application of *k*-means for different clusters number relative to azurin structural analysis (Figure 7). It is clear that there is no sharp cut corresponding to an optimal number of clusters, that is a typical result for the *k*-means application in cases of disjoint sets of nodes (spatial points) where the optimal number corresponds to the number of well-parted groups; moreover, $M_{(\text{km})}$ never exceeds 0.65 even with a relatively large number of clusters, pointing to a loose module identification by *k*-means.

In the case of the molecular protein structures, the points (residues) are homogeneously distributed into the space, thus looking for an optimal number of geometrical clusters does not make any sense. All in all we can say only spectral clustering, concentrating on an apparently poorer information (contact graph instead of the actual geometrical coordinates representation) is able to catch the functionally relevant structure partition.

The "topology first" approaches, like spectral clustering, have an additional advantage with respect to purely geometrical methods; namely, they derive from the nature of the information transfer across the molecule, following the path of between-residue contacts, following the track of the seminal work of Skourtis and Beratan.[67] This is particularly clear in the case of azurin. In the past years, thanks to its small size and to the presence of a single copper center, azurin has been used as a

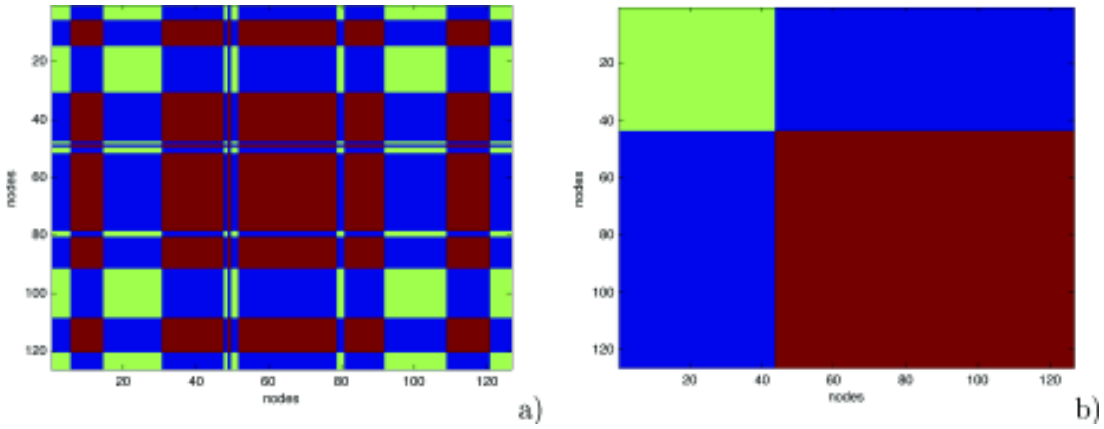


Figure 5. Color map of azurin clustering, comparison between the two methods: (a) spectral clustering; (b) *k*-means.

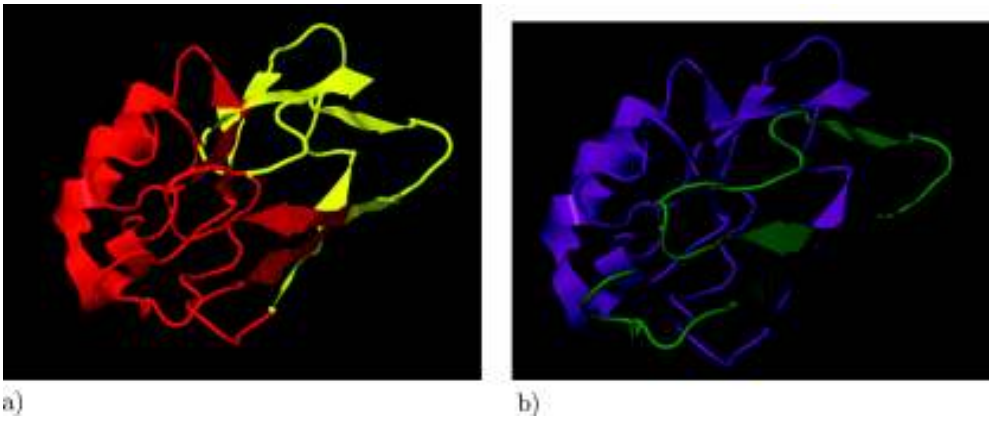


Figure 6. Azurin structure, the different colors refer to different clusters: (a) spectral clustering; (b) *k*-means.

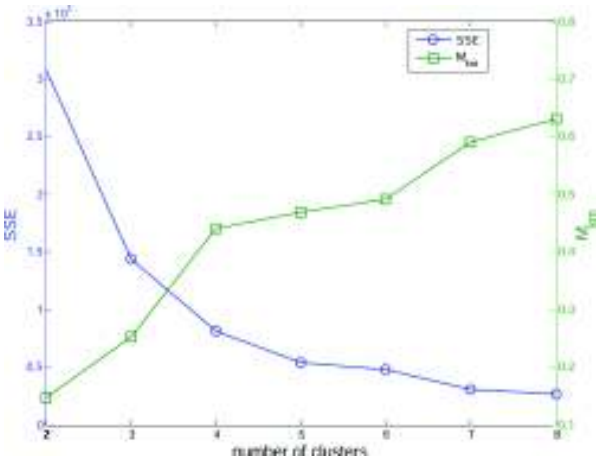


Figure 7. SSE and *M* profiles for *k*-means application for different number of clusters for azurin structural analysis: modularity does not reach a plateau, even for a relatively high number of clusters.

model protein to study the electron transfer pathway in redox protein systems.[68,69] In particular, pulse radiolysis measurements[65] have provided evidence that electrons can flow from the "periphery" of the protein (Cys 3) up to the Cu-bound residue (Cys 112) through a preferential route, which involves the following residues: Thr 30, Val 31, Trp 48, Val 49, Phe 111. It has been demonstrated[70,71] that the transmission path is continuous with the only exception of the Val 31−Trp 48 step, where the electron is hypothesized to move through an electric arc. As a matter of fact, we were able to exactly reconstruct the hypothesized mechanism in topological terms: a noncovalent bond exists between all the steps involving not contiguous residues (contiguous ones are trivially linked by the backbone). The only exception again is Val 31−Trp 48, as predicted by the model. It is worth noting that the Val 49− Phe 111 step involves two residues very far in sequence (62 aminoacids of distance) that occurs extremely rarely (7.4% of total edges refers to contacts between residues far in sequences 60−70 units). Moreover, all the involved residues show negative *z* values and relatively high *P* values, pointing to their connection character.

