Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

Modules Identification in Protein Structures: The Topological and Geometrical Solutions

Setareh Tasdighian,[†] Luisa Di Paola,*,[‡] Micol De Ruvo,[§] Paola Paci,[§] Daniele Santoni,[⊥] Pasquale Palumbo,[⊥,§] Giampiero Mei,[⊥] Almerinda Di Venere,[∥] and Alessandro Giuliani[∥]

[†]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium

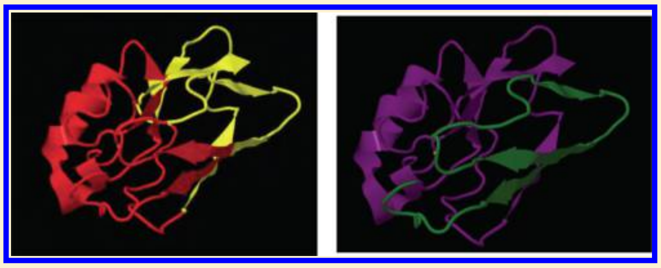
[‡]Faculty of Engineering, Università CAMPUS BioMedico, Via A. del Portillo, 21, 00128 Roma, Italy

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

[⊥]Department of Experimental Medicine and Surgery, University of Rome "Tor Vergata", via Montpellier 1, 00133 Rome, Italy

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

S *Supporting Information*



ABSTRACT: The identification of modules in protein structures has major relevance in structural biology, with consequences in protein stability and functional classification, adding new perspectives in drug design. In this work, we present the comparison between a topological (spectral clustering) and a geometrical (*k*-means) approach to module identification, in the frame of a multiscale analysis of the protein architecture principles. The global consistency of an adjacency matrix based technique (spectral clustering) and a method based on full rank geometrical information (*k*-means) give a proof-of-concept of the relevance of protein contact networks in structure determination. The peculiar "small-world" character of protein contact graphs is established as well, pointing to average shortest path as a mesoscopic crucial variable to maximize the efficiency of within-molecule signal transmission. The specific nature of protein architecture indicates topological approach as the most proper one to highlight protein functional domains, and two new representations linking sequence and topological role of aminoacids are demonstrated to be of use for protein structural analysis. Here we present a case study regarding azurin, a small copper protein implied in the *Pseudomonas aeruginosa* respiratory chain. Its pocket molecular shape and its electron transfer function have challenged the method, highlighting its potentiality to catch jointly the structure and function features of protein structures through their decomposition into modules.

■ INTRODUCTION

Modularity is a central concept in Biology; as a matter of fact, the greater part of biological investigation for centuries has been devoted to looking for the most meaningful clusterization of structures and systems in tissues, organs, and physiological regulation circuits. The main objective of Systems Biology is the exploitation of the modular architecture of biological regulation at different organization scales, as explicitly stated by Denis Noble[1] that equates Systems Biology research to time-honored Physiology studies. The intuition that global behavior of the biological systems can be predicted by a clear picture of the constituting modules is present, for instance, in the task of identification of modules in a complex network of interacting proteins (PPI).[2] This approach relies on the translation of biological systems into graphs: a networklike structure whose elements are nodes and their mutual interactions are expressed as edges connecting them. In protein science, the search for the optimal decomposition of 3D structure into functionally meaningful modules is of utmost importance.[3]

Much work has been devoted to the definition and analysis of the amino-acid residue contact networks. Since the Holm and Sander work on protein comparison by interresidue distance matrices,[4] the field had a huge flourishing with interest in contact networks, describing intramolecular bonds between residues in a protein structure.[5−34] The recent indication of a possible reconstruction of the global 3D information of a protein molecule on the sole basis of its contact map[35] is only one of the myriad of proofs (e.g., refs 12, 14, 19, and 36−39) of the importance of the protein-as-contact-graph paradigm that promises to play the role structural formula played in organic chemistry as the most synthetic and immediate representation able to recover an information-rich picture of the biomolecular system under analysis.

From our viewpoint, the translation of a protein three-dimensional structure into a contact network is a rigorous and

Received: April 10, 2013
Published: December 1, 2013

ACS Publications © 2013 American Chemical Society

simple starting point to search for modularity principles embedded into protein structures.

Modularity detection has evolutionary and physiological implications: as a matter of fact, protein domains are highly conserved throughout a whole proteome, keeping the same function in different protein structures.[40] Thus, the detection of functional regions (domains) has a relevant outcome in the phylogenetic analysis of the protein families. The automatic identification of these domains in protein structures relies on methods (CATH,[40] SCOP,[41] and FSSP[42]) based on secondary and supersecondary structural elements recognition.[43]

Modules identification in protein networks is strictly related to the topological role played by different nodes;[20,44] this purely topological characterization (a residue is defined on the sole basis of its contact pattern in the network) allowed for the identification of key residues in folding process,[45] holding a strong correlation between the strongly connected residue (hub) depletion and the lethality of the corresponding mutation.[17]

The identification of hubs in a network is generally based on the estimation of node "centrality" by different measurement paradigms.[46,47] The higher the betweenness, the stronger the role played in network robustness, so a removal of a high-centrality node is supposed to result in an abrupt network structure modification.

Roughly speaking, a node (amino-acid) endowed with high betweenness centrality is a node by which a lot of "shortest paths" (i.e., the most efficient paths linking one residue to another along the contact network) pass by. The basic analogy in this case is with a transmission network: if a node crucial for maintaining the most efficient (shortest) paths between different regions of the system is perturbed, we can expect a general detrimental effect on the entire system. One possible way to implement this general philosophy of network transmission efficiency (that has its biochemical counterpart in phenomena like allostery[14] and folding[48]) is by shifting from the perspective of paths linking different nodes to the separate consideration of between- and within-module communication. A simple analogy with the partition between highways (between modules) and local streets (within modules) could be of use to introduce this point.

The so-called cartography of Guimerà−Amaral,[49] a method based on complex network clustering, provides at the same time node classification according to their topological role in terms of between- and within-module connections (cartography) and the module (cluster) identification. Previous observations by our group[17,22] highlighted the fact any protein molecule, independently of its general shape and size, gave rise to very similar node role distribution in the Guimerà−Amaral cartography, provided the clustering was "optimal" in a global statistical sense (well-separated clusters), suggesting a sort of "optimal wiring" for a protein system. This feature allows us to check the reliability of different clustering procedures to analyze protein structures by exploring such maps (we call these "dentist's chairs" for their shape).

In this work we focused on the comparison between a geometrical clustering, taking into consideration the physical Euclidean distance between different residues coming from their actual three-dimensional coordinates in space (k-means on residues described by their X, Y, Z coordinates[50]) and a topological clustering, based on the discrete contact matrix in which each residue is described only in terms of its contacts (Shi−Malik spectral clustering[51]). We will show how the two clustering procedures, as applied on a data set of allosteric and nonallosteric proteins in their bound (holo) and free (apo) forms, gave rise to very consistent partitions. This gives a proof-of-concept of the isomorphism between contact networks and three-dimensional configuration. The application of both spectral and k-means clustering methods and consequent comparison is novel and first applied in this work.

