Node Degree Distribution in Amino Acid Interaction Networks

Omar Gaci, Stefan Balev

Node Degree Distribution in Amino Acid Interaction Networks

Omar GACI
Le Havre University
LITIS Laboratory
Le Havre, France
Email: omar.gaci@univ-lehavre.fr

Stefan BALEV
Le Havre University
LITIS Laboratory
Le Havre, France
Email: stefan.balev@univ-lehavre.fr

***Abstract*—A protein interaction network 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 study the properties of these networks. In particular, we are interested in the degree distribution and mean degree of the vertices. The results presented in this paper constitute the first steps of a new network approach to the protein folding problem.**

***Keywords*-interaction network; protein structure; scale-free 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 three-dimensional 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.

In this study, we treat proteins as networks of interacting amino acid pairs [3]. In particular, we consider the subgraph induced by the set of amino acids participating in the secondary structure also called Secondary Structure Elements (SSE). We term this graph SSE interaction network (SSE-IN). We carry out a study to identify the node distribution relying on a dataset composed by more than 18000 proteins.

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 three general network models defined by their cumulative degree distribution. In section V we compare protein interaction networks to a general model and empirically characterize them based on a dataset. 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.

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 [5].

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

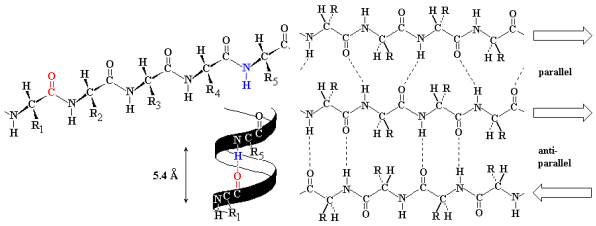

Figure 1. Left: an α-helix illustrated as ribbon diagram, there are 3.6 residues per turn corresponding to 5.4 Å. Right: A β-sheet composed by three strands.

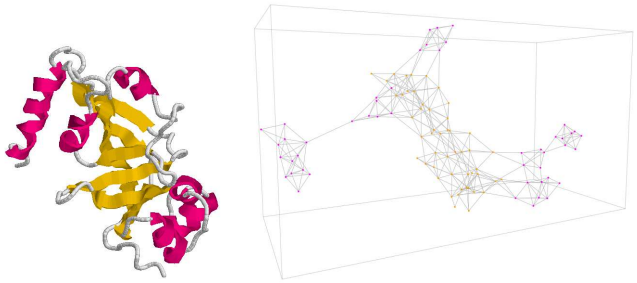

Figure 2. Protein 1DTP (left) and its SSE-IN (right).

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).

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) [4], 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 [13]. 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 [8] 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.

In [15], [6] 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.

As we will see in the next section, there are three main models of interaction networks, extensively studied and whose properties are identified. The purpose of our work is to identify specific properties which associate the proteins SSE-IN with a general network model. Based on such a pattern description of SSE-IN, one can plan the study of their formation, dynamics and evolution.

IV. Interaction Networks

Many systems, both natural and artificial, can be represented by networks, that is by sites or vertices bound by links [16]. The study of these networks is interdisciplinary because they appear in scientific fields like physics, biology, computer science or information technology. The purpose of these studies is to explain how elements interact inside the network and what are the general laws which govern the observed network properties.

From physics and computer science to biology and the social sciences, researchers have found that a broad variety of systems can be represented as networks, and that there

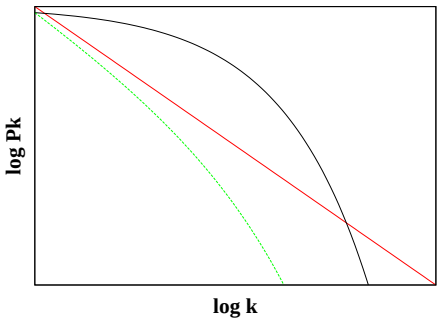

Figure 3. Degree distribution for each the three models described by Amaral [2]. The red line follows a power law, a function with a relatively "fat tail" as for scale-free networks. The green line corresponds to truncated scale-free networks because it describes a power law regime followed by a sharp cut-off. The black curve has a fast decaying tail, typically exponential, and corresponds to single-scale networks.

is much to be learned by studing these networks [7]. Indeed, the study of the Web [1], of social networks [17] or of metabolic networks [14] are contribute to put in light common non-trivial properties to 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 phenomenom acting on their constitution and formation [18].

One of the most important network properties is the degree distribution of vertices. A degree of a vertex is the number of edges incident to it. The mean degree of a network is the mean of the degrees of all vertices. For a network with n vertices and m edges the mean degree is $z = 2m/n$. We will note by p_k the ratio of vertices having degree k (or the probability that a vertex has a degree k). The values p_k define the degree distribution of a network. The cumulative degree distribution $P_k = \sum_{i=k}^{\infty} p_k$ is the probability for a vertex to have a degree at least k.

The random graphs of Erdős and Rényi [9], [10] are the most studied network model. They have Poisson degree distribution. However, many real networks have different degree distributions. Amaral et al [2] have studied networks that can be classified it three groups according to the shape of their cumulative degree distribution, see Fig. 3. First, scale-free networks are those with power law distribution $p_k \sim k^{-\alpha}$ or $P_k \sim k^{-(\alpha-1)}$, a function which decreases polynomially with k. The second class are single scale networks with exponential degree distribution $P_k \sim e^{-k/\alpha}$. This distribution decreases exponentially, much faster than the previous. The third class are broad-scale or truncated scale-free networks with distribution

$$P_k \sim k^{-(\alpha-1)} e^{-k/\alpha} \quad (1)$$

This distribution is somewhere between the previous two, a power law regime followed by a sharp exponential cutoff.

Table I
STRUCTURAL FAMILIES STUDIED. WE CHOOSE ONLY FAMILIES WHICH CONTAIN MORE THAN 100 PROTEINS, FOR A TOTAL OF 18294 PROTEINS. WE HAVE WORKED WITH THE SCOP 1.73 CLASSIFICATION.

Class	Number of families	Number of proteins
All α	12	2968
All β	17	6372
α/β	18	5197
$\alpha+\beta$	16	3757

The common feature of these classes is that most of the vertices have low degree and there exists a small number of high degree nodes. The last are called hubs and play important role for the connectivity of the whole network.

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 v1.73 classification. We have computed the measures from the previous section for the four mains classes of the hierarchical classification SCOP (see Tab. 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.

To compute the cumulative degree distribution for protein SSE-IN (denoted P_k, see Eq. 1), we divided our dataset into two parts whose first one is composed with 20% of the total studied proteins. Then, we fit our specific sub dataset with a function expressed as follows

$$P_k = a\ k^{-b} exp^{-k/c}$$

We have realised a numerical approximation using the method of least squares. Once, we have obtained the coefficients for our sub dataset, we apply the apply them for the rest of the studied protein SSE-IN, a sample of our results is presented on Fig. 4. We can remark that the curves describe a power law regime followed by the sharp cut-off. The power law function is expressed as following:

$$p(k) = 213.413k^{-\alpha},\ \text{where}\ \alpha = 3.2 \pm 0.6$$

while the distribution is approximed by the next function:

$$P_k = 1.48347k^{0.962515} exp^{-k/2.12615}$$

We observe the same result for all studied proteins, that is a cumulative degree distribution approximated by the function P_k. Here, we discuss about characteristics or conditions which involve a such a behavior.

First, we are interested in the degree distribution and mainly its shape, see Fig. 5. We can see that degree

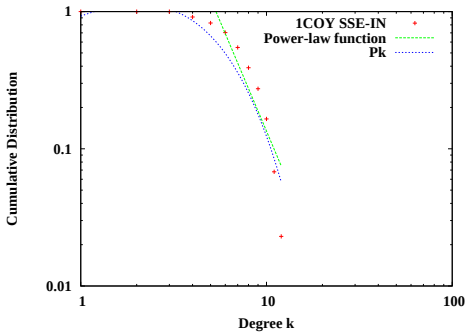

Figure 4. Cumulative degree distribution for protein 1COY SSE-IN.

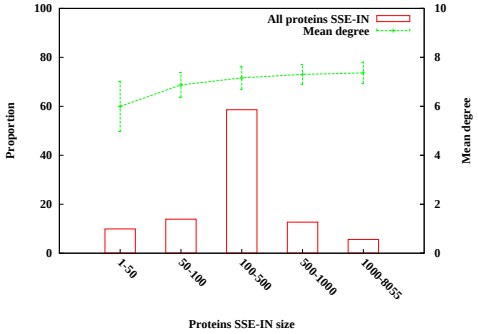

Figure 6. Mean degree distribution according to protein SSE-IN size. It evolves with values enough close, between 5 and 8.

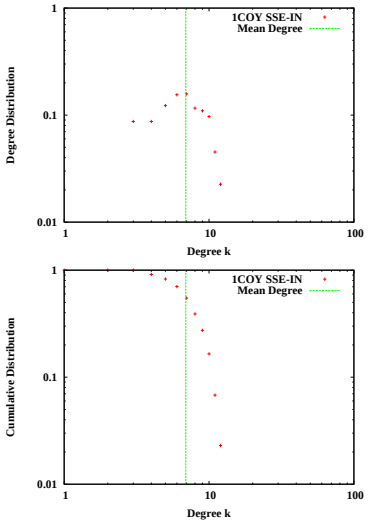

Figure 5. Degree and cumulative distribution for 1COY SSE-IN. They decrease for degree values greather than the mean degree.

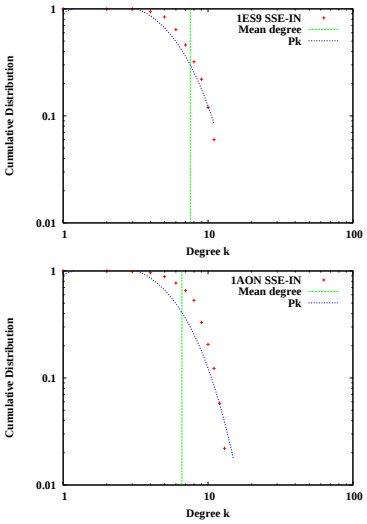

Figure 7. Cumulative degree distribution of 1SE9 and 1AON SSE-IN whose size equal 50 and and 4988. Despite their important size difference, their mean degree stay close and worth respectively 6.6 and 7.5.

distribution follows a Poisson distribution whose peak is reached for a degree near z. This result provides precision about how the vertices are connected within SSE-IN. It implies that the degree of the vertices is homogenous. In other words, a major part of them has a connectivity enough close to the mean degree. Consequently, the cumulative distribution depends on the mean degree value which acts as a threshold beyond which it decreases as an exponential since it's approximed *via* P_k.

Second, we study how the mean degree evolves through all SSE-IN. Its distribution, see Fig. 6, indicates a relative weak variation according to the size. Even if two protein SSE-IN have size ratio around 10 or 100, their mean degree ratio is esimated to 1.05 or 1.15 and remains in the same scale order.

To illustrate the mean degree homogeneity we choose two proteins, namely 1SE9 and 1AON with sizes respectively 50 and 4998. Their size ratio is approximately 100. Even if the mean degrees are slightly different, the distributions are very similar (see Fig. 7).

To recapitulate, we show that the mean degree values constitute a threshold for protein SSE-IN cumulative degree distribution. For degrees lower than the mean degree it decreases slowly and after this threshold its decrease is fast compared to an exponential one, as shown Fig. 4,5,7.

Consequently, we find a way to approximate all proteins SSE-IN cumulative degree distribution by the function P_k which can be adjusted. This function describes a power law regime followed by a sharp cut-off which arises for degree values exceeding the mean degree. Proteins SSE-IN are so truncated scale-free networks.

Since the mean degree plays the role of a threshold beyond which the cumulative degree distribution decreases exponentially, it is interesting to study its evolution with the size of the network. Fig. 6 shows that the mean degree increases very slightly with the size of the network. Even

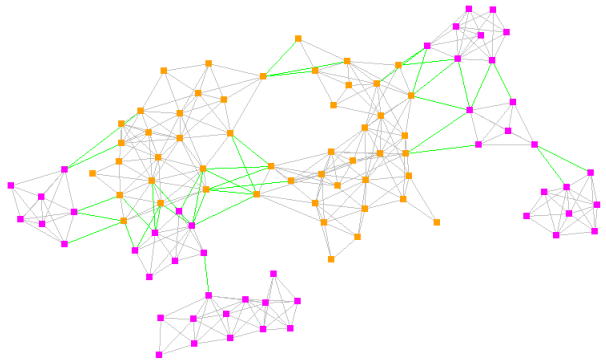

Figure 8. SSE-IN of 1DTP protein. The edges connecting different SSE are green.

for networks with size ratio of 100, the mean degree ratio is only 1.15. As an example, see Fig. 7.

Whatever the size of the network is, we observe that the mean degree is always between 5 and 8. This mean degree interval is a common property characterizing all SSE-IN. In order to explain this property, let us consider the structure of our networks. They are composed of densely connected subgraphs corresponding to secondary structure elements (see Fig. 8). The number of edges connecting different subgraphs is relatively small, but these edges are the most important, since they correspond to interactions determining the tertiary structure.

We start by computing the mean degree in each SSE subgraph. The results are shown on Fig. 9. We can see that the mean degree evolution at microscopic level is almost the same as at macroscopic level (compare to Fig. 6). Independently of the SSE size and type, the mean degree of each SSE subgraph, z_{SSE} is always bounded:

$$z_{\min} < z_{\text{SSE}} < z_{\max} \tag{2}$$

when the size of the network is more than 10. In the general case $z_{\min} = 5$ and $z_{\max} = 8$, but when we consider a specific SSE size and type, finer bounds can be found (see Fig. 9).

Now let us consider a whole protein. Suppose that it contains s secondary structure elements and let the element i has n_i vertices and m_i edges, $i = 1, \ldots, s$. Then the total number of vertices is $n = \sum_{i=1}^{s} n_i$ and the total number of edges is $m = \sum_{i=1}^{s} m_i + m_{\text{inter}}$, where m_{inter} is the number of edges connecting vertices from different SSEs. Let $r = m_{\text{inter}}/m$ be the ratio of inter-SSE edges. Then:

$$\frac{m}{n} = \frac{\sum_{i=1}^{s} m_i + m_{\text{inter}}}{\sum_{i=1}^{s} n_i} = \frac{\sum_{i=1}^{s} m_i}{\sum_{i=1}^{s} n_i} + r\frac{m}{n} \tag{3}$$

and hence for the mean degree z we have

$$z = \frac{2m}{n} = \frac{2}{1-r}\frac{\sum_{i=1}^{s} m_i}{\sum_{i=1}^{s} n_i} \tag{4}$$

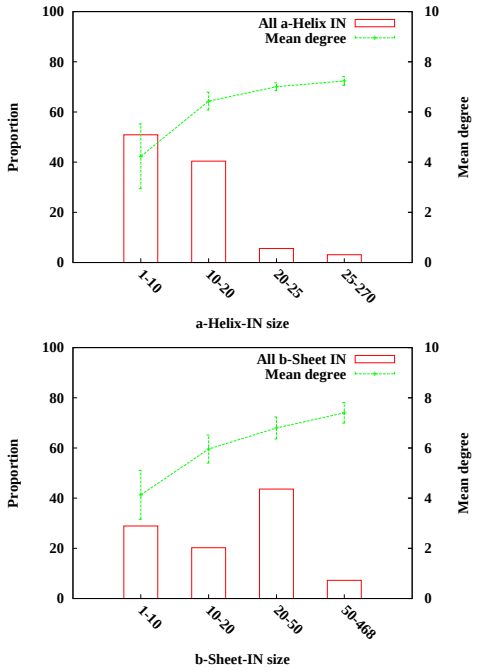

Figure 9. SSE subgraphs size distribution and mean degree as a function of the size.

On the other hand, from (2) it follows that

$$\frac{z_{\min}}{2} n_i < m_i < \frac{z_{\max}}{2} n_i, \; i = 1, \ldots, s \tag{5}$$

By summing up the last equation we obtain

$$\frac{z_{\min}}{2} < \frac{\sum_{i=1}^{s} m_i}{\sum_{i=1}^{s} n_i} < \frac{z_{\max}}{2} \tag{6}$$

which together with (4) gives

$$\frac{z_{\min}}{1-r} < z < \frac{z_{\max}}{1-r} \tag{7}$$

The last equation gives finer bounds on the mean degree. It shows that the bounds on z depend not only on the bounds on z_{SSE}, but also on the ratio of inter-SSE edges. A higher proportion of inter-SSE edges shifts up the bounds. Proteins with bigger size have more SSEs and hence more links between different SSEs. This explains the increase of the mean degree with the size of the networks. Fig. 10 shows that the number of inter-SSE edges is quite variable, but it never exceeds 20%. It is the consequence of the excluded volume effect, since the number of residues that can physically reside within a given radius is limited. This last property explains why the mean degree is homogenous.

VI. Conclusion

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

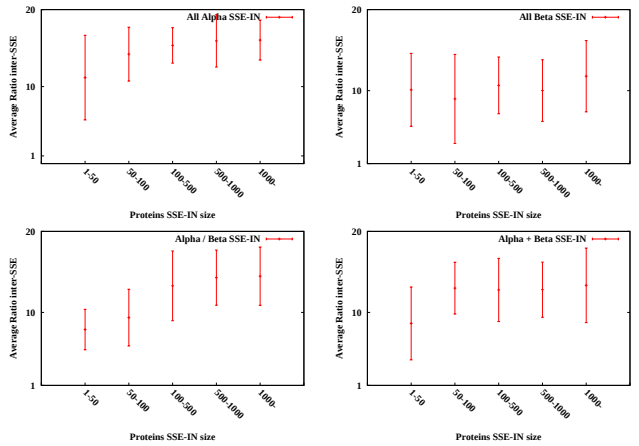

Figure 10. Ratio of inter-SSE edges (r) as a function of the network size for the four classes of studied proteins.

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.

In this paper we show that we can approximate all proteins SSE-IN cumulative degree distribution by a unique function. This function describes a power law regime followed by a sharp cut-off which arises for degree values exceeding the mean degree. Proteins SSE-IN are so truncated scale-free networks. This node distribution implies that there exist amino acids whose degree is marginal (greater than the mean degree).

A short term perspective is to study the hubs notably to understand how they appear in the folded protein. By this way, we would be able to understand if the nature of a hub depends on its position in the protein sequence or it depends on other chimical parameters.

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 studied [11], [12], 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.

REFERENCES

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

[2] L.A.N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci USA.*, 97(21), 2000.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

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

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

[11] O. Gaci and S. Balev. Hubs identification in amino acids interaction networks. In *Proceedings of the 7th ACS/IEEE International Conference on Computer Systems and Applications*, 2009. 7 pages.

[12] O. Gaci and S. Balev. The small-world model for amino acid interaction networks. In *Proceedings of the IEEE AINA 2009, workshop on Bioinformatics and Life Science Modeling and Computing*, 2009. 6 pages.

[13] 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.

[14] 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.

[15] 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.

[16] S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[17] 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.

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