■ CONCLUSIONS

The most relevant conclusion of our work is the strong consistency between global geometrical information, carried by protein structure, and its representation in terms of adjacency matrix. This allows to think of protein contact networks as a sort of structural formula for macromolecules.[72]

The two graphical representations presented here, clustering color map and dentist's chair, promise to be a complement to the usual ribbon diagrams to identify physiologically relevant aminoacid residues in protein structures.

Considering protein structures as contact networks allows to give a rational and consistent definition of modularity, going from local (average clustering coefficient) to global (percentage of information explained by clusters) by the mediation of the average shortest path asp. The maximization of communication efficiency across the entire molecule by minimizing the average shortest path can be considered as a crucial architectural principle governing protein structures. Tracing back this mesoscopic feature by network cartography to single-residue roles in the global wiring promises to be a precious tool to locate the relevant residues in a protein molecular system.

■ ASSOCIATED CONTENT

Supporting Information

Method applied on an additional data set, composed of different categories of structures, classified on the basis of their structural properties. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: l.dipaola@unicampus.it.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Noble, D. *Exp Physiol* **2008**, *1*, 16−26.
(2) Agarwal, S.; Deane, C.; Porter, M.; Jones, N. *PLoS Comput Biol* **2001**, *17*, e1000817.
(3) Baron, M.; Norman, D.; Campbell, I. *Trends Biochem. Sci.* **1991**, *16*, 13−17.
(4) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123−138.
(5) Aftabuddin, M.; Kundu, S. *Phys. A* **2006**, 895−904.
(6) Bagler, G.; Sinha, S. *Phys. A* **2005**, *346*, 27−33.
(7) Barah, P.; Sinha, S. *Pramana* **2008**, *71*, 369−378.
(8) Bartoli, L.; Fariselli, P.; Casadio, R. *Phys Biol* **2008**, *4*, L1−L5.
(9) Brinda, K.; Surolia, A.; Vishveshwara, S. *Biochem. J.* **2005**, *391*, 1−15.
(10) Brinda, K.; Vishveshwara, S. *Biophys. J.* **2005**, *89*, 4159−4170.
(11) Brinda, K. V.; Vishveshwara, S.; Vishveshwara, S. *Mol. Biosyst.* **2010**, *6*, 391−398.
(12) Csermely, P.; Sandhu, K.; Hazai, E.; Hoksza, Z.; Kiss, H.; Miozzo, F.; Veres, D.; Piazza, F.; Nussinov, R. *Curr. Protein Peptide Sci.* **2012**, *13*, 19−33.
(13) Dehmer, M.; Barbarini, N.; Varmuza, K.; Graber, A. *BMC Struct. Biol.* **2010**, *10*, 1−17.
(14) De Ruvo, M.; Giuliani, A.; Paci, P.; Santoni, D.; Di Paola, L. *Biophys. Chem.* **2012**, *165*−*166*, 21−29.
(15) Di Paola, L.; Paci, P.; Santoni, D.; De Ruvo, M.; Giuliani, A. *J. Chem. Inf. Model.* **2012**, *52*, 474−482.
(16) Doncheva, N.; Klein, K.; Domingues, F.; Albrecht, M. *Trends Biochem. Sci.* **2011**, *36*, 179−182.
(17) Giuliani, A.; Di Paola, L.; Setola, R. *Curr. Proteomics* **2009**, *6*, 235−245.
(18) Greene, L.; Highman, V. *J. Mol. Biol.* **2003**, *334*, 781−791.
(19) Gromiha, M. *J. Chem. Inf. Model.* **2009**, *49*, 1130−1135.
(20) Gurso, A.; Keskin, O.; Nussinov, R. *Biochem. Soc. Trans.* **2008**, *36*, 1398−1403.
(21) Krishnan, A.; Giuliani, A.; Zbilut, J.; Tomita, M. *J. Proteome Res.* **2007**, *6*, 3924−3934.
(22) Krishnan, A.; Zbilut, J. P.; Tomita, M.; Giuliani, A. *Curr. Protein Peptide Sci.* **2008**, *9*, 28−38.
(23) Kundu, S. *Phys. A* **2005**, *346*, 104−109.
(24) Mekenyan, O.; Bonchev, D.; Trinajstic, N. *Int. J. Quantum Chem.* **1980**, *18*, 369−380.
(25) Kim, D.; Park, K. *BMC Bioinf.* **2011**, *12*, 1471−2105.
(26) Plaxco, K.; Simons, K.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985−994.
(27) Sathyapriya, R.; Vishveshwara, S. *Proteins* **2007**, *68*, 541−550.
(28) Sengupta, D.; Kundu, S. *Phys. A* **2012**, *391*, 4266−4278.
(29) Tan, L.; Zhang, J.; Jiang, L. *J. Biol. Phys.* **2009**, *35*, 197−207.
(30) Vendruscolo, M.; Dokholyan, N.; Paci, E.; Karplus, M. *Phys. Rev. E* **2002**, *65*, 061910.
(31) Vijayabaskar, M.; Vishveshwara, S. *Biophys. J.* **2010**, *99*, 3704−3715.
(32) Vishveshwara, S.; Brinda, K.; Kannan, N. *J. Theor. Comp. Chem.* **2002**, *1*, 1−25.
(33) Vishveshwara, S.; Ghosh, A.; Hansia, P. *Curr. Protein Peptide Sci.* **2009**, *10*, 146−160.
(34) Giuliani, A.; Di Paola, L.; Paci, P.; De Ruvo, M.; Arcangeli, C.; Santoni, D.; Celino, M. Proteins as Networks: Usefulness of Graph Theory in Protein Science. In *Advances in Protein and Peptide Science*; Dunn, B., Ed.; Bentham, 2012; in revision.
(35) Chen, J.; Shen, H. *Curr. Bioinf.* **2012**, *7*, 116−124.
(36) Tsai, C.; del Sol, A.; Nussinov, R. *J. Mol. Biol.* **2008**, *378*, 1−11.
(37) del Sol, A.; Araúzo-Bravo, M.; Amoros, D.; Nussinov, R. *Genome Biol.* **2007**, *8*, R92.
(38) Nussinov, R.; Tsai, C.; Csermely, P. *Trends Pharmacol. Sci.* **2011**, *32*, 686−693.
(39) Csermely, P. *Trends Biochem. Sci.* **2008**, *33*, 569−576.
(40) Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. *Structure* **1997**, *5*, 1093−1109.
(41) Murzin, A.; Brenner, S.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536−540.
(42) Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; Vriend, G. *Protein Sci.* **1992**, *1*, 1691−1698.
(43) Csaba, G.; Birzele, F.; Zimmer, R. *BMC Struct. Biol.* **2009**, *9*, 23−33.
(44) Jeong, H.; Mason, S.; Barabàsi, A.; Oltvai, Z. *Nature* **2001**, *411*, 41−42.
(45) Vendruscolo, M.; Paci, E.; Dobson, C.; Karplus, M. *Nature* **2001**, *409*, 641−645.
(46) Barabási, A. L.; Oltvai, Z. N. *Nat. Rev.* **2004**, *5*, 101−113.
(47) Koschutzki, D. Network Centralities. In *Analysis of Biological Networks*; Junker, B., Schreiber, F., Eds.; Wiley Series on Bioinformatics, Computational Techniques and Engineering; Wiley VCH, 2008; pp 65−84.
(48) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. *FASEB J.* **2011**, *26*, 868−881.
(49) Guimerà, R.; Sales-Pardo, M.; Amaral, L. A. N. *Nat. Phys.* **2006**, *3*, 63−69.
(50) Lloyd, S. *IEEE Trans. Inf. Theory* **1982**, *28*, 129−137.
(51) Shi, J.; Malik, J. *IEEE Trans. Pattern Anal.* **2000**, *22*, 888−905.
(52) Yamada, T.; Hiraoka, Y.; Gupta, T.; Chakrabarty, A. *Cell Cycle* **2004**, *3*, 752−755.
(53) Gabellieri, E.; Bucciantini, M.; Stefani, M.; Cioni, P. *Biophys. Chem.* **2011**, *159*, 287−293.
(54) Mei, G.; Di Venere, A.; Malvezzi Campeggi, F.; Gilardi, G.; Rosato, N.; De Matteis, F.; Finazzi-Agrò, A. *Eur. J. Biochem.* **1999**, *265*, 619−626.
(55) Nar, H.; Messerschmidt, A.; Huber, R. *J. Mol. Biol.* **1991**, *221*, 765−772.
(56) Gilardi, G.; Mei, G.; Rosato, N.; Canters, G.; Finazzi-Agrò, A. *Biochemistry* **2004**, *33*, 1425−1432.
(57) Bahar, I.; Jernigan, R. *J. Mol. Biol.* **1997**, *266*, 195.
(58) Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. *J. Intell. Inf. Syst.* **2001**, *17*, 107−145.
(59) Rand, W. *J. Am. Stat. Soc.* **1971**, *66*, 846−850.
(60) Jain, A. *Pattern Recog. Lett.* **2009**, *31*, 651−666.
(61) Paci, P.; Di Paola, L.; Santoni, D.; De Ruvo, M.; Giuliani, A. *Curr. Proteomics* **2012**, *9*, 160−166.
(62) Humphries, M.; Gurney, K. *PLoS ONE* **2008**, *3*, e0002051.
(63) Watts, D. J.; Strogatz, S. H. *Nature* **1998**, *393*, 440−442.
(64) Csermely, P.; Korcsmáros, T.; Kiss, H.; London, G.; Nussinov, R. *Pharmacol. Therapeut.* **2013**, *138*, 333−408.

