Open Access

Characterizing Protein Shape by a Volume Distribution Asymmetry Index

Nicola Arrigo[1], Paola Paci[2], Luisa Di Paola[1*], Daniele Santoni[3], Micol De Ruvo[1], Alessandro Giuliani[5] and Filippo Castiglione[4]

[1]*Università Campus Biomedico, 00128 Rome, Italy*

[2]*CNR-Institute of Systems Analysis and Computer Science "Antonio Ruberti", Bio Math Lab, 00185 Rome, Italy*

[3]*CNR-Institute of Systems Analysis and Computer Science "Antonio Ruberti", 00185 Rome, Italy*

[4]*CNR-Institute for Computing Applications "Mauro Picone", National Research Council, 00185 Rome, Italy*

[5]*Department of Environment and Health, Istituto Superiore di Sanità, 00161 - Rome, Italy*

Abstract: A fully quantitative shape index relying upon the asymmetry of mass distribution of protein molecules along the three space dimensions is proposed. Multidimensional statistical analysis, based on principal component extraction and subsequent linear discriminant analysis, showed the presence of three major 'attractor forms' roughly correspondent to rod-like, discoidal and spherical shapes. This classification of protein shapes was in turn demonstrated to be strictly connected with topological features of proteins, as emerging from complex network invariants of their contact maps.

Keywords: Protein shape, protein contact network, topological indices, principal component analysis.

1. INTRODUCTION

It is commonly stated that the activity of a protein is somewhat encoded into its shape [1]. A rough classification of proteins on the basis of their shape, identifies two distinct classes: globins (near spherical molecules) and scleroproteins (rod-like or fibrous). Fibrous proteins are for the major part mainly structural elements (for instance, collagen in the connective tissue); on the other hand, globins are apt to many different tasks, often subdued to the presence of specific interaction sites located on the protein surface [2].

The concept of molecular shape is somewhat elusive: the identification of quantitative descriptors for the molecular structure is, thus, a potentially very interesting avenue of research [3].

Several methods have been proposed to characterize proteins shape [4, 5]: so far, shape analysis has been limited to protein surface representation, assuming believing that surface as the privileged view given is a key factor, because is the region where it is where biologically meaningful interactions take place [6]. Actually, geomitetric shape has been often defined with reference to a finite set of points, a space curve, or a surface [7], instead of considering the overall volume of a molecule, that is specifically the purpose of this work, building upon previous results in which we demonstrated both the lack of any marked separation between protein internal and external milieu and the basic fractal structure of protein fold (Di Paola *et al*. JCIM).

Literature provides several different approaches to describe the molecular surfaces: among these, Van der Waals surface(VdW) refers to the union of atoms (modeled as balls) according to their van der Waals radiuses [8]; the Solvent AccessibleSurface (SAS), originally proposed by Lee and Richard [9], is the surface traced out by the center of a probe sphere (typically a water molecule) rolling on top of the VdW surface: in this way, the overall protein molecule hindrance comprises also the hydration shell. The Solvent Excluded Surface (SES) is the result of the SAS erosion by the same probe [10, 11]. For a graphic representation, see Fig. (**1**).

Hopfinger developed a useful method for small molecules named 'molecular shape analysis' [12], based on the comparison of electrostatic fields, later adapted by Arteca and Mezey to define shape descriptors of macromolecules [13].

Some authors have focused on the detection of protrusions and cavities of known input structures, provided by shape descriptors.

The Connelly function [10], the most used and known, is derived as follows: in any point on the surface, a sphere is centered, having a diameter as large as a water molecule. If the fraction of the sphere volume within the SES volume (see dashed sphere in Fig. **1**) is smaller than 0.5, the surface is considered as locally convex, otherwise concave.

Formally, for any point x $\in$ M , let us consider the ball B (r, x)centered at x with radius r : if S (r, x) = δB (r, x) is the boundary of B (r, x) and SI the portion of S (r, x) contained within the surface, the Connolly function fr: M → R is defined as:

*Address correspondence to this author at the Università Campus Biomedico, via Alvaro del Portillo 21, 00128 Rome, Italy;
Tel: ++39 06 225419634; Fax: ++39 0622541456;
E-mail: l.dipaola@unicampus.it#

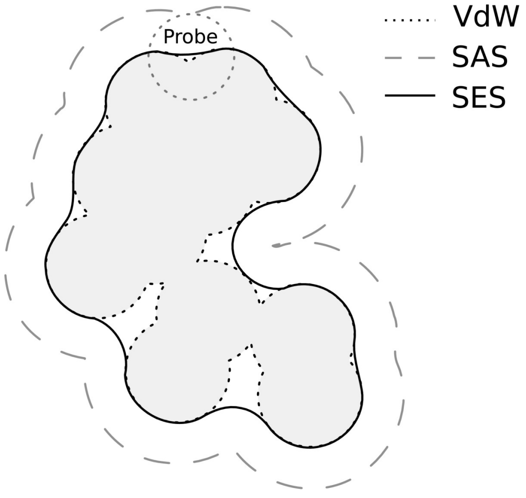

Fig. (1). VdW, SAS and SES molecular surfaces [9].

$$f_r(x) = \frac{Area\,(S_I)}{r^2} \qquad \#$$

High values of f_r (x) indicate that the surface around x is largely concave, while low values point to a prevalent convexity around x. Røgen and Sinclair introduced protein shape descriptors based on backbone [14].

Although curvature-based methods (as Connelly's) well identify points located at local protrusions and cavities, they all depend on a pre-fixed value r (the neighborhood size); in many cases, it is desirable that the function value can also give some clues about the length scale of the conformational feature the function refers to.

All these models have a strong 'theoretical flavor' and are concentrated on the molecule surface shape. On the contrary, we adopted a mainly statistical bottom-up approach in order to derive a coarse-grain, but easily computable and free from *a priori* constraints shape index. At odds with surface-based approach, the proposed index is based on the volume distribution of the atoms along the three axes of the space.

The starting point of this work is that the most interesting geometrical templates in structural biochemistry are the sphere, the disc and the cylinder; thus, we decided to rely upon the amount of symmetry of the volume distribution on the three dimensional space so to develop a global index catching the relative 'spherical', 'discoidal' or 'cylinder' character of the studied structure.

A data set spanning the entire range of variation of protein shapes, from perfect sphere to almost perfect cylinders, was developed in order to check by means of a correlative approach based on Principal Component Analysis (PCA) [15], the consistency of the proposed index with relevant size, geometry and topology related properties of protein structures.

The demonstrated ability of the proposed method not only to discriminate different shapes but to discover the shape variability typical of a functional protein class (membrane proteins) confirms the relevance of volume based shape representation.

2. METHODS

In this work, we perform an analysis of the three-dimensional protein structures along the canonical axes, as reported in PDB files, containing the relevant information about biomolecular structures.

As a first step, we identify the Center of Mass (CM) of the molecule, reducing each amino acid residue to the correspondent α -carbon. In the case of the sphere, the center of mass coincides with its geometrical center and the distance from the CM to the surface of the molecule is identical along each of the three dimensions. In the case of the disc, the CM represents its center, two dimensions have an almost identical elongation and the third is not relevant; in the case of the cylinder, there is just one relevant dimension and the CM is located approximately at the middle point of the cylinder main axis.

Once identified the CM, the maximal distance of α–carbons from the CM is computed along the three axes; the three values R_x ,R_y , R_z represent the radius of hypothetical spheres (Fig. **2**), whose volume provides indication concerning the length of the object along a specific direction:

$$R_x = max(x_{CDM} - x_i)$$

$$R_y = max(y_{CDM} - y_i)$$

$$R_z = max(z_{CDM} - z_i)$$

The corresponding equivalent spherical volumes are then computed:

$$V_x = \frac{4}{3}\pi R_x^3$$

$$V_y = \frac{4}{3}\pi R_y^3$$

$$V_z = \frac{4}{3}\pi R_z^3$$

In the case of the sphere, V_x= V_y= V_z , whereas for a generic non-spherical molecule, these three volumes differ from each other.

Let's now introduce a shape space, in which each protein is identified by a vector ρ defined as follows:

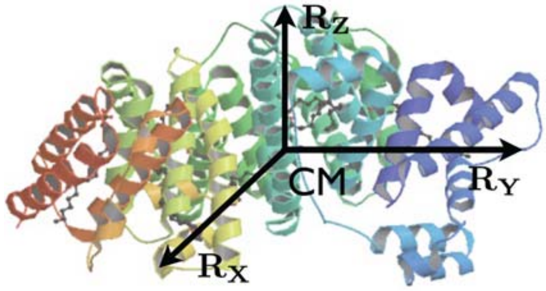

Fig. (2). Radiuses R_x, $R_y$and $R_z$in the case of HSA (PDB code 1E7I).

$$\rho = \left[\frac{V_x}{V_x + V_y + V_z}, \frac{V_y}{V_x + V_y + V_z}, \frac{V_z}{V_x + V_y + V_z}\right]$$

The reference shapes correspond to the following points in the shape space (Fig. **3**):

- Sphere: $S = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$;
- Disc: $D_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$; $D_2 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$; $D_3 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$;
- Rod-like: $R_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$; $R_2 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$; $R_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$.

The tetrahedron represented in Fig. (**3**) is the space of possible protein molecular shapes. Clearly, a nearly spherical molecule is represented by a point closely located to S; on the contrary, the largest distance from S accounts for rod-like proteins. Thus we use the ratio between the actual distance of the protein from S with the maximum distance correspondent to perfect rod-like shape; this maximum distance is:

$$d_1 \equiv \overline{SR_1} = \sqrt{\left(\frac{1}{3} - 1\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2} = \frac{\sqrt{6}}{3}$$

#

On the other hand, the distance related to disc template is:

$$d_2 \equiv \overline{SD_1} = \sqrt{\left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{1}{3}\right)^2} = \frac{\sqrt{6}}{6} = \frac{d_1}{2}$$

#

Thus, let us define a normalized distance $\xi = \frac{d}{d_1}$, being the distance of a generic point in the shape space from S; according to ξ values, molecules can be classified as follows:

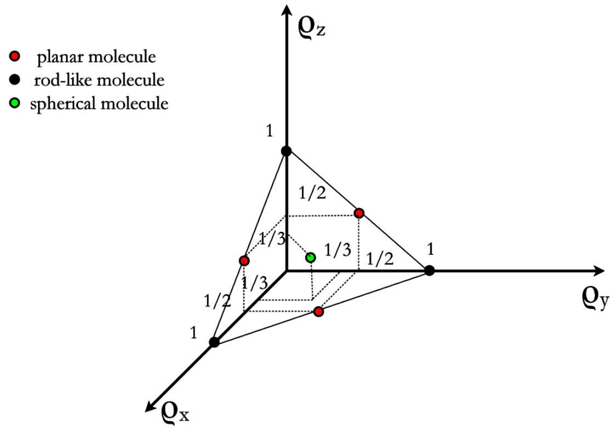

Fig. (3). Shape space: each point of the space represents a molecule in terms of its own shape; green, red and black dots refer to spheres, discs and rods, respectively.

$$\begin{cases} 0 \le \xi < 0.4 & \textit{spherical} \\ 0.4 \le \xi < 0.6 & \textit{discoidal} \\ 0.6 \le \xi \le 1 & \textit{rod-like} \end{cases}$$

#

where ξ is an asymmetry index, given its character of departurefrom perfect symmetry in space.

To prove the effectiveness of our index, we tested it on a 40 proteins data set, half of which chosen amongglobular shapes and half among fibrous.

In Table **1** the values of ξ are reported along with the classification consistent with our proposal and the number of residues(nodes). 3D structures of three proteins belonging to spherical, discoidal and rod-like groups are shown in Fig. **4 a**), **b**), **c**) respectively.

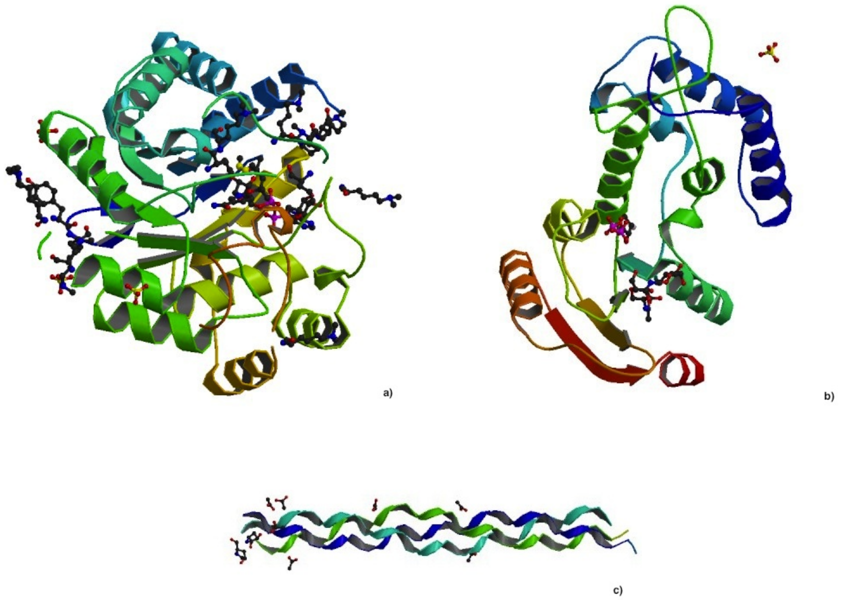

Fig. (4). Protein reference sample structures : **a**) 3ln3 (globular): putative reductase; **b**) 3kou (planar): cyclic ADP ribose hydrolase; **c**) 1cgd (rod-like): collagen peptide.

Table 1. Protein Data Set: PDB Codes are Reported in the First Column; Nodes = Number of Residues; ξ is the Asymmetry Index; Shape = Class

PDB ID	Nodes	ξ	Shape
1cgd#	60 #	1#	Rodlike#
1cag#	58#	1#	Rodlike#
3qc7#	172#	0,99#	Rodlike#
1h6w#	161#	0,94#	Rodlike#
2kt9#	116 #	0,76#	Rodlike#
1qqk#	254#	0,73#	Rodlike#
1emw#	88 #	0,71#	Rodlike#
1gr3#	132#	0,65#	Rodlike#
3dmw#	87#	0,65#	Rodlike#
3hon#	55 #	0,63#	Rodlike#
3p46#	39#	0,57#	Discoidal#
3js7#	174 #	0,55 #	Discoidal#
3hal#	262#	0,54#	Discoidal#
3jqu#	173#	0,52#	Discoidal#
1azz#	727#	0,51#	Discoidal#
1v1h#	610#	0,51#	Discoidal#
2dkm#	104#	0,50#	Discoidal#
2v53#	292#	0,47#	Discoidal#
2vl5#	869#	0,42 #	Discoidal#
3kou #	482 #	0,41 #	Discoidal#
1bkv #	68 #	0,41 #	discoidal#
2otp #	387 #	0,37 #	Spherical#
3qae #	371 #	0,36 #	Spherical#
3lqb #	207 #	0,32 #	Spherical#
1ao6 #	1556 #	0,30 #	Spherical#
2xvq #	1135 #	0,21 #	Spherical#
2vuf #	1105 #	0,21 #	Spherical#
2xw0 #	1135 #	0,20 #	Spherical#
3gbl #	97 #	0,18 #	Spherical#
3jxp #	307 #	0,16 #	Spherical#
2y3z #	351 #	0,13 #	Spherical#
3po6 #	263 #	0,12 #	Spherical#
3ln3 #	331 #	0,11 #	Spherical#
3p43 #	126 #	0,11 #	Spherical#
2q2m #	152 #	0,10 #	Spherical#

Table 1 Contd.....

PDB ID	Nodes	ξ	Shape
2jtd #	112 #	0,10 #	Spherical#
3p4h #	142 #	0,08 #	Spherical#
1uz2 #	158 #	0,05 #	Spherical#
3q1q #	112 #	0,05 #	Spherical#
3npo #	169 #	0,02 #	Spherical#

In order to put into perspective the proposed asymmetry index, we introduce some topological descriptors, based on a protein structure representation in terms of inter-residue contact graphs [16].

As a matter of fact, the 3D crystal structure of a protein canbe translated into a contact matrix among α -carbons that in turn can be considered as a network with α-carbons as nodes and the contacts between them as edges. This kind of formalization isextremely useful to study protein properties at all [17-19].

Starting from the spatial position of α -carbons, in the PDB files, the mutual residue distance matrix **d** = $\{d_{ij}\}$ is computed : the generic element d_{ij} is the Euclidean distance in the 3D space between the i-th and j-th residue, holding the primary structure ordering. A link is established between two residues if their mutual distance lies in the range 4 − 8Å; the contact graph adjacency matrix A =$\{a_{ij}\}$ is therefore defined as:

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

#

Some topological descriptors can be extracted from **A** [20]:

- *N* : number of nodes (residues) in the graph;
- *E* : number of edges connecting the graph nodes;
- *density* : ratio between the actual number of edges *E* and the maximum value;
- *N* (*N*− 1)/ 2, corresponding to the complete graph;
- *avdegree* : the average of node degrees k_i , where $k_i = \sum_{j=1}^{N} a_{ij}$ is the number of links involving the i-th node;
- *avshortpath* : the shortest path is the minimum number oflinks connecting two residues; this value, averaged over allthe residue pairs, is the average shortest path;
- *diameter* : the longest shortest path;
- *avcluscoeff*: the clustering coefficient $C_i = \sum_{j,m\in N,\, j\neq m} \frac{2 \quad a_{ij}a_{jm}a_{mi}}{k_i \cdot (k_i - 1)}$ is a measure of connectivity on a local scale, for the i-th node: it measures the connectivity of the sub-graph made of nodes con-

nected to thei-th node. C_i averaged over the whole set of nodes is the *avcluscoeff*.

3. RESULTS

We computed the asymmetry ξ and the seven above mentioned topological properties for each protein of the data set. In order to evaluate the correlation of ξ with the other parameters, a multivariate data analysis is required. To this aim, we applied PCA to the data matrix, containing all the computed properties for each protein in the data set.

The presence of a specific component highly correlated withξ (PC2) is a consequence of the selection of a data set spanning the entire range from spherical to rod-like structures. On the other hand, protein size as measured by N is the main order parameter shaping the data set (PC1).

Results are reported in Table **2** in terms of component loadings, i. e., of correlation coefficients between principal components(PCs) and original variables. A high absolute value of the correlation coefficient (loading) between a variable and a component is used as guide for the structural interpretation of the extracted components.

PCA highlighted a three component solution as explaining the by far most important (and reasonably signal-like) part of information correspondent to the 86% of total variance, with PC1 explaining the 47% of variability, while PC2 and PC3 accounting for 25% and 14% respectively.

Not surprisingly, the first component (PC1) corresponds toprotein size: the number of nodes *N*, as well as the number of links *E*, are strongly related to this component.

Both contact density and clustering coefficient negatively scale with size, confirming previous results, [19]. As shown in Fig. (**5**), density neatly scales with size (here, number of nodes).

The second component (PC2) identifies the 'general shape',since asymmetry has the highest correlation. It has to be stressed that *diameter* and *avcluscoeff* bring considerable contributions to PC2, suggesting that topology influences general shape.

In the case of PC3, the only relevant descriptor is *avdegree*. Therefore, this component is a topologic invariant, since it is neither influenced by size nor by general shape.

Afterwards, we repeated the analysis taking out asymme-

Table 2. Principal Component (PC) Pattern

	PC1	PC2	PC3
ξ(AS)	-0.37	0.80	-0.20
nodes (*N*)	0.92	-0.02	-0.07
avdegree	0.13	0.27	0.94
edges (*E*)	0.92	-0.02	-0.004
density	-0.7	0.24	-0.30
avshortpath	0.82	0.52	-0.13
diameter	0.69	0.64	-0.17
avcluscoeff	-0.55	0.76	0.20

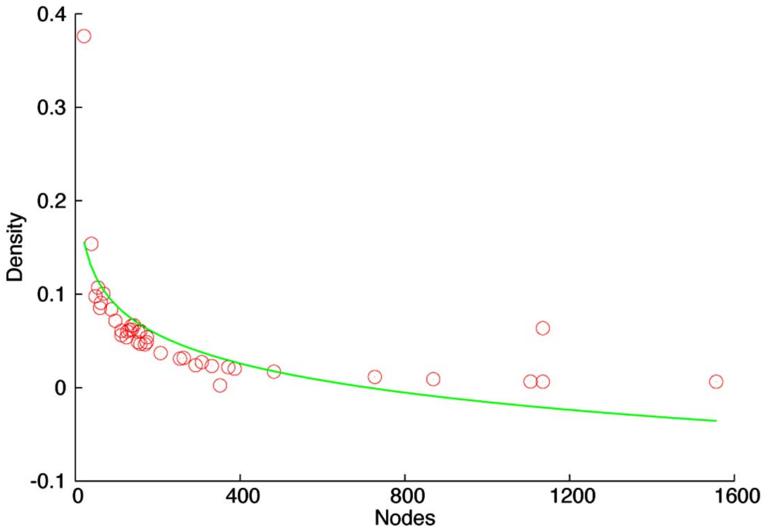

Fig. (5). Effect of protein size on density (open red circles). The data follows a power low behavior (green line).

try, thus evaluating the ability of sole topologic features to predict protein shape.

The ‘reduced’ principal components (PC1r - PC3r, i.e., those components generated without the explicit contribution of ξ)are indeed able to perform a very significant classification of the three groups of rod-like, discoidal, and spherical proteins, as reported in Table **3**, where the classification matrix based on linear discriminant analysis based PC1r-PC3r is reported.

As evident from Table **3**, the discriminant analysis allows for an 84.4% of correct classification.

The efficacy of this discrimination can be appreciated in Fig. (**6**), reporting the space spanned by the first two reduced space components, where the space is approximately subdivided into three regions, correspondent to the three classes.

The ability of the components to separate the three classes is a proof-of-concept of the fact ξ is consistent with other features of protein organization (as those used for generating PC1r-PC3r). As it can be observed in Table **4**, PC1r still represents size, but *avshortpath* is now concentrated on PC1r.

Furthermore, *avcluscoeff* is now the most sensitive descriptor to PC2r, being clusterization linked to shape; on the contrary, *avdegree* does not change appreciably, confirming it isa general invariant property. The ‘reduced’ component space explains the 88% of total variability as follows: PC1r = 52%,PC2r = 21% and PC3r = 15%.

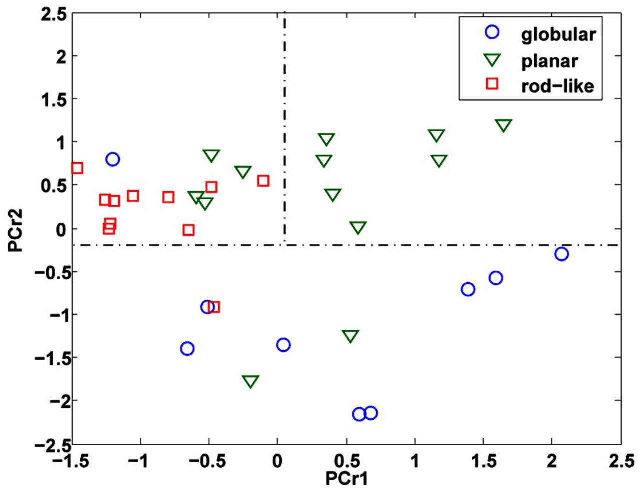

Fig. (6). Cartographic representation of PC2 vs PC1 on the basis of protein shape index ξ.

In order to check the relevance of the proposed index with an independent data set, asymmetry index was computed on a sample made of three different classes of proteins: globins, membrane and fibrous proteins. While ’globin’ and ’fibrous’ classifications refer directly to the protein shape, the denomination ’membrane’ has to do only with the location of the molecule in the cell. According to the presence in membrane proteins of a part of the structure in the form of an elongated (mainly alpha helix) patch inside the membrane,

Table 3. Linear Discriminant Analysis Results. Topology is Able to Predict General Shape

	Estimated Shape			
Original shape	rod	disc	sphere	Total
rod	11	0	3	14
%	78.57	0.00	21.43	100
disc	0	8	1	9
%	0.00	88.89	11.11	100
sphere	1	1	9	11
%	9.09	9.09	81.82	100

Table 4. PCr Component Loadings

	PC1r	**PC2r**	**PC3r**
nodes (N)	0.91	-0.12	0.04
avdegree	0.14	0.53	-0.82
edges (E)	0.91	-0.11	-0.03
density	-0.67	0.26	0.45
avshortpath	0.87	0.38	0.23
diameter	0.76	0.49	0.32
avcluscoeff	-0.46	0.86	0.08

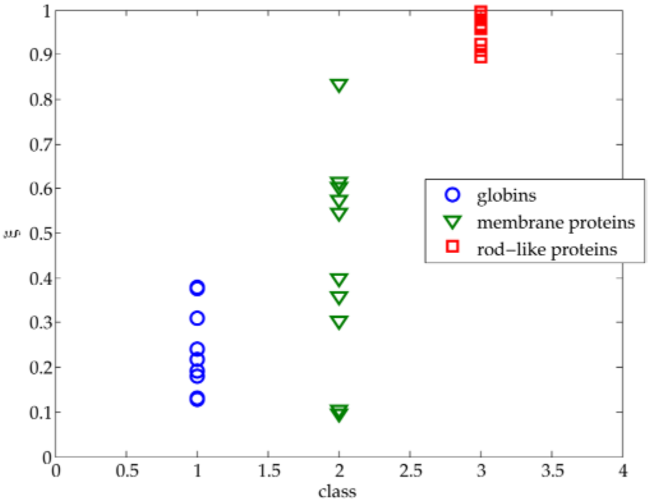

Fig. (7). Structural class of proteins discriminated on the basis of the asymmetry index.

we expect that membrane proteins must lie in between ’globin’ and ’fibrous’ shapes as for their asymmetry index values. In the meantime, we do expect membrane proteins do have an higher variability with respect to the two other classes as for their asymmetry values. Here, see Fig. (**7**), we report result for the shape index for the above three classes of proteins; blue dots are proteins sharing the same globin-fold pattern, resulting in a spheroidal structure; green triangles are membrane proteins known to have widely different shapes with a slight prevalence for elongated forms (at least for the membrane embedded part of the structure), red squares correspond to fibrous proteins, having an elongated, rod-likemolecular shape. As shown in figure, fibrous protein segregate in the upper part of the figure, with asymmetry index close to maximum (Mean = 0.96, Std. Dev. = 0.03) ; globins, that are approximately spherical, locate, as expected, on the bottom, in a wider area with respect to rod-like structures (Mean= 0.24, Std.Dev. =0.09). Membrane proteins, finally, spread out in the wide central part of the figure (Mean = 0.44, Std.Dev. = 0.24),consistently with their morphological variability going hand-in-hand with a tendency toward elongated shape as for their membrane-embedded part. The above results were highly statistically significant for both mean (Students t-test) and variance (F-test) pairwise comparisons. Fibrous vs. membrane comparison scored a t value =6.8 (p ¡ 0.0001) and an F value = 44.35 ($p \leq 0.0001$); globin vs membrane comparison scored a t value= 2.53 ($p \leq 0.03$) and an F value = 6.05 ($p \leq 0.02$), eventually globin vs. fibrous comparison scored a t-value = 21.2 ($p \leq 0.0001$) and an F-value = 7.34 ($p \leq 0.008$). The ability of the index not only to discriminate between different classes but to account for the internal variance of the membrane proteins is afurther proof of their possible use as a simple quantitative shapeindex to study diff erent protein folds.

4. CONCLUSION

As previously suggested by Holm and Sander [1], the generation of a principal component space based on the mutual correlation of different shape features allows for the identification of ‘attractor shapes’ acting as ideal templates rationalizing the apparently wild variety of protein forms. In this work, the same strategy was adopted in order to validate a global shape index allowing for a quantitative appreciation of the position of a given structure in the continuum spanning from very asymmetric fibrous structured to approximately globular shapes.

The possibility to discriminate the pertaining of a given molecule to the ‘rod-like’, ‘discoidal’ and ‘spherical’ attractors by the components of a feature space, not explicitly taking into account the proposed index, was a proof-of-concept of both the existence of such attractors and the consistency of the asymmetry descriptor here defined. Focusing on the quantification of symmetry, in order to build a general shape descriptor is notonly one of the many possible choices. In contrast, symmetry, as aptly explained by Goodsell and Olson[2], is a crucial property for rationalizing structure, function and even evolution history of protein molecules. Here it is sufficient to remind the role played by protein internal structural symmetries in allosteric effects, folding and cell localization [2] and the importance of detecting sequence-based symmetries, for both the modeling of sequence-structure relations and the protein evolution by gene duplication [21].

Our results point to the possibility to sketch a quantitative formalization of a so far largely qualitative concept as protein form, that could have very relevant outcomes in protein science.

This hope is substantiated by the strong, and still largely unexploited, link between general shape information and graph theoretical properties of protein contact networks.

CONFLICTS OF INTEREST

None Declared

ACKNOWLEDGMENTS

We thank the **“**Consorzio interuniversitario per le Applicazioni di Supercalcolo Per Universita e Ricerca**”** (CASPUR) for computing resources and support.

REFERENCES

[1] L. Holm and C .Sander, “Mapping the protein universe”, *Science,* vol. 273, no. 5275 pp. 595-602, 1996.

[2] D. Goodsell and A. Olson, “Structural symmetry and protein function*”, Annu. Rev. Biophys. Biomol. Struct.,* vol. 29*,* pp. 105-153, 2000.

[3] E. Callaway, “The shape of protein structures to come”, *Nature,*vol.449, no.4164, p. 765, 2007.

[4] W. Taylor, A. May, N. Brown and A. Asz´odi, “Protein structure: geometry, topology and classification”, *Rep. Prog. Phys*., vol. 64 , p.517, 2001.

[5] J. Ponomarenko, H. Bui, W. Li, N. Fusseder, P. Bourne, A. Sette and B. Peters, “Ellipro: a new structure-based tool for the prediction of antibody epitopes”, *B.M.C. Bioinformatics,*vol. 9, p. 514, 2008.

[6] V. Natarajan, P. Koehl, Y. Wang and B. Hamann, “Visual analysis of biomolecular surfaces”, *Math Vis*., vol.5, pp.237-255,2008.

[7] Y. Wang, “*Geometric and Topological Methods in Protein Structure Analysis*”, Ph.D. thesis, Department of Computer Science, Duke University,USA, 2004.

[8] A. Bondi, “Van der waals volumes and radii”, *J. Phys. Chem*., vol. 68, no. 3, pp. 441-451, 1964.

[9] B. Lee and F. Richards, "The interpretation of protein structures: Estimation of static accessibility", *J. Mol. Biol.,* vol. 55, no. 3, p. 379, 1971.

[10] M. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids", *Science*, vol. 221, pp. 709-713, 1983.

[11] J. Giard, J. Ambroise, J. Gala and B. Macq, "Regression applied to protein binding site prediction and comparison with classification", BMC. *Bioinformatics*, vol.10, p. 276, 2009.

[12] A. Hopfinger, "Theory and application of molecular potential energy fields in molecular shape analysis: a quantitative structure-activity relationship study of 2,4-diamino-5-benzylpyrimidines as dihydrofolate reductase inhibitors", *J. Med. Chem.,*vol. 26, no.7, pp. 990-996, 1983.

[13] M. Connelly, "Shapes of small molecules and proteins". Available at: http://www.netsci.org/Science/Compchem/feature14.html

[14] P. Røgen and R. Sinclair, "Computing a new family of shape descriptors for protein structures", *J. Chem. Inf. Comput. Sci*., vol. 43, pp. 1740-1747, 2003.

[15] R. Preisendorfer, "*Principal Component Analysis in Meteorology and Oceanography*", Elsevier: Amsterdam, 1988.

[16] A. Giuliani, L. Di Paola and R. Setola, "Proteins as networks: A mesoscopic approach using haemoglobin molecule as case study", *Curr. Proteomics*, vol.6, pp. 235-245, 2009.

[17] A. Giuliani, J. Zbilut and M. Tomita," Network scaling invariants help to elucidate basic topological principles of proteins", *J. Proteome Res.,* vol. 6 ,no.10, pp. 3924-3934, 2007.

[18] G. Bagler and S. Sinha,"Assortative mixing in protein contact networks and protein folding kinetics", *Bioinformatics*, vol. 23, no.14, pp.1760-1767, 2007.

[19] J. Zbilut, G. Chua, A. Krishnan, C. Bossa, K. Rother, C. Webber and A. Giuliani, "A topologically related singularity suggests a maximum preferred size for protein domains", *Proteins,* vol. 66, no. 3, pp.621-629, 2007.

[20] M. Newman, *Networks: An Introduction*, Oxford University Press: USA, 2010.

[21] X. Ji and H. C. Y. Xiao, "Hidden symmetries in the primary sequences of beta-barrel family", *Comput. Biol. Chem.,* vol. 3, no.1, pp. 61–63, 2007.

Received: November 21, 2011 Revised: December 20, 2011 Accepted: December 23, 2011