JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Proteins as Sponges: A Statistical Journey along Protein Structure Organization Principles

Luisa Di Paola,[†] Paola Paci,[‡] Daniele Santoni,[¶] Micol De Ruvo,[§] and Alessandro Giuliani*[,∥]

[†]Università Campus Biomedico, via Alvaro del Portillo 21, 00128 Rome, Italy

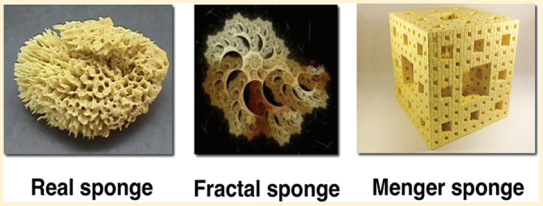
[‡]CNR-Institute of Systems Analysis and Computer Science (IASI), BioMathLab, viale Manzoni 30, 00185 Rome, Italy

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

[§]Università Campus Biomedico, via Alvaro del Portillo 21, 00128 Rome, Italy

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

ABSTRACT: The analysis of a large database of protein structures by means of topological and shape indexes inspired by complex network and fractal analysis shed light on some organizational principles of proteins. Proteins appear much more similar to "fractal" sponges than to closely packed spheres, casting doubts on the tenability of the hydrophobic core concept. Principal component analysis highlighted three main order parameters shaping the protein universe: (1) "size", with the consequent generation of progressively less dense and more empty structures at an increasing number of residues, (2) "microscopic structuring", linked to the existence of a spectrum going from the prevalence of heterologous (different hydrophobicity) to the prevalence of homologous (similar hydrophobicity) contacts, and (3) "fractal shape", an organizing protein data set along a continuum going from approximately linear to very intermingled structures. Perhaps the time has come for seriously taking into consideration the real relevance of time-honored principles like the hydrophobic core and hydrophobic effect.



INTRODUCTION

In their extremely innovative and provocative paper, Banerij and Ghosh[1] explicitly state "A student of protein structure is constantly reminded of several myths prevalent in this paradigm. He (she, at any rate) studies that the globular proteins are so compactly packed that their interior mimics that of solids, but finds it a bit irreconcilable with reports of inhomogeneus packing in protein interior and the presence of cavities therein". The authors continue, claiming that the elusive character of the presence of the time honored "hydrophobic core" is the main driver of folding and suggest a much more realistic fractal folding of proteins, giving rise to objects more similar to sponges than to densely packed spheres.

Building upon the above statements and our previous experience demonstrating an exponential scaling of density of contacts rapidly fading away with increasing size,[2] in this work we approach a large scale statistical study to substantiate the correlations between different "shape" viewpoints of actual protein structures.

The concept of shape is one of the most fundamental (and consequently most elusive) concepts in science. The intuitive (while geometrically rigorous) definition of shape deals with the fulfillment of certain constraints linking the different dimensions of a given entity. Thus, a circle shape is defined by the fulfillment, by a set of points, of the constraint of an invariant distance from a special point called the center, while a triangular shape corresponds to a 180° sum of the internal angles formed by a set of three incident segments. When we move away from the world of regular geometrical entities, defining shapes becomes much less simple, and a wide spectrum of possible quantitative descriptors of the shape of natural objects comes into life. The case of proteins, with their diverse beautiful and intermingled three-dimensional architectures is paradigmatic of this multiplicity. Several methods have been proposed to characterize proteins shape, almost all of these methods focused on the proteins' surface representations, this was motivated by the fact that surface geometry plays the most relevant biological role because it delineates the interface between the molecule and its environment, i.e., the region where physiologically meaningful interactions take place.[3] As a consequence, protein shape has been often defined with reference to a finite set of points, a space curve, or a surface.[4] Among the wide variety of approaches proposed to describe a molecular surface, van der Waals surface (VdW) refers to the union of the atoms (modeled as balls) according to their van der Waals radii. The solvent accessibile surface (SAS) is a measure for quantitatively determining the interaction tendencies of a protein, delineated by the center of a probe sphere (typically a water molecule) rolling on top of the VDW surface. By removing a layer of solvent radius depth from the SAS model, the molecular surface (MS) can be obtained (Figure 1).

In these models, the surface of the molecule is depicted as a polyhedron, triangular facets link a triplet of surface atoms

Received: October 26, 2011
Published: January 11, 2012

   dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

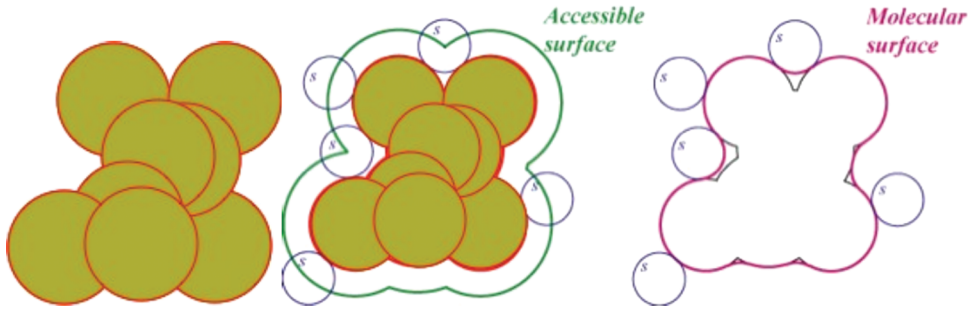


Figure 1. VDW, SAS, and MS surfaces.

blocking a probe sphere, whereas an edge links two atoms that allow the same probe sphere to roll from one blocking position to another. The vertices of the shape are exactly the atoms with a strictly positive accessible surface area.[4]