(65) Farver, O.; Skov, L.; van de Kamp, M.; Canters, G.; Pecht, I. *Eur. J. Biochem.* **1992**, *210*, 399−403.
(66) Bottini, S.; Bernini, A.; De Chiara, M.; Garlaschelli, D.; Spiga, O.; Dioguardi, M.; Vannuccini, E.; Tramontano, A.; Niccolai, N. *Comput. Biol. Chem.* **2013**, *43*, 29−34.
(67) Onuchic, J.; Beratan, D.; Winkler, J.; Gray, H. *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 349−377.
(68) Langen, R.; Chang, I.; Germanas, J.; Richards, J.; Winkler, J.; Gray, H. *Science* **1995**, *268*, 1733−1735.
(69) Regan, J.; Di Bilio, A.; Langen, R.; Skov, L.; Winkler, J.; Gray, H.; Onuchic, J. *Chem. Biol.* **1995**, *2*, 489−496.
(70) Mikkelsen, K.; Shoc, L.; Nar, H.; Farver, O. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5443−5445.
(71) Onuchic, J.; Beratan, D.; Winkler, J.; Gray, H. *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 349−377.
(72) Di Paola, L.; De Ruvo, M.; Paci, P.; Santoni, D.; Giuliani, A. *Chem. Rev.* **2013**, *113*, 1598−1613.

Bibliography

[1] Chavarrìa-Krauser A, Jager W, and Schurr U. Primary root growth: a biophysical model of auxin - related control. *Funct Plant Biol*, 32:849–862, 2005.

[2] S Artavanis-Tsakonas, M D Rand, and R J Lake. Notch signaling: cell fate control and signal integration in development. *Science*, 284(5415):770–6, April 1999.

[3] R E Baker and P K Maini. Travelling gradients in interacting morphogen systems. *Math. Biosci.*, 209(1):30–50, September 2007.

[4] Mohammed H Baluch. convection equation over a finite domain. 7(January):285–287, 1983.

[5] L R Band and J R King. Multiscale modelling of auxin transport in the plant-root elongation zone. *J. Math. Biol.*, 65(4):743–85, October 2012.

[6] L. R. Band, D. M. Wells, J. a. Fozard, T. Ghetiu, a. P. French, M. P. Pound, M. H. Wilson, L. Yu, W. Li, H. I. Hijazi, J. Oh, S. P. Pearce, M. a. Perez-Amador, J. Yun, E. Kramer, J. M. Alonso, C. Godin, T. Vernoux, T. C. Hodgman, T. P. Pridmore, R. Swarup, J. R. King, and M. J. Bennett. Systems Analysis of Auxin Transport in the Arabidopsis Root Apex. *Plant Cell*, March 2014.

[7] Rafael A Barrio, José Roberto Romero-Arias, Marco A Noguez, Eugenio Azpeitia, Elizabeth Ortiz-Gutiérrez, Valeria Hernández-Hernández, Yuriria Cortes-Poza, and Elena R Álvarez-Buylla. Cell patterns emerge from coupled chemical and physical fields with cell proliferation dynamics: The arabidopsis thaliana root as a study system. *PLoS Comput Biol*, 9(5):e1003026, May 2013.

[8] Tobias I. Baskin, Benjamin Peret, Frantisek Baluška, Philip N. Benfey, Malcolm Bennett, Brian G. Forde, Simon Gilroy, Ykä Helariutta, Peter K. Hepler, Ottoline Leyser, Patrick H. Masson, Gloria K. Muday, Angus S. Murphy, Scott Poethig, Abidur Rahman, Keith Roberts, Ben Scheres, Robert E. Sharp, and Chris Somerville. Shootward and rootward: peak terminology for plant polarity. *Trends in Plant Science*, 15(11):593–594, Oct 2010.

[9] Eva Benková and Jan Hejátko. Hormone interactions at the root apical meristem. *Plant Mol. Biol.*, 69(4):383–96, March 2009.

[10] Rishikesh P Bhalerao and Malcolm J Bennett. The case for morphogens in plants. 5(11):939–943, 2003.

[11] Marta Del Bianco, Leonardo Giustini, and Sabrina Sabatini. Spatiotemporal changes in the role of cytokinin during root development. 2013.

[12] A Bielach, J Duclercq, P Marhavy, and E Benkova. Genetic approach towards the identification of auxin-cytokinin crosstalk components involved in root development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1595):1469–1478, Jun 2012.

[13] Anthony Bishopp, Satu Lehesranta, Anne Vate, Hanna Help, Sedeer El-showk, and Ben Scheres. Report Phloem-Transported Cytokinin Regulates Polar Auxin Transport and Maintains Vascular Pattern in the Root Meristem. pages 927–932, 2011.

[14] M B Bitonti and A Chiappetta. *Progress in Botany 72*, volume 72 of *Progress in Botany*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[15] Tobias Bollenbach. *Formation of morphogen gradients*. PhD thesis, Technischen Universität Dresden, 2005.

[16] Renee M Borges. Plasticity comparisons between plants and animals: Concepts and mechanisms. *Plant Signal. Behav.*, 3(6):367–75, June 2008.

[17] Géraldine Brunoud, Darren M Wells, Marina Oliva, Antoine Larrieu, Vincent Mirabet, Amy H Burrow, Tom Beeckman, Stefan Kepinski, Jan Traas, Malcolm J Bennett, and Teva Vernoux. A novel sensor to map auxin response and distribution at high spatio-temporal resolution. *Nature*, 482(7383):103–6, February 2012.

[18] Heidi M Cederholm, Anjali S Iyer-Pascuzzi, and Philip N Benfey. Patterning the primary root in Arabidopsis. *Wiley Interdiscip. Rev. Dev. Biol.*, 1(5):675–91, 2012.

[19] Angela H Chau, Jessica M Walter, Jaline Gerardin, Chao Tang, and Wendell A Lim. Designing synthetic regulatory networks capable of self-organizing cell polarization. *Cell*, 151(2):320–332, Oct 2012.

[20] A Chavarría-Krauser and M Ptashnyk. Homogenization of long-range auxin transport in plant tissues. *Nonlinear Analysis: Real World Applications*, 11(6):4524–4532, Sep 2010.

[21] Natalya M St Clair. *Pattern Formation in Partial Differential Equations*. PhD thesis, Scripps College, 2006.

[22] Natalie M Clark, Maria a de Luis Balaguer, and Rosangela Sozzani. Experimental data and computational modeling link auxin gradient and development in the Arabidopsis root. *Front. Plant Sci.*, 5(July):328, January 2014.

[23] Steven Clark. Plant Development. *Cell*, 114(1):11–12, July 2003.

[24] E J Crampin, E a Gaffney, and P K Maini. Reaction and diffusion on growing domains: scenarios for robust pattern formation. *Bull. Math. Biol.*, 61:1093–1120, 1999.

[25] Alfredo Cruz-ramírez, Sara Díaz-trivi no, Ikram Blilou, Verônica A Grieneisen, Rosangela Sozzani, Christos Zamioudis, Pál Miskolczi, Jeroen Nieuwland, Pankaj Dhonukshe, Juan Caballero-pérez, Beatrix Horvath, Yuchen Long, Ari Pekka

Mähönen, Hongtao Zhang, Jian Xu, James A H Murray, Philip N Benfey, Laszlo Bako, Athanasius F M Marée, and Ben Scheres. A Bistable Circuit Involving SCARECROW-RETINOBLASTOMA Integrates Cues to Inform Asymmetric Stem Cell Division. *Cell*, 150(5):1002–1015, 2013.

[26] Gibbon J D. The chain rule in partial differentiation.

[27] Christian Dahmann, Andrew C Oates, and Michael Brand. Boundary formation and maintenance in tissue development. *Nat. Rev. Genet.*, 12(1):43–55, January 2011.

[28] Ive De Smet, Steffen Vanneste, Dirk Inzé, and Tom Beeckman. Lateral root initiation or the birth of a new meristem. *Plant Mol. Biol.*, 60(6):871–87, April 2006.

[29] Raffaele Dello Ioio, Carla Galinha, Alexander G Fletcher, Stephen P Grigg, Attila Molnar, Viola Willemsen, Ben Scheres, Sabrina Sabatini, David Baulcombe, Philip K Maini, and Miltos Tsiantis. A PHABULOSA/cytokinin feedback loop controls root growth in Arabidopsis. *Curr. Biol.*, 22(18):1699–704, September 2012.

[30] Raffaele Dello Ioio, Francisco Scaglia Linhares, and Sabrina Sabatini. Emerging role of cytokinin as a regulator of cellular differentiation. *Curr. Opin. Plant Biol.*, 11(1):23–7, February 2008.

[31] Raffaele Dello Ioio, Francisco Scaglia Linhares, Emanuele Scacchi, Eva Casamitjana-Martinez, Renze Heidstra, Paolo Costantino, and Sabrina Sabatini. Cytokinins determine Arabidopsis root-meristem size by controlling cell differentiation. *Curr. Biol.*, 17(8):678–682, April 2007.

