

Investigating the structure of protein similarity networks both on sequence and structural level

IOANNIS VALAVANIS, GEORGE SPYROU, KONSTANTINA NIKITA

Abstract-- The current work presents a study concerning similarity networks for a well defined and used dataset of proteins, constructed using sequence and structural derived similarity criteria. The analysis is made on the initial set of proteins and the subsets of proteins that yield to fully connected networks using both similarity criteria. Several parameters describing the networks are reported, whereas the values of clustering coefficient and characteristic path length classify both types of fully connected networks into the class of small-world networks. This reveals the grouping of proteins in clusters and the existence of random edges in the networks due to similarity transition among proteins.

Index Terms—characteristic path length, clustering coefficient, proteins, similarity network, fold

I. INTRODUCTION

OUR world can be described as the world of networks, where each complex system is defined by a number of vertices, the elements that construct the system, and a number of edges that describe the interactions for all pairs of the elements. Biological and chemical systems, social interacting species, computer networks and the Internet are only some paradigms of networks [1]. Two extreme structures of a network are the regular and the random networks. The regular network (Fig. 1, a) is highly clustered, that is it has a high density of connections between nearby vertices, but it has long path length: one needs to pass in average from a great number of edges in order to get from one vertex to another. Random networks (Fig. 1, c) are unclustered and random edges between vertices give the property of a short path length. When a few edges between not nearby vertices are superimposed on regular networks, we get a structure between the two

Manuscript received February 20, 2007. This work was supported in part by the Federal Scholarships Foundation of Greece. Ioannis Valavanis (email: ivalavan@biosim.ntua.gr, tel. +302107722968) and Konstantina Nikita (email: knkita@cc.ece.ntua.gr) are with Biomedical Simulations and Imaging Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str., 15780, Zografos, Athens, Greece. George Spyrou (corresponding author, email: gspyrou@bioacademy.gr, tel. +302106597151) is with Biomedical Research Foundation of the Academy of Athens, Soranou Efessiou 4, 115 27 Athens, Greece.

extremes which is called a small-world network (SWN) (Fig. 1, b) [2]. SWNs are highly clustered and have a short path length, combination of properties that suggest

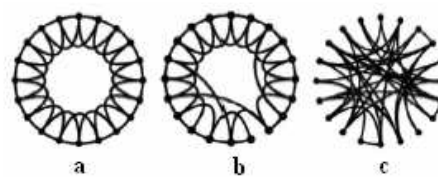


Fig. 1. Graph depictions of a regular network (a), a small-world network (b) and a random network (c)

that they can be used to describe real-world interacting systems [3]. The regularity or randomness of a network can be described by two parameters: i) the characteristic path length, L , which is the average over the minimum number of connections that must be traversed to connect all possible pairs of vertices and ii) the clustering coefficient, C , which is the average degree of clustering in the neighborhoods of all vertices. A random network has small L and C , a regular network has great L and C , while a SWN has small or intermediate L and great C [5].

Proteins, among other real-world complex systems, have been described already as small-world networks. In [4]-[5], folded proteins are transformed to networks where amino acids correspond to vertices and edges occur when two amino acids are closer than a distance threshold. The networks obtained possess small-world properties. In [6], a similarity network using representative proteins of 585 folds as vertices was constructed and each vertex is connected with another given a structural similarity of the corresponding proteins. The constructed network belongs to the class of small-world network.

In the current study, similarity networks using an initial non-redundant set of proteins which belong to four categories (α , β , α/β , $\alpha+\beta$) on class level and 27 categories on fold level according to SCOP [8] are constructed. The networks are constructed using similarity criteria based both on the euclidian distance between feature vectors that describe the protein sequences and a score of the structural alignment of the proteins. Several parameters are calculated to describe the networks, which are finally classified in one of the categories of random, small world or regular networks.

II. PROTEIN DATASET

A well defined and used as reference dataset of 311 proteins ([7], [9]) was used in order to construct the protein similarity networks. This dataset is a non-redundant set of proteins that belong to 27 folds according to SCOP [8] and contains proteins with more than 35% of sequence identity for aligned subsequences longer than 80 residues [7]. The 27 folds represent all major structural classes (α , β , α/β , $\alpha+\beta$). Since the dataset was originally used for processing protein sequences, a search for the corresponding structures in Protein Data Bank [10] and a pattern matching between the sequences contained in the set and the whole .pdb files were utilized within the current study. This yielded to a total of 296 protein substructures (out of the 311 protein sequences), due to .pdb files with missing information on atom coordinates. The distribution of proteins sequences and structures of the dataset in the 27 folds and four structural classes is presented in Table I.

TABLE I
PROTEINS OF 27 SCOP FOLDS USED IN THE STUDY

Fold	N _{seq.}	N _{struc.}
α class		
Globin-like	13	13
Cytochrome <i>c</i>	7	6
DNA-binding 3- helical bundle	12	12
4-helical up-and-down bundle	7	7
4-helical cytokines	9	7
Alpha;EF-hand	6	6
β class		
Immunoglobulin-like β -sandwich	30	27
Cupredoxins	9	9
Viral coat and capsid protein	16	16
ConA-like lectins/glucanases	7	6
SH3-like barrel	8	8
OB-fold	13	13
Trefoil	8	7
Trypsin-like proteases	9	8
Lipocalines	9	9
α/β class		
(TIM)-barrel	29	28
FAD (also NAD)-binding motif	11	11
Flavodoxin-like	11	11
NAD(P)-binding Rossmann-fold	13	13
P-loop containing nucleotide	10	9
Thioredoxin-like	9	9
Ribonuclease H-like motif	10	9
Hydrolases	11	11
Periplasmic binding protein-like	11	11
$\alpha+\beta$ class		
β -grasp	7	7
Ferredoxin-like	13	11
Small inhibitors,toxins,lectins	13	12

N_{seq.} and N_{struc.} are the number of sequences and structures of each fold used in the current study.

III. METHODOLOGY

A. Construction of Similarity Networks

The protein similarity networks are constructed using each of the proteins as vertex in the network, whereas an edge occurs between two proteins given that a certain similarity criterion is met. Similarity networks are constructed separately using similarity criteria on sequence and structural level. In the case of sequences, the similarity between two proteins is calculated based

on 125 sequence derived features concerning amino acid composition (20 features), predicted secondary structure (21 features), hydrophobicity (21 features), normalized van der waals volume (21 features), polarity (21 features) and polarizability (21 features) [7]. The exact way these features are extracted is described in [11]. In this study, the values of features were normalized in [0-1] and the euclidian distances between the 125-dimensional feature vectors for all possible pairs ($311 \times 310 / 2$ in total) of proteins were calculated. An edge occurs between two vertices (proteins) given the distance is less than a predefined threshold. This threshold was manually set to 1.6 so as the constructed network is sparse (~9% of all possible edges appear), whereas the min, mean and max value of all distances were found 0.75, 2.45 and 5.39 respectively. In case of structures, the DALI program [12] for structural alignment of proteins was utilized. A total number of $296 \times (296 - 1) / 2$ pairwise structural alignments were executed and the Z-score that DALI outputs was used as quantitative measure of the structural similarity. The criterion $Z \geq 2$ was used to ensure that significant similarity is found between two structures, as done in other studies ([6], [13]), and an edge between two proteins appears in the structural similarity network given that this criterion is met. The similarity network obtained was found to be sparse as well (~8% of all possible edges appear).

Both similarity networks obtained using the initial sets of proteins (311 sequences or 296 structures) were not found to be fully connected, indicating that there are vertices that could not be connected with all others through a finite path. Thus, networks were reconstructed after removing the isolated vertices (either existing as orphan vertices or in isolated groups) in order to get fully connected similarity networks.

B. Analysis of Similarity Networks

In order to analyze a network of N vertices and K edges several parameters are calculated to

describe the network: i.e. the degree k , that is the number of neighbors of a vertex averaged over all vertices ($k=2K/N$) or the level of sparsity of the network, that is the fraction of all possible edges that appear in the network ($S=2K/N(N-1)$). The fraction of all paths between vertices that are finite (100% for a fully connected network) is another measure of the connectivity of the network, whereas the regularity or randomness of a network can be described by the characteristic path length L and the clustering coefficient C : L is the average over the minimum number of connections that must be traversed to connect vertices i and j :

$$L = \frac{2}{N(N-1)} \sum_{j>i} \lambda_{ij} \quad (1)$$

where λ_{ij} is the minimal path between vertices i and j , and C is average value of local connectivity in the neighborhood of a vertex:

$$C = \frac{1}{N} \sum_i \frac{K_i}{N_i(N_i-1)/2} \quad (2)$$

where N_i the number of neighbors of a vertex i and K_i the number of edges among the neighbor vertices of vertex i . L and C values can be calculated for the random or regular network with the same number of vertices N and edges K , classifying a network as regular, normal or SWN:

$L_{\text{random}} = \ln(N)/\ln(k)$, $C_{\text{random}} = k/N$, $L_{\text{regular}} = N(N+k-2)/[2k(N-1)]$ and $C_{\text{regular}} = 3(k-2)/[4(k-1)]$ [5].

IV. RESULTS AND DISCUSSION

Initially, a sequence similarity network using all available protein sequences (311 in number) and a structural similarity network using all available protein structures (296 in number) were constructed. A graphic depiction of their symmetric adjacency matrices are presented in Fig. 2a and b respectively, where a dot (.) represents an edge between two vertices i and j . The indexing of proteins along x, y axes follows the order of the folds they belong to as presented in Table I, so as proteins belonging to the same fold or class are allocated on the same segments of x, y axes. It

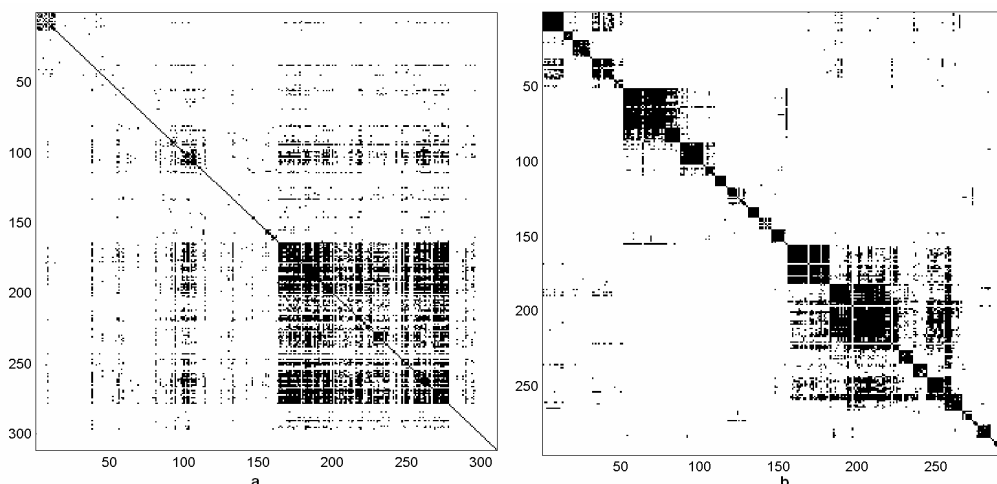


Fig. 2. Graph depictions of the symmetric adjacency matrix for a) the sequence similarity matrix using all protein sequences (311x311) and b) the structural similarity matrix using all protein structures (296x296)

can be observed that the adjacency matrices tend to be denser in regions near the diagonal thus showing the clustering of proteins in fold or classes. This is very obvious in the case of structures (Fig. 2b), as fold and class assignment derives from the study of protein structures, but can also be inferred in the case of sequences (Fig. 2a), especially in class α/β ($164 \leq i, j \leq 278$). However, there are many protein sequences that appear as single dots along the axis and are actually vertices with no neighbours in the network (orphan vertices). Parameters S , k and fraction of finite paths were calculated for both networks and are reported in Table II. It is observed that despite the sparse networks ($S=9.3\%$ and $S=8.4\%$ for sequences and structures, respectively) and the low average network degrees ($k=28.9$ and $k=24.9$ for sequences and structures, respectively), a great fraction of all possible paths are finite (47.3% and 76.2% for sequences and structures, respectively). This shows the connectivity that is achieved on both networks due to random edges between clusters (dots located far from the region of diagonal). A fraction of finite paths equal to 76.2% in the sparse network of protein structures can be assigned to the existence of evolutionary pathways on the structural level.

Fully connected similarity networks were constructed in a next step by removing the isolated vertices from the largest connected cluster of the initial set of sequences or structures. Thus, a total of 97 sequences (out of the 311), 93 orphan vertices and 2 isolated pairs of vertices, were

removed from the initial sequence similarity network, while only 38 out of 296 (~13%), 13 orphans and 25 in isolated groups with up to 13 members, were removed in order to get a full connected structure similarity network. Proteins corresponding to vertices that were removed can be the result of serious alterations during biological evolution that kept them out of the big connected group of proteins on sequence and structural level. Similar results for the case of a network of structures of representative folds were also found in [6]. Parameters S , k are reported for the full connected similarity networks in Table II. The networks are found less sparse and of greater degree than the initial networks, while all paths to connect all possible pairs of vertices are finite as expected in a full connected network. Furthermore, the parameters L and C were calculated for both networks and the networks with the same number of edges and vertices that belong to the class of random and regular networks, as well (Table II). It can be observed that the similarity networks have a high clustering coefficient ($C=72.6\%$ for both networks) and an intermediate value for L ($L=2.414$ and $L=3.339$ for sequence and structure similarity network, respectively) when compared to L calculated for a random and a regular network. Thus, it is inferred that the similarity networks possess SWN properties and belong to this class of networks. The high clustering coefficient is due to the structure or sequence similarity in a group of sequences or structures belonging to same fold or class, while the intermediate characteristic path lengths occur due to the continuous transition of structural or sequence similarity among proteins. Similar results were reported in [6], which studied a network of structures of representative folds. It can be noted that L calculated for structure similarity network is relatively smaller than L obtained for sequence similarity network (compared to respective L_{regular} and L_{random}), showing that transition of similarity is mainly done on the level of structure. This is also shown by the great fraction of finite paths (~76%) in the initial structure similarity network and

TABLE II
PARAMETERS DESCRIBING THE CONSTRUCTED NETWORKS

Similarity Network	N	S	k	Fraction of finite paths	L	C	L_{random}	C_{random}	$L_{regular}$	$C_{regular}$
Sequence level										
Using all initial protein sequences	311	9.3	28.9	47.3	-	-	-	-	-	-
Full connected network	214	19.7	41.9	100	2.414	72.6	1.446	19.6	3.045	73.2
Structure level										
Using all initial protein structures	296	8.4	24.9	76.2	-	-	-	-	-	-
Full connected network	258	10.9	28.0	100	3.339	72.6	1.685	10.9	5.089	72.2

N , S (%), k , L and C (%) are the number of vertices, the sparsity, the degree, the characteristic path length and the clustering coefficient, respectively. L_{random} , C_{random} (%), $L_{regular}$ and $C_{regular}$ (%) are the values of the characteristic path length and the clustering coefficient for random or regular network with same number of vertices and edges. The fraction of finite paths over all possible paths (%) is reported, as well.

the small number of structures that were removed to get a fully connected network.

V. CONCLUSION

Similarity networks of proteins were constructed using sequence or structural similarity criteria and shown to express both clustering of sequences and structures in groups, i.e. classes and folds, and sequence or structural similarity transition. Fully connected networks were obtained by removing isolated vertices from the initial networks of sequences and structures and were assigned to the class of small-world networks using the values of characteristic path length and clustering coefficient. Finally, similarity transition appears more on the level of structure as a result of the evolutionary pathways on this level.

REFERENCES

- [1] Y. Bar-Yam, "Dynamics of Complex Systems", Addison-Wesley, Reading, MA, 1997.
- [2] D.J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks" Nature, 393, 440-442, 1998.
- [3] S.H. Strogatz, "Exploring complex networks", Nature, 410, 268-276, 2001
- [4] L.H. Greene, V.A. Higman, "Uncovering Network Systems Within Protein Structures", Journal of Molecular Biology, 334, 781-791, 2003.
- [5] M. Vendruscolo, N.V. Dokholyan, E. Paci, M. Karplus M, "Small-world view of the amino acids that play a key role in protein folding", Physical Review E, 061910, 2002.
- [6] Z.B. Sun, X.W. Zou, W. Guan, Z.Z. Jin, "The architectonic fold similarity network in protein fold space", European Physical Journal B, 49, 127-134, 2006
- [7] C.H.Q. Ding, I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks", Bioinformatics, 7, 349-358, 2001.
- [8] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, "Scop: a structural classification of proteins database for the investigation of sequences and structures", Journal of Molecular Biology, 247, 536-540, 1995.
- [9] K. Marsolo, S. Parthasarathy, C. Ding, "A Multi-Level Approach to SCOP Fold Recognition", In Proceedings of the Fifth IEEE Symposium on Bioinformatics and Bioengineering, 57-64, 2005.
- [10] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, 28, 235-242, 2000.
- [11] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, "Prediction of protein folding class using global description of amino acid sequence", In Proceedings of National Academy of Science of USA, 92, 8700-8704, 1995.
- [12] L. Holm, C. Sander, "Protein structure comparison by alignment of distance matrices", Journal of Molecular Biology, 233, 123-138, 1994.
- [13] G. Getz, A. Starovolsky, E. Domany, "F2CS: FSSP to CATH and SCOP prediction server", Bioinformatics 20, 13, 2150-2152, 2004.