The statistical analysis of different modularity descriptors allowed us to recognize some general principles of protein architecture. Once stated the statistical superposition of the two methods, we demonstrated, in a specific study, how the topological method overcomes the geometrical one with respect to functional modules identification.

Moreover, we tested both methods on a case study, regarding the azurin, a small copper protein with a remarkable pocket shape. We demonstrated that the nontrivial shape, able to decouple pure geometry and topology, provides a discrimination case for the two methods, putting into light the potentiality of the topological approach (spectral clustering) in catching jointly structural and functional features of the protein molecules.

Reminding the contact matrix represents a dramatic collapse of information with respect to X, Y, Z space, this result points to the peculiar role played by contacts topological metrics in protein function.

■ MATERIALS AND METHODS

Protein Data Set. Statistical analysis was performed on a protein data set comprising 5 different protein molecules, for a total of 11 protein structures, with different allosteric properties; we analyzed both the apo and the holo forms (listed in Table 1), to detect major descriptors, able to catch the allosteric nature of their function. Their biological role and chemicophysical properties are described thoroughly in ref 14. This choice was dictated by the need to have a consistent, albeit relatively small, set of proteins with common shared properties. The crucial role that between modules communication exerts in allosteric systems and its variation between holo and apo forms[14] are relevant reasons that prompted us to focus on the present data set.

Table 1. Protein Data Set

	PDB ID	
protein name	apo	holo
calcium binding proteins		
calbindin D_{9k}	1CLB	2BCA
parvalbumin (PV)	2NLN	1TTX
recoverin (RC)	1IKU	1JSA
human hemoglobin (Hb_{O2}, Hb_{CO})	2DN2	1GZX (O_2), 1BBB (CO)
human serum albumin (Ab)	1AO6	1E7I (stearic acid)

The specific case study taken into account to investigate the peculiar properties of geometrical and topological approaches refers to the azurin structure (PDB code 1AZU, from *Pseudomonas aeruginosa*). Azurin is a small (14.6 kDa) redox protein that works as a mediator in the electron transfer system of denitrifying bacteria. Its ability to induce apoptosis in mammalian cells[52] has raised a growing interest in studying its binding capability to the mammalian tumor suppressor p53,[53] since it could act as a new therapeutic agent against cancer cells.[52] Azurin tertiary structure[54] is characterized by an eight strand, b-barrel motif,[55] which surrounds an hydrophobic cavity containing a buried tryptophan residue, whose peculiar fluorescence dynamics allowed a number of unique structural features of the molecule to be revealed. For instance, high pressure measurements revealed that even at 3000 bar the hydrophobic core is not accessible to water molecules, thanks to the rather though and rigid scaffolding, which guarantees the protein a relevant conformational stability.[54] Indeed, enlarging the volume within the tryptophan cavity by site-directed mutagenesis[56] enhanced the protein flexibility, leading to a less stable structure.[54]

The presence of a remarkable pocket shape involving residues in the middle of the sequence decouples topological and geometrical views of the molecular structure so providing a very important test to distinguish topological and geometrical description of protein structure.

Definition of Protein Contact Network. The protein contact network is an undirected, unweighted graph; it is built on the basis of the distance matrix **d**, whose generic element d_{ij} records the Euclidean distance between the *i*th and the *j*th residue (measured between the corresponding α carbons). Then, we established a cutoff for inter-residue distance ranging within $\mathcal{I} = [4-8]$ Å accounting for intramolecular noncovalent interactions;[57] thus, the corresponding unweighted protein structure graph was built up, whose adjacency matrix $\mathbf{A} = \{A_{ij}\}$ is formally defined as follows:

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

Once **A** has been defined, different topological descriptors can be directly computed:

1. adeg (average degree): the node degree k_i represents the number of contacts the aminoacid is involved into, and it is defined as

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

where *N* is the overall number of nodes; the average degree is the mean of k_i over the whole node set. The node degree is a direct measure of a single node relevance in the global wiring, while the average value is an indicator of the overall intramolecular interaction strength that, last, corresponds to protein stability and stiffness.

2. acc (average clustering coefficient): the clustering coefficient C_i is defined as the measure of how much the neighbors of a single node are close to each other:

$$C_i = \sum_{j,m \in N, j \neq m} \frac{2A_{ij}A_{jm}A_{mi}}{k_i(k_i - 1)} \tag{2}$$

Residues with high clustering coefficients are likely to be crucial for the protein structure stability, since they play a key role in keeping the community on by the establishment of mutual relationships.

The average clustering coefficient is the mean of C_i over all nodes and represents a mean value of the local wiring strength in the protein structure.

3. asp (average shortest path): the shortest path sp_{ij} indicates the minimum number of links that connect the *i*th and *j*th residues; the average value over the whole set of residues pairs is called asp, that is a measure of the interresidue communication across the whole network. The minimization of asp is crucial for proteins, especially in the case of concerted motions, like those occurring in allosteric transitions.[14,37]

Clustering Algoritms. Clustering procedures are aimed at partitioning the whole network into communities (modules) that present specific intracommunity and intercommunity connectivity patterns; in this respect, the intramodule and intermodule connectivities are represented respectively by the two following parameters:[49]

- Within-module *z*-score

$$z_i = \frac{k_i - \overline{k_{si}}}{\sigma_{si}} \tag{3}$$

k_i is the overall node *i*th degree, k_{si} is the average value of the within-module degree and σ_{si} is its corresponding standard deviation of k_{si} distribution.

- Participation coefficient that describes the attitude of the node to connect to other nodes outside of its own module

$$P_i = 1 - \left(\frac{k_{si}}{k_i}\right)^2 \tag{4}$$

According to the *P* and *z* values, Guimerà et al.[49] established a cartography for the nodes, based on their role in terms of intra- and intermodule connections, on the basis of *P*/*z* values of residues (Table 2).

Table 2. Guimerà−Amaral Role Cartography

role	*z*	*P*
R1: ultraperipheral node	<2.5	<0.05
R2: peripheral node	<2.5	0.05 < P < 0.625
R3: nonhub connectors	<2.5	0.625 < *P* < 0.8
R4: nonhub kinless nodes	<2.5	>0.8
R5: provincial hubs	>2.5	<0.3
R6: connector hubs	>2.5	0.3 < *P* < 0.75
R7: kinless hubs	>2.5	>0.75

P/*z* curves have a special shape for protein structures,[22] called the dentist's chair, that is absent in other biological networks.[2] We recognized that a noticeable threshold for *P* is 0.75: indeed, nodes with $P = 0.75$ play a really special role in their community, since they share exactly half of their links with nodes belonging to their own cluster, while the remaining half is spent to establish connections with nodes belonging to other communities (from eq 4, for $P = 0.75$, $(k_{si}k_i)^2 = 0.25$, thus $k_{si} = 0.5k_i$). In this respect, they represent a frontier between the inner part of the community, with a strong homeland structure, and the outside world. On the other hand, nodes with $P > 0.75$ are mostly devoted to establish connections with other communities than to participate to their own community stability and structure: for this reason, we report the percentage of both classes of nodes (see Figure 1). Hereinafter, we are indicating with $P_{0.75}$ the percentage of nodes having a value of $P = 0.75$ (enclosed in the area sketched in Figure 1) and denoting with $P_{>0.75}$ the percentage of those whose *P* value exceeds 0.75.

