View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Strathclyde Institutional Repository

Strathprints Institutional Repository

Estrada, E. and , New Professor's Fund - University of Strathclyde (Funder) (2010) *Universality in protein residue networks.* Biophysical Journal, 98 (5). pp. 890-900. ISSN 0006-3495

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (http://strathprints.strath.ac.uk/) and the content of this paper for research or study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to Strathprints administrator: mailto:strathprints@strath.ac.uk

http://strathprints.strath.ac.uk/

Estrada, E. (2010) Universality in protein residue networks. Biophysical Journal . ISSN 0006-3495 (In Press)

http://strathprints.strath.ac.uk/14358/

This is an author produced version of a paper published in Biophysical Journal . ISSN 0006-3495. This version has been peer-reviewed but does not include the final publisher proof corrections, published layout or pagination.

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (http://strathprints.strath.ac.uk) and the content of this paper for research or study, educational, or not-for-profit purposes without prior permission or charge. You may freely distribute the url (http://strathprints.strath.ac.uk) of the Strathprints website.

Any correspondence concerning this service should be sent to The Strathprints Administrator: eprints@cis.strath.ac.uk

Universality in Protein Residue Networks

Ernesto Estrada[1]

Department of Mathematics and Statistics, Department of Physics and Institute of Complex Systems, University of Strathclyde, Glasgow G11XQ, UK

ABSTRACT Residue networks representing 595 nonhomologous proteins are studied. These networks exhibit universal topological characteristics as they belong to the topological class of modular networks formed by several highly interconnected clusters separated by topological cavities. There are some networks which tend to deviate from this universality. These networks represent small-size proteins having less than 200 residues. We explain such differences in terms of the domain structure of these proteins. On the other hand, we find that the topological cavities characterizing proteins residue networks match very well with protein binding sites. We then investigate the effect of the cutoff value used in building the residue network. For small cutoff values, less than 5Å, the cavities found are very large corresponding almost to the whole protein surface. On the contrary, for large cutoff value, more than 10.0 Å, only very large cavities are detected and the networks look very homogeneous. These findings are useful for practical purposes as well as for identifying "protein-like" complex networks. Finally, we show that the main topological class of residue networks is not reproduced by random networks growing according to Erdös-Rényi model or the preferential attachment method of Barabási-Albert. However, the Watts-Strogatz (WS) model reproduces very well the topological class as well as other topological properties of residue network. We propose here a more biologically appealing modification of the WS model to describe residue networks.

[1] E-mail: Ernesto.estrada@strath.ac.uk

INTRODUCTION

Complex networks are ubiquitous in many fields of science, studying systems which range from biology to social sciences (1-3). In a complex network the elements of the system are represented by nodes and the interactions between these elements are represented by links. Several charcateristic features have been observed in these systems, such as "small-worldness" (4), "scale-freeness" (5), fractality (6), characteristic motifs (7) and mixing patterns (8).

These concepts of network theory have been applied to the study of protein structure with promising results (9). In a seminal paper Vendruscolo et al. (10) constructed networks corresponding to protein structures and showed by the first time that they display "small-world" features. The nodes corresponding to key residues were observed to play the role of "hubs" in the network of interactions stabilizing the structure of the transition states. In a further work Atilgan et al. (11) proposed to represent proteins as residue networks, showing that residues in folded proteins display "small-world" topology. In these networks the nodes represent C_α or C_β atoms of the amino acids and two nodes are connected if they are separated at a geometric distance lower than certain cutoff value. These authors have found that their results are independent of the use of different cutoff values (11). In another work, Bagler and Sinha (12) used the same representation of Atilgan et al. (11) and found that proteins, regardless of their structure class, show small-world properties. They also gave some insights about the modular structure of proteins by using various network parameters. On the other hand, Brinda and Vishveshwara (13) used a parametrized measure of the strength of interaction to decide whether two nodes representing protein residues are connected in the network. They observed that these networks exhibit a complex degree distribution with combinations of Gaussian-like, sigmoidal, and exponential/power-law decay for different interaction cutoffs (13). In a variation of these representation methods, Greene and Higman (14) considered each amino acid as a node, and the links were established between two nodes, if any two atoms of the amino acids are separated at less than 5 Å of each other. They found again that these networks exhibit small-world, single-scale, and at some degree, scale-free properties.

