A General Model for Amino Acid Interaction Networks.

Omar Gaci, Stefan Balev

A General Model for Amino Acid Interaction Networks

Omar GACI and Stefan BALEV

***Abstract*—In this paper we introduce the notion of protein interaction network. This is a graph whose vertices are the protein's amino acids and whose edges are the interactions between them. Using a graph theory approach, we identify a number of properties of these networks. We compare them to the general small-world network model and we analyze their hierarchical structure.**

***Keywords*—interaction network, protein structure, small-world network.**

I. Introduction

Proteins are biological macromolecules participating in the large majority of processes which govern organisms. The roles played by proteins are varied and complex. Certain proteins, called enzymes, act as catalysts and increase several orders of magnitude, with a remarkable specificity, the speed of multiple chemical reactions essential to the organism survival. Proteins are also used for storage and transport of small molecules or ions, control the passage of molecules through the cell membranes, etc. Hormones, which transmit information and allow the regulation of complex cellular processes, are also proteins.

Genome sequencing projects generate an ever increasing number of protein sequences. For example, the Human Genome Project has identified over 30,000 genes which may encode about 100,000 proteins. One of the first tasks when annotating a new genome, is to assign functions to the proteins produced by the genes. To fully understand the biological functions of proteins, the knowledge of their structure is essential.

In their natural environment, proteins adopt a native compact form. This process is called folding and is not fully understood. The process is a result of interactions between the protein's amino acids which form chemical bonds. In this paper we identify some of the properties of the network of interacting amino acids. We believe that understanding these networks can help to better understand the folding process.

The rest of the paper is organized as follows. In section II we briefly present the main types of amino acid interactions which determine the protein structure. In section III we introduce our model of amino acid interaction networks. Section IV presents two general network models, random graphs and small-world networks. In section V we compare protein interaction networks to the general models and empirically characterize them based on all protein structures available in PDB. We show how the properties of these networks are related to the structure of the corresponding proteins. Finally, in section VI we conclude and give some future research directions.

Le Havre University, LITIS EA 4108, BP 540, 76058 Le Havre - France, email: {Omar.Gaci, Stefan.Balev}@univ-lehavre.fr

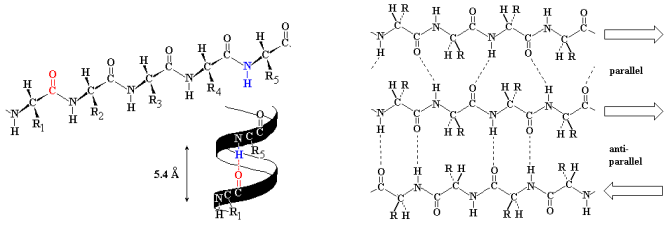

Fig. 1. Left: an α-helix illustrated as ribbon diagram, there are 3.6 residues per turn corresponding to 5.4 Å. Right: A β-sheet composed by three strands.

II. Protein structure

Unlike other biological macromolecules (e.g., DNA), proteins have complex, irregular structures. They are built up by amino acids that are linked by peptide bonds to form a polypeptide chain. We distinguish four levels of protein structure:

- The amino acid sequence of a protein's polypeptide chain is called its primary or one-dimensional (1D) structure. It can be considered as a word over the 20-letter amino acid alphabet.
- Different elements of the sequence form local regular secondary (2D) structures, such as α-helices or β-strands.
- The tertiary (3D) structure is formed by packing such structural elements into one or several compact globular units called domains.
- The final protein may contain several polypeptide chains arranged in a quaternary structure.

By formation of such tertiary and quaternary structure, amino acids far apart in the sequence are brought close together to form functional regions (active sites). The reader can find more on protein structure in [4].

One of the general principles of protein structure is that hydrophobic residues prefer to be inside the protein contributing to form a hydrophobic core and a hydrophilic surface. To maintain a high residue density in the hydrophobic core, proteins adopt regular secondary structures that allow non covalent hydrogen-bond and hold a rigid and stable framework. There are two main classes of secondary structure elements (SSE), α-helices and β-sheets (see Fig 1).

