Network Analysis of the Protein Chain Tertiary Structures of Heterocomplexes+

Jing-Jing Li[1], De-Shuang Huang[1,*], Tat-Ming Lok[2], Michael R. Lyu[3], Yi-Xue Li[4] and Yun-Ping Zhu[5]

[1]*Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, 230031, P. R. China;* [2]*Information Engineering Dept., The Chinese University of Hong Kong, Shatin, Hong Kong;* [3]*Computer Science & Engineering Dept, The Chinese University of Hong Kong, Shatin, Hong Kong;* [4]*Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Science, Shanghai, 200031, P. R. China;* [5]*Beijing Institute of Radiation Medicine, Taiping Road 27, Beijing 100850,China*

Abstract: In this paper, the tertiary structures of protein chains of heterocomplexes were mapped to 2D networks; based on the mapping approach, statistical properties of these networks were systematically studied. Firstly, our experimental results confirmed that the networks derived from protein structures possess small-world properties. Secondly, an interesting relationship between network average degree and the network size was discovered, which was quantified as an empirical function enabling us to estimate the number of residue contacts of the protein chains accurately. Thirdly, by analyzing the average clustering coefficient for nodes having the same degree in the network, it was found that the architectures of the networks and protein structures analyzed are hierarchically organized. Finally, network motifs were detected in the networks which are believed to determine the family or superfamily the networks belong to. The study of protein structures with the new perspective might shed some light on understanding the underlying laws of evolution, function and structures of proteins, and therefore would be complementary to other currently existing methods.

1. INTRODUCTION

Complex systems, such as the World Wide Web (WWW), social insect colonies, telecommunication systems and protein-protein interaction networks, can be modeled as complex networks, or mathematical graphs [1,2], in which each component is denoted as a vertex and the relationships between two components are represented as edges connecting the corresponding vertices. Real-world networks typically exhibit small-world properties [3], by which is meant that a high degree of local clustering is observed; also, the average shortest path length between any two vertices scales logarithmically with the network size, which cannot be observed in either regular lattices or random graphs [1,4].

Protein structures are complex systems *per se*, with several tens, hundreds or even thousands of residues, interacting with each other to help stabilize the tertiary structures so that specific functions can be realized *in vivo*. In this sense, the network modeling approach is suitable for characterizing and analyzing protein structures, in which residues correspond to vertices of the networks, and interaction (or any other type of relationship) between residues is represented as an edge linking the corresponding nodes. Vendruscolo *et al.* [5] constructed networks for a set of protein structures; through investigating the value of *betweenness* for each node of the networks, some "key nodes", acting as nucleation centers for protein folding, were identified. Atigan *et al.* [6] discovered that the average shortest path lengths are strongly correlated with residue fluctuations. Dokholyan *et al.* [7] found that the network's topological properties are crucial for the protein to have the kinetic ability to fold. Recently, based on a set of enzyme chains, Amitai *et al.* [8] reported that, by using the closeness value of each node in the network, it could be predicted which residues in the protein structure are functional. Besides, protein structure flexibility [9], recurring structural patterns [10] and side-chain clusters [11] can all be studied using the graph representation approach.

In the following discussion, we will present our initial investigation of a set of non-redundant protein chains of heterocomplexes using the network modeling approach. Through investigating the relationship between network average degree and the network size, a function can be summarized to estimate the number of residue contacts of the protein chains effectively. By analyzing the average clustering coefficient for the nodes having the same degree in the network, we can show that the architectures of the networks and protein structures analyzed are hierarchically organized. Finally, network motifs have been detected in the networks; these are believed to determine the general family or superfamily the networks belong to.

This remainder of this paper is organized as follows. Section 2 gives an overall introduction of the data we made use of and some crucial definitions. In Section 3, the experimental results are demonstrated to evaluate our approach. Finally, some concluding remarks and directions for future research are included in Section 4.

*Address correspondence to this author at the Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, 230031, P. R. China; Email: dshuang@iim.ac.cn

+*This work was supported by the National Science Foundation of China (Nos. 60472111 and 30570368) and Chinese Human Liver Proteome Project (N. 2004BA711A21).*

0929-8665/06 $50.00+.00

2. METHOD

2.1 Data Collection

The data analyzed in our experiment is originally from P. Aloy and R. B. Russell [12]. Problematic chains and chains sharing high homologies are eliminated from the data they collected. After this filtering procedure, 424 non-redundant protein chains were retained; each chain shares less than 30% sequence identity with other chains in the data set. The structures of these chains were obtained from the Protein Data Bank (PDB) [13]. A list of the protein chains we studied is available at http://www.dmresearch.net/hetero_list.htm.

2.2 Representation of Protein Tertiary Structures

In the experiment, a reduced representation of protein structures is employed, namely, each residue is represented by its C-α atom; thus in three-dimensional space, the structures of protein chains are represented simply as the connectivity of the C-α atoms of each residue [14]. In addition, we define two residues to be in contact if the distance between their C-α atoms is less than a threshold [15], for which the value of 8 ångstroms is adopted.

2.3 Definitions

To convert the tertiary structures into networks, we first define the edges of the networks to be the contacts of residues (more precisely, the contacts of their C-α atoms); in other words, if two residues are in contact in the tertiary structure, an edge connecting the two corresponding nodes will appear in the network.

Complex networks are essentially mathematical graphs; in mathematical terms, the residue contact network built in our experiment is a graph with a vertex set V of residues and an edge set $E = \{(x, y) | (x, y \in V) \cap (0 < \|x-y\| \leq 8)\}$, where $\| \ \|$ is the Euclidean distance (l norm) between the residues x and y in the three-dimensional structure of the protein, meaning that two nodes in the graph are adjacent if the distance of two residues is less than or equal to 8 ångstrom. Note that the graphs mentioned are naturally unweighted, which is equivalent to imposing a weight of 1 on all the edges and nodes.

One of the protein chains in our dataset is shown below to clarify the conversion procedure. Fig. **1** presents the tertiary structure of the B chain of 1a0h; its reduced representation is shown in Fig. **2**, and the network derived from the simplified structure is exhibited in Fig. **3**.

2.4 Network Properties

Supposing that there are a total of N vertices in one network, the network is equivalent to a two-dimensional array A with the size of N^2N; each element A_{ij} in the array is defined as follows:

$$A_{ij} = \begin{cases} 1, \text{ if node } i \text{ is connected with node } j \\ 0, \text{ otherwise} \end{cases} \quad \text{Eq. (1)}$$

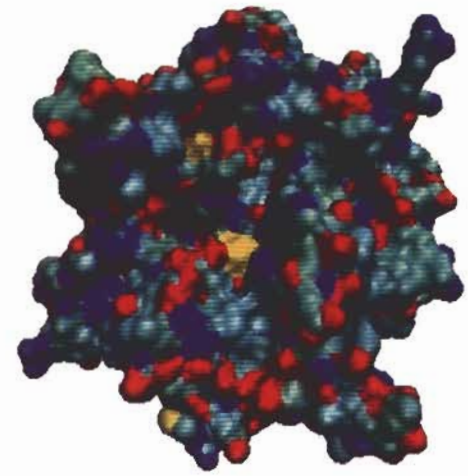

Figure 1. The three-dimensional structure of chain B of the protein (PDB code: 1a0h); the colors represent different amino acid residues. The figure was generated by VMD [16].

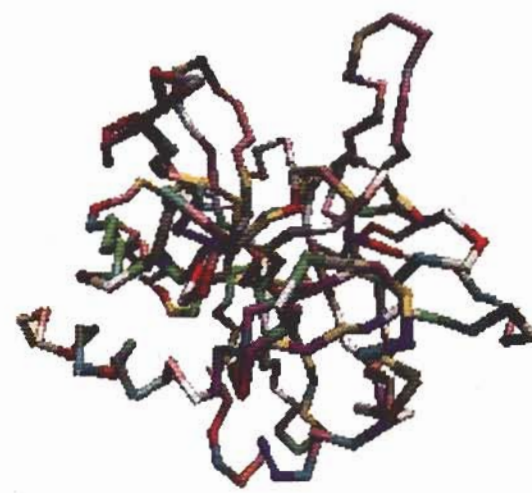

Figure 2. The reduced representation of the same chain (C-α trace); the colors represent different amino acid residues. The figure was generated by VMD [16].

Figure 3. The residue contact network derived from the tertiary structure of the same chain. The figure was generated by Ospray [17].

Traditionally, this 2D array A is known as the adjacency matrix in Graph Theory.

Networks are typically quantified by several numerical measures so that comparisons and analyses can be made directly; all the required properties can be derived from the adjacency matrix.

Some important properties used in this paper are defined as follows:

(1) The degree K of a vertex i is the number of nodes that are directly connected to i; thus,

$$K_i = \sum_{j=1}^{N} A_{ij} \quad \text{Eq. (2)}$$

(2) The average degree of the network can be written as:

$$K = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} \quad \text{Eq. (3)}$$

which is the average over all the degree of each node in the network.

(3) The degree distribution $P_i(k)$ of the network refers to the probability that the vertex i has degree k. Degree distribution is a widely used criterion to classify a specific network into some category; for example, the degree distributions of typical random networks [18,19] and small-world networks (SWN) [3] have a Poisson distribution, while the degree distribution of scale-free networks [20] usually follows the power law, namely, $P_i(k)\ k^{-\gamma}$,where $\gamma$is a positive number.

(4) The clustering coefficient C [3] of a vertex i is the value measuring the probability that two nodes of the network are adjacent if they have a common neighbor i. Thus,

$$C_i = \frac{\frac{1}{2} \sum_{m=1}^{N} \sum_{n=1}^{N} A_{im} A_{in} A_{mn}}{\left\{ \sum_{j=1}^{N} A_{ij} \right\}^{2}} \quad \text{Eq. (4)}$$

(5) Similarly, the average clustering coefficient C of the network is defined as:

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i \quad \text{Eq. (5)}$$

(6) The characteristic path length of a network is the average minimum number of connections that must be traversed to link any residue pair i and j. It is usually considered a useful measure to reflect the efficiency of information spread and communication among vertices in a network.

(7) Network motifs are patterns that are over-represented in a network. As R. Milo's pivotal work [21,22] shows, they are the basic building blocks of complex networks, and thus define the universal family or superfamily of networks.

3. EXPERIMENTAL RESULTS

(1) General Properties of the Networks

We first confirmed that the residue contact networks derived from the chains of heterocomplexes possess small-world properties. The degree distribution is shown in Fig. **4**, from which it is evident that the metric exhibits a Poisson distribution.

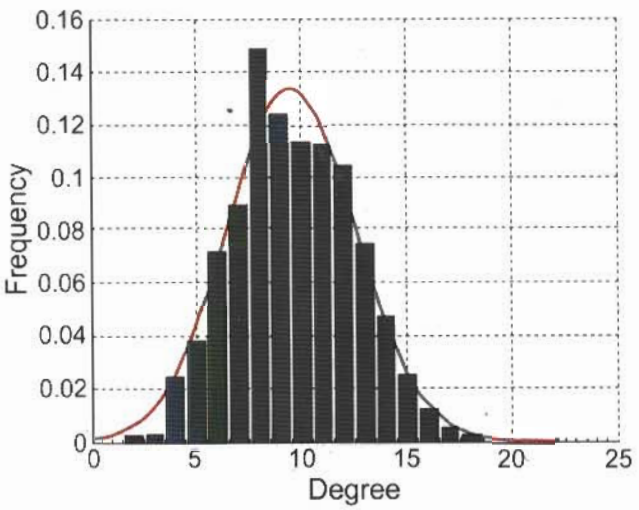

Figure 4. Degree distribution of the residue contact networks.

The characteristic path length calculated for the networks is 4.787±1.537, and the average clustering coefficient of the networks is 0.5726±0.038. To prove that the networks possess small-world properties, we can compare them with those derived from regular lattices and random graphs [5]. For random graphs whose size N and average degree K are identical with the ones we analyzed, the characteristic path length and clustering coefficient are 2.31±0.277 and 0.07±0.05598 respectively; as for regular lattices, the characteristic path length and clustering coefficient are 12.1±8.725 and 0.6549±0.026 respectively. This strongly suggests that the model of the networks in our experiments is a compromise between the regular lattice model and the random graph model, with intermediate values for the characteristic path length and clustering coefficient; thus, as described by Vendruscolo *et al.* [5] and Atilan *et al.* [6], the relatively high clustering coefficient compared with random graphs and low characteristic path length compared with regular lattices indicate that the networks derived from the chains of heterocomplexes indeed have small-world network properties.

(2) Average Degree Versus Network Size

The relationship between the average degree and the size of the proteins in our dataset was studied. Interestingly, we discovered that the increase in the average degree is quite regular with respect to the accretion of the network size (which is equivalent to the size of the protein). The relation-

ship is shown in Fig. **5**. In order to quantify the relationship, a curve fitting technique was used to derive a function $f(x) = 9.113e^{0.0002475x} - 7.288e^{-0.02674x}$ for estimating the average degree of a network from its size (denoted by ^{x}).

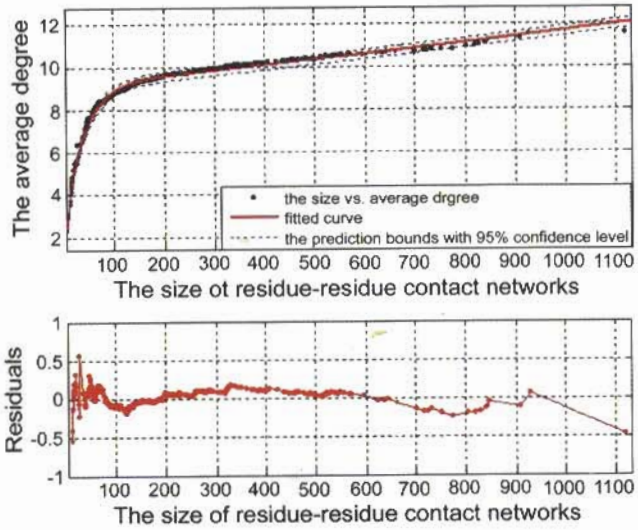

Figure 5. Average degree vs. network size. The fitted curve and the prediction bounds are also shown in the upper plot, and the fitting residuals are presented in the lower plot.

We note that the definition of average degree given in equation (3) is exactly equivalent to the following:

$$K = \frac{2|E|}{|V|} \quad \text{Eq. (6)}$$

where E, V are the sets of edges and vertices respectively; $|E|$ and $|V|$ are the numbers of edges and vertices in the network. Since the average degree can be estimated by $f(x)$, as we mentioned above, the following formula is always true (note here) $x = |V|$:

$$K = \frac{2|E|}{|V|} = f(x) \quad \text{Eq. (7)}$$

Therefore, the number of the edges of the network can be calculated using the following formula:

$$|E| = \frac{1}{2}xf(x) = \frac{1}{2}x(9.113e^{0.0002475x} - 7.288e^{-0.02674x}) \quad \text{Eq. (8)}$$

According to the definition of network we stated earlier, the number of edges of a network is equal to the number of residue contacts in the protein tertiary structures. Therefore, the number of residue contacts can be estimated accurately from the protein size (number of residues) alone.

The following figure (Fig. **6**) presents the estimated results using this approach:

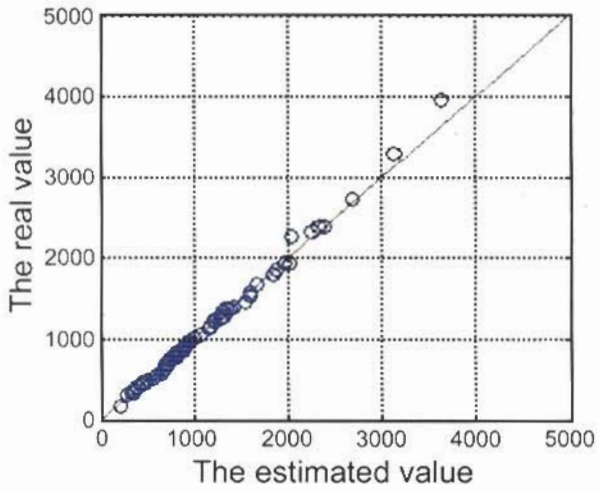

Figure 6. Correspondence between estimated and actual residue contact numbers.

(3) Average Clustering Coefficient for Vertices Having the Same Degree

The average clustering coefficient $C(k)$ over all the nodes having degree k was also calculated. The results are presented in the following figure (Fig. **7**).

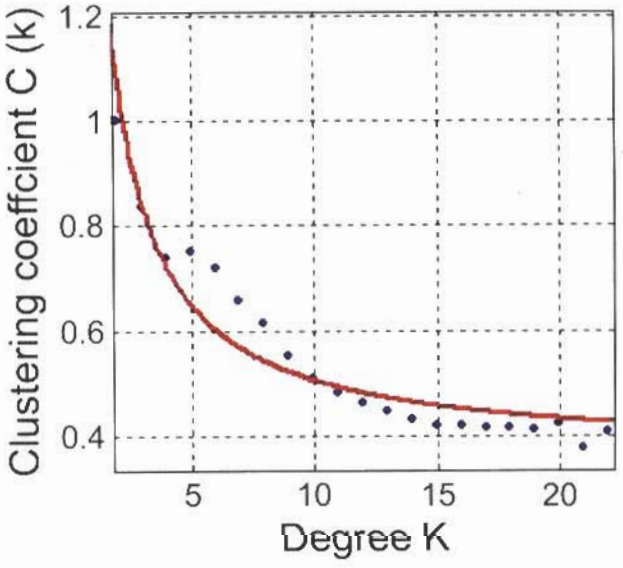

Figure 7. The average clustering coefficient for nodes having the same degree (the red line in the figure is a fitted curve based on the calculated results).

As has frequently been pointed out [23-25], if it is found that $C(k)$ is proportional to $\frac{1}{k}$ in a network, it typically indicates that the network is highly hierarchically organized. Fig. **7** shows that our network exhibits this property, indicating that the residue contact networks in our experiment possess hierarchical architectures, which, in turn, could explain why

the clustering coefficients of the networks are much higher than those of random graphs.

Furthermore, the hierarchical architectures of the networks imply that the networks are, in general, composed of several physical modules [26], or clusters, within which vertices are intensively connected, while nodes belonging to different modules are sparsely linked. A possible reason for the protein networks exhibiting this architecture is that the modules contain many functional related residues, which perform the same function simultaneously; the intense connection among the residues in a module would be likely to promote the efficiency of their communication.

(4) Network Motif Detection

Network motifs play an important role in the analysis of networks, since they are believed to be the simple building blocks of networks, and their character defines the general categories (families or superfamilies) networks belong to. Hence, identifying the prevailing network motifs can greatly facilitate the comparison of the evolution, structure, organization and function of different networks falling into the same family or superfamily.

Essentially, motifs are over-represented graphlets of a network compared with that in random graphs. In our analysis, we have considered 3-node or 4-node graphlets only; graphlets with more than 4 nodes are not taken into account in this paper because of the unfeasible computational task they represent. To discover the motifs, as shown by R. Milo *et al.*, numerous random graphs with key properties very similar to the those of real network are generated. In order to ensure the comparison is fair, only those recurring patterns with a Z-score greater than 2 are selected as motifs of the network.

In our analysis of the network derived from protein chains of heterocomplexes, the program *Mfinder* (provided by R. Milo *et al.* [11]) was implemented. The motifs detected are as follows:

Three-node motif:

Four-node motifs:

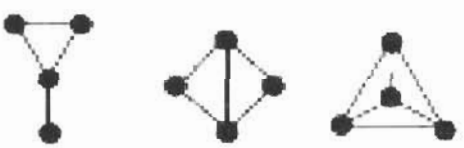

Figure 8. Motifs of the networks.

In total, there are 8 patterns for 3 and 4-node graphlets in an undirected graph. It is remarkable that only the four kinds of graphlets shown above occurred frequently. Considering the fact that the clustering coefficient (which reflects the possibility that two nodes are adjacent if they have a common neighbor) is relatively high, as we mentioned earlier, it is no surprise that the 3-node motif are triangular. As for the 4-node motifs, one conclusion that could be drawn is that, in the three-dimensional space, the structures of the protein chains we analyzed are naturally tightly packed, so that each residue would naturally be in contact with many other residues in its neighborhood.

4. DISCUSSION AND CONCLUDING REMARKS

We have presented our initial work on analyzing the structure of complex networks derived from the protein chain tertiary structures of heterocomplexes. From this perspective, some novel network characteristics have been revealed.

Networks, or mathematical graphs, are undoubtedly an effective approach for analyzing complex systems, not just the protein structures we considered here. Using this approach, various intrinsic properties of the system can be discerned, greatly facilitating our understanding of the networks of interest. Previous studies have shown that the characteristic path length is an effective parameter for investigating the underlying principles of protein structures [6-8]. However, we cannot ignore the other properties of the network if we hope to understand the protein structure more deeply.

In future research work, further attention could be given to the quantification of the derived networks, in order to reflect the intrinsic character of the protein structures more completely.

ACKNOWLEDGEMENTS

We are grateful to Dr. Zhi Liang of University of Science and Technology of China (USTC) for helpful discussion of the network motifs, Dr. Piero Fariselli of University of Bologna via Irnerio, Italy and Dr. Michael Tress of the Protein Design Group (PDG), Centro Nacional de Biotecnologia, Spain for their generous provision of the experimental data and their valuable suggestions about the data; Miss Xiao Gu of Intelligent Computation Lab, Institute of Intelligent Machines (IIM), Chinese Academy of Sciences (CAS) for solving some coding problems; and Jun Shi of IIM/CAS for providing us with the host space to store all the information about the experimental data.

REFERENCES

[1] Strogatz, S. H. (**2001**). Exploring complex networks. *Nature, 410*, 268–276.

[2] Albert, R. and Barabasi, A.-L. (**2002**). Statistical mechanics of complex networks. *Rev. Mod. Phys., 74*, 47–97.

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

[4] Newman, M. E. J. (**2000**). Models of the small world. *J. Stat. Phys., 101*, 819-841.

[5] Vendruscolo, M., Dokholyan, N. V., Paci, E. and Karplus, M. (**2002**). Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E, 65*, 061910

[6] Atigan, A. R., Akan, P. and Baysal. C. (**2004**). Small-world communication of the residues and significance for protein dynamics. *Biophys. J., 86*, 85-91.

[7] Dokholyan, N. V., Li, L., Ding, F. and Shakhnovich, E.I. (**2002**). Topological determinants of protein folding. *Proc. Nat. Acad. Sci., 99*, 8637-8641.

[8] Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Venger, D. N. I. and Pietrokovski, S. (**2004**). Network analysis of protein structures identifies functional residues. *J. Mol. Biol., 344*, 1135-1146.

[9] Jacobs, D. J., Rader, A. J., Kuhn, L. A. and Thorpe, M. F. (**2001**). Protein flexibility predictions using graph theory. *Proteins: Struct. Funct. Genet., 44*, 150–165.

[10] Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N. and Sarawagi, S. (**2003**). Functional sites in protein families uncovered

via an objective and automated graph theoretic approach. *J. Mol. Biol., 326*, 955–978.

[11] Kannan, N. & Vishveshwara, S. (**1999**). Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol. 292*, 441–464.

[12] Aloy, P and Russell, R. B. (**2002**). Interrogating protein-interaction networks through structural biology. *Proc. Nat. Acad. Sci. USA, 99*, 5896-5901.

[13] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (**2000**). The protein data bank. *Nucleic Acids Res., 28*, 235-242.

[14] Fariselli, P., Pazos, F., Valencia, A. and Casadia, R. (**2002**). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem., 269*, 1356-1361.

[15] Fariselli, P., Olmea, O., Valencia, A. and Casadio, R. (**2001**). Prediction of contact maps with neural networks and correlated mutations. *Protein Eng., 14*, 835-843.

[16] Humphrey, W., Dalke, A. and Schulten, K. (**1996**). VMD - Visual Molecular Dynamics. *J. Molec. Graphics, 14*, 33-38.

[17] Breitkreutz, B. J., Stark, C. and Tyers, M. (**2003**). Osprey: A Network Visualization System. *Genome Biol., 4(3)*:R22.

[18] Erdos, P. and Renyi, A. (**1959**). On random graphs. *Publicationes Mathematicae, 6*, 290–297.

[19] Erdos, P. and Renyi, A. (**1960**). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci., 5*, 17–61.

[20] Barabasi, A.-L. and Albert, R. (**1999**). Emergence of scaling in random networks. *Science, 286 (5439)*, 509–512.

[21] Milo, R., Shen-Orr, S. S., Itzkovitz, S., Kashtan, N.,Chklovskii, D. and Alon, U. (**2002**). Network motifs: simple building blocks of complex networks. *Science, 298*, 824–827.

[22] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U. (**2004**). Superfamilies of evolved and designed networks. *Science, 303*, 1538–1542.

[23] Ravasz, E. and Barabasi, A.L. (**2003**). Hierarchical organization in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 67*, 026112.

[24] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A.-L. (**2002**). Hierarchical organization of modularity in metabolic networks. *Science, 297*, 1551–1555.

[25] Barabasi A.–L., Dezso, Z., Ravasz, E., Yook, Z.–H. and Oltvai, Z. N. (2004). Scale-free and hierarchical structures in complex networks. *Stiges Proceedings on complex networks*.Spirin, V. and Mirny, L. A. (**2003**). Protein complexes and functional modules in molecular networks. *Proc. Nat. Acad. Sci. USA, 100*, 12123-12128.

Received: August 20, 2005 Revised: October 06, 2005 Accepted: October 10, 2005