Concerning the representation of proteins as networks da Silveira et al. (15) have carried out a comparative analysis between the cutoff-dependent and cutoff-free methods used to represent protein networks. They have found that the cutoff value of 7 Å "emerges as an important distance parameter", because at this distance "all contacts are complete and legitimate (not occluded)". Consequently, they concluded that in the strict range up to 7 Å, the cutoff-dependent approach "revealed to be simpler, more complete, and reliable technique" than the other approaches in use. Also importantly, they have shown that representing amino acids by C_α atoms may introduce bias for cutoffs below 6.8 Å, which supports the use of the cutoff value of 7 Å.

On the other hand, Bartoli et al. (16) have analyzed 1753 non-redundant protein structures and have shown that the small-world behaviour of inter-residue contact graphs is conditioned by the backbone connectivity. They have concluded that the characteristic path length L and clustering coefficient C in which the small-world concept is based, are not useful quantities for protein fingerprinting. Bartoli et al. (16) have used a randomization procedure from which L and C are indistinguishable from those of real proteins. They have verified that these results are independent of the selected protein representations, residue composition and protein secondary structures.

The main objective of the current work is to explore beyond general statistical properties of complex protein networks, such as small-worldness or scale-freeness. In fact, we show here that proteins represented as residue networks belong to a universal topological class, which can be unambiguously characterized from a mathematical point of view. This universal class corresponds to modular networks in which some highly interconnected nodes are separated from each other by topological cavities. We show that these cavities are potentially binding sites in proteins and that the domain structure of proteins plays an important role in this structural organization.

METHODS

Proteins as residue networks

Spatial residue networks are built by using the Cartesian coordinates of the protein reported in the protein data bank (PDB) (17). We use here the dataset prepared by Atilgan et al. (11), who studied 595 proteins with less than 25% of sequence homology. This dataset was selected previously by Fariselli and Casadio (18). We have obtained from Atilgan et al. (11) the dataset of these 595 proteins represented as residue networks.

The nodes of these residue networks represent the amino acids of the protein, centred at their C_β atoms, with the exception of glycine for which C_α is used. In order to connect the nodes of the network it is considered a cutoff radius r_C, which represents an upper limit for the separation between two residues in contact. The distance between two residues is measured by taking the distance between C_β atoms of both residues. Then, when the inter-residue distance is equal or less than r_C both residues are considered to be interacting. In this case, the corresponding nodes in the residue network are connected.

The elements of adjacency matrix of the residue network are obtained by

$$A_{ij} = \begin{cases} H(r_C - r_{ij}) & i \neq j \\ 0 & i = j \end{cases}$$

where $H(x > 0) = 1$ and $H(x \leq 0) = 0$. Then, a protein is represented by the graph $G = (V, E)$, where V represents the set of amino acids and E represents the set of interactions between then. The residue networks studied here were constructed by using a cutoff of 7 Å. Atilgan et al. have verified that the general conclusions concerning the topology of these networks are not affected when a cutoff value of 8.5 Å is used instead (11). As a matter of example, the residue network of the protein with PDB code 1ash is represented in Fig. 1.

Insert Fig. 1 about here.

Topological classification of networks

Let us consider a subset of amino acids $S \subseteq V$. The number of interactions between an amino acid in this subset S with another amino acid which is not in this set, is named the boundary of S and it is denoted by $|\partial S|$. We select the subset of amino acids S to be at most half the number of residues in the protein. Then, we can use the following measure to account for how efficiently connected the residue network is (19)

$$\phi(G)=\inf\left\{\frac{|\partial S|}{|S|}, S\subseteq V, 0<|S|\leq\frac{|V|}{2}<+\infty\right\}, \quad (1)$$

which is known as the *expansion or isoperimetric constant* of the network.

The higher the expansion constant $\phi(G)$ the better the connectivity of the network (20), which means that the number of interactions that must be removed to separate the protein into isolated parts is relatively high in comparison with the number of amino acids in the protein. These networks are said to have good expansion (GE) properties. It is known that a network has GE if the gap between the first and second eigenvalues of the adjacency matrix ($\Delta\lambda=\lambda_2-\lambda_1$) is sufficiently large (20). When $\phi(G)=o(1)$ it means that the number of links inside the subset S is approximately the same as the number of links going out from it. This means that high expansion implies high homogeneity of the network. In the following we explain a method that permits to classify networks into universal classes according to their structural homogeneity, which is known as the *spectral scaling method* (21, 22).