On a general perspective, the distribution of residues in the *P*/*z* plane gives a mesoscopic view of the protein from the viewpoint of clusters (modules): the higher the percentage of nodes with $P > 0.75$, the wider the module's concertated motions; the higher $P = 0.75$ residues number, the larger the between-module contact surface.

Another very informative clustering representation is what we call the clustering color map that projects the results of topological network analysis on the sequence space (Figure 2).

The modularity *M* is a global parameter, able to catch the presence of well identifiable clusters, i.e. how well the separation

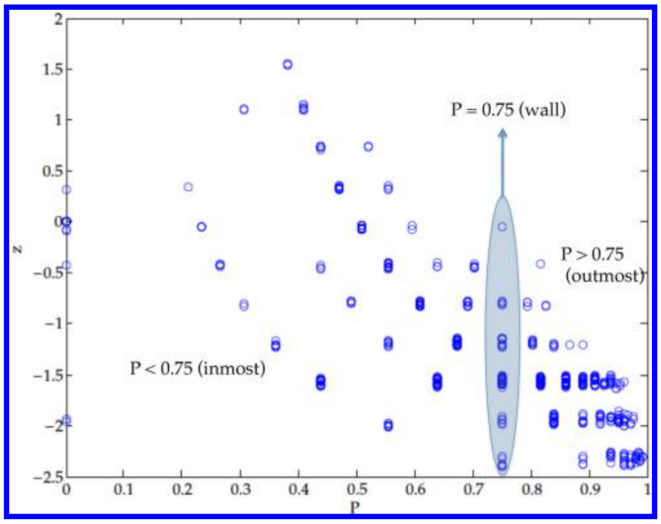


Figure 1. Dentist's chair map. The different roles of nodes on the basis of *P* values are shown in Table 2. The map refers to the hemoglobin structure (PDB code 1HBB) partitioned into 4 clusters.

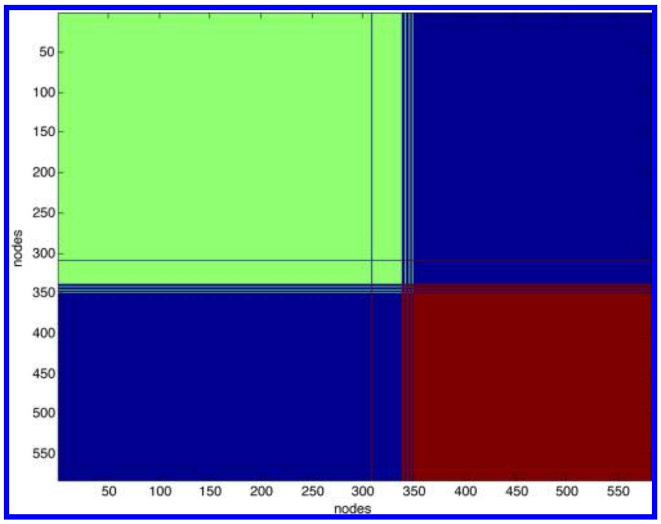


Figure 2. Clustering color map. The reports the map coming from human serum albumin structure (PDB code 1E7I) decomposition into two modules. The blue background represents (*i*, *j*) residue couples belonging to different clusters, whereas the couples pertaining to the same cluster are coded with the same color (other than blue). Coordinates of the map correspond to the protein sequence, thus the residues belonging to a given cluster can be easily identified accordingly. The interruptions of the continuity between sequence and cluster location correspond to long-range contacts in the molecule, putting into contact far in sequence residues; the corresponding residues show high value of *P*, pointing to their character of intermodule communication.

of amino-acids in modules is supported by their actual locations; it is defined as:[58]

$$M = \frac{\text{between-cluster variance}}{\text{total variance}} = \frac{\sum_{i=1}^{N} {r_{i,S_i}}^2}{\sum_{i=1}^{N} r_i^2} \tag{5}$$

where $r_{i,Si}$ represents the distance of each residue from the protein center of mass, when the residues coordinates correspond to those of its own cluster center of mass; r_i is the distance of a node from the center of mass of the whole structure: so, the higher *M*, the higher the system's modularity. *M* corresponds to an R^2 ordinary statistics, its numerator being the variance explained by the model, whereas its denominator is the total variance. With a number of clusters *N* we get a trivial unitary value for *M* (no intracluster variance). When approaching an optimal partition number, *M* approaches to a plateau.[58] We compute *M* with respect to spectral (Shi–Malik $M_{(sc)}$) and *k*-means ($M_{(m)}$) clustering algorithms.

Finally, to estimate the superposition of the two algorithms, we introduce the Rand index *R*;[59] it is generally defined in terms of number of pairs that change mutually the clusters they belong, so it is a direct measurement of the percentage of elements that are (or are not) in the same cluster for a method A and are not (or are) any more in the same cluster as for the method B. Specifically, it is defined as

$$\text{Rand} = 1 - [\text{number of pairs that are (are not)} \\ \text{in the same cluster(A/B)}] / \binom{N}{2} \tag{6}$$

The closer Rand is to unity, the more the two methods match.

Shi–Malik (Spectral Clustering). Spectral Clustering is one of the most well-known algorithms for clustering.[51] The algorithm benefits from some concepts in the linear algebra related to graph Laplacians and similarity matrixes.

We have performed the algorithm application on the adjacency matrix **A**, by defining the corresponding Laplacian matrix **L** defined as:

$$\mathbf{L} = \mathbf{A} - \mathbf{D} \tag{7}$$

where **D** is the degree matrix, defined as the diagonal matrix whose generic non-null element corresponds to the *i*th residue degree ($D_{ii} = k_i$).

The next step has been to calculate the eigenvalues of **L** and put them in an ascending order. For mathematical reasons, the first eigenvalue is always equal to zero, then, $\lambda_1 = 0 < \lambda_2 < ... < \lambda_N$. In this way, the corresponding eigenvectors are ordered accordingly: $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_N$. The second eigenvector $\boldsymbol{v}_2$, named Fiedler's vector is used to split the set into two clusters (and, further, each resulting subset into two smaller subsets iteratively) according to the sign of its corresponding components.

The method presents advantages and drawbacks: the spectral clustering algorithm is easy to implement and stable under perturbation of the data set. It is an exact method, so it is not sensible to the initial conditions choice; however, it requires the computation of the Laplacian eigenvector for each step, that is a burdening computational step. Moreover, the method is not recommended for large graphs, since spectral graph partitioning falls in the class of NP-hard problems, having a high-order complexity with respect to the number of set elements.

k-Means. *k*-Means is an unsupervised learning and partitioning clustering algorithm, targetted at partitioning a data set into *k* clusters.[60] In the first step, the algorithm chooses *k* data points as centroids and then assigns each data point to its nearest centroid. In the next step *k*-means calculates a value called sum square error (SSE), for each cluster, defined as

$$\text{SSE}_k = \sum_{i=1}^{n_k} (x_i - c_k)^2 \tag{8}$$

where n_k is the number of data points in the *k*th cluster; x_i represents the coordinates of the generic *i*th node in the *k*th cluster, whose centroid (the center of mass of the cluster nodes) position is c_k. Next, *k*-means tries to vary the choice of centroid to find out the minimum possible value for the SSE_k and finally

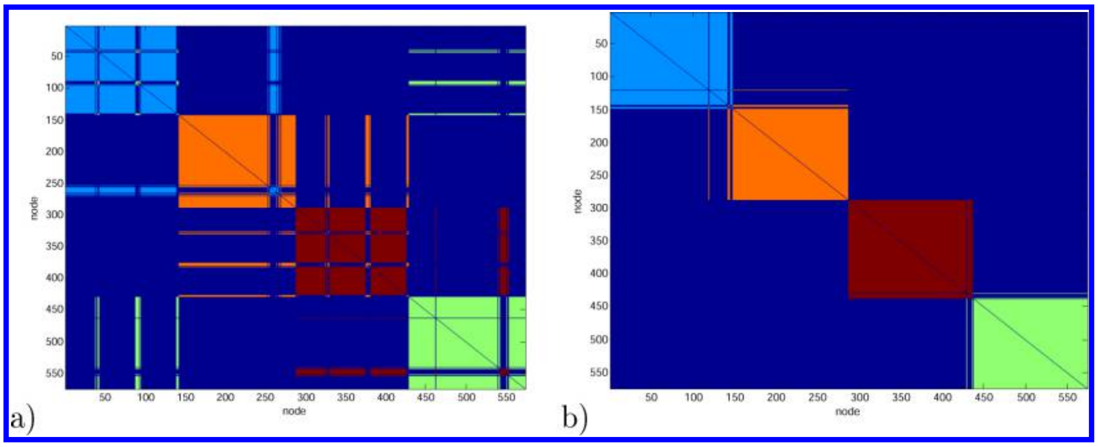


Figure 3. Clustering color map of hemoglobin residues: (a) spectral clustering; (b) *k*-means. The general similarity of the two partitions is noteworthy. Spectral clustering (a) is endowed with horizontal (vertical) lines starting from the cluster core and going forward; they represent the residues with high *P* values, establishing between-cluster links.[61] The geometrical character of *k*-means, on the other hand, results into a sharp transition partition along the sequence, tightly linked to residue spatial position.

Table 3. Topological Descriptors for the Protein Data Set[a]

PDB code	K_{OPT}	$\mathrm{Rand}_{(K_{\mathrm{OPT}})}$	adeg	acc	asp	N/K_{OPT}	$P^{(\mathrm{sc})}_{0.75}$	$P^{(\mathrm{sc})}_{>0.75}$	$M_{(\mathrm{sc})}$	$M_{(k\mathrm{m})}$	A/H
					hemoglobin						
1HBB	4	0.9047	3.9390	0.2744	6.3922	143.5	10.10	57.14	0.6083	0.6083	A
1BBB	4	0.9278	3.9059	0.2688	6.4042	143.5	10.98	58.89	0.6047	0.6050	H
1GZX	4	0.9282	3.9321	0.2734	6.3970	143.5	10.63	56.62	0.6161	0.6159	H
					albumin						
1AO6	2	0.9794	3.6419	0.2402	10.7505	291	19.46	61.42	0.4407	0.4407	A
1E7I	2	0.8960	3.5069	0.2425	8.0870	291	24.05	63.40	0.3424	0.3395	H
					calbindin						
1CLB	4	0.8782	3.2667	0.3083	3.1168	18.75	16	68	0.5216	0.5031	A
2BCA	4	0.9052	3.48	0.3087	3.0897	18.75	9.33	56	0.5705	0.5925	H
					parvalbumin						
2NLN	8	0.8787	4.2037	0.3435	3.1722	13.5	11.11	66.67	0.7756	0.8036	A
1TTX	8	0.8714	3.5688	0.2845	3.4373	13.5	7.34	64.22	0.7470	0.7537	H
					recoverin						
1IKU	8	0.8588	3.5638	0.2624	4.3864	23.5	14.36	68.62	0.7454	0.7819	A
1JSA	8	0.8693	3.2819	0.2555	4.7849	23.5	17.55	69.68	0.7651	0.8882	H

[a] *P* and *M* values refer to the Shi–Malik algorithm application, K_{OPT} is estimated by the *k*-means application, and A/H refers to the apo/holo form.

updates the centroid of the *k*th cluster to the data point which gives the minimum SSE_k.

When all the centroids are updated (no unit changes its centroid at the next iteration), *k*-means calculates a more general SSE value for all the data set, defined as

$$\mathrm{SSE} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_i - c_k)^2$$

If SSE is less than a given threshold, the algorithm stops; otherwise, it goes back to the first step and starts over by choosing different initial centroids.

To choose the correct number of clusters, one has to run *k*-means over and over again for different values of *k* and calculate the SSE in each case. The optimal number of clusters would be the one corresponding to the minimum SSE. One of the central issue of the *k*-means method deals is the choice of the initial centroids and with how this choice affects the final results of the algorithm.

The method presents some pros and cons: *k*-means is fast and easy to implement. On the other hand, the final clustering results depend on the initial choice of the centroids. Moreover, it has to be run it several times to determine the right number of clusters. Finally, if the number of clusters *k* is defined, it takes a time that increases largely with the number of clusters but depends less strongly on the number of elements. So, it represents a good choice in the case of few clusters partition among a large number of elements, as in the case of modules identification in protein contact network.

■ RESULTS AND DISCUSSION

Statistical Analysis. The results of the topological analysis and clustering are sketched out in Figure 3, where all descriptors refer to each protein of the data set.

To get a direct picture of the cluster partition, in Figure 3 a color map representation of clusters for spectral clustering and *k*-means is shown, relative to the partition into four clusters of the hemoglobin structure (PDB code 1HBB).

Looking at results reported in Table 4 (coming from the analysis of the raw data of Table 3) it is worth noting the concordance between the spectral and the *k*-means clustering as measured by the Rand index Rand[59] is very high. Rand has an average value of 0.90; it never goes below 0.86 and has a maximal value of 0.98. Keeping in mind that the Rand maximum value

Table 4. Statistical Analysis of Topological Descriptors

	mean	std dev	minimum	maximum
K_{OPT}	5.09091	2.42712	2	8
N	329.81818	239.70015	75	282
$\mathrm{Rand}_{(K_{OPT})}$	0.89979	0.03511	0.85880	0.97940
adeg	3.66279	0.29652	3.26670	4.20370
acc	0.27838	0.03122	0.24020	0.34350
asp	5.45618	2.44072	3.08970	10.75050
N/K_{OPT}	102.18182	108.83865	13.5	291
$P^{(sc)}_{0.75}$	13.71909	5.06835	7.34	24.05
$P^{(sc)}_{>0.75}$	62.78727	5.09175	56	69.68
$M_{(sc)}$	0.61249	0.14079	0.34240	0.34240
$M_{(km)}$	0.63022	0.16559	0.33950	0.88820

(complete concordance of the two classifications) is equal to 1, we can safely state that the apparently dramatic collapse of information going from actual geometric coordinates of residues to the corresponding adjacency matrix maintains the relevant mesoscopic features (domains) of macromolecule structure.

Having safely established the basic coherence between the spectral and *k*-means clustering procedures (image in light of the preservation of the relevant spatial information by the adjacency matrix), we can go more in depth into the different modular character of the studied proteins as estimated by different descriptors. For this reason, we compute the Pearson correlation coefficient *r* between different descriptors so to get a consistent picture of protein modularity. $M_{(sc)}$ and $M_{(km)}$ measure the proportion of variance explained by clusterization (spectral and *k*-means respectively, see Materials and Methods), so they keep track of the amount of "modularity" of the studied protein correspondent to the optimal (K_{OPT}) number of clustering. For simple statistical reasons, the value of both $M_{(sc)}$ and $M_{(km)}$ grows with K_{OPT} ($r(K_{OPT}, M_{(sc)}) = 0.94$ and $r(K_{OPT}, M_{(km)}) = 0.94$ in our data set), but the important point is that spectral and *k*-means clusterizations give the same estimation of the relative modularity of the analyzed proteins, with a mutual correlation near to unity ($r(M_{(km)}, M_{(sc)}) = 0.98$). This correlation between the estimation of modularity relative to the two methods does not depend on number of clusters (K_{OPT}); as a matter of fact, when separated out of the common effect of K_{OPT}, the correlation coefficient between $M_{(km)}$ and $M_{(sc)}$ is still very high ($r(M_{(km)}, M_{(sc)})_{K_{OPT}} = 0.84$).

The partial correlation technique allows us to discover some other interesting features about the definition of a modularity metrics. In this work we obtained three different viewpoints on "degree of modularity":

1. a global one ($M_{(sc)}$ and $M_{(km)}$) computed over the entire molecule;
2. a local one measured at the level of single residue (acc);
3. a mesoscopic, indirect one, correspondent to asp.

The relation between asp and modularity asks for further clarification. We can imagine two opposite scenarios for the relation between the characteristic length of between residues shortest paths (asp) and the degree of clusterization:

Scenario 1: If a clear and well discriminated clusterization does exist (let us keep in mind the original statistical definition of a well formed cluster is "a set of elements for which the intracluster distances are much lower than the between-cluster distances"), the between-cluster separation (as a mountain range between two cities) is a hurdle to the establishment of short-cuts between residues pertaining to different clusters, so implying a POSITIVE correlation between asp characteristic length and global modularity ($M_{(sc)}$).

Scenario 2: On the other hand, if a high number of "local paths" is present (this means high values of acc, local clustering measure) due to a strong clusterization (cluster compactness comes from the richness of edges linking residues pertaining to the same cluster), it is sufficient to establish relatively few "short cuts" between clusters to serve the entire cluster community and optimize between-cluster communication. In this case, the clusterization degree has an opposite effect with respect to scenario 1, and asp should have a NEGATIVE correlation with global modularity ($M_{(sc)}$). In other words, the relative importance of the two drivers of modularity (within-cluster compactness and between-cluster separation) is at the basis on the two possible effects of modularity on the between-residues communication efficiency, as expressed by asp. Scenario 2 has to do with the "small-world" character of networks:[62] small-world networks are those graphs occupying a middle ground between regular and random networks, showing high local clustering of elements, like regular networks, but also short path lengths between elements, like random networks. Small-worldness implies emergent features very different from both random and regular networks, such as a strong resilience to damage and a collective dynamics with few coherent modes.[63] Which of the two above architectural principles are typical of protein graphs wiring can be immediately decided by the mutual relations holding between acc, asp, and $M_{(sc)}$, i.e., between-modularity degree at local, mesoscopic, and global scale. The direct Pearson pairwise correlation coefficients between asp, acc, and $M_{(sc)}$ are reported in Table 5.

Table 5. Correlation Matrix between Different Modularity Descriptors

	asp	acc	$M_{(sc)}$
asp	1	−0.77	−0.64
acc	−0.77	1	0.41
$M_{(sc)}$	−0.64	0.41	1

It is worth noting the prevalence of scenario 2 implying an optimized choice of between clusters paths so to minimize asp: the higher the global modularity (as measured by $M_{(sc)}$), the shorter the characteristic path length ($r(\mathrm{asp}, M_{(sc)}) = -0.64$, $p = 0.03$), thus it can be argued local compactness (acc) is used by protein network wiring as a communication efficiency enhancer ($r(\mathrm{asp, acc}) = -0.77$, $p = 0.006$). There is no strong and significant direct relation between local clustering and global clustering ($r(M_{(sc)}, \mathrm{acc}) = 0.41$, $p = 0.2067$); this interpretation is strengthened by the computed values of partial correlation between $M_{(sc)}$ and acc, demonstrated by the common relation with asp. The correlation drops from 0.41 to a counterintuitive (but practically null) $r(\mathrm{acc}, M_{(sc)})_{\mathrm{asp}} = -0.16$; on the other hand, the correlation coefficient between asp and acc, separated out of $M_{(sc)}$ remains practically invariant $r(\mathrm{acc, asp})_{M(sc)} = 0.72$, $p = 0.019$) and the same happens as for the relation between asp and $M_{(sc)}$ separated out of the common relation with acc ($r(\mathrm{asp}, M_{(sc)})_{\mathrm{acc}} = 0.55$). Thus, it is confirmed that the mesoscopic character of asp mediating the local residue scale (acc) and global ($M_{(sc)}$) network modularity, pointing to the crucial role played by characteristic path length (asp) in protein functionality. Only through of the mediation of modules organization, the local (acc) and global ($M_{(sc)}$) structures are put into relation.

The point of between-module efficient communication will be further demonstrated to be the main determinant of the specific differences between topological and geometrical approaches when analyzing in depth azurin molecules. Moreover the link between most efficient paths and actual inside molecule signal transmission will be demonstrated in the case of electron transfer.

We get some more clues about the residues maintaining the global network communication efficiency through the analysis of dentist-chair graphs: the variable $P_{0.75}$ identifies residues having exactly the same number of connections within its own cluster and with neighboring cluster residues (see Materials and Methods). There is an excess of these "communicating residues" that represent on average the 13% of aminoacids. The $P = 0.75$ threshold parts the dentist-chair into two zones: on the left we have the proper "well-clustered" population of residues (intra-cluster connections outnumber intercluster ones) responsible for "keeping alive" the modular properties of the molecule, whereas on the right side we find the residues whose intercluster relations outnumber intracluster ones. As expected, Rand is negatively correlated with $P_{>0.75}$ ($r(\text{Rand}, P_{>0.75}) = -0.64$, $p = 0.03$) given these residues fuzzify the modular structure decreasing the spectral-geometrical correlation. We hypothesize "frontier residues" are the most crucial ones for cluster communication, while over-the-frontier residues roughly correspond to the residues Csermely and Nussinov indicate as "creative elements", endowed with a crucial role in protein dynamics by allowing the system to attain different configurations.[39,64] The statistical analysis of our data set allowed to sketch some peculiar properties of protein modular organization consistently with the small-world character of contact graphs. Well-formed domains are not maximally separated and distinct parts of the molecule, but modules whose identity must go hand-in-hand with richness of between-clusters contacts so to optimize the efficiency of signal transmission through the entire molecule. The coexistence of cluster identity and between-modules connections is a conundrum for k-means clustering based on geometry, but not for topological spectral clustering; for this reason in the next section, we will abandon the general perspective to go in depth into the analysis of a single protein structure.

Case Study. Figure 3 shows a different character of the k-means and spectral clustering solutions: while k-means strictly follows sequence order, spectral clustering highlights the peculiar role of residues in charge of intermodule communication. These residues appear in the as the disarranged lines in the spectral clustering profile, interrupting the strict superposition between structural cluster and sequence position. Figure 3 refers to hemoglobin, whose residues (nodes) mismatching the cluster location in the two methods are lying at the interlocking interfaces between the folded chains. The hemoglobin case, anyway, is too straightforward and in some sense trivial: the symmetric tetrameric structure is well characterized by both methods, even though some details are different. Thus we moved on a specific case study, regarding the azurin (arsenate reductase), a small, monomeric copper protein, whose structure presents a pocket enclosing the prosthetic group (copper; Figure 4).

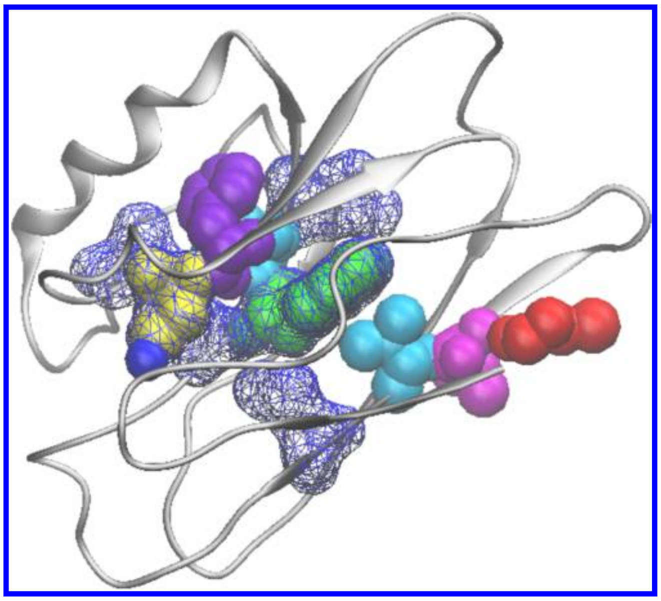


Figure 4. Ribbon structure of azurin. The model shows from right to left (filled van der Waals) the preferential pathway obtained in pulse radiolysis measurements from Cys 3 to Cu.[65] The residues are in blue, and empty wireframes are the aminoacids that form the hydrophobic cavity of azurin.[66]

We tested the two clustering methods as for detection of the pocket structure, and results are shown in Figures 5 and 6 (partitioned into two clusters). In this case, the two methods mismatch largely, showing two different pictures: while k-means simply splits the sequence into the two terminal regions (Figure 5), the result for spectral clustering tells of a strongly entangled structure, where the sequence position and the cluster location are loosely related. However, still a partition in two gross sequence emerges, the first (green zone in color map of Figure 5a) comprising residues located in both terminals, and a central cluster, including the copper pocket (red zone in the color map of Figure 5a). In Figure 6a, the same result is reported in the usual ribbonlike representation where the clusters correspond to different colors on the sequence. It is evident that in Figure 6a that the copper pocket comprises only one cluster (red), while both terminals fall in the other cluster (yellow region). On the contrary, k-means clustering results in a pocket splitting in two different clusters that, in turn, correspond to the two terminals (purple and green regions in Figure 6b); differences in clustering are mirrored in a low Rand index (0.52) as well.

Additionally, we report the SSE and $M_{(km)}$ profiles for the application of k-means for different clusters number relative to azurin structural analysis (Figure 7). It is clear that there is no sharp cut corresponding to an optimal number of clusters, that is a typical result for the k-means application in cases of disjoint sets of nodes (spatial points) where the optimal number corresponds to the number of well-parted groups; moreover, $M_{(km)}$ never exceeds 0.65 even with a relatively large number of clusters, pointing to a loose module identification by k-means.

In the case of the molecular protein structures, the points (residues) are homogeneously distributed into the space, thus looking for an optimal number of geometrical clusters does not make any sense. All in all we can say only spectral clustering, concentrating on an apparently poorer information (contact graph instead of the actual geometrical coordinates representation) is able to catch the functionally relevant structure partition.

The "topology first" approaches, like spectral clustering, have an additional advantage with respect to purely geometrical methods; namely, they derive from the nature of the information transfer across the molecule, following the path of between-residue contacts, following the track of the seminal work of Skourtis and Beratan.[67] This is particularly clear in the case of azurin. In the past years, thanks to its small size and to the presence of a single copper center, azurin has been used as a

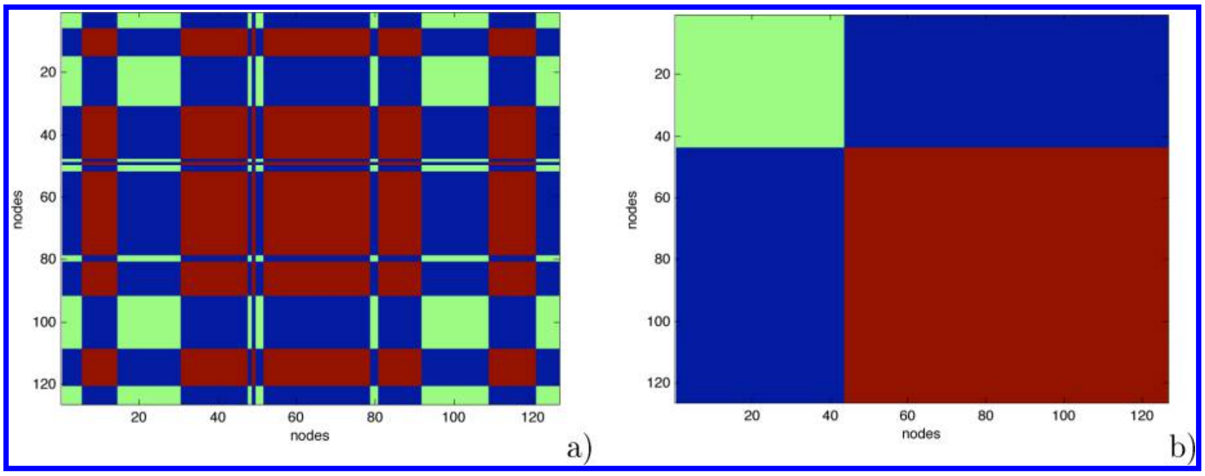


Figure 5. Color map of azurin clustering, comparison between the two methods: (a) spectral clustering; (b) *k*-means.

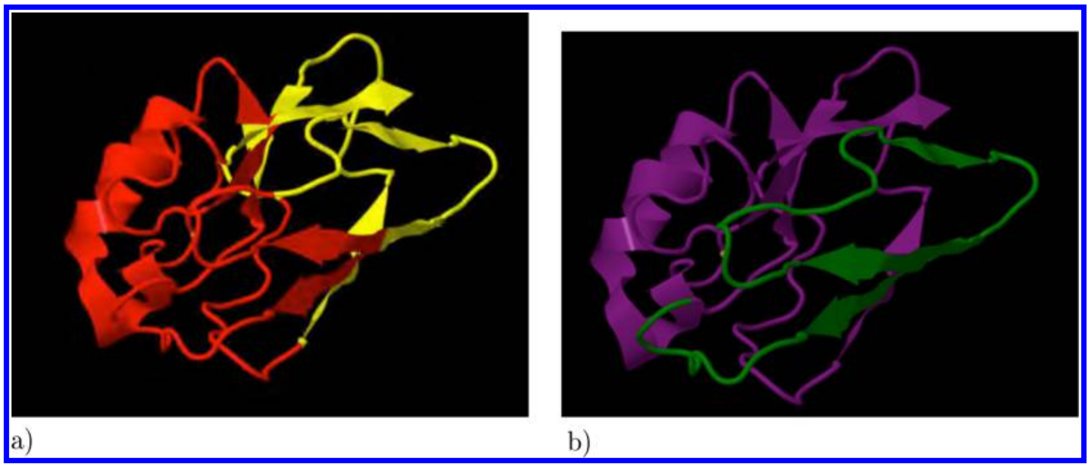


Figure 6. Azurin structure, the different colors refer to different clusters: (a) spectral clustering; (b) *k*-means.

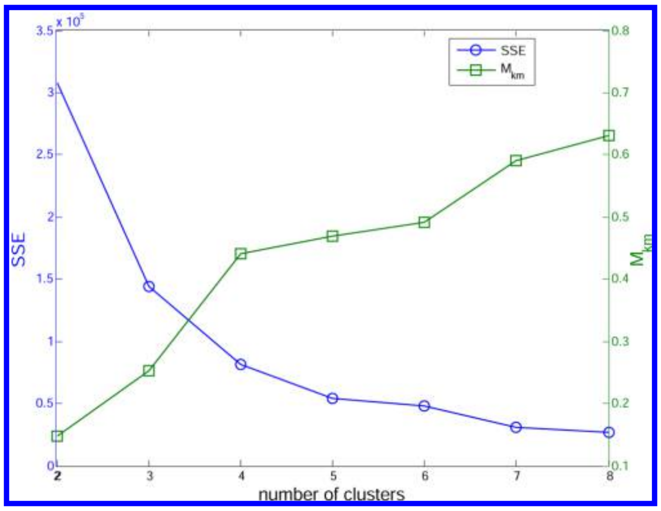


Figure 7. SSE and *M* profiles for *k*-means application for different number of clusters for azurin structural analysis: modularity does not reach a plateau, even for a relatively high number of clusters.

model protein to study the electron transfer pathway in redox protein systems.[68,69] In particular, pulse radiolysis measurements[65] have provided evidence that electrons can flow from the "periphery" of the protein (Cys 3) up to the Cu-bound residue (Cys 112) through a preferential route, which involves the following residues: Thr 30, Val 31, Trp 48, Val 49, Phe 111. It has been demonstrated[70,71] that the transmission path is continuous with the only exception of the Val 31−Trp 48 step, where the electron is hypothesized to move through an electric arc. As a matter of fact, we were able to exactly reconstruct the hypothesized mechanism in topological terms: a noncovalent bond exists between all the steps involving not contiguous residues (contiguous ones are trivially linked by the backbone). The only exception again is Val 31−Trp 48, as predicted by the model. It is worth noting that the Val 49− Phe 111 step involves two residues very far in sequence (62 aminoacids of distance) that occurs extremely rarely (7.4% of total edges refers to contacts between residues far in sequences 60−70 units). Moreover, all the involved residues show negative *z* values and relatively high *P* values, pointing to their connection character.

■ CONCLUSIONS

The most relevant conclusion of our work is the strong consistency between global geometrical information, carried by protein structure, and its representation in terms of adjacency matrix. This allows to think of protein contact networks as a sort of structural formula for macromolecules.[72]

The two graphical representations presented here, clustering color map and dentist's chair, promise to be a complement to the usual ribbon diagrams to identify physiologically relevant aminoacid residues in protein structures.

Considering protein structures as contact networks allows to give a rational and consistent definition of modularity, going from local (average clustering coefficient) to global (percentage of information explained by clusters) by the mediation of the average shortest path asp. The maximization of communication efficiency across the entire molecule by minimizing the average shortest path can be considered as a crucial architectural principle governing protein structures. Tracing back this mesoscopic feature by network cartography to single-residue roles in the global wiring promises to be a precious tool to locate the relevant residues in a protein molecular system.

■ ASSOCIATED CONTENT

Supporting Information

Method applied on an additional data set, composed of different categories of structures, classified on the basis of their structural properties. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: l.dipaola@unicampus.it.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Noble, D. *Exp Physiol* **2008**, *1*, 16−26.

(2) Agarwal, S.; Deane, C.; Porter, M.; Jones, N. *PLoS Comput Biol* **2001**, *17*, e1000817.

(3) Baron, M.; Norman, D.; Campbell, I. *Trends Biochem. Sci.* **1991**, *16*, 13−17.

(4) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123−138.

(5) Aftabuddin, M.; Kundu, S. *Phys. A* **2006**, 895−904.

(6) Bagler, G.; Sinha, S. *Phys. A* **2005**, *346*, 27−33.

(7) Barah, P.; Sinha, S. *Pramana* **2008**, *71*, 369−378.

(8) Bartoli, L.; Fariselli, P.; Casadio, R. *Phys Biol* **2008**, *4*, L1−L5.

(9) Brinda, K.; Surolia, A.; Vishveshwara, S. *Biochem. J.* **2005**, *391*, 1−15.

(10) Brinda, K.; Vishveshwara, S. *Biophys. J.* **2005**, *89*, 4159−4170.

(11) Brinda, K. V.; Vishveshwara, S.; Vishveshwara, S. *Mol. Biosyst.* **2010**, *6*, 391−398.

(12) Csermely, P.; Sandhu, K.; Hazai, E.; Hoksza, Z.; Kiss, H.; Miozzo, F.; Veres, D.; Piazza, F.; Nussinov, R. *Curr. Protein Peptide Sci.* **2012**, *13*, 19−33.

(13) Dehmer, M.; Barbarini, N.; Varmuza, K.; Graber, A. *BMC Struct. Biol.* **2010**, *10*, 1−17.

(14) De Ruvo, M.; Giuliani, A.; Paci, P.; Santoni, D.; Di Paola, L. *Biophys. Chem.* **2012**, *165*−*166*, 21−29.

(15) Di Paola, L.; Paci, P.; Santoni, D.; De Ruvo, M.; Giuliani, A. *J. Chem. Inf. Model.* **2012**, *52*, 474−482.

(16) Doncheva, N.; Klein, K.; Domingues, F.; Albrecht, M. *Trends Biochem. Sci.* **2011**, *36*, 179−182.

(17) Giuliani, A.; Di Paola, L.; Setola, R. *Curr. Proteomics* **2009**, *6*, 235−245.

(18) Greene, L.; Highman, V. *J. Mol. Biol.* **2003**, *334*, 781−791.

(19) Gromiha, M. *J. Chem. Inf. Model.* **2009**, *49*, 1130−1135.

(20) Gurso, A.; Keskin, O.; Nussinov, R. *Biochem. Soc. Trans.* **2008**, *36*, 1398−1403.

(21) Krishnan, A.; Giuliani, A.; Zbilut, J.; Tomita, M. *J. Proteome Res.* **2007**, *6*, 3924−3934.

(22) Krishnan, A.; Zbilut, J. P.; Tomita, M.; Giuliani, A. *Curr. Protein Peptide Sci.* **2008**, *9*, 28−38.

(23) Kundu, S. *Phys. A* **2005**, *346*, 104−109.

(24) Mekenyan, O.; Bonchev, D.; Trinajstic, N. *Int. J. Quantum Chem.* **1980**, *18*, 369−380.

(25) Kim, D.; Park, K. *BMC Bioinf.* **2011**, *12*, 1471−2105.

(26) Plaxco, K.; Simons, K.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985−994.

(27) Sathyapriya, R.; Vishveshwara, S. *Proteins* **2007**, *68*, 541−550.

(28) Sengupta, D.; Kundu, S. *Phys. A* **2012**, *391*, 4266−4278.

(29) Tan, L.; Zhang, J.; Jiang, L. *J. Biol. Phys.* **2009**, *35*, 197−207.

(30) Vendruscolo, M.; Dokholyan, N.; Paci, E.; Karplus, M. *Phys. Rev. E* **2002**, *65*, 061910.

(31) Vijayabaskar, M.; Vishveshwara, S. *Biophys. J.* **2010**, *99*, 3704−3715.

(32) Vishveshwara, S.; Brinda, K.; Kannan, N. *J. Theor. Comp. Chem.* **2002**, *1*, 1−25.

(33) Vishveshwara, S.; Ghosh, A.; Hansia, P. *Curr. Protein Peptide Sci.* **2009**, *10*, 146−160.

(34) Giuliani, A.; Di Paola, L.; Paci, P.; De Ruvo, M.; Arcangeli, C.; Santoni, D.; Celino, M. Proteins as Networks: Usefulness of Graph Theory in Protein Science. In *Advances in Protein and Peptide Science*; Dunn, B., Ed.; Bentham, 2012; in revision.

(35) Chen, J.; Shen, H. *Curr. Bioinf.* **2012**, *7*, 116−124.

(36) Tsai, C.; del Sol, A.; Nussinov, R. *J. Mol. Biol.* **2008**, *378*, 1−11.

(37) del Sol, A.; Araúzo-Bravo, M.; Amoros, D.; Nussinov, R. *Genome Biol.* **2007**, *8*, R92.

(38) Nussinov, R.; Tsai, C.; Csermely, P. *Trends Pharmacol. Sci.* **2011**, *32*, 686−693.

(39) Csermely, P. *Trends Biochem. Sci.* **2008**, *33*, 569−576.

(40) Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. *Structure* **1997**, *5*, 1093−1109.

(41) Murzin, A.; Brenner, S.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536−540.

(42) Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; Vriend, G. *Protein Sci.* **1992**, *1*, 1691−1698.

(43) Csaba, G.; Birzele, F.; Zimmer, R. *BMC Struct. Biol.* **2009**, *9*, 23−33.

(44) Jeong, H.; Mason, S.; Barabàsi, A.; Oltvai, Z. *Nature* **2001**, *411*, 41−42.

(45) Vendruscolo, M.; Paci, E.; Dobson, C.; Karplus, M. *Nature* **2001**, *409*, 641−645.

(46) Barabási, A. L.; Oltvai, Z. N. *Nat. Rev.* **2004**, *5*, 101−113.

(47) Koschutzki, D. Network Centralities. In *Analysis of Biological Networks*; Junker, B., Schreiber, F., Eds.; Wiley Series on Bioinformatics, Computational Techniques and Engineering; Wiley VCH, 2008; pp 65−84.

(48) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. *FASEB J.* **2011**, *26*, 868−881.

(49) Guimerà, R.; Sales-Pardo, M.; Amaral, L. A. N. *Nat. Phys.* **2006**, *3*, 63−69.

(50) Lloyd, S. *IEEE Trans. Inf. Theory* **1982**, *28*, 129−137.

(51) Shi, J.; Malik, J. *IEEE Trans. Pattern Anal.* **2000**, *22*, 888−905.

(52) Yamada, T.; Hiraoka, Y.; Gupta, T.; Chakrabarty, A. *Cell Cycle* **2004**, *3*, 752−755.

(53) Gabellieri, E.; Bucciantini, M.; Stefani, M.; Cioni, P. *Biophys. Chem.* **2011**, *159*, 287−293.

(54) Mei, G.; Di Venere, A.; Malvezzi Campeggi, F.; Gilardi, G.; Rosato, N.; De Matteis, F.; Finazzi-Agrò, A. *Eur. J. Biochem.* **1999**, *265*, 619−626.

(55) Nar, H.; Messerschmidt, A.; Huber, R. *J. Mol. Biol.* **1991**, *221*, 765−772.

(56) Gilardi, G.; Mei, G.; Rosato, N.; Canters, G.; Finazzi-Agrò, A. *Biochemistry* **2004**, *33*, 1425−1432.

(57) Bahar, I.; Jernigan, R. *J. Mol. Biol.* **1997**, *266*, 195.

(58) Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. *J. Intell. Inf. Syst.* **2001**, *17*, 107−145.

(59) Rand, W. *J. Am. Stat. Soc.* **1971**, *66*, 846−850.

(60) Jain, A. *Pattern Recog. Lett.* **2009**, *31*, 651−666.

(61) Paci, P.; Di Paola, L.; Santoni, D.; De Ruvo, M.; Giuliani, A. *Curr. Proteomics* **2012**, *9*, 160−166.

(62) Humphries, M.; Gurney, K. *PLoS ONE* **2008**, *3*, e0002051.

(63) Watts, D. J.; Strogatz, S. H. *Nature* **1998**, *393*, 440−442.

(64) Csermely, P.; Korcsmáros, T.; Kiss, H.; London, G.; Nussinov, R. *Pharmacol. Therapeut.* **2013**, *138*, 333−408.

(65) Farver, O.; Skov, L.; van de Kamp, M.; Canters, G.; Pecht, I. *Eur. J. Biochem.* **1992**, *210*, 399−403.

(66) Bottini, S.; Bernini, A.; De Chiara, M.; Garlaschelli, D.; Spiga, O.; Dioguardi, M.; Vannuccini, E.; Tramontano, A.; Niccolai, N. *Comput. Biol. Chem.* **2013**, *43*, 29−34.

(67) Onuchic, J.; Beratan, D.; Winkler, J.; Gray, H. *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 349−377.

(68) Langen, R.; Chang, I.; Germanas, J.; Richards, J.; Winkler, J.; Gray, H. *Science* **1995**, *268*, 1733−1735.

(69) Regan, J.; Di Bilio, A.; Langen, R.; Skov, L.; Winkler, J.; Gray, H.; Onuchic, J. *Chem. Biol.* **1995**, *2*, 489−496.

(70) Mikkelsen, K.; Shoc, L.; Nar, H.; Farver, O. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5443−5445.

(71) Onuchic, J.; Beratan, D.; Winkler, J.; Gray, H. *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 349−377.

(72) Di Paola, L.; De Ruvo, M.; Paci, P.; Santoni, D.; Giuliani, A. *Chem. Rev.* **2013**, *113*, 1598−1613.