Anthony Hopfinger developed a "molecular shape analysis" on the basis of the comparison of electrostatic fields of small organic molecules[5] extended to proteins by Arteca.[6] The Hopfinger–Arteca paradigm gives a physical flavor (electrostatic forces) to the purely topological character of shape. In particular, shape descriptors have been exploited to individuate protrusions and cavities of a known input structure. The consideration of shape we adopted in this work, instead of being surface-based, concentrates on the idea of shape as a "three-dimensional distribution of mass", neglecting the explicit hidden/exposed separation of surface-based approaches. This separation, as we will explain, is implicitly taken into consideration in our work by two indexes (corrHBMJ and corrHBKD) corresponding to the quantification of internal/external asymmetry of hydrophobicity distribution in terms of the Pearson correlation coefficient between distance from the center of mass and relative hydrophobicity of residues. In her fundamental work on cell shapes,[7] Carole Heckman explicitly states "Earlier work suggested that certain restrictions apply to the geometric configuration that can be assumed by cells. Such restrictions were indicated by the finding that the values of certain shape descriptors were highly correlated with one another. This was surprising because these descriptors appeared to measure dissimilar geometrical properties of the cell. The present research confirms that the high levels of correlation are due to geometrical constraints on cell shape". This is a striking "back-to-the-root" simplification of the concept of shape: a given form is nothing more and nothing less than a set of constraints imposed on different shape descriptors (analogously to the circle and triangle examples quoted before), consequently the shape of complex objects like proteins (or cells in the case of Heckman) can be interpreted as a correlation structure imposed to a set of diverse shape descriptors. This is the classical way of spectroscopy. After all, a particular molecule (or mixture) is identified as a set of peaks, where the correlation is imposed to an initially independent set of descriptors (different wavelengths intensities) by the particular composition of the analyzed sample. Consistent to this "form-as-correlation" idea, the statistical paradigm we assumed was the principal component analysis in which the major fluxes of correlation of the studied descriptors correspond to the organization principles of protein structures.[8−10] This approach allows for a bottom-up, data-driven description of the main structural and topological determinants of the protein structures not imposing any particular theoretical frame on the data.[11−13] In this work, we demonstrate that the hydrophobic core is a pertinent concept only for very small proteins being linked to the microscopic (few residues) but not macroscopic protein structural organization, giving a proof-of-concept to both the Banerji–Ghosh hypothesis and to our previous contact density scaling results.[1,2] Local structure (single contacts) was demonstrated to be the place for the :hydrophobicity-coupling" of classical folding models.[14] This local organization was captured by a component of its own independent from global structure organization. Fractal dimension was independent from the above-sketched global and local organizations, while being connected with another component of shape description, namely, the relative fibrous/globular shape of proteins. This tripartite description of protein topology and structure into "size", "local contacts", and "shape complexity" was demonstrated to be strictly linked to the formation of cavities and promises to be of use for the elucidation of interesting aspects of protein physiology and dynamics.

RESULTS AND DISCUSSION

Descriptive Statistics. Before entering the "correlation business" to model the emerging relations linking the variables, it is worth considering the univariate distribution of each descriptor (Table 1) to get a dimensional idea of the variance/

Table 1. Simple Statistics

variable	*N*	mean	std dev	minimum	maximum
corrHBKD	911	−0.085	0.083	−0.548	0.196
corrHBMJ	911	−0.1408	0.1041	−0.526	0.089
N	911	613.698	656.130	50	6894
E	911	403023	1317913	1225	23760171
AS	911	0.306	0.163	0.003	0.938
R_G	911	12.775	5.138	4.76	48.197
R_{Gh}	911	12.233	5.287	3.973	47.878
R_{Gp}	911	13.062	5.076	4.915	48.343
MFD	911	2.356	0.403	0.930	4.148
H	911	0.982	0.044	0.805	1.209
D	911	1.361	0.134	1.008	2.306
HBAKD	911	0.005	0.038	−0.152	0.141
HBAMJ	911	0.045	0.035	−0.14	0.173
DBA	911	0.156	0.07	−0.118	0.373

invariance of each index in the considered data set. Obviously, we expect a relevant range of variation candidates for a given descriptor to become a major order parameter of the data set (i.e., a measure potentially effective to discriminate different protein architectures), while a relatively invariant descriptor indicates a measure that is almost identical across protein structures. Thus, while potentially important for its common

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

role in all proteins, it is not endowed with classification ability. Our PCA-based approach clearly favors the "wide range variation" descriptors. As for hydrophobic core, we can safely say, according to ref 1, that is a myth or in any case a very "low power" concept to look at protein structures at large. Both Kyte−Doolitle and Myiazawa−Jernigan[15] based the Pearson correlation between hydrophobicity and the center of the protein (coorHBKD and corrHBMJ) very low on average (−0.08 and −0.14, respectively), accounting for 0.64% and 0.20% of residue localization in protein structures, respectively. It is worth noting the big spread (especially for corrHBKD) of this index going from frankly negative, consistent with the existence of hydrophobic core values (minimum = −0.548), to paradoxically positive (maximum = 0.196) values.

Size variables were designed in order to get the largest possible variation range to allow a significant scaling of the descriptors with protein dimensions. Thus, the smallest protein of our collection is 50 residues long, while the larger one has N = 6894. This range of variation, for algebraic reasons, is still greater for E. The fact these two size descriptors vary over three (four in the case of E) orders of magnitude makes them less manageable as synthetic size measures with respect to radius of gyration variables (R_G, R_{Gh}, R_{Gp}). These variables are constrained into much more manageable variation ranges (approximately from 4 to 50). This mathematical well conditioning makes R_G variables the most correlated with the emerging "size component" (PC1) with which they will be shown to have almost unitary correlation coefficients (loadings, Table 2).

Table 2. Loading Matrix of Components[a]

variables	PC1	PC2	PC3	PC4
corrHBKD	**0.604**	−0.278	0.323	**-0.441**
corrHBMJ	**0.767**	−0.258	0.271	−0.171
N	**0.870**	0.172	−0.357	0.039
E	**0.702**	0.180	**-0.411**	0.101
AS	0.202	−0.203	**0.654**	**0.34**
R_G	**0.970**	0.060	−0.041	0.071
R_{Gh}	**0.974**	0.040	−0.037	0.068
R_{Gp}	**0.966**	0.070	−0.043	0.071
MFD	−0.296	0.258	**-0.729**	−0.152
H	−0.135	**-0.702**	−0.204	0.400
D	−0.224	**0.419**	0.337	−0.236
HBAKD	0.026	**0.483**	0.137	**0.649**
HBAMJ	0.030	**0.865**	0.314	−0.022
DBA	0.112	0.030	0.007	**-0.356**

[a]Bolded values represent the loadings of the variables most relevant for interpretation of each component.

As for the general shape variables, AS goes from perfect spherical symmetry (minimum = 0.003) to a straight linear structure (maximum = 0.94), spanning the entire theoretical asymmetry space with a small bias toward more globular shapes (mean = 0.30). This wide range of variation candidates AS as a possible order parameter of the data set. Fractal dimension (MFD) has an average value of 2.36 that is in strict concordance with previous results.[1] We inserted in our data set (despite the general invariance of this descriptor, getting a standard deviation of 0.40, that confirms the existence of a typical "fractal signature" of protein molecules) some extreme variants going from a totally linear object (minimum = 0.93) to an extremely intermingled structure (maximum = 4.15). It is worth noting that the lowest outliers are in some way artificial constructs (1mow, MFD = 0.93, artificial endonuclease; 1gk4, MFD = 1.10, Vimentin coil), while the paradoxical four dimensions object at the MFD maximum (2pn8, MFD = 4.15, thiredoxin peroxidase) gives rise to an extremely complex quaternary structure that can be hardly recognized as a single unitary object, being more similar to a "magic circle" of relatively loosely connected elements. In Figure 2, these extreme

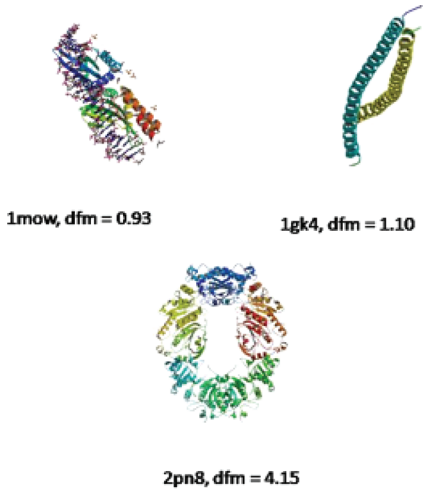


Figure 2. Example of extreme protein structures.

structures are reported. The "normal" range of variation (95% confidence limits over the data set) of MFD goes from 1.71 to 3.01, consistent with the general meaning of fractal dimension.

Topology descriptors were based on the proteins considered as contact networks, where α-carbons of the residues represent the network nodes and their thresholded (see Material and Methods) mutual distances the edges.[16] Besides DBA, that is based solely on graph invariants without any reference to the chemico−physical characterization of nodes (residues), these descriptors all redound to the influence of hydrophobicity on the probability of contacts. In the cases of HBAKD and HBAMJ, the departure from complete independence between hydrophobicity and contact probability is expressed by correlation coefficient metrics. Zero value correlation points to the complete independence of hydrophobicity and contact probabilities, while a negative value implies a tendency toward dissimilar hydrophobicity contacts, and a positive value implies a tendency toward similar hydrophobicity contacts. It is worth noting a substantial independence between hydrophobicity and contact probability with a very minor tendency of Myiazawa−Jernigan (HBAMJ) of preferring similar contacts over HBAKD, reminiscent of the character of statistical potential of the MJ index.[15] Notwithstanding the almost perfect "average" independence of hydrophobicity and contact probability on the entire data set, both HBAKD and HBAMJ display a nontrivial range of variation going approximately from −0.15 to 0.15 that will give rise to the second principal component (PC2), thus representing the second most relevant order parameter shaping protein structural organization. Similar considerations hold for H and D descriptors, in which the perfect independence corresponds to unitary value. In the case of DBA, we are dealing with a pure "architectural" principle related to the tendency of high degree (high number of contacts) nodes to be connected to each other (negative values of DBA,

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

again over a correlation coefficient metrics having −1 and +1 as theoretical extremes) or to be actively kept apart (positive values). Again, we have a substantial average symmetry of the two choices (mean DBA = 0.16) but with a multiplicity of architectural solutions going from DBA = −0.11 to DBA = 0.37. In the presence of a strong common compact core we expect a positive DBA, while a negative DBA is a marker of a very distributed architecture with almost independent "local cores" for different parts of the structure.

Correlation Structure and the Emerging of General Organizational Principles. Having described in the previous paragraph the univariate structure of the different descriptors of the data set focusing attention on both the location (mean values, "typical protein" view) and variability (standard deviation, range, "ordination of the protein set" view), we now shift to the most relevant part of the work: the emerging of consistent correlation fluxes (principal components) shaping the entire data set collecting the consistent variation across multiple descriptors. The fact that principal components are each other independent by construction[17] and globally provide the maximum parsimonious representation of the data set reassures us these components represent the basic independent factors describing protein molecules configurations. The interpretation of the loading matrix (being the loadings of the correlation coefficients of the original variables with the components) allows us to sketch an interpretation of the meaning of the extracted components. This interpretation will in turn be verified by its ability to predict "external variables" that did not enter into the component construction and that in this case were the two eps and eps_1 scores. The original 14 dimensions variable space, when analyzed by means of PCA, collapsed to a four component bona fide signal solution,[13,18] globally explaining the 71% of total variability, with a by far most relevant order parameter (PC1) responsible for the 37.4% of variance. The scaling of eigenvalues with component number is reported in Figure 3, where the threshold between

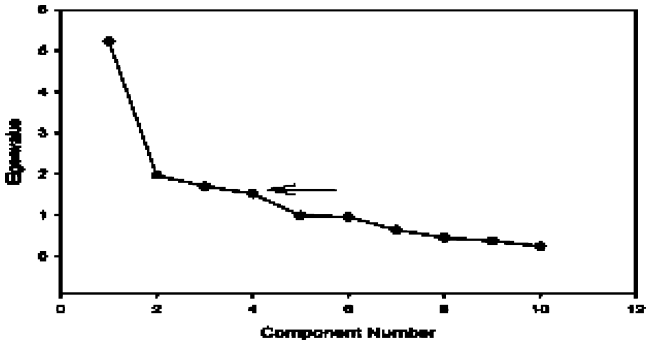


Figure 3. Saling of eigenvalues with component number. The arrow marks the threshold between "estimated signal" and "noise floor", substantiating our choice of four components to be analyzed.

"estimated signal" and "noise floor" is marked by an arrow. The loading matrix is reported in Table 2; the loadings of the variables most relevant for the interpretation of each component are bolded.

In Table 3, the percentage of variation explained by each component is reported. The near to maximal correlation of PC1 with size variables allows us to identify this component as "protein size". The fact that size is the main order parameter shaping the data set is a consequence of the presence of similar

Table 3. Percentage of Explained Variance by Each Component

PC1	PC2	PC3	PC4
37.4	14.1	12.2	8.2

construction principles common to all protein molecules. It is worth noting physical size (R'_Gs) is much more relevant (near unit loadings) for the component meaning than indirect indexes, as N or E (edges). This is an emerging property of the PCA solution reassuring us of the "well conditioning" of the component space that is not driven by the presence of outliers

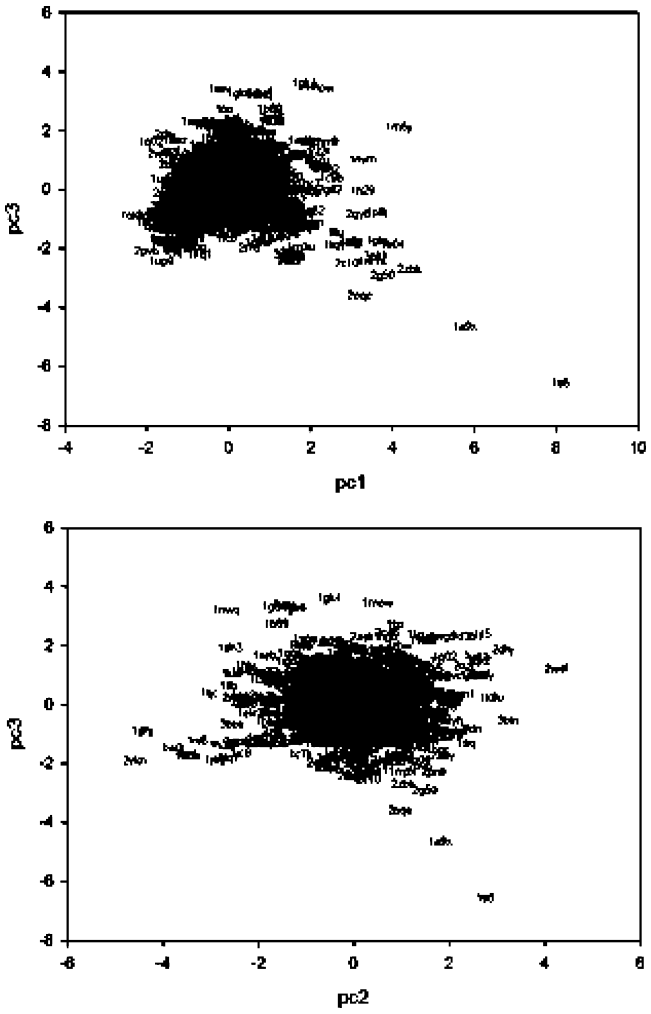


Figure 4. Two views of the component space are reported. PC1-PC3 in the upper panel and PC2-PC3 in the lower panel.

but by a consistent scattering of the entire data set in the component space. This is confirmed by a simple look at Figure 4, where two views (PC1-PC3 and PC2-PC3 planes) of the component space are reported. The (relatively few) outliers go together with a globally continuous distribution of the component scores of the entire protein data set. As for the size component (PC1), it is worth noting both corrHBKD and corrHBMJ have a relevant positive loading on the component. This comes from the fact the "hydrophobic core" hypothesis is only tenable for very small proteins, while it loses any relevance (with the consequent converging of corrHBKD and corrHBMJ to zero) at increasing size. This interpretation is

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

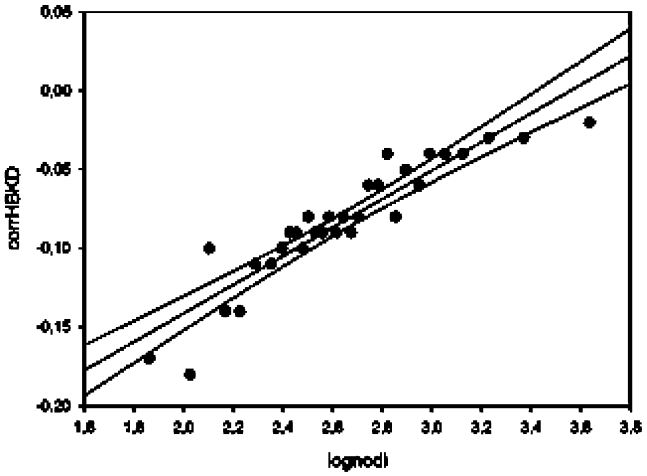


Figure 5. Average corrHBKD value is plotted vs average log(N). Here, corrHBKD = −0.32 + 0.091 × log(N), and the correlation between corrHBKD and the size is r = 0.92.

given a proof-of-concept in Figure 5, where the relation between size (expressed as logarithm of the number of residues) and corrHBKD is reported. In order to normalize the scattering of corrHBKD present in small molecules, we applied to the data a smoothing procedure. The protein data set was ordered along increasing N and partitioned into nonoverlapping sets of 30 molecules each. For each set, the average corrHBKD value is plotted versus the average log(N), obtaining a striking correlation between corrHBKD and size (r = 0.92) confirming our hypothesis. This result allows us to hypothesize that the origin of the "myth" of the "hydrophobic core" dates back to an age where the protein molecules whose structure was known was largely biased toward the "small proteins" end. The fading away of a closely packed hydrophobic core with size is consistent with the demonstrated progressive decrease in the density of contacts with the generation of less dense (and consequently with an increasing number of voids) architectures for larger proteins.[2]

The second component (PC2) explains the 14.1% of total variance and is clearly linked to topology hydrophobicity-based variables (H, D, HBAKD, HBAMJ). Basically, PC2 orders protein structures along an axis going from the prevalence of "opposite polarity" (low component scores) to a prevalence of "same polarity" (high component scores) contacts. As we stressed in the previous paragraph, the extremes of positive/negative assortativity go from −0.15 to 0.15 (with a theoretical range going from −1 to +1); nevertheless; this difference is sufficient to give rise to a relevant component of protein organization, pointing to a different relative balance of microscopic forces (hydrophobic effect, hydrogen bonds, electrostatic forces, etc.) intervening in different proteins folding mechanisms. This different relevance of physical forces, in our opinion, deserves a more detailed investigation given the implicit unitary character of force fields approximation used in simulation and theoretical studies. In any case, this local topology component is independent from both size and general shape (PC3) that in turn explains 12.2% of total variance. AS and MFD are the two variables leading PC3 with linear structures (high asymmetry), being less complex than globular structures (AS and MFD enter PC3 with opposite loadings). All in all, PC3 is a "general shape" or "relative complexity" index with low scores (high MFD) pointing to very complex structures and high scores to less complex architectures.

We were not able to attach a clear meaning to PC4; the loading profile points to very specific architectural constraints possibly linked to class-specific features. It is worth noting that DBA has a non-null loading only on this component.