Let $EE_{odd}(i)$ be the subgraph centrality of the amino acid i (23). The subgraph centrality represents the weighted participation of a node in subgraphs containing at least one cycle of odd length (23). It has the following spectral formula $EE_{odd}(i)=\sum_{j=1}[\varphi_j(i)]^2\sinh(\lambda_j)$, where $\varphi_j(i)$ is the entry of the j th eigenvector of the adjacency matrix corresponding to the i th amino acid and λ_j is the corresponding eigenvalue. The subgraph centrality can be expressed as follow (21)

$$EE_{odd}(i)=[EC(i)]^2\sinh(\lambda_1)+\sum_{j=2}[\gamma_j(i)]^2\sinh(\lambda_j), \quad (2)$$

where we have written $EC(i)$ for the principal eigenvector $\varphi_1(i)$ and λ_1 corresponds to the principal (Perron-Frobenius) eigenvalue of the network.

Let us assume that the network has GE properties such that $\lambda_1>>\lambda_2$ (20). Then, we can consider that

$$[EC(i)]^2\sinh(\lambda_1)>>\sum_{j=2}[\gamma_j(i)]^2\sinh(\lambda_j).$$

Consequently, we can approximate the odd-subgraph centrality as,

$$EE_{odd}(i)\approx[EC(i)]^2\sinh(\lambda_1). \quad (3)$$

This means that the principal eigenvector of the network is directly related to the subgraph centrality in GENs according to the following spectral power-law scaling relationship:

$$EC(i)\propto A[EE_{odd}(i)]^{\eta}. \quad (4)$$

Here, $A\approx[\sinh(\lambda_1)]^{-0.5}$ and $\eta\approx 0.5$. This expression can be written in a log-log scale as (21, 22):

$$\log[EC(i)]=\log A+\eta\log[EE_{odd}(i)]. \quad (5)$$

Consequently, in a homogeneous network a log-log plot of $EC(i)$ *vs.* $EE_{odd}(i)$ displays a perfect straight line fit with slope $\eta \approx 0.5$ and intercept $\log A$. Topologically non-homogeneous networks will display large deviations from this perfect fit as a consequence of their modularity.

Let us consider the homogeneous case, in which a network displays perfect spectral scaling, such that we can calculate the eigenvector centrality by using the following expression,

$$\log EC^{Homo}(i) = 0.5 \log EE_{odd}(i) - 0.5 \log[\sinh(\lambda_1)]. \quad (6)$$

Now, let us consider the deviations from this perfect fit. We can account for these deviations from *perfect homogeneity* by measuring the departure of the points from the straight line respect to $\log EC^{Homo}(i)$ (21):

$$\Delta \log EC(i) = \log \frac{EC(i)}{EC^{Homo}(i)} = \log \left\{ \frac{[EC(i)]^2 \sinh(\lambda_1)}{EE_{odd}(i)} \right\}^{0.5}. \quad (7)$$

Using $\Delta \log EC(i)$ we can classify any network into any of four possible topological classes. The first class correspond to the case when $\Delta \log EC(i) \cong 0$ for all nodes in the network and it corresponds to the homogeneous networks. In the second class $\Delta \log EC(i) \leq 0$ for all nodes, which indicates that the network contains cavities in its structure (see further explanation). In class III $\Delta \log EC(i) \geq 0$ for all nodes, which indicate the existence of a core-periphery structure of the network. Finally, the class IV occurs when $\Delta \log EC(i) \leq 0$ for some nodes and $\Delta \log EC(i) > 0$ for the rest. In Fig. 2 we give four model networks which represent these universal topological classes of networks together with their respective spectral scaling plots.

Insert Fig. 2 about here.

In a previous report (24) we have studied real-world complex networks and we have quantified the degree of deviation of the nodes from the ideal spectral scaling by accounting for the mean square error of all points with positive and negative deviations in the spectral scaling, respectively:

$$\xi^{+} = \sqrt{\frac{1}{N_{+}} \sum\nolimits_{+} \left(\log \frac{\gamma_1(i)}{\gamma_1^{Ideal}(i)} \right)} \text{ and } \xi^{-} = \sqrt{\frac{1}{N_{-}} \sum\nolimits_{-} \left(\log \frac{\gamma_1(i)}{\gamma_1^{Ideal}(i)} \right)}$$

where $\sum_{+}$ and $\sum_{-}$ are the sums carried out for the N_{+} points having $\Delta \log \gamma_1(i) > 0$ and for the N_{-} having $\Delta \log \gamma_1(i) < 0$, respectively. In this previous work (24) we have plotted the values of these deviations for 60 real-world networks representing biological, ecological, socio-economical, technological and informational systems, where we observed that the four topological classes are populated by real-world networks (24).