An α-helix adopts a right-handed helical conformation with 3.6 residues per turn with hydrogen bonds between C'=O group of residue n and NH group of residue $n+4$.

A β-sheet is build up from a combination of several regions of the polypeptide chain where hydrogen bonds can form

between C'=O groups of one β strand and another NH group parallel to the first strand. There are two kinds of β-sheet formations, anti-parallel β-sheets (in which the two strands run in opposite directions) and parallel sheets (in which the two strands run in the same direction).

From this first division, a more detailed classification can be done. The most frequently used ones are SCOP, Structural Classification Of Proteins [13], and CATH, Class Architecture Topology Homology [14]. They are hierarchical classifications of proteins' structural domains. A domain corresponds to a part of a protein which has a hydrophobic core and not much interaction with other parts of the protein.

A. SCOP

The SCOP classification is built manually from structural information. The process of classification starts by the division into domains of a protein. The protein is then classified on four levels, from the more general to the more specific :

1) *Class*: There are 4 main classes (see above) and 7 others with very small number of members. A class regroups proteins whose the secondary structure composition is similar.
2) *Fold*: The secondary structure composition, the spatial arrangement and the connexions are similar.
3) *Superfamily*: The structures and the functions tend to be similar.
4) *Family*: Proteins have at least 30% of their sequence identical or have very similar functions and structures.

In 2008, the SCOP classification has identified 1086 folds.

III. Amino Acid Interaction Networks

The 3D structure of a protein is determined by the coordinates of its atoms. This information is available in Protein Data Bank (PDB) [3], which regroups all experimentally solved protein structures. Using the coordinates of two atoms, one can compute the distance between them. We define the distance between two amino acids as the distance between their C_α atoms. Considering the C_α atom as a "center" of the amino acid is an approximation, but it works well enough for our purposes. Let us denote by N the number of amino acids in the protein. A contact map matrix is a $N \times N$ 0-1 matrix, whose element (i, j) is one if there is a contact between amino acids i and j and zero otherwise. It provides useful information about the protein. For example, the secondary structure elements can be identified using this matrix. Indeed, α-helices spread along the main diagonal, while β-sheets appear as bands parallel or perpendicular to the main diagonal [10]. There are different ways to define the contact between two amino acids. Our notion is based on spacial proximity, so that the contact map can consider non-covalent interactions. We say that two amino acids are in contact iff the distance between them is below a given threshold. A commonly used threshold is 7 Å and this is the value we use.

Consider a graph with N vertices (each vertex corresponds to an amino acid) and the contact map matrix as incidence matrix. It is called contact map graph. The contact map graph is an abstract description of the protein structure taking into account only the interactions between the amino acids. Now let us consider the subgraph induced by the set of amino acids participating in SSE. We call this graph SSE interaction network (SSE-IN) and this is the object we study in the present paper. The reason of ignoring the amino acids not participating in SSE is simple. Evolution tends to preserve the structural core of proteins composed from SSE. In the other hand, the loops (regions between SSE) are not so important to the structure and hence, are subject to more mutations. That is why homologous proteins tend to have relatively preserved structural cores and variable loop regions. Thus, the structure determining interactions are those between amino acids belonging to the same SSE on local level and between different SSEs on global level. Fig. 2 gives an example of a protein and its SSE-IN.

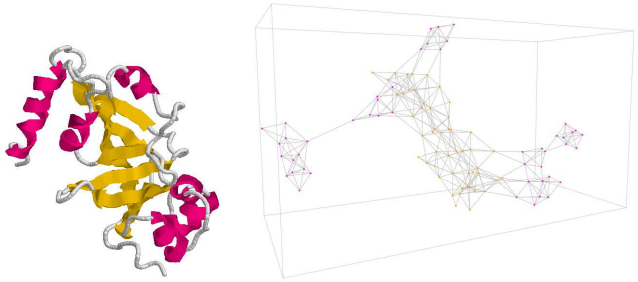

Fig. 2. Protein 1DTP (left) and its SSE-IN (right).

In [12], [5], [2], [7] the authors rely on similar models of amino acid interaction networks to study some of their properties, in particular concerning the role played by certain nodes or comparing the graph to general interaction networks models. Thanks to this point of view the protein folding problem can be tackled by graph theory approaches.

IV. General Models of Networks

Many systems, both natural and artificial, can be represented by networks, that is, by sites or vertices bound by links. The study of these networks is interdisciplinary because they appear in scientific fields like physics, biology, computer science or information technology. These studies are lead with the aim to explain how elements interact with each other inside the network and what are the general laws which govern the observed network properties.

From physics and computer science to biology and social sciences, researchers have found that a broad variety of systems can be represented as networks, and that there is much to be learned by studying these networks. Indeed, the studies of the Web [6], [1], of social networks [16] or of metabolic networks [11] contribute to put in light common non-trivial properties of these networks which have *a priori* nothing in common. The ambition is to understand how the large networks are structured, how they evolve and what are the phenomena acting on their constitution and formation.

In this section we present two classes of interaction networks by describing their specific properties. We introduce some empirical measures which will be used in the next section in order to study SSE-INs.

A. Random Graphs

The random graph models are one of the oldest network models, introduced in [15] and further studied in [8], [9]. These works identify two different classes of random graphs, called $G_{n,m}$ and $G_{n,p}$ and defined by the following connection rules:

- $G_{n,m}$ regroups all graphs with n vertices and m edges. To generate a graph sampled uniformly at random from the set $G_{n,m}$, one has to put m edges between vertex pairs chosen randomly from n initially unconnected vertices.
- $G_{n,p}$ is the set of all graphs consisting of n vertices, where each vertex is connected to others with independent probability p. To generate a graph sampled randomly, one has to begin with n initially unconnected vertices and join each pair by an edge with probability p.

In $G_{n,m}$ the number of edges is fixed whereas in $G_{n,p}$ the number of edges can fluctuate but its average is fixed. When n tends to be large the two models are equivalent.

Definition 1: The degree of a vertex v, k_v, is the number of edges incident to v. The mean degree, z, of a graph G is defined as follows:

$$z = \frac{1}{n}\sum_{v \in V} k_v = \frac{2m}{n} = p(n-1)$$

B. Small-world Networks

This network model was introduced in [18] as a model of social networks. It has been since adopted to treat phenomena in physics, computer science or social sciences. The model comes from the observation that many real-world networks have the following two properties:

1) The small-world effect, meaning that most pairs of vertices are connected by a short path through the network. This phenomenon has two explanations. First, the concept of "shortcuts" through a network allows to join two distant vertices by a small number of edges [17]. Second, the concept of "hubs", vertices whose connectivity is higher than others provide bridges between distant vertices because most vertices are linked to them.
2) High "clustering", meaning that there is a high probability that two vertices are connected one to another if they share the same neighbor.

To determine if a network is a small-world, one can use the measures described below and compare them to the corresponding measures of a random graph.

Definition 2: The characteristic path length [17], denoted L, of a graph G is the median of the means of the shortest path lengths connecting each vertex v to all other vertices. More precisely, let $d(v,u)$ be the length of the shortest path between two vertices v and u and let $\overline{d}(v)$ be the average of $d(v,u)$ over all $u \in V$. Then the characteristic path length is the median of $\{\overline{d}(v)\}$.

This definition applies when the graph consists of single connected component. However, the SSE-IN we consider in the next section may have several connected components. In this case, when we calculate the mean of the shortest path lengths $\overline{d}(v)$ we take into account only the vertices u which are in the same connected component as v.

TABLE I
STRUCTURAL FAMILIES STUDIED FOR THE SMALL-WORLD PROPERTIES. WE CHOOSE ONLY FAMILIES WHICH COUNT MORE THAN 100 PROTEINS, FOR A TOTAL OF 18294 PROTEINS.

Class	Family Number	Protein Number
All α	12	2968
All β	17	6372
α/β	18	5197
$\alpha+\beta$	16	3757

Since the mean and the median are practically identical for any reasonably symmetric distribution, the characteristic path length of a random graph is the mean value of the shortest path lengths between any two vertices. The characteristic path length of a random graph with mean degree z is

$$L_{\text{RG}} = \frac{\log n}{\log z}$$

It increases only logarithmically with the size of the network and remains therefore small even for large systems.

Definition 3: The local clustering coefficient [17], C_v, of a vertex v with k_v neighbors measures the density of the links in the neighborhood of v.

$$C_v = \frac{|E(\Gamma_v)|}{\binom{k_v}{2}}$$

where $|E(\Gamma_v)|$ is the number of edges in the neighborhood of v and $\binom{k_v}{2}$ is the number of all possible edges in this neighborhood. The clustering coefficient C of a graph is the average of the local clustering coefficients of all vertices:

$$C = \frac{1}{n}\sum_{v \in V} C_v$$

The clustering coefficient of a random graph with mean degree z is

$$C_{\text{RG}} = \frac{z}{n-1}$$

Watts and Strogatz [18] defined a network to be a small-world if it shows both of the following properties:

1) Small world effect: $L \sim L_{\text{RG}}$
2) High clustering: $C \gg C_{\text{RG}}$

V. EXPERIMENTAL RESULTS

The first step before studying the proteins SSE-IN is to select them according to their SSE arrangements. Thus, a protein belongs to a SCOP fold level iff all its domains are the same. We have worked with the SCOP 1.7.3 files. We have computed the measures from the previous section for the four mains classes of the hierarchical classification SCOP (see Table I). Thus, each class provides a broad sample guarantying more general results and avoiding fluctuations. Moreover, these four classes contain proteins of very different sizes, varying from several dozens to several thousands amino acids in SSE.

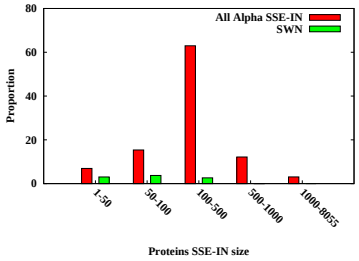

Fig. 3. Size distribution of proteins SSE-IN and small-world networks ratio for All α class.

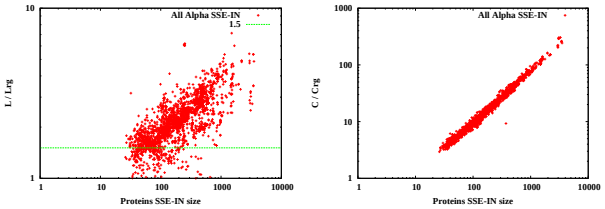

Fig. 4. L/L_{RG} (left) and C/C_{RG} (right) ratios as a function of SSE-IN size.

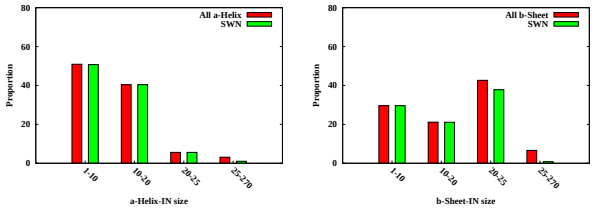

Fig. 5. Size distribution of α-helix (left) and β-sheet (right) interaction networks. Almost all of them are small worlds.

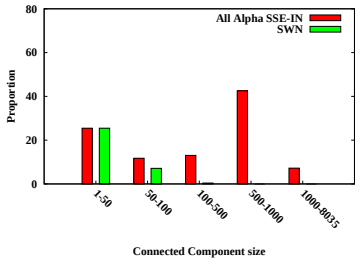

Fig. 6. Distribution of connected component size, SWN appear when size does not exceed 100.