Modeling with Components: The Prediction of Cavities and Empty Spaces. The protein structure picture emerging from the above results can be metaphorically imagined as a sponge with increasing space for cavities at increasing size and complexity of the molecule. This is actually what happens with real sponges that in turn can be considered as fractal, large-scale invariant, objects. Figure 6 reports a real

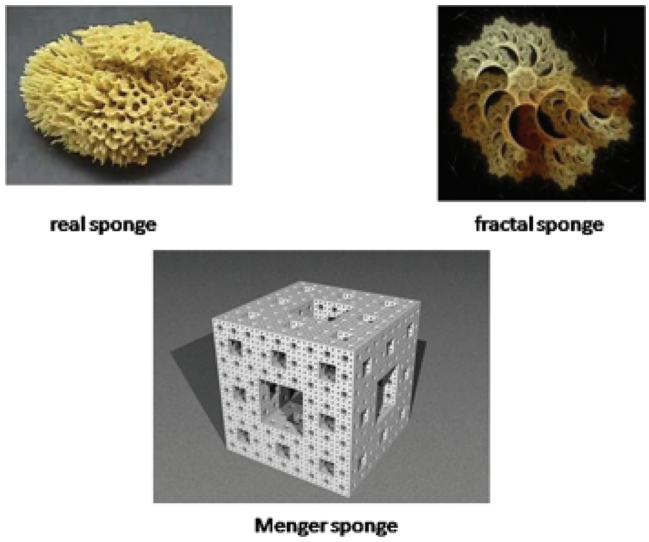


Figure 6. Natural and mathematical sponges.

sponge from the Mediterranean Sea, a so-called "fractal sponge" computationally generated, and the so-called "Menger sponge",[19] a mathematical 3-D object, which is a 3-D generalization of the unidimensional Cantor set, obtained by application of the same iterative void/dense rule at different scales. According to the "proteins as sponges" model, we expect PC1 and PC3 (i.e., size and fractal dimension) must be able to efficiently model the eps (or eps_1) variables quantifying the relative amount of voids (cavities) in the molecular volume. This comes from the consideration that the number of voids increases with both size (PC1) and fractal complexity (PC3). On the same heading, the microscopic topology−hydrophobic organization principles collected in PC2 must not have any role in modeling voids. This was actually the case; eps and eps_1 were maximally correlated between them (Pearson r = 0.96), giving a proof-of-concept of the substantial robustness of void estimation with different measurement paradigms. PC1 and PC3 were both significantly correlated to eps_1 (Pearson r = 0.69 and 0.41 for PC1-eps_1 and PC3-eps_1, respectively), while PC2 scored a very low −0.14 Pearson r value with eps_1. When we modeled eps_1 by both PC1 and PC3 (whose effect in the estimation of eps_1 can be supposed as purely additive given the components are each other linearly independent by construction) using a linear multiple regression paradigm, we obtain a striking Pearson correlation r = 0.80 between

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

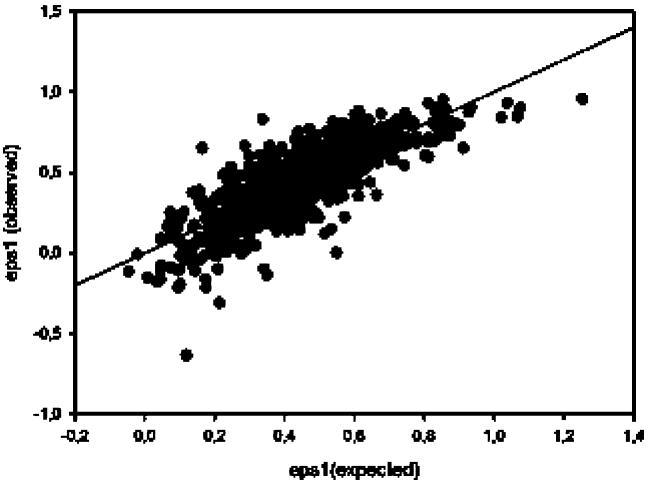


Figure 7. Each point represents a protein with the expected eps_1 on the horizontal axis (given by $eps_1 = 0.44 + 0.15 \times PC1 + 0.09 \times PC3$) and the observed eps_1 on the vertical axis. The correlation between observed eps_1 and expected eps_1 is $r = 0.804$ ($p < 0.0001$).

estimated and observed eps_1. Figure 7 reports model fitting and the correspondent linear equation linking voids to PC1 and PC3 scores. PC1 has a larger role than PC3 in void estimation, as expected by the prevailing role of size over other order parameters in shaping our data set. On the other hand, we need both the PC1 and PC3 "organizing principles" to efficiently model the portion of empty volume of the different proteins.

Table 4. Variables Describing Each Protein Structure

size	hydrophobic core	general shape	topology
N	corrHBKD	AS	*H*
E	corrHBMJ	MFD	*D*
R_G			HBKD
R_{Gh}			HBMJ

CONCLUSION

The effective prediction of empty space volume of 911 proteins by means of the extracted components gives a proof-of-concept of the relevance of the organizing principles we derived from PCA. Our results allow for a refinement of the general concepts used for describing proteins and cast doubts on firmly accepted concepts as the existence of hydrophobic core and the presence of common energetic principles for protein folding. A puzzling result of our analysis is the emerging of a component (PC2) ordering proteins from a preference for heterophilyc (low values of PC2) to homophilic (high values of PC2) contacts between adjacent residues. If only a single alternative was the preferred one, shared by all protein systems, we should not observe any component related to this mechanism (principal components follow the directions of maximal variance of the data set) on the contrary, the second most relevant organizing principle after size is the nature of local contacts. Using an architectural metaphor we can equate PC1 (size) and PC3 (shape) to global features of a building (its dimensions and form respectively) while PC2 has to do with the nature of the forces keeping it together at the microscopic/mesoscopic level (bricks disposition, junctions, arches...). These components, exactly like in architecture, are each other independent and while we have a clear appreciation of what size and general shape are, we are still in trouble to have a decent rationalization of the emerging fact the relative importance of physical forces driving folding (at the basis of the local contacts) can vary among different proteins. On the other hand the general statements set forth by Banerij and Ghosh[1] are largely confirmed by our analysis that in turn was able to confirm and give an explicit quantification to the fractal scaling hypothesis of the authors.[1] On another heading, the fractal sponge-like structure of proteins with water filled cavities asks for different paradigm of protein dynamics and physiology.

MATERIALS AND METHODS

We started from a repository of 1000 sequences, provided by the most widely used protein structure classification system

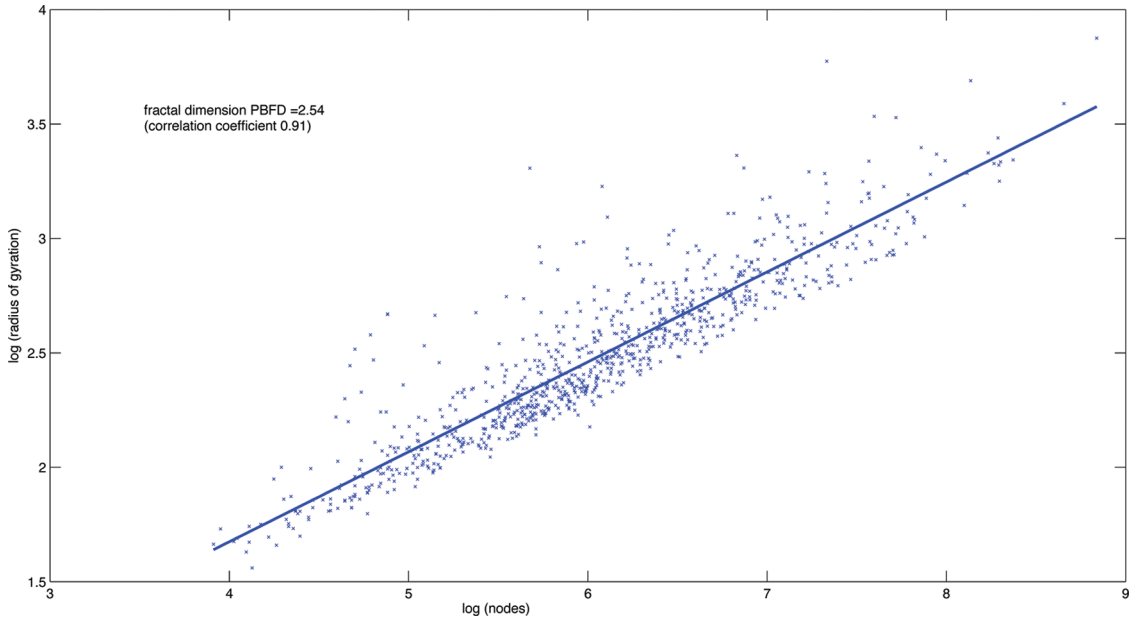


Figure 8. Scaling of protein radius of gyration R_g with the number of residue *N* (protein backbone fractal dimension).

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

CATH (SHREC'10, http://www.loria.fr/mavridis/SHREC_10); the superfamilies are randomly selected from CATH v3.3.

For each sequence, we extracted the corresponding PDB code, on which the analysis has been performed. The presence of a wide range of CATH classes guaranteed the consistency of analysis. The overall data set comprises 911 protein structures because several sequences belong to the same quaternary protein structure.

Each protein structure was described by means of 14 variables that were in turn submitted to a principal component analysis (PCA) and by two external variables consisting in two indexes (eps, eps_1) proportional to the "amount of cavitation" of the structures that were modeled by the extracted components to give a "proof-of-concept" of the components interpretation. The variables can be roughly classified into a-priori "groups" as shown in Table 4.

Topological parameters refer to the protein contact graphs that have been attained starting from the structural information embedded into PDB files, i.e., the spatial positions for all atoms in protein molecules.[16] We extracted the position of α-carbons representing the whole residue location; the distance matrix $\mathbf{d} = \{d_{ij}\}$ has been computed, the generic element d_{ij} being the Euclidean distance in the 3-D space between the i-th and j-th residues (holding the sequence order). Once the spatial distances are known, the corresponding contact graph has been obtained. Residues are the graph nodes, and links represent between residue contacts. A contact is established when a between residue distance is in the range 4−8 Å. Thus, the corresponding adjacency matrix $\mathbf{A} = \{a_{ij}\}$ is defined as

$$a_{ij} = \begin{cases} 1 & \text{if } d_{ij} \in [4-8] \\ 0 & \text{if } d_{ij} \notin [4-8] \end{cases}$$

The consideration of the protein structure as such a simplified graph was demonstrated to be very effective allowing for reconstructing the basic features of protein configuration in space as secondary structure,[20] proteins subunits and structural domains,[16,21,22] active sites,[23] residues relevant for protein folding kinetics and general stability of proteins,[24,25] allosteric signal pathways,[26] and many others.

As follows, a detailed description for each variable is provided.

- N: number of residues
- E: number of possible edges in the case of fully connected network, $E = N(N-1)/2$
- R_G: radius of gyration of the whole protein structure, defined as

$$R_G = \sqrt{\frac{1}{2}\frac{\sum_{i=1}^{N} m_i r_i^2}{\sum_{i=1}^{N} m_i}} \quad (1)$$

where m_i is the molecular mass of the i-th residue, and r_i is the corresponding distance (referred to the α-carbon position) from the center of mass, whose coordinates are the mass-weighted average of the α-carbons values.

- R_{Gh} and R_{Gp}: radius of gyration of hydrophobic and polar residues, respectively, defined similarly to eq 1, but it includes only the hydrophobic and polar residues contribution[27]
- corrHBKD and corrHBMJ: the hydrophobic core notion relies on a nonuniform distribution of hydrophobic residues that should stand closer to the center of mass of the whole protein structure. To test this hypothesis, we computed the Pearson correlation coefficients between two vectors, one reporting the distance from the center of mass for each residue and the other the corresponding hydrophobicity value (we applied this method to the Kyte−Doolittle hydrophobicity, corrHBKD, and Miyazawa−Jernigan scores, corrHBMJ); a strong negative correlation between hydrophobicity and distance from the center of mass would be proof of the core presence.
- AS: asymmetry. This is the departure from equality of the major displacements on the X,Y,Z of protein structures; high values of AS point to the existence of a major "elongation" axis and thus to linear (fibrous) protein; low values point to symmetric and thus very globular structures. AS is defined such that it is comprised between 0 and 1, with the values close to 0.5 corresponding to discoidal molecules.
- MFD: protein fractal nature has been recognized as a useful tool to interpret structural data and molecule function. On the other hand, many scaling laws apply to different protein properties (for a detailed review of protein structure fractal nature see ref 28). Mass fractal dimension (MFD), also known as Haussdorff dimension, refers to the inner structure of proteins. It has been computed by evaluating the mass comprised within spheres of varying radii into the protein 3-D structure; if the structure is not a compact, yet a fractal object, the following dependence shows up

$$M \sim R^{\mathrm{MFD}} \quad (2)$$

The value of MFD = 3 corresponds to compact objects. MFD has been computed measuring the mass of residues contained within spheres of different radii, R, centered on the protein center-of-mass. Specifically, we chose 20 different radiuses equally spaced in the range of $[\overline{d}_{CM} - 10 \times \delta, \overline{d}_{CM} + 10 \times \delta]$, $\overline{d}_{CM}$ being the mean value of the distances of residues from the center-of-mass of the whole protein structure, with $\delta = \overline{d}_{CM}/400$ corresponding to the steps of different values of R .

In our work, we state proteins are more similar to sponges than to compact spheres; this is something more than a metaphor. A well-known mathematical model for fractal sponges is the Menger or the Menger−Sierpinsky sponge.[29] It is built up by applying an infinite number of operations of volume subtraction. Starting from a cube whose faces are divided into nine squares, the primitive cube is parted into 27 smaller cubes. Those placed at the center of every face (6) plus that located at the very center of the cube are eliminated. Then, the initial volume has been reducted of a fraction 20/27. By repeating this operation for each resulting cube, a sponge-like object is obtained. The mass fractal dimension and the void fraction, or porosity, corresponding to the n-th stage are therefore

$$d_f^{(n)} = \frac{\log 20^n}{\log 3^n} = \log_3 20 \simeq 2.73 \quad (3)$$

$$\mathrm{eps}^{(n)} = 1 - \left(\frac{20}{27}\right)^n \quad (4)$$

It is worth noting that sponge fractal dimension is in the same order of magnitude as proteins. Moreover, this mathematical object allows for a rigorous definition of porosity that

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

is exactly what we did in our study. We computed the MFD for each single protein, but a very similar scaling behavior could emerge also analyzing the whole protein population. In this case, the radius of gyration R_G of the protein scales with the number of residues N as[30]

$$R_G \sim N^{\nu} \tag{5}$$

ν is a fractional scaling exponent that depends on the residue−solvent interactions. In a "good solvent", protein residues interact preferentially with the solvent molecules rather than with each other. The protein structures are then stretched into the solvent environment, and the corresponding value for ν is 3/5.

The fractional exponent ν is found to be the inverse of the protein backbone fractal dimension (PBFD) describing the scaling relationship between polymer length and number of residues, interpreted as rulers of fixed unitary length.

Analyzing protein backbone scaling (eq 5) on the whole set of 911 proteins, in order to determine the protein backbone fractal dimension, we obtained PBFD = 2.54 (Figure 8). The corresponding protein fractal dimension PBFD = 2.54 is in good agreement with both literature data[30] and our MFD computations. This gives further strength to our sponge-like model demonstrating that a large part of the residue is in contact with solvent.

- H and D: protein graphs describe the connectivity of residues (nodes) notwithstanding of the specific chemico−physical nature of each residue. In order to correlate topological descriptor to chemico−physical properties, combined descriptors are required, accounting for both classes of properties (topological and chemico−physical ones). Given a key physical property (for instance, hydrophobicity), if nodes show an attitude to preferentially established links with other similar nodes, the network is named dyadic; otherwise, if nodes preferentially link to dissimilar ones, the network is said antidyadic.[31]

  Let n_1 and n_0 denote, respectively, the number of node possessing or not a specific property (discretized hydrophobicity, in the case of point); e_{10} and e_{11} represent the number of edges connecting homologous or heterelogous nodes, respectively. The heterophilicity score H is then defined as

$$H = \frac{e_{10}}{e_{10,r}} \tag{6}$$

where $e_{10,r}$ is the expected value in the case of uniform distribution of the property among nodes, that depends on E. It is finally found

$$e_{10,r} = E \times n_1 \times (N - n_1) \tag{7}$$

Analogously, as for the homologous contacts, it is defined the dyadicity D as

$$D = \frac{e_{11}}{e_{11,r}} \tag{8}$$

and the corresponding value for uniform distribution is

$$e_{11,r} = E \times \frac{n_1 \times (n_1 - 1)}{2} \tag{9}$$

- Assortativity: this is another index for the proneness of nodes to connect to other nodes possessing similar features.[32] It is computed as the Pearson correlation coefficient between the two vectors containing a selected property for pairs of incident nodes in a network. For instance, in some networks, high-degree nodes preferentially connect to other high-degree nodes (assortative networks), whereas in other types of networks high-degree nodes preferentially connect to low-degree ones (disassortative networks). For the first class of networks, the correspondent assortativity index is close to 1, while it is close to −1 for the second class of networks.

We extended the analysis also to hydrophobicity related assortativity. Specifically, we adopted two kinds of scores defined as follows

- HBAKD, HBAMJ: hydrophobic-based assortativity correspondent to the Pearson correlation coefficient between incident residue hydrophobicity lists; hydrophobicity scores are based on two scales: Kyte−Doolittle and Miyazawa−Jernigan.
- DBA: degree-based assortativity that is Pearson correlation coefficient between incident residues computed over their respective degree (namely, the number of contacts each residue engages in the 3-D structure)
- eps and eps_1: porosity or void fraction is a parameter featuring the structure of porous, fractal media.[33] On the molecular scale, this definition applies to porous biomacromolecular structures;[28] hence, we defined two slightly different void fractions for the protein volume

$$\begin{aligned} \text{eps} &= \frac{V_f}{V} \\ &= 1 - \frac{V_{\text{residue}}}{V};\ \text{eps}_1 \\ &= 1 - \frac{V_{\text{residue}}}{V_1} \end{aligned} \tag{10}$$

where V_f is the free volume within the protein structure, being the complementary part of V_{residue} (volume occupied by residue atoms with respect to the overall protein volume). The definition of the two parameters slightly differs only for the overall protein volume definition: V represents the average volume between the three spherical volumes $[V_X, V_Y, V_z]$, evaluated at three diameters corresponding to the maximum distance between residues in the three coordinates; V_1 corresponds to spherical volumes $[V_X', V_Y', V_z']$ related to the maximum distance along the three spatial directions of residues from the center of mass.

Principal component analysis: PCA was computed by SAS software in terms of the eigenvectors of the pairwise correlation matrix of the descriptors. The use of a correlation matrix instead of the covariance matrix implies a normalization of the data that in this case is mandatory given the variables have heterogeneous dimensions and range of variation. Noise floor and the consequent selection of meaningful components comes from a visual scree test. As we aptly demonstrated in ref 34, even a very minor component in terms of explained variance can have a signal-like character. The only way to decide about the signal character of a component can be only "semantic" and not "syntactic". Shortly, if a very small component is found to correlate with an external variable not explicitly put into analysis, this implies it carries relevant information and cannot be considered as pure noise. This feature is of crucial

dx.doi.org/10.1021/ci2005127 | J. Chem. Inf. Model. 2012, 52, 474−482

importance (and routinely used) for the analysis of large scale high-throughput microarray data, where the presence of an overwhelming first principal "size" component corresponds to the characteristic transcriptome of the tissue of origin and relegates the usable part of information for sample discrimination to very minor components explaining a few percent of the total variability (for a very accurate and thorough explanation of this approach see ref 35). The demonstrated correlation between extracted components and external variables (eps and eps_1) reassure us about the signal-like character of extracted components, while the substantial identity (correlation coefficients around 0.9) between the components of the general data set and the components relative to reduced sets (data not shown) is proof of the robustness of the obtained solution.

AUTHOR INFORMATION

Corresponding Author

*E-mail: alessandro.giuliani@iss.it.

ACKNOWLEDGMENTS

We thank the "Consorzio interuniversitario per le Applicazioni di Supercalcolo Per Università e Ricerca" (CASPUR) for computing resources and support. Work partially granted by Italy–U.S.A. Cooperation Project "Systems Effects of Chemicals: Evaluation of High-Throughput Profiling".

REFERENCES

(1) Banerji, A.; Ghosh, I. *PloS One* **2009**, *4*, e7361.
(2) Zbilut, J.; Chua, G.; Krishnan, A.; Bossa, C.; Rother, K.; Webber, C.; Giuliani, A. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 621–629.
(3) Natarajan, V.; Koehl, P.; Wang, Y.; Hamann, B. Visual Analysis of Biomolecular Surfaces. In *Mathematical Methods for Visualization in Medicine and Life Sciences*; Linsen, L., Hagen, H., Hamann, B., Eds.; Mathematics and Visualization Series; Springer Verlag: Berlin, 2007.
(4) Wang, Y. Ph.D. Thesis, Department of Computer Science, Duke University, 2004.
(5) Hopfinger, A. *J. Med. Chem.* **1983**, *26*, 990–996.
(6) Arteca, G. *Biopolymers* **1993**, *33*, 1829–1841.
(7) Heckman, C. *Cytometry* **1990**, *11*, 771–783.
(8) Colafranceschi, M.; Colosimo, A.; Zbilut, J.; Uversky, V.; Giuliani, A. *J. Chem. Inf. Model.* **2005**, *45*, 183–189.
(9) Emberly, E.; Mukhopadhyay, R.; Tang, C.; Wingreen, N. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 91–98.
(10) Bloom, J.; Drummond, D.; Arnold, F.; Wilke, C. *Mol. Biol. Evol.* **2006**, *23*, 1751–1761.
(11) Benigni, R.; Giuliani, A. *Am. J. Physiol.* **1994**, *266*, R1697–R1704.
(12) Christie, O. *Chemom. Intell. Lab. Syst.* **1995**, *29*, 177–188.
(13) Preisendorfer, R. *Principal Component Analysis in Meteorology and Oceanography*, Elsevier; Amsterdam, 1988.
(14) Dill, K. *Biochemistry* **1990**, *29*, 7133–7155.
(15) Vajda, S.; Sippl, M.; Novotny, J. *Curr. Opin. Struct. Biol.* **1997**, *7*, 222–228.
(16) Giuliani, A.; Di Paola, L.; Setola, R. *Curr. Proteomics* **2009**, *6*, 235–245.
(17) Cotta-Ramusino, M.; Benigni, R.; Passerini, L.; Giuliani, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 248–254.
(18) Jackson, D. *Ecology* **1993**, *74*, 2204–2214.
(19) Takeda, M.; Kirihara, S.; Miyamoto, Y.; Kazuaki, S.; Honda, K. *Phys. Rev. Lett.* **2004**, *92*, 093902.
(20) Webber, C.; Giuliani, A.; Zbilut, J.; Colosimo, A. *Protein: Struct., Funct., Bioinf.* **2001**, *44*, 292–303.
(21) Giuliani, A; Zbilut, J.; Tomita, M. *J. Proteome Res.* **2007**, *6*, 3924–3934.
(22) Kannan, N.; Vishveshwara, S. *J. Mol. Biol.* **1999**, *292*, 441–464.
(23) Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. *J. Mol. Biol.* **2004**, *344*, 1135–1146.
(24) Gromiha, M. M.; Selvaraj, S. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235–277.
(25) Krishnan, A.; A; Giuliani, A. J. Z.; Tomita, M. *PLoS ONE* **2008**, *3*, e2149.
(26) Sol, A. D.; Fujihashi, H.; Amoros, D.; Nussinov, R. *Mol. Syst. Biol.* **2006**, *2*, 0019.
(27) Alves, N.; Aleksenko, V.; Hansmann, U. *J. Phys: Condens. Matter* **2005**, *17*, S1595.
(28) Enright, M. B.; Leitner, D. M. *Phys. Rev. E* **2005**, *71*, 011912.
(29) Russ, J. *Fractal Surfaces*; Plenum Press: New York, 1994.
(30) Dewey, G. *Fractals in Molecular Biophysics*; Oxford University Press, New York, 1998.
(31) Park, J.; Barabasi, A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17916–17920.
(32) Newman, M. *Phys. Rev. E* **2006**, *74*, 056108.
(33) Elias-Kohav, T.; Moshe, S.; Avnir, D. *Chem. Eng. Sci.* **1991**, *46*, 2787–2798.
(34) Giuliani, A.; Colosimo, A.; Benigni, R.; Zbilut, J. *Phys. Lett. A* **1998**, *247*, 47–52.
(35) Roden, J.; King, B.; Trout, D.; Mortazavi, A.; Wold, B.; Hart, C. *BMC Bioinf.* **2006**, *7*, 194.