RESULTS AND DISCUSSION

Topological classes of residue networks

We have investigated the spectral scaling for the 595 residue networks studied in this paper. As a warming up example we study the protein 1ash. In this case all residues deviate

negatively from the perfect spectral scaling. The value of the mean square error of negative deviations for this protein is $\xi^- = 0.498$ and that for positive deviations is exactly zero. We have found that 95% of the proteins studied exhibit negative and no positive deviations from the perfect spectral scaling. This means that 564 out of 595 proteins have ξ^+ exactly equally to zero. However, the average of the mean square error of negative deviations for these proteins is $\langle \xi^- \rangle = 0.996$. The other 31 proteins has small positive deviations from the perfect scaling, having $\langle \xi^+ \rangle = 0.0011$ and negative deviations $\langle \xi^- \rangle = 1.550$. In Fig. 3 we plot the logarithms of values of ξ^+ and ξ^- of these 595 residue networks. We have incorporated 60 complex networks representing a variety of biological, ecological, informational, social and technological systems and we classify each of them into one of the four topological classes according to our previous report (24). The following cutoff values have been used for classification purposes, class I: $\xi^- < 0.02$ and $\xi^+ < 0.02$; class II: $\xi^- > 0.02$ and $\xi^+ < 0.02$; class III: $\xi^- < 0.02$ and $\xi^+ > 0.02$; class IV: $\xi^- > 0.02$ and $\xi^+ > 0.02$.

Insert Fig. 3 about here.

From Fig. 3 we immediately see that 99.5% of the proteins represented by their residues networks are in the class II. There are only three proteins which are classified in the class IV. These findings mean that most of the 595 proteins studied, which have between 54 and 1021 residues and less than 25% of homology in their sequences, can be modelled as networks in which several highly connected clusters are separated from each other by forming structural cavities, as represented by the model graph B given in Fig. 2. In other words, protein residue networks exhibit universal topological characteristics.

In order to explain why protein residue networks are in class II let us consider a residue located in one of the highly dense modules, e.g., a node in one of the external corners of the grid represented by the model graph B given in Fig. 2. If we walk around this node we observe a high density of connections in its neighbourhood. However, if we enlarge our walk to traverse the whole network we observe that this high connectivity is not kept as soon as we arrive at the central cavity in the network (21, 22). Our measurement of local connectivity is the subgraph centrality, which is larger than expected from the homogeneity line of the spectral scaling. This means that the points representing the amino acids of the residue network are placed to the right of the line representing the perfect scaling. An equivalent view is obtained by considering the eigenvector centrality, which is our global measure of connectivity. In this case the global connectivity is lower than expected from the homogeneity line placing the points below the straight line of the spectral scaling (21, 22). A discussion about the use of other local and global measures in biological networks can be found in (25).

The eigenvectors corresponding to positive eigenvalues of the adjacency matrix of a network give a partition of the network into clusters of tightly connected nodes. On the contrary, the eigenvectors corresponding to negative eigenvalues make partitions in which nodes are not close to those which they are linked, but rather with those with which they are not linked. In other words, the nodes will be close to other nodes which have similar patterns of connections with other sets of nodes, i.e., nodes to which they are structurally equivalent. In the case of the eigenvectors corresponding to positive eigenvalues the nodes corresponding to larger components tend to form *quasi-cliques*. That is, clusters in which every two nodes tend to interact with each other. On the contrary, for eigenvectors corresponding to negative

eigenvalues, nodes tend to form *quasi-bipartites*, i.e., nodes are partitioned into disjoint subsets with high connectivity between sets but low internal connectivity. It has been shown analytically (24) that for class II networks

$$\left|\sum_{+}\left[\varphi_j(i)\right]^2 \sinh\left(\lambda_j\right)\right| > \left|\sum_{-}\left[\varphi_j(i)\right]^2 \sinh\left(\lambda_j\right)\right|,$$

where $\sum_{+}$ and $\sum_{-}$ designate the sums corresponding to positive and negative eigenvalues for $j \geq 2$. This means that the networks of class II are dominated by partitions into quasi-cliques more than into quasi-bipartites. In words, these networks are characterized by two or more clusters of highly interconnected nodes which display a low inter-cluster connectivity.

The tiny fraction of residue networks in class IV displays a mixture of positive and negative deviations. They are characterized by a combination of both quasi-cliques and quasi-bipartites, without the predominance of either structure over the other. On the one hand, the central nodes connecting the highly interconnected clusters display larger connectivity to all other nodes in the network than the one expected from their local cliquishness, i.e., they display positive deviations from the perfect scaling. On the other hand, the nodes on one side of the graph are not well-connected to the nodes on the other side despite they are internally highly connected. Consequently, these nodes display negative deviations from the perfect scaling.

Topological analysis of protein classes

Despite the fact that 99.5% of the protein residue networks are classified in the class II it can be seen in Fig. 4A that there is a continuous transition between classes II and IV. In other words, there is not a clear gap separating proteins in class II from those in class IV. However, 95% of all proteins are clearly in class II having negative deviations from the perfect scaling and $\xi^+ = 0$. The other fraction of proteins can be considered as borderline cases between both classes (see Fig. 3).

The most significant difference between these two groups of proteins is given by the size of the proteins contained in each class. We have already remarked that the proteins studied here have between 54 and 1021 amino acids, with an average size of 254 residues. The average size of protein networks in class II having $\xi^+ = 0$ is 261 residues. However, those protein networks for which $\xi^+ \neq 0$ have average size of 126 residues, which is significantly lower than the average size for the whole data set of proteins. When the values of ξ^+ and ξ^- of all proteins are plotted versus the number of residues in Fig. 4A we obtain a clear illustration of the size dependence of the topologies found for proteins.

Insert Fig. 4 about here.

As can be seen in Fig. 4A most proteins for which $\xi^+ \neq 0$ have less than 200 residues. The only one exception is the protein with PDB code 1aa6, which has 697 amino acids. The probability of finding a protein having $\xi^+ \neq 0$ is 63.6% (7 out of 11) for proteins with no more than 75 residues. This probability drops to 29.5% (13 out of 44) for proteins with less than 100 residues and to 14.2% (25 out of 176) for proteins with no more than 150 amino acids. This probability is only 0.4% for proteins having more than 240 residues.

In order to understand these results we are going to consider the domain structure of proteins. A domain is a part of the protein which has a compact three-dimensional structure

and often can be independently stable and folded. Then, in a residue network those amino acids which are in the same domain tend to form highly interconnected clusters, such as the ones represented by the corners of the model network B in Fig. 2. However, the number of interactions between two different domains in a multi-domain protein is relatively low compared to the intra-domain interactions. Thus, we can think that most of the structure of class II residue networks can be explained by the multi-domain organization that these proteins have. This hypothesis is supported by the findings that most of domains found in proteins have between 50 and 150 residues (26). Then, most of small proteins have only one domain, while larger proteins tend to be combinations of such domains (26). This explains why practically all residue networks with more than 240 amino acids are clearly in class II. The structure of those proteins having $\xi^+ \neq 0$ can be explained by considering the organization of the different elements of their secondary structure, e.g., helices and sheets. These elements can form small clusters, which are then interconnected in the form given by the graph D in Fig. 2 with more or less predominance of the quasi-cliques or quasi-bipartite structures.

In Fig. 4B we plot the negative deviations from perfect scaling ξ^- versus the number of residues in the proteins studied. There are two characteristics of this plot that can also be related to the domain structure of proteins. The first is that the variability in the values of ξ^- is significantly larger for small proteins than for larger ones. For instance, for proteins having no more than 200 residues the deviations range from zero to almost 6, $0 < \xi^- < 6$. However, for proteins over 500 amino acids the deviations are concentrated in the range $1 \leq \xi^- < 3$. The larger variability in the deviations of small proteins could be due to the before mentioned fact that most of protein domains have between 50 to 150 residues. Then, small proteins are formed by one domain only. Because the dataset studied contains less than 25% similarity in the sequence of proteins it is expected that two small proteins are formed by different domains with high probability. For instance, some of these small proteins are mainly-α, while others are mainly-β. The differences in the structures of these domains are reflected in the spectral scaling of their residue networks giving rise to the observed variability.

In the case of larger proteins it has also been reported that 64.3% of proteins having more than 200 residues are formed by more than one domain (27). It has been reported previously that some of these multidomain proteins display degree of folding which are average of the folding of the different domains forming the protein (28). Then, we can consider multidomain proteins as combinations of the small ones in such a way that their topological properties are the average of the ones observed for one-domain proteins.