As described in the previous section, a small-world network (SWN) is characterized by two main properties, the small world effect and a high clustering. These measures are computed by evaluating the ratio between a graph and a random graph whose mean degree and size are the same. We consider that a protein SSE-IN is small world iff $L/L_{\mathrm{RG}} \leq 1.5$ and $C/C_{\mathrm{RG}} \geq 2$.

Fig. 3 shows the size distribution of one class of studied SSE-IN (other classes providing similar results, we limit ourselves to one plot). We can see that near 60% of the proteins have SSE-IN of size between 100 and 500 amino acids. The small-world properties are satisfied mainly when the size of the network does not exceed 500 amino acids and there are about 13.74% small-world networks among all studied SSE-IN.

Fig. 4 explains the reason for this low rate. One can see that although highly clustered, most SSE-IN do not satisfy the first small-world property.

To explain the results presented on Fig. 4, note that the mean degree z is not very different from one SSE-IN to another and is generally independent from the size. When the mean degree is fixed, the characteristic path length of a random graph grows like $\log n$ and its clustering coefficient has n^{-1} behavior. We can see that there is no clear relationship between the characteristic path length and the SSE-IN size. There are proteins with close sizes but very different path lengths. However, in general it grows faster than logarithmically with the size. The figure also shows that SSE-IN are very highly clustered. The C/C_{RG} has clearly linear behavior, hence the clustering coefficient (like the mean degree) is independent from the size.

In the rest of this section we lead an upward analysis of the network following its hierarchical structure. We start by low-level subnetworks corresponding to single SSE. At intermediate level, we are interested in SSE-IN topological constitution, notably connected components. Finally, we go back to macroscopic level to put in highlight the conditions under which a given SSE-IN is small world.

The first step consists in describing the subnetworks corresponding to secondary structure elements, α-helices and β-sheets. We observe that these secondary structure motifs are practically all small-world networks, whatever their size (see Fig. 5). The explanation of this fact is the high residue density and the big number of interactions in these compact structures.

The second step concerns the study of connected components. Their size distribution (see Fig. 6) shows that when their size is higher than a threshold, here estimated at 100 residues, the small-world properties are no longer satisfied. To explain this fact, let us consider the edges whose extremities belong to different SSEs (see Fig. 7). These “shortcuts” represent the interactions between different SSEs and they determine the tertiary protein structure. They provide short paths between different network regions and bigger number of shortcuts implies smaller characteristic path length. The ratio of shortcuts is shown on Fig. 8. We can see that the small-world networks have almost the same ratio of shortcuts as the other networks. Although the ratio grows slightly after size of 100, this is not sufficient to decrease the characteristic path length. Consequently, a connected component is a small world network only if its size does not exceed 100 residues.

The third step relies on the previous observations and generalizes them on protein level. Computing the average connected component size, we observe that this size does not exceed 100 for the small-world networks (see Fig. 9). Thus, average connected component size less than 100 is a necessary condition for SSE-IN to be small world. Once again, the shortcut ratio is comparable for the small-world networks and the other networks (see Fig. 9). This ratio is bounded and does not exceed 25%, its the consequence of the excluded volume effect, since the number of residues that can physically

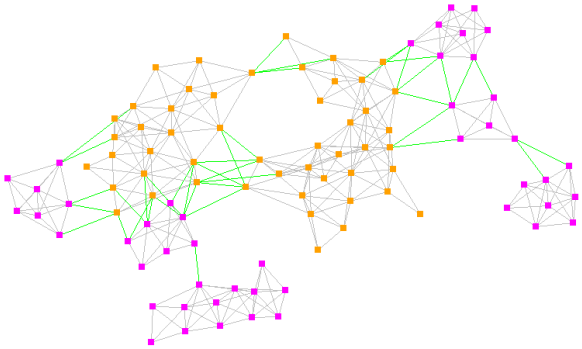

Fig. 7. Protein 1DTP SSE-IN. Shortcut edges are plotted in green.

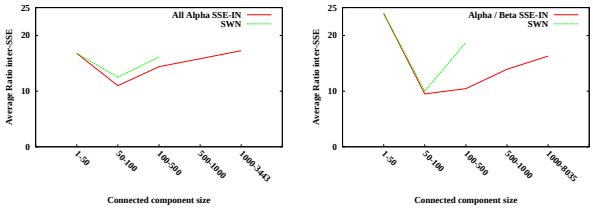

Fig. 8. Shortcut edges / all edges ratio as a function of connected component size. Left: All α class, right: α/β class

reside within a given radius is limited.

VI. Conclusion and Perspectives

In this paper we introduce the notion of interaction network of amino acids of a protein (SSE-IN) and study some of the properties of these networks. The main advantage of this model is that it allows to cope with different biological problems related to protein structure using graph theory tools. Ignoring details, such as the type and the exact position of each amino acid, this abstract and compact description allows to focus on the interactions' structure and organization. We have shown that the subnetworks corresponding to secondary structure elements satisfy the properties of the small-world network model. Small-world networks are widely studied and their properties are well identified. These properties can give insight on the formation of SSEs. On the other hand, the links between these subnetworks describe the interactions between different SSEs, which determine the tertiary protein structure.

A short term perspective is to give a finer characterization of shortcut edges in order to better understand how different SSEs are linked to each other. As a long term perspective, the characterization we propose constitutes a first step of a new approach to the protein folding problem. The properties identified here, but also other properties we plan to study, can give us an insight on the folding process. They can be used to guide a folding simulation in the topological pathway from unfolded to folded state.

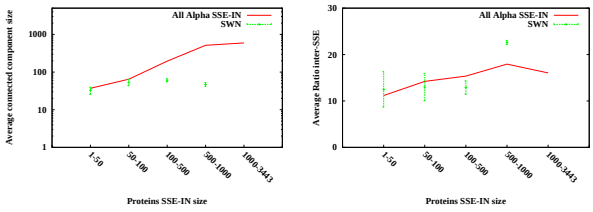

Fig. 9. Average connected component size (left) and shortcut ratio (right) as a function of protein SSE-In size.

References

[1] R. Albert, H. Jeong, and A.-L. Barabási. The diameter of the world wide web. *Nature*, 401:130–131, 1999.

[2] A. R. Atilgan, P. Akan, and C. Baysal. Small-world communication of residues and significance for protein dynamics. *Biophys J*, 86(1 Pt 1):85–91, January 2004.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[4] C. Branden and J. Tooze. *Introduction to protein structure*. Garland Publishing, 1999.

[5] K. V. Brinda and S. Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophys J*, 89(6):4159–4170, December 2005.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, 33(1-6):309–320, 2000.

[7] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich. Topological determinants of protein folding. *Proc Natl Acad Sci U S A*, 99(13):8637–8641, June 2002.

[8] P. Erdõs and A. Rényi. On random graphs I. *Publicationes Mathematicae.*, 6:290–297, 1959.

[9] P. Erdõs and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 7:17, 1960.

[10] A. Ghosh, K. V. Brinda, and S. Vishveshwara. Dynamics of lysozyme structure network: probing the process of unfolding. *Biophys J*, 92(7):2523–2535, April 2007.

[11] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000.

[12] U. K. Muppirala and Z. Li. A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues. *Protein Eng Des Sel*, 19(6):265–275, June 2006.

[13] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of the protein database for the investigation of sequence and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[14] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH - a hierarchic classification of protein domain structures. *Structure.*, 5:1093–1108, 1997.

[15] R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bull. Math. Biophys.*, 13:107111, 1951.

[16] S. Wasserman and K. Faust. *Social network analysis : methods and applications* , volume 8 of *Structural analysis in the social sciences*. Cambridge University Press, Cambridge, 1994.

[17] D. J. Watts. *Small Worlds*. Princeton University Press, Princeton, New Jersey, 1999.

[18] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature.*, 393:440–442, 1998.