[32] Raffaele Dello Ioio, Kinu Nakamura, Laila Moubayidin, Serena Perilli, Masatoshi Taniguchi, Miyo T Morita, Takashi Aoyama, Paolo Costantino, and Sabrina Sabatini. A genetic framework for the control of cell division and differentiation in the root meristem. *Science*, 322(5906):1380–4, November 2008.

[33] Jan Dettmer and Jiří Friml. Cell polarity in plants: when two do the same, it is not the same.... *Curr. Opin. Cell Biol.*, 23(6):686–96, December 2011.

[34] Pankaj Dhonukshe. Cell polarity in plants: Linking PIN polarity generation mechanisms to morphogenic auxin gradients. *Commun. Integr. Biol.*, 2(2):184–90, March 2009.

[35] José R. Dinneny and Martin F. Yanofsky. Vascular Patterning: Xylem or Phloem? *Curr. Biol.*, 14(3):R112–R114, February 2004.

[36] L Dolan, K Janmaat, V Willemsen, P Linstead, S Poethig, K Roberts, and B Scheres. Cellular organisation of the Arabidopsis thaliana root. *Development*, 119(1):71–84, September 1993.

[37] L Dolan, P Linstead, C Kidner, K Boudonck, X F Cao, and F Berger. Cell fate in plants. Lessons from the Arabidopsis root. *Symp. Soc. Exp. Biol.*, 51:11–7, January 1998.

[38] Lionel Dupuy, Jonathan Mackenzie, and Jim Haseloff. Coordination of plant cell division and expansion in a simple morphogenetic system. *Proc. Natl. Acad. Sci. U. S. A.*, 107(6):2711–6, February 2010.

[39] Sedeer El-Showk, Raili Ruonala, and Ykä Helariutta. Crossing paths: cytokinin signalling and crosstalk. *Development*, 140(7):1373–83, April 2013.

[40] C Feller, JP Gabriel, C Mazza, and F Yerly. Pattern formation in auxin flux. *Journal of Mathematical Biology*, 68(4):879–909, 2014.

[41] Andrew P French, Michael H Wilson, Kim Kenobi, Daniela Dietrich, Ute Voss, Susana Ubeda-Tomas, Tony P Pridmore, and Darren M Wells. Identifying biological landmarks using a novel cell measuring image analysis tool: Cell-o-Tape. *Plant Methods*, 8(1):7, 2012.

[42] Jirí Friml. Auxin transport - shaping the plant. *Curr. Opin. Plant Biol.*, 6(1):7–12, February 2003.

[43] Ying Fu, Ying Gu, Zhiliang Zheng, Geoffrey Wasteneys, and Zhenbiao Yang. Arabidopsis interdigitating cell growth requires two antagonistic pathways with opposing action on cell morphogenesis. *Cell*, 120:687–700, Mar 2005.

[44] Schwank G and Basler K. Regulation of organ growth by morphogen gradients. pages 1–16, Dec 2010.

[45] Philip Garnett, Arno Steinacher, Susan Stepney, Richard Clayton, and Ottoline Leyser. Computer simulation: the imaginary friend of auxin transport biology. *Bioessays*, 32(9):828–35, September 2010.

[46] A Gierer and H Meinhardt. A Theory of Biological Pattern Formation. *Kybernetik*, 1972.

[47] M H M Goldsmith. The Polar Transport of Auxin. *Annu. Rev. Plant Physiol.*, 28(1):439–478, June 1977.

[48] Mary-Paz González-García, Josep Vilarrasa-Blasi, Miroslava Zhiponova, Fanchon Divol, Santiago Mora-García, Eugenia Russinova, and Ana I Caño Delgado. Brassinosteroids control meristem size by promoting cell cycle progression in Arabidopsis roots. *Development*, 138(5):849–59, March 2011.

[49] Peter V Gordon, Christine Sample, Alexander M Berezhkovskii, Cyrill B Muratov, and Stanislav Y Shvartsman. Local kinetics of morphogen gradients. *Proc. Natl. Acad. Sci. U. S. A.*, 108(15):6157–62, April 2011.

[50] V A Grieneisen, J Xu, A F Marée, P Hogeweg, and B Scheres. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature*, 449:1008–13, 2007.

[51] VA Grieneisen and B Scheres. Morphogengineering roots: comparing mechanisms of morphogen gradient formation. *BMC Syst.* 2012.

[52] Renze Heidstra and Sabrina Sabatini. Plant and animal stem cells: similar yet different. *Nat. Rev. Mol. Cell Biol.*, 15(5):301–12, May 2014.

[53] Martin Howard. How to build a robust intracellular concentration gradient. *Trends Cell Biol.*, 22(6):311–7, June 2012.

[54] Marta Ibañes, Yasuhiko Kawakami, Diego Rasskin-Gutman, and Juan Carlos Izpisúa Belmonte. Cell lineage transport: a mechanism for molecular gradient formation. *Mol. Syst. Biol.*, 2:57, January 2006.

[55] Victor B Ivanov and Joseph G Dubrovsky. Longitudinal zonation pattern in plant roots: conflicts and solutions. *Trends in Plant Science*, pages 1–7, Nov 2012.

[56] Crampin E J. Pattern formation in reaction–diffusion models with non uniform growth. *Bull Math Biol*, 64:1–23, Jul 2002.

[57] Keni Jiang and Lewis J Feldman. Regulation of root apical meristem development. *Annu. Rev. Cell Dev. Biol.*, 21(1):485–509, Nov 2005.

[58] JM Jin. The Finite Difference Method. *Theory Comput. Electromagn. Fields*, 1972.

[59] Brian Jones, Sara Andersson Gunneråa s, Sara V Petersson, Petr Tarkowski, Neil Graham, Sean May, Karel Dolezal, Göran Sandberg, and Karin Ljung. Cytokinin regulation of auxin synthesis in Arabidopsis involves a homeostatic feedback loop regulated via auxin and cytokinin signal transduction. *Plant Cell*, 22:2956–2969, 2010.

[60] Henrik Jönsson, Marcus G Heisler, Bruce E Shapiro, Elliot M Meyerowitz, and Eric Mjolsness. An auxin-driven polarized transport model for phyllotaxis. *Proc. Natl. Acad. Sci. U. S. A.*, 103(5):1633–8, January 2006.

[61] Hironaka K and Morishita Y. Encoding and decoding of positional information in morphogen-dependent patterning. *Current Opinion in Genetics & Development*, pages 1–9, Nov 2012.

[62] Anna Kicheva, Anna Kicheva, Michael Cohen, and James Briscoe. Developmental Pattern Formation : Insights from Physics and Biology. 210, 2012.

[63] Jürgen Kleine-Vehn and Jirí Friml. Polar targeting and endocytic recycling in auxin-dependent plant development. *Annu. Rev. Cell Dev. Biol.*, 24:447–73, January 2008.

[64] Johannes F Knabe, Maria J Schilstra, and Chrystopher Nehaniv. Evolution and Morphogenesis of Differentiated Multicellular Organisms : Autonomously Generated Diffusion Gradients for Positional Information. (8):321–328, 2008.

[65] Shigeru Kondo and Takashi Miura. Reaction-diffusion model as a framework for understanding biological pattern formation. *Science*, 329(5999):1616–20, September 2010.

[66] Eric M Kramer. PIN and AUX/LAX proteins: their role in auxin accumulation. *Trends Plant Sci.*, 9(12):578–82, December 2004.

[67] Eric M. Kramer, Heidi L. Rutschow, and Sturm S. Mabie. AuxV: A database of auxin transport velocities. *Trends Plant Sci.*, 16(9):461–463, 2011.

[68] P Krupinski and H Jonsson. Modeling auxin-regulated development. *Cold Spring Harbor Perspectives in Biology*, 2(2):a001560–a001560, Feb 2010.

[69] Marta Laskowski, Verônica a Grieneisen, Hugo Hofhuis, Colette a Ten Hove, Paulien Hogeweg, Athanasius F M Marée, and Ben Scheres. Root system architecture from coupling cell shape to auxin transport. *PLoS Biol.*, 6(12):e307, December 2008.

[70] M Lawson, B Drawert, Mustafa Khammash, Linda Petzold, and Tau-Mu Yi. Spatial stochastic dynamics enable robust cell polarization. *PLOS Comp Biol*, pages 1–38, Apr 2013.

[71] Daniel R Lewis and Gloria K Muday. Measurement of auxin transport in Arabidopsis thaliana. *Nat. Protoc.*, 4(4):437–451, 2009.

[72] Ottoline Leyser and Stephen Day. *Wiley: Mechanisms in Plant Development - Ottoline Leyser, Stephen Day*. Wiley-Blackwell, 2002 edition, 2002.

[73] V. A Likhoshvai, N. A Omel'yanchuk, V. V Mironova, S. I Fadeev, E. D Mjolsness, and N. A Kolchanov. Mathematical model of auxin distribution in the plant root. *Russ J Dev Biol*, 38(6):374–382, Nov 2007.

[74] Karin Ljung. Auxin metabolism and homeostasis during plant development. *Development*, 140(5):943–50, March 2013.

[75] Karin Ljung, Anna K Hull, John Celenza, Masashi Yamada, Mark Estelle, and Jennifer Normanly. Sites and Regulation of Auxin Biosynthesis in Arabidopsis Roots. 17(April):1090–1104, 2005.

[76] Imperial College London. The laboratory of plant morphogenesis.

[77] Mikaël Lucas, Yann Guédon, Christian Jay-Allemand, Christophe Godin, and Laurent Laplaze. An auxin transport-based model of root branching in arabidopsis thaliana. *PLoS ONE*, 3(11):e3673, Nov 2008.

[78] Ari Pekka Mähönen. *Cytokinins Regulate Vascular Morphogenesis in the Arabidopsis thaliana root Institute of Biotechnology and Department of Biological and Environmental Sciences Division of Genetics Faculty of Biosciences and Viikki Graduate School in Biosciences*. 2005.

[79] Ari Pekka Mähönen, Kirsten Ten Tusscher, Riccardo Siligato, Ondřej Smetana, Sara Díaz-Triviño, Jarkko Salojärvi, Guy Wachsman, Kalika Prasad, Renze Heidstra, and Ben Scheres. PLETHORA gradient formation mechanism separates auxin responses. *Nature*, August 2014.

[80] JM McDonough. Lectures on Computational Numerical Analysis of partial Differential Equations. 2007.

[81] Hans Meinhardt. Models of biological pattern formation : common in plant and animal development.

[82] Elliot M Meyerowitz. Plants compared to animals: the broadest comparative study of development. *Science*, 295(5559):1482–5, February 2002.

[83] Victoria V Mironova, Nadezda a Omelyanchuk, Guy Yosiphon, Stanislav I Fadeev, Nikolai a Kolchanov, Eric Mjolsness, and Vitaly a Likhoshvai. A plausible mechanism for auxin patterning along the developing root. *BMC Syst. Biol.*, 4(3):98, January 2010.

[84] Takashi Miura and Philip K Maini. Periodic pattern formation in reaction – diffusion systems : An introduction for numerical simulation Introduction : Periodic pattern formation. pages 112–123, 2004.

[85] Laila Moubayidin, Riccardo Di Mambro, Rosangela Sozzani, Elena Pacifici, Elena Salvi, Inez Terpstra, Dongping Bao, Anja van Dijken, Raffaele Dello Ioio, Serena Perilli, Karin Ljung, Philip N Benfey, Renze Heidstra, Paolo Costantino, and Sabrina Sabatini. Spatial coordination between stem cell activity and cell differentiation in the root meristem. *Dev. Cell*, 26(4):405–15, August 2013.

[86] Laila Moubayidin, Serena Perilli, Raffaele Dello Ioio, Riccardo Di Mambro, Paolo Costantino, and Sabrina Sabatini. The rate of cell differentiation controls the Arabidopsis root meristem growth phase. *Curr. Biol.*, 20(12):1138–43, June 2010.

[87] Ndivhuwo Mphephu. Numerical Solution of 1-D Convection-Diffusion-Reaction Equation. (October), 2013.

[88] Patrick Müller, Katherine W Rogers, Shuizi R Yu, Michael Brand, and Alexander F Schier. Morphogen transport. *Development*, 140(8):1621–38, April 2013.

[89] D. Muraro, H.M. Byrne, J.R. King, and M.J. Bennett. Mathematical modelling plant signalling networks. *Math. Model. Nat. Phenom*, 7(2):32–48, Jul 2012.

[90] Daniele Muraro, Helen Byrne, John King, and Malcolm Bennett. The role of auxin and cytokinin signalling in specifying the root architecture of Arabidopsis thaliana. *J. Theor. Biol.*, 317:71–86, January 2013.

[91] Moritaka Nakamura, Christian S Kiefer, and Markus Grebe. Planar polarity, tissue polarity and planar morphogenesis in plants. *Current Opinion in Plant Biology*, 15(6):593–600, Dec 2012.

[92] B Neta. PARTIAL DIFFERENTIAL EQUATIONS - LECTURE NOTES, 2012.

[93] Ekaterina S Novoselova, Victoria V Mironova, Nadya a Omelyanchuk, Fedor V Kazantsev, and Vitaly a Likhoshvai. Mathematical modeling of auxin transport in protoxylem and protophloem of Arabidopsis thaliana root tips. *J. Bioinform. Comput. Biol.*, 11(1):1340010, March 2013.

[94] Wartlick O, Kicheva A, and González-Gaitán M. Morphogen gradient formation. *Cold Spring Harb Perspect Biol*, page 1:a00125, 2009.

[95] Y Oda and H Fukuda. Initiation of cell wall pattern by a rho- and microtubule-driven symmetry breaking. *Science*, 337(6100):1333–1336, Sep 2012.

[96] Paul Overvoorde, Hidehiro Fukaki, and Tom Beeckman. Auxin Control of Root Development. pages 1–16, 2010.

[97] D. W. Peaceman and H. H. Rachford, Jr. The Numerical Solution of Parabolic and Elliptic Differential Equations. *J. Soc. Ind. Appl. Math.*, 3(1):28–41, March 1955.

[98] W. A Peer, J. J Blakeslee, H Yang, and A. S Murphy. Seven things we think we know about auxin transport. *Molecular Plant*, 4(3):487–504, May 2011.

[99] Serena Perilli, Riccardo Di Mambro, and Sabrina Sabatini. Growth and development of the root apical meristem. *Curr. Opin. Plant Biol.*, 15(1):17–23, February 2012.

[100] Serena Perilli, José Manuel Perez-Perez, Riccardo Di Mambro, Cristina Llavata Peris, Sara Díaz-Triviño, Marta Del Bianco, Emanuela Pierdonati, Laila Moubayidin, Alfredo Cruz-Ramírez, Paolo Costantino, Ben Scheres, and Sabrina Sabatini. RETINOBLASTOMA-RELATED protein stimulates cell differentiation in the Arabidopsis root meristem by interacting with cytokinin signaling. *Plant Cell*, 25(11):4469–78, November 2013.

[101] Serena Perilli and Sabrina Sabatini. Analysis of root meristem size development. *Methods Mol. Biol.*, 655:177–87, January 2010.

[102] Sara V Petersson, Annika I Johansson, Mariusz Kowalczyk, Alexander Makoveychuk, Jean Y Wang, Thomas Moritz, Markus Grebe, Philip N Benfey, Göran Sandberg, and Karin Ljung. An auxin gradient and maximum in the Arabidopsis root apex shown by high-resolution cell-specific analysis of IAA distribution and synthesis. *Plant Cell*, 21(6):1659–68, June 2009.

[103] JJ Petricka, CM Winter, and PN Benfey. Control of Arabidopsis Root Development. *Annu. Rev. Plant Biol.*, pages 563–590, 2012.

[104] M. P. Pound, a. P. French, D. M. Wells, M. J. Bennett, and T. P. Pridmore. CellSeT: Novel Software to Extract and Analyze Structured Networks of Plant Cells from Confocal Images. *Plant Cell*, 24(April):1353–1361, 2012.

[105] James F Price. Lagrangian and Eulerian Representations of Fluid Flow : Kinematics and the Equations of Motion. 2006.

[106] Przemyslaw Prusinkiewicz and Adam Runions. Computational models of plant development and form. *New Phytol.*, 193:549–569, 2012.

[107] Marie-Christine Ramel and Caroline S Hill. The ventral to dorsal BMP activity gradient in the early zebrafish embryo is determined by graded expression of BMP ligands. *Dev. Biol.*, 378(2):170–82, June 2013.

[108] Luciano Rezzolla. Numerical Methods for the Solution of Hyperbolic Partial Differential Equations. 2005.

[109] Katherine W Rogers and Alexander F Schier. Morphogen gradients: from generation to interpretation. *Annu. Rev. Cell Dev. Biol.*, 27:377–407, January 2011.

[110] Catharine J Roussel and Marc R Roussel. Reaction-diffusion models of development with state-dependent chemical diffusion coefficients. *Prog. Biophys. Mol. Biol.*, 86(1):113–60, September 2004.

[111] Tim Rudge and Jim Haseloff. A Computational Model of Cellular Morphogenesis in Plants. pages 78–87.

[112] Heidi L Rutschow, Tobias I Baskin, and Eric M Kramer. The carrier AUXIN RESISTANT (AUX1) dominates auxin flux into Arabidopsis protoplasts. *New Phytol.*, pages 1–9, July 2014.

[113] K Růžička and M Šimášková. Cytokinin regulates root meristem activity via modulation of the polar auxin transport. *Proc. ...*, 2009.

[114] Beemster G T S and Baskin T I. Analysis of cell division and elongation underlying the developmental acceleration of root growth in arabidopsis thaliana. *Plant Physiol*, 116:1515–1526, 1998.

[115] S Sabatini, D Beis, H Wolkenfelt, J Murfett, T Guilfoyle, J Malamy, P Benfey, O Leyser, N Bechtold, P Weisbeek, and B Scheres. An auxin-dependent distal organizer of pattern and polarity in the Arabidopsis root. *Cell*, 99(5):463–72, November 1999.

[116] Tsvi Sachs. Cell polarity and tissue patterning in plants. *Development*, 113(Supplement_1):83–93, January 1991.

[117] Hitoshi Sakakibara. Cytokinins: activity, biosynthesis, and translocation. *Annu. Rev. Plant Biol.*, 57:431–49, January 2006.

[118] Luca Santuari, Emanuele Scacchi, Antia Rodriguez-Villalon, Paula Salinas, Esther M N Dohmann, Géraldine Brunoud, Teva Vernoux, Richard S Smith, and Christian S Hardtke. Positional information by differential endocytosis splits auxin response to drive Arabidopsis root meristem growth. *Curr. Biol.*, 21(22):1918–23, November 2011.

[119] B. Scheres, H. Wolkenfelt, V. Willemsen, M. Terlouw, E. Lawson, C. Dean, and P. Weisbeek. Embryonic origin of the Arabidopsis primary root and root meristem initials. *Development*, 120(9):2475–2487, September 1994.

[120] A. Shimotohno, N. Sotta, T. Sato, M. De Ruvo, A. F. M. Maree, V. A. Grieneisen, and T. Fujiwara. Mathematical modelling and experimental validation of spatial distribution of boron in the root of Arabidopsis thaliana identify high boron

accumulation in the tip and predict a distinct root tip uptake function. *Plant Cell Physiol.*, pages pcv016–, February 2015.

[121] Stanislav Y Shvartsman and Ruth E Baker. Mathematical models of morphogen gradients and their effects on gene expression. *Wiley Interdiscip. Rev. Dev. Biol.*, 1(5):715–30, 2012.

[122] Richard S Smith, Soazig Guyomarc'h, Therese Mandel, Didier Reinhardt, Cris Kuhlemeier, and Przemyslaw Prusinkiewicz. A plausible model of phyllotaxis. *Proc. Natl. Acad. Sci. U. S. A.*, 103(5):1301–6, January 2006.

[123] Michalina Smolarkiewicz and Pankaj Dhonukshe. Formative cell divisions: principal determinants of plant morphogenesis. *Plant Cell Physiol.*, 54(3):333–42, March 2013.

[124] Chris Somerville and Maarten Koornneef. A fortunate choice: the history of Arabidopsis as a model plant. *Nat. Rev. Genet.*, 3(11):883–9, November 2002.

[125] Arno Steinacher. *Self-organisation of auxin transport in plant cells Mathematical modelling of auxin / proton dynamics at a single cell level*. PhD thesis, University of Sheffield, 2011.

[126] Helen Strutt and David Strutt. Long-range coordination of planar polarity in Drosophila. *Bioessays*, 27(12):1218–27, December 2005.

[127] Jana Svacinova, Ondrej Novak, Lenka Plackova, Rene Lenobel, Josef Holik, Miroslav Strnad, and Karel Dolezal. A new approach for cytokinin isolation from Arabidopsis tissues using miniaturized purification: pipette tip solid-phase extraction. *Plant Methods*, 8:17, 2012.

[128] William D. Teale, Ivan a. Paponov, Franck Ditengou, and Klaus Palme. Auxin and the developing root of Arabidopsis thaliana. *Physiol. Plant.*, 123(2):130–138, February 2005.

[129] William D Teale, Ivan A Paponov, and Klaus Palme. Auxin in action: signalling, transport and the control of plant growth and development. *Nat. Rev. Mol. Cell Biol.*, 7(11):847–59, November 2006.

[130] Huiyu Tian, Krzysztof Wabnik, Tiantian Niu, Hanbing Li, Qianqian Yu, Stephan Pollmann, Steffen Vanneste, Willy Govaerts, Jakub Rolcík, Markus Geisler, Jirí Friml, and Zhaojun Ding. WOX5-IAA17 feedback circuit-mediated cellular auxin response is crucial for the patterning of root stem cell niches in Arabidopsis. *Mol. Plant*, 7(2):277–89, February 2014.

[131] A M Turing. The Chemical Basis for Morphogenesis. In *Philos. Trans. R. Soc. London.*, pages 37–72. 1952.

[132] Susana Ubeda-Tomás, Gerrit T S Beemster, and Malcolm J Bennett. Hormonal regulation of root growth: integrating local activities into global behaviour. *Trends Plant Sci.*, 17(6):326–31, June 2012.

[133] Claes Uggla, Thomas Moritz, Goran Sandberg, and Bjorn Sundberg. Auxin as a positional signal in pattern formation in plants. 93(August):9282–9286, 1996.

[134] David M Umulis and Hans G Othmer. Scale invariance of morphogen-mediated patterning by flux optimization.

[135] Steffen Vanneste and Jirí Friml. Auxin: a trigger for change in plant development. *Cell*, 136(6):1005–16, March 2009.

[136] Jean-Pierre Verbelen, Tinne De Cnodder, Jie Le, Kris Vissenberg, and František Baluška. The Root Apex of Arabidopsis thaliana Consists of Four Distinct Zones of Growth Activities. *Plant Signal. Behav.*, 1(6):296–304, 2006.

[137] Teva Vernoux, Géraldine Brunoud, Etienne Farcot, Valérie Morin, Hilde Van den Daele, Jonathan Legrand, Marina Oliva, Pradeep Das, Antoine Larrieu, Darren Wells, Yann Guédon, Lynne Armitage, Franck Picard, Soazig Guyomarc'h, Coralie Cellier, Geraint Parry, Rachil Koumproglou, John H Doonan, Mark Estelle, Christophe Godin, Stefan Kepinski, Malcolm Bennett, Lieven De Veylder,

and Jan Traas. The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Mol. Syst. Biol.*, 7:508, January 2011.

[138] Petra Žádníková and Rüdiger Simon. How boundaries control plant development. *Curr. Opin. Plant Biol.*, 17:116–25, February 2014.

[139] O Wartlick, P Mumcu, a Kicheva, T Bittig, C Seum, F Jülicher, and M González-Gaitán. Dynamics of Dpp signaling and proliferation control. *Science*, 331(6021):1154–9, March 2011.

[140] Detlef Weigel and Gerd Jürgens. Stem cells that make stems. *Nature*, 415(6873):751–4, February 2002.

[141] T Werner, E Nehnevajova, I Kollmer, O Novak, M Strnad, U Kramer, and T Schmulling. Root-specific reduction of cytokinin causes enhanced root growth, drought tolerance, and leaf mineral enrichment in arabidopsis and tobacco. *THE PLANT CELL ONLINE*, 22(12):3905–3920, Dec 2010.

[142] Lewis Wolpert. Positional information revisited. pages 3–12, 1989.

[143] Lewis Wolpert. The Progress Zone Model for specifying Positional Information. 870:869–870, 2002.

[144] X Zheng, N. D Miller, D. R Lewis, M. J Christians, K.-H Lee, G. K Muday, E. P Spalding, and R. D Vierstra. Auxin up-regulated f-box protein1 regulates the cross talk between auxin transport and cytokinin signaling during plant root growth. *PLANT PHYSIOLOGY*, 156(4):1878–1893, Aug 2011.