The second characteristic of the plot given in Fig. 4B is the existence of a lower bound in the negative deviations from perfect scaling. This lower bound is a function of the protein size. For instance, for proteins having less than 100 residues there are no negative deviations below $\xi^- = 0.25$. This bound is $\xi^- \approx 0.45$ for proteins with 200 residues and it is larger than 1 for proteins with more than 800 amino acids. This threshold can be approximated by a straight line as can be seen in Fig. 4B. The existence of this lower bound can be explained by the fact that the number of domains per protein steadily increases with the size of the proteins. For instance, 42.9% of proteins with more than 350 residues have more than 2 domains and 38.5% of those with more than 500 amino acids have more than 3 domains (27). Then, it is intuitive to think that the deviation from homogeneity in the residue networks increases with the number of domains due to the larger inter-domain cavities created.

Now, we want to add some words about the existence of cavities in class II protein networks. In graph theory a hole is defined as a *chordless cycle*, which is a cycle of length at least 4 such that there is not link between two non-consecutive nodes (29). It is easy to check that any of the four universal topological classes of networks contains holes in their structures. In Fig. 2 it is easily recognizable that model networks in class II, III and IV contain chordless cycles of length 4. Then, a cavity in networks of class II cannot be identified with a network hole. Instead we propose the following definition of a topological cavity:

Definition: A *topological cavity* is a hole, i.e., a chordless cycle of length at least 4, which separates at least two highly connected clusters in the network.

The problem of determining whether a network has a chordless cycle of length k is NP-complete (30). The problem has been solved for fixed k in $O\left(n^k\right)$ (31) and some improvement exists which solves the problem in $O\left(n^{k-3}T\right)$, where $T \approx n^{2.376}$ (32). In addition, for the identification of a topological cavity these holes should separate two or more highly connected clusters in the network. Here we are not concerned with the development and/or application of these algorithms to detect holes or topological cavities in residue networks. Instead, we are going to illustrate the existence of such cavities in a residue network and its possible implications for the study of protein structures. Our aim is to stimulate other researchers in the search for algorithms to detect topological cavities and holes in residue networks.

For the sake of illustration we study the residue network for the protein with PDB code 1ash. We have found the presence of a hole of length 15 in this network, which is formed by the residues 27, 30, 33, 40, 43, 59, 60, 62, 64, 67, 71, 92, 95, 96 and 101. That is, these residues form a cycle of length 15 in which no two of them are joined by a link which does not belong to the cycle. As can be seen in this figure this hole is separating several highly connected regions of the residue network and can be considered as a network cavity according to the definition given here. We have to recognize that no quantitative method whatsoever has been developed for the identification of these structures in complex networks in general or in residue networks in particular. Thus, the identification of this particular cavity for illustration purposes here has been carried out using a *brute-force* search.

When this topological cavity is placed on the three-dimensional structure of the protein we can see that it practically corresponds to the binding site of this protein. According to the information provided by the database PDBSum (33) the residues having contacts with the ligand are 30, 33, 40, 43, 44, 60, 64, 67, 68, 71, 95, 96, 101, 103, 108, and 140. This means that there is 71% of overlapping between this list of residues in the binding site and the topological cavity found in the residue network. The overlapping is calculated as twice the number of residues in the interception of the two lists divided by the number of residues in the union of both lists. This example clearly illustrates the relationship between topological cavities and potential binding sites in proteins.

The current work is not aimed, however, to the development of mathematical or computational methods for identifying binding sites in proteins, for which excellent methods already exist (34, 35). As previously stated we are aimed here with the identification of topological properties that residues networks share in a universal way, such as the organization of their nodes/links in class II type of structure. This universal topological property adds to other properties like the small-worldness and Poisson degree distributions previously reported (11) for residue networks and they together allow us to understand the

organizational principles of protein architectures. However, because the cavities are important characteristics of class II networks we are going to analyze some of their general features. First, we are going to use the nomenclature of *voids*, *pockets* and *depressions* for the three types of cavities studied previously in (34). In a residue network representation of a protein a void can be identified by the current approach as a *topological cavity* if, and only if, the distance r_i between pairs of atoms in the interior of the cavity is larger than the cutoff value used, $r_i > r_C$. For pockets and depressions being recognized as topological cavities the distance r_m between the two atoms forming the mouth of the site should be smaller than the cutoff value used $r_m < r_C$ as well as $r_i > r_C$ for internal atoms. Then, there are obvious situations in which the current approach is not able to identify some kinds of binding sites. The situation is not different, however, to other methods used for this purpose. For instance, in the case of small binding sites, such as those for metals, a network representation with a relatively large cutoff value does not identify the corresponding binding pockets. What is interesting here is that the topological cavities defined in this work are present as a universal characteristic of residue networks for the cutoff value and network definition used in this work.

In the following we are going to analyze the influence of the cutoff value in the topological structure of residue networks. In particular we study the protein with PDB code 1ash for cutoff values 5.0 Å, 7.0 Å, 10.0 Å and 15.0 Å. We recall that the cutoff value 7.0 Å is the one used in the current work for all residue networks. When a value 5.0 Å is used only C_β atoms of the different residues which are separated at most 5.0 Å apart are connected. This includes only 146 links connecting nearest-neighbor residues in the protein backbone plus 40 other inter-residue interactions. Consequently, the resulting residue network displays very low density (the average degree is only 2.68). The spectral scaling of this version of the residue network is characteristic of class IV networks in which there are holes together with some core-periphery structures. The holes in this version of the network are quite large. For instance, there is a 53-nodes hole formed by residues 32-34, 55-58, 60-78, 85-96, 101-103 and 107-118. In Fig. 5A we display the spectral scaling for this network and a visualization of the previously mentioned hole as a surface in the protein structure. As can be seen this hole corresponds to practically the whole external surface of the protein. This confirms our previous statement that using small cutoff values is equivalent to consider very small probe spheres to explore the protein, in which all small cavities are considered describing all rugosities of the protein surface. Then, if we are interested in studying only the total surface of a protein it is recommended to use very small cutoff values, e.g., $r_C < 5.0\,\text{Å}$. For the sake of comparison, we illustrate the spectral scaling and surface of the 15-nodes hole in the same protein obtained for a cutoff value 7.0 Å (Fig. 5B).

Insert Fig. 5 about here.

When the cutoff value is augmented to 10.0 Å only large holes are detected. The residue network display a class II topological structure, but the spectral scaling displays considerable less dispersion than for the case of 7.0 Å (Fig. 5C). In this case the 15-nodes hole previously detected for 7.0 Å is shrank to one having 9 nodes only, which is formed by residues 30, 33, 40, 43, 34, 37, 71, 95 and 101. In Fig. 5C we can see that these residues are located in the wider part of the binding site for 1ash. The same trend is observed when we study the network built by using cutoff 15.0 Å (Fig. 5D). In this case the spectral scaling corresponds to a class I network due to the high density observed in the network. We recall that in this case all C_β atoms separated at less than 15.0 Å are connected. Consequently, only

very large holes are detected, which in this case are of very small size, such as the one formed by residues 40, 60, 62, 71 and 96. This approach of using very large cutoff values is only recommended when very large cavities need to be detected in the protein. These results confirm the previous report of da Silveira et al. (15) about the use of a cutoff value equal to 7.0 Å to represent proteins as residue networks.

Random models and protein topological classes

We have previously shown that proteins represented by residue networks mainly belong to a universal topological class. Only 5% of proteins primarily of small size are in a different topological class. Ravasz et al. (37) have modelled hierarchical networks by replicating a core set of nodes and links. However, as Bagler and Sinha have remarked (12), "proteins grow linearly first, and then this polypeptide chain organizes itself in a modular way at different levels". It was already mentioned that Bartoli et al. (13) have found that the small-world properties of these residue networks are well replicated by random models, which excludes the use of the average path length and clustering as protein fingerprints. Then, what is the situation with the current universal property found for residue networks. Is the universal class of proteins reproduced by some of the best known models of network growing? We are going to analyze these questions now.

Here we investigate whether the random network growing models of Erdös-Rényi (ER) (38) and Barabási-Albert (BA) model (5) are able to reproduce the structural classes populated by protein residue networks. In both models each random network starts with *m* nodes and new nodes are added consecutively in such a way that a new node is connected to exactly *m* of the already existing nodes, which are chosen randomly. The new edges are attached according to the probability distribution used, e.g., Poisson distribution for the ER and the preferential attachment mechanism, power-law degree distribution, in the BA model. We have studied random networks generated by these two growing mechanisms having $n = 1000$ nodes by changing systematically the value of *m* from 2 to 8, giving rise to networks with average degrees, $\langle k \rangle$ between 4 and 16. For every value of *m* we have generated 100 random networks. Then, we have averaged the values of ξ^- and ξ^+ for every value of *m*. In Fig. 6A we plot the values of $\ln\left(\xi^- + 10^{-5}\right)$ versus $\ln\left(\xi^+ + 10^{-5}\right)$ for the random networks generated using both growing models. In the same plot we have placed the values corresponding to 595 proteins studied in this work.

Insert Fig. 6 about here.

As can be seen in this figure neither ER nor BA growing mechanisms are able to reproduce the topological properties of residue networks. None of these random networks replicate, even partially, the structure of 99.5% of proteins in which highly interconnected clusters are separated by topological cavities. The characteristic of class IV protein residue networks, in which a tiny fraction of small-size proteins appears, are only partially reproduced by random networks at low average degree regimes. It is clear from this figure that as the average degree grows both kinds of random networks exhibit good expansion properties typical of class I networks. These results are supported by previous theoretical findings indicating that for $\langle k \rangle \geq 3$ these random networks are expanders with high probability (39, 40). In summary, the topological organization exhibited by protein residue networks is of more complex nature than the ones reproduced by random growing mechanisms, such as ER and preferential attachment. This finding should be added to the well-known fact that these two models do not reproduce important structural properties of

residue networks. For instance, networks created with the ER model lacks internal structures reflected by its very poor clustering and the BA model does not reproduce the Poisson distribution of these networks (11).

In view of the previous findings it should be more appropriate to consider other random models that reproduce better the structural organization of residue networks. One of such models is that of Watts and Strogatz (WS) to generate "small-world" networks by starting from regular ones (4). This model starts with a ring of n nodes in which each node is symmetrically connected to its $2m$ nearest neighbours. Then, for every node, each link connected to a clockwise neighbour is rewired with probability p. This rewiring process usually creates shortcuts between distant nodes in the ring. At the very beginning when we are in presence of the n-nodes ring there is a central hole in the graph. Thus, the initial step in this process is a network belonging to class II, which also has large clustering but lacks the small average shortest path length that characterizes small-world networks like the protein residue ones. Then, we studied here the spectral scaling for networks created by using the WS model for networks having 300 nodes and $0 \le p \le 1$. The results are illustrated in Fig. 6B, where we can see that the networks generated by WS model are in class II for values of $p < 0.5$. For probabilities beyond this value the networks are in class IV and go to class I when $p = 1.0$. As can be seen in this figure the networks obtained for values of $p \approx 0.01$ display large clustering coefficient similar to the original ring but they have average path length that scales as the logarithm of the number of nodes and in addition they display class II topological characteristics. In other words, they are small-world class II networks, which are topological characteristics shared by protein residue networks.

The fact that the WS networks duplicate some important properties of residue networks should not be considered as an indication that this model can give insights about the way in which proteins fold into their 3D structures. For instance, the WS model starts by considering a ring with each node symmetrically connected to its $2m$ nearest neighbours. This violates the principle that "proteins grow linearly first, and then this polypeptide chain organizes itself in a modular way at different levels" (12). The presence of this ring is the main cause for the existence of the hole in the networks generated by this model. Secondly, in the WS model the rewiring process takes place for any link in the network, while in a protein those links representing the protein backbone cannot be rewired. In fact, in the WS model for large number of nodes it is usually to find disconnected networks, which is not ever the case in protein folding process. Then, a more realistic model of residue network "formation" is given by the following modified WS (MWS) model

i) starts with a linear chain of n nodes in which every link is colored blue,
ii) connect node i of the linear chain to $i+k$, e.g., $k=2$, and color every of these links in red,
iii) for every node, each red link is rewired with probability p.

We are not going to analyze this model in detail. However, our preliminary results indicate that the networks generated by using MWS model are both small-world and class II graphs, which encourage us to study these networks in more detail in further works. In addition, this growing process fulfils the requirement of starting from a linear chain that then organizes itself in a modular way (12).

CONCLUSIONS

Proteins represented by their cutoff-dependent residue networks exhibit universal topological characteristics. Residue networks belong primarily to the topological class of modular networks formed by several highly connected clusters separated by topological cavities. We call the attention that not every modular network fulfils the requirement of being a class II network. A very small fraction of protein residue networks deviates from this behaviour and are characterized by having some central core surrounded by a periphery of small clusters and cavities. These networks represent mainly small-size proteins having less than 200 residues. In general, small proteins display larger variability in their deviations from homogeneity than larger proteins. These differences can be accounted for by the domain structure of these proteins. Finally, we have shown that the topological cavities characterizing proteins residue networks are intimately related to protein binding sites. The cutoff value for the construction of the network has been analyzed and we have seen that the spectral scaling method is appropriate to distinguish between protein-like and non protein-like networks. In addition, different cutoff values can be used to model different kind of characteristics in proteins, such as molecular surface (small cutoff values) or cavities of different sizes. We have also analyzed some random models for mimicking the properties of residue networks. Among then, the Watts-Strogatz model reproduces very well the topological class and small-worldness observed for residue networks. We have proposed here a modification of this model that is able to build protein-like networks keeping in mind the restrