Proteins: From Structural Classification to Amino Acid Interaction Networks.

Omar Gaci, Stefan Balev

Proteins: From Structural Classification to Amino Acid Interaction Networks

O. GACI
LITIS Laboratory
University of Le Havre
France

S. BALEV
LITIS Laboratory
University of Le Havre
France

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. Some of them are common to all proteins, while others depend on the structure arrangement. The last group of properties allows to characterize structural classes, defined by CATH or SCOP, in the terms of interaction network properties.*

Keywords: proteins, amino acid, interaction network, structural classification

1 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.

There exist different classifications of proteins according to their structure, such as CATH [15] and SCOP [12]. Proteins from the same class have similar structures and most often, similar functions. In this paper we show that structure classes can also be defined in the terms of the properties of amino acid networks.

2 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 world 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].

Based on the local organization of the secondary structure elements (SSE), proteins are divided in the following four classes [11]:

- All α, proteins have only α-helix secondary structure.
- All β, proteins have only β-strand secondary structure.
- α/β, proteins have mixed α-helix and β-strand secondary structure.
- $\alpha+\beta$, proteins have separated α-helix and β-strand secondary structure.

From this first division, a more detailed classification can be done. The most frequently used ones are SCOP, Structural Classification Of Proteins [12], and CATH, Class Architecture Topology Homology [15]. 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.

2.1 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 2007, the SCOP classification has identified 971 fold classes.

2.2 CATH

The CATH classification is maintained by both manual and automatic methods. Like SCOP, it is hierarchical and has 4 main levels and three additional levels concerning the similarity of protein sequences. The first 4 levels are the following :

1. *Class*: Proteins are grouped according to secondary structure composition and interaction between them. There are four classes: mainly α, mainly β, mixed α-β and all the rest.
2. *Architecture*: The secondary structure organization is the same.
3. *Topology*: Regroups structures whose foldings in terms of numbers, order and connexions of secondary structure are similar.
4. *Homologous surperfamily*: Domains which have structure and function very similar.

In 2007, CATH contained 1084 topology families.

3 Models and methods

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. 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. Fig. 1 gives an example of a protein and its SSE-IN. Note that the positions of the vertices in the graph do not correspond to the amino acid positions in the space. The graph is presented in this way only for visualization purposes.

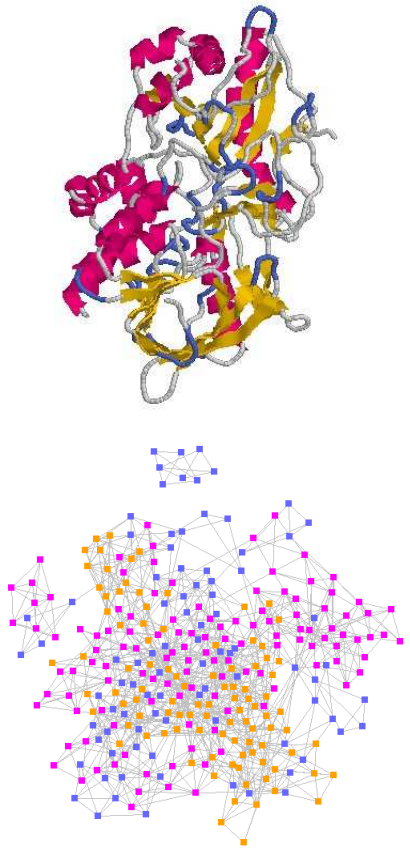

Fig. 1. Protein 1COY and its SSE-IN

The purpose of our work is to offer a graph theory interpretation of the hierarchical protein classifications. Consequently, when a protein belongs to a hierarchical level according to its biological properties then one can say also that the protein SSE-IN belongs to the same level. The SSE-IN is then characterized by graph theory properties to understand its behavior and the way is has formed. Thanks to this point of view, the protein folding problem can be tackled by the study of interaction networks.

4 Interaction networks

Many systems, both natural and artificial, can be represented by networks, that is, by sites or vertices bound by links [18]. 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 the 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 studing these networks [1]. Indeed, the study of the Web [16], of social networks [17] or of metabolic networks [10] 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 phenomena acting on their constitution and formation [20].

In this section we present some measures that we use to describe proteins' SSE-IN. Among these measures, there are simple ones, the most frequently used, but also more subtle, which allow a more precise discrimination between interaction networks.

4.1 Diameter and mean distance

The distance in a graph $G = (V, E)$ between two vertices $u, v \in V$, denoted by $d(u, v)$, is the length of the shortest path connecting u and v [7]. If there is no path between u and v, we suppose that $d(u, v)$ is undefined.

A graph diameter, D, is the longest shortest path between any two vertices of a graph [7]:

$$D = \max\{d(u, v) : \ u, v \in V\}$$

The mean distance is defined as the average distance between each couple of vertices:

$$\overline{d}_G = \frac{2}{n(n-1)} \sum_{u,v \in V} d(u, v)$$

4.2 Density and mean degree

A degree of a vertex u, k_u, is the number of edges incident to u. The mean degree, $\overline{k}_G$, of a graph G is definied as follows:

$$\overline{k}_G = \frac{1}{n} \sum_{u \in V} k_u = \frac{2m}{n}$$

The density, denoted δ, is defined as the ratio between the number of edges in a graph and the maximum number of edges which it could have:

$$\delta(G) = \frac{2m}{n(n-1)} \sim \frac{2m}{n^2}$$

The density of a graph is a number between 0 and 1. When the density is close to one, the graph is called dense, when it is close to zero, the graph is called sparse [6].

4.3 Degree distribution

If n_k is the number of vertices having degree k, then the degree distribution is given by the next formula:

$$p_k = \frac{n_k}{n}$$

The cumulative degree distribution [2, 8] is defined as follows:

$$P_k = \sum_{k'=k}^{\infty} p_{k'}$$

The power law distribution is defined as follows [14]:

$$P_k \sim \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-(\alpha-1)}$$

This distribution decreases in a polynomial way so that the number of vertices with weak degree is important, while a small number of vertices have high degree (see Fig. 2). The last are called "hubs", that is, sites which have a large connectivity through the network.

The degree distribution can also follow a Poisson law meaning that a small number of vertices have few links, a large number of vertices have a moderate number of incident edges, and a small number of vertices have a large number of incident edges (see Fig. 3). The Poisson distribution law is expressed as following:

$$p_k \sim e^{-\lambda} \frac{\lambda^k}{k!}$$

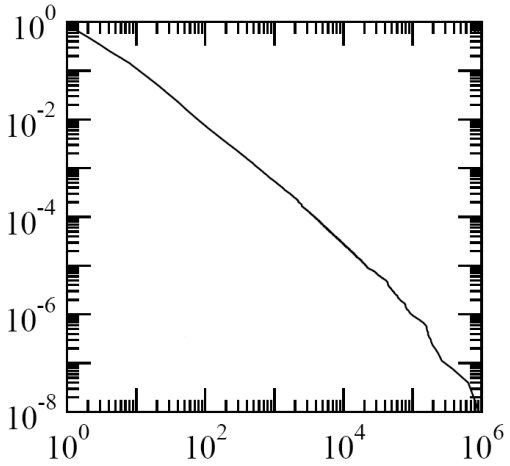

Fig. 2. Cumulative distribution following power law

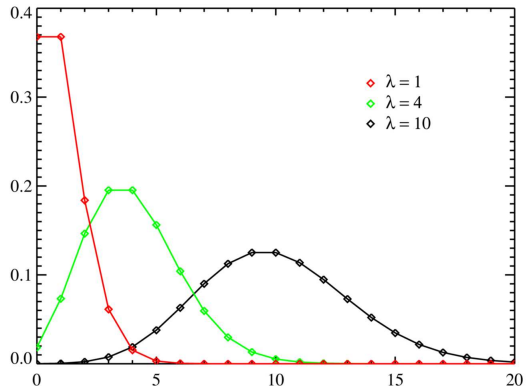

Fig. 3. Poisson distribution. Each curve has a peak close to $k = \lambda$ near the mean degree $\overline{k}_G$

The degree distribution analysis is an important characteristic of networks because it involves their internal organisation [2]. According to the kind of distribution followed, particularly if it's a power law, an interaction network can belong to a general model like scale-free model [3, 9] or small-world model [20].

4.4 Clustering coefficients

Watts and Strogatz proposed a measure of clustering [19] and defined it as a measure of local vertices density, thus for each node v, the local clustering around its neighbourhood is defined in the following way:

$$C_v = \frac{1}{2} k_v (k_v - 1)$$

The clustering coefficient is a ratio between the number of edges and the maximum number of possible edges in the vertice neighbourhood. If we extend the previous defintion to the entire graph, the clustering is given by the expression:

$$C_{\text{local}} = \frac{1}{n} \sum_{v \in V} \frac{\text{number of connected neighbour pairs}}{C_v}$$

The last defintion is mainly local because for each node, it involves only its neighbourhood.

The global clustering was studied by Newman et al. [13] and can be mesuared by the following formula:

$$C_{\text{global}} = \frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triplets of vertices}}$$

A triangle is formed by three vertices which are all connected and a triplet is constituted by three nodes and two edges. The global clustering coefficient C_{global} is the mean probability that two vertices that are neighbors of the same other vertex will themselves be neighbors.

5 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 CATH topology level or a SCOP fold level iff all its domains are the same. We have worked with the CATH v3.1.0 and SCOP 1.7.1 files. We have computed the measures from the previous section for three families of each hierarchical classification, namely SCOP and CATH (see Table 1). We have chosen these three families by classification, in particular because of their huge protein number. Thus, each family provides a broad sample guarantying more general results and avoiding fluctuations. Moreover, these six families contain proteins of very different sizes, varying from several dozens to several thousands amino acids in SSE.

Name	Type	Class	Proteins
Rossmann fold	CATH	$\alpha\ \beta$	2576
TIM Barrel	CATH	$\alpha\ \beta$	1051
Lysozyme	CATH	Mainly α	871
Globin-like	SCOP	All α	733
TIM β/α-barrel	SCOP	α/β	896
Lysozyme-like	SCOP	$\alpha+\beta$	819

Table 1. Families studied, mainly due to their protein number

5.1 Diameter and mean distance

Table 2 shows the average diameter for each one of the studied families. We observe very close diameters between *TIM Barrel* and *TIM beta/alpha-barrel* and also between *Lysozyme* and *Lysozyme-like* families. This is explained by the fact that each pair of families contains almost the same proteins, in other worlds, *Lysozyme* topology in CATH is the equivalent of *Lysozyme-like* fold level in SCOP.

Name	D
Rossmann fold	18.84
TIM Barrel	19.83
Lysozyme	12.81
Globin-like	15.65
TIM beta/alpha-barrel	20.09
Lysozyme-like	12.85

Table 2. Average diameter for each family

Figure 4 shows the distribution of the diameter values for two of the studied families. We observe that the distribution follows roughly a Poisson law. These results confirm that the mean diameter is a suitable property to discriminate families between them.

The diameter being an upper bound of distances in interaction networks, we expect that the mean distance $\overline{d}_G$ will be lower than D. Table 3 confirms this. Again, we observe very close values between the equivalent SCOP and CATH families for the reasons discussed above. But we can also see that different families have values which allow discrimination between them based on this parameter. It is interesting to note that the ratio $D/\overline{d}_G$ is about 2.5 for all the families. The last property is a characterization of all proteins' SSE-IN.

Name	$\overline{d_G}$
Rossmann fold	7.26
TIM Barrel	7.79
Lysozyme	4.99
Globin-like	6.64
TIM beta/alpha-barrel	7.86
Lysozyme-like	5.03

Table 3. Average of mean distances for each family

5.2 Density and mean degree

As defined earlier, the density measures the ratio between the number of available edges and the number of all possible edges. Results presented in Table 4 show that the two families *TIM Barrel* and *TIM beta/alpha-barrel* have the minimum density. It has a consequence on their SSE-IN topology. When the density is low, the network is less connected and consequently, the diameter and the average distance are higher. Comparing these results to Tables 1 and 2 one can see the inversely proportional relation between density in one hand, and diameter and average distance on the other.

Name	$\delta(G)$
Rossmann fold	0.033
TIM Barrel	0.030
Lysozyme	0.038
Globin-like	0.034
TIM beta/alpha-barrel	0.029
Lysozyme-like	0.042

Table 4. Average density for each family

The mean degree, $\overline{k}_G$ is presented in Table 5. The observed values are close enough from one family to another. That is why the mean degree is not discriminating property, but rather a property characterizing all proteins' SSE-IN.

Name	$\overline{k}_G$
Rossmann fold	7.20
TIM Barrel	7.17
Lysozyme	6.82
Globin-like	7.69
TIM beta/alpha-barrel	7.15
Lysozyme-like	6.81

Table 5. Average of mean degrees for each family

5.3 Degree distribution

We compute the cumulative degree distribution for all proteins SSE-IN of studied families. A sample of our results is presented on Figure 5. We can remark that the curves follow a power law distribution and can be approximated by the following power-law function:

$$p(k) = 141.29k^{-\alpha}, \text{ where } \alpha = 2.99 \pm 0.6$$

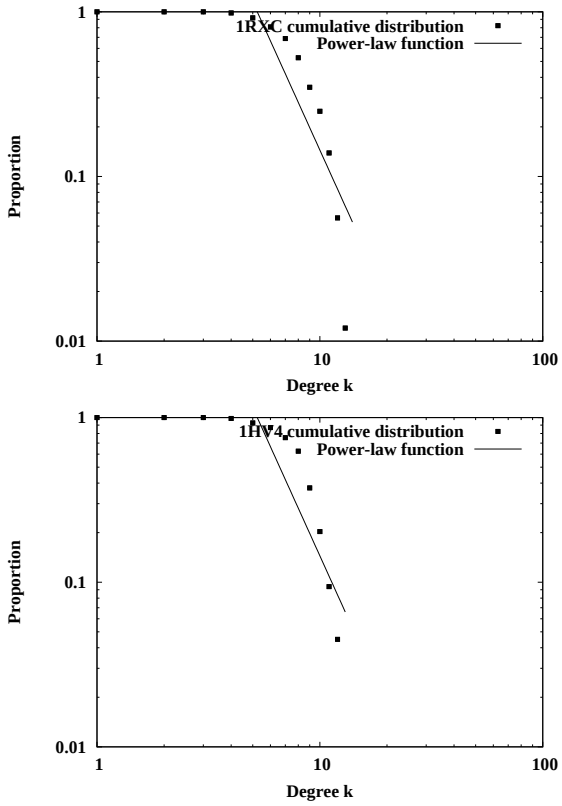

Fig. 5. Cumulative degree distribution for 1RXC from Rossman fold, top, and 1HV4 from TIM beta/alpha-barrel, bottom.

We observe the same results for all studied proteins. To explain this phenomenon, we have to rely on two facts. First, the mean degree of all proteins SSE-IN is nearly constant (see Table 5). Second, the degree distribution, see Figure 6, follows a Poisson distribution whose peak is reached for a degree near $\overline{k}_G$. These two facts imply that for degree lower than the peak the cumulative degree distribution decreases slowly and after the peak its decrease is fast compared to an exponential one. Consequently, all proteins SSE-IN studied have a similar cumulative degree distribution which can be approximated by a unique power-law function.

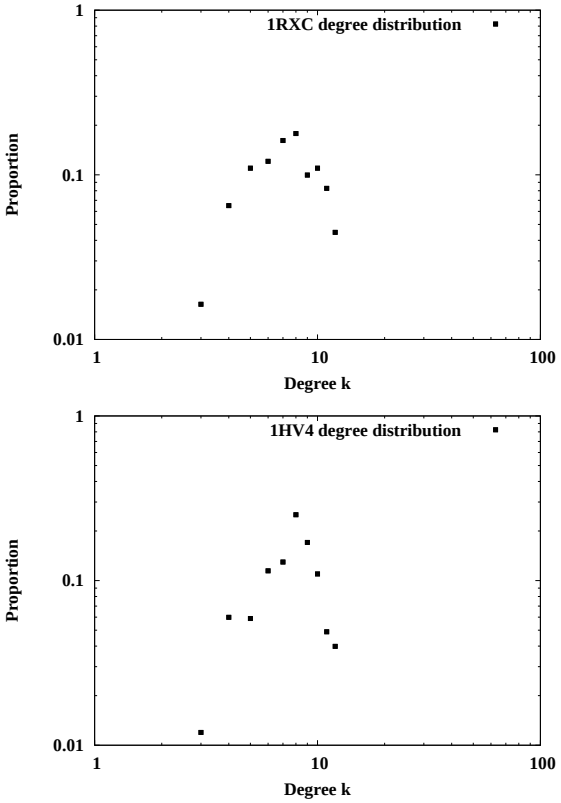

Fig. 6. Degree distribution for 1RXC from Rossman fold, top, and 1HV4 from TIM beta/alpha-barrel, bottom.

5.4 Clustering coefficients

The local clustering C_{local} measures the fraction of pairs of a vertex's neighbors and the global clustering C_{global} gives the probability that among three vertices at least two are connected. The results presented in Table 6 show that the clustering coefficients are close for different families and cannot be correlated to density values. Consequently, the neighbour density remains independent of the previously studied properties.

Name	C_{local}	C_{global}
Rossmann fold	0.63	0.56
TIM Barrel	0.64	0.57
Lysozyme	0.65	0.58
Globin-like	0.63	0.57
TIM beta/alpha-barrel	0.64	0.57
Lysozyme-like	0.66	0.58

Table 6. Clustering coefficients for each family

6 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. We give different means to describe a protein structural family by characterizing their SSE-IN. Some of the properties, like diameter and density, allow to discriminate two distinct families, while others, like mean degree and power law degree distribution, are general properties of all SSE-IN. Thus, proteins having similar structural properties and biological functions will also have similar SSE-IN properties. In this way our model allows us to draw a parallel between biology and graph theory.

The characterization we propose constitutes a first step of a new approach to the protein folding problem. The properties we identified, both general and specific, 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.

Another perspective is to study more deeply the general properties of SSE-IN, in particular degree distribution, and associate them to more general models, such as scale-free or small-world networks, whose behavior and evolution are well known.

References

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

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

[3] A.-L Barabási and R. Albert. Emergence of scaling in random networks. *Science.*, 286:509–512, 1999.

[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] T. F. Coleman and J. J. Moré. Estimation of sparse jacobian matrices and graph coloring problems. *SIAM Journal on Numerical Analysis.*, 20:187–209, 1983.

[7] R. Diestel. *Graph Theory*. Springer Verlag, Princeton, New Jersey, 2000.

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

[9] K-.I. Goh, Kahng B., and D. Kim. Universal behavior of load distribution in scale-free networks. *Phys. Rev.*, 87, 2001.

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

[11] M. Levitt and C. Chothia. Structural patterns in globular proteins. *Nature.*, 261:552–558, 1976.

[12] 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.

[13] M. E. J. Newman. The strucuture of scientific collaboration networks. *Proc. Natl. Acad. Sci USA.*, 98:404–409, 2001.

[14] M. E. J. Newman. The strucuture and function of networks. *Computer Physics Communications.*, 147:40–45, 2002.

[15] 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.

[16] R. Albert and H. Jeong and A.L. Barabási. The Diameter of the World Wide Web. *Nature*, 401:130–131, 1999.

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

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

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

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