CHEMICAL REVIEWS

Review

pubs.acs.org/CR

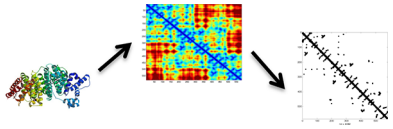
Protein Contact Networks: An Emerging Paradigm in Chemistry

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



Downloaded via INDIAN INST OF TECH GANDHINAGAR on September 19, 2024 at 05:15:10 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

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
Published: November 27, 2012

 © 2012 American Chemical Society dx.doi.org/10.1021/cr3002356 | Chem. Rev. 2013, 113, 1598−1613

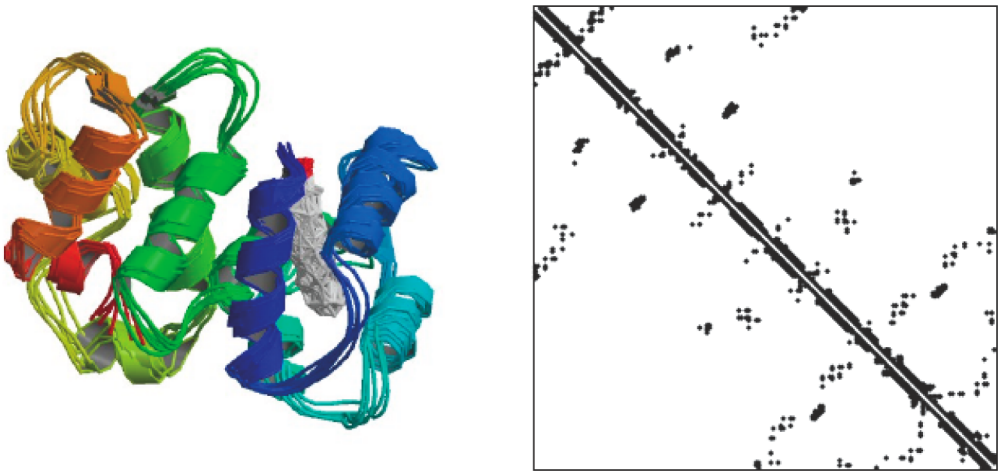


Figure 1. Recoverin 3D structure (left) and correspondent adjacency matrix.

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.* 2013, 113, 1598−1613

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).

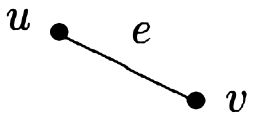


Figure 2. Example of an undirect graph comprising two nodes and an edge.

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. 2013, 113, 1598−1613

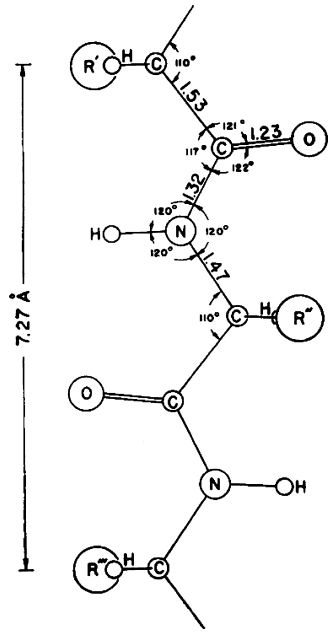


Figure 3. Geometry of the peptide bond: the upper threshold of 8 Å, commonly introduced in the analysis of PCNs, roughly corresponds to two peptide bond lengths.[36]

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

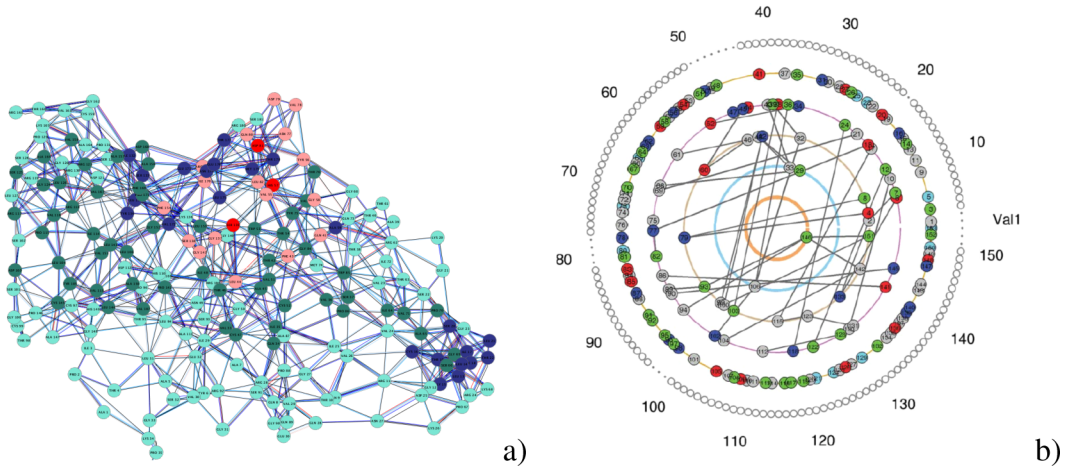


Figure 4. Graph protein formulas: (a) contact map[53] and (b) wheel diagram.[27]

dx.doi.org/10.1021/cr3002356 | *Chem. Rev.* 2013, 113, 1598−1613

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. 2013, 113, 1598−1613

$$P_i = 1 - \sum_{s=1}^{n_M} \left(\frac{k_{is}}{k_i} \right)^2 \quad (2.4)$$

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.

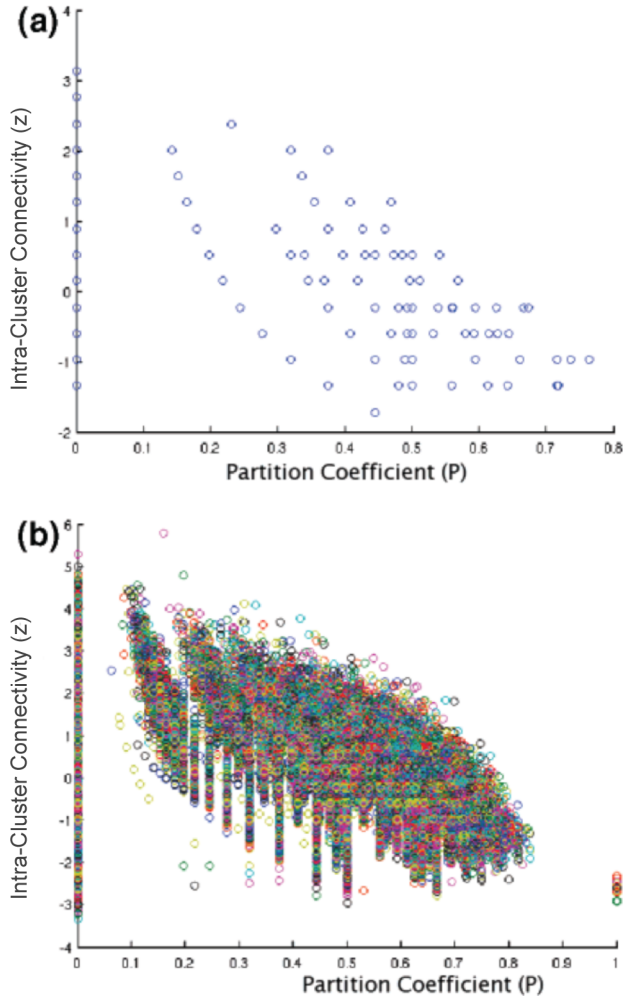


Figure 5. P vs z plot (dentist's chair) (a) for a single protein, where each point identifies a residue, and (b) the superposition of structures analysis of 1420 proteins[58].

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

dx.doi.org/10.1021/cr3002356 | Chem. Rev. 2013, 113, 1598–1613

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. 2013, 113, 1598−1613

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).

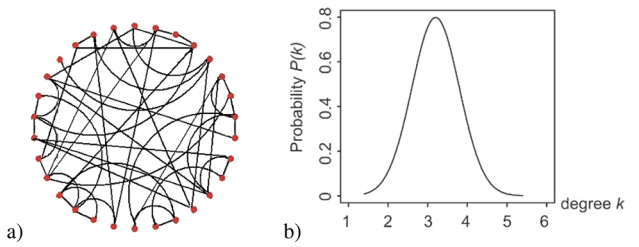


Figure 6. Random graph: (a) a sample picture, where most nodes have three or four links, and (b) the bell-shaped degree distribution.

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

dx.doi.org/10.1021/cr3002356 | Chem. Rev. 2013, 113, 1598−1613

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).

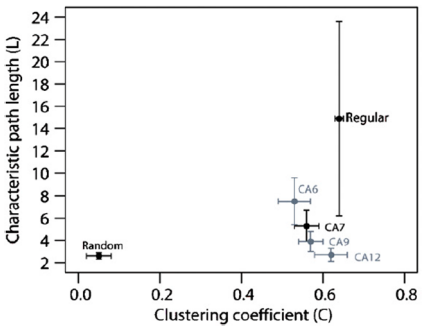


Figure 7. Characteristic path length vs Clustering Coefficient (Figure 3 in ref 104): sample protein classes are labeled as CA#, the label "random" refers to collection of random graphs, whereas "regular" points to periodic lattices.

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):

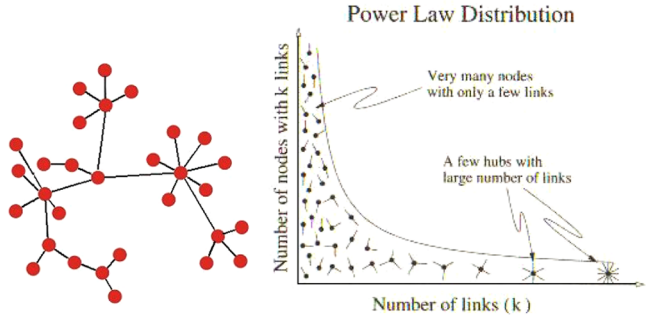


Figure 8. Scale-free networks: (a) a sample scale-free networks, in which few nodes have many links, and (b) the degree distribution of the scale-free graph power law.[106]

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

dx.doi.org/10.1021/cr3002356 | Chem. Rev. 2013, 113, 1598−1613

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]

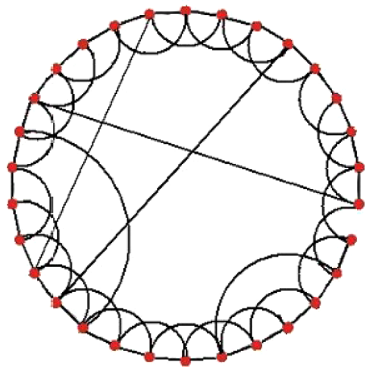


Figure 9. An example of a small-world network: most nodes are linked only to their immediate neighbors, while few edges generate shortcuts between distant regions of the network.

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

dx.doi.org/10.1021/cr3002356 | Chem. Rev. 2013, 113, 1598−1613

(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

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:

dx.doi.org/10.1021/cr3002356 | Chem. Rev. 2013, 113, 1598−1613

(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.

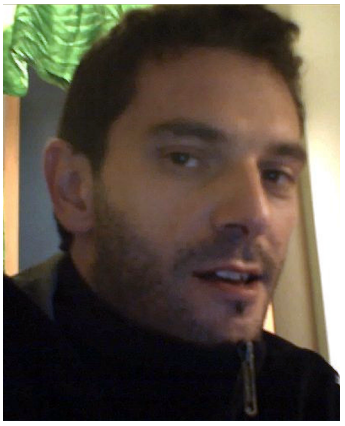
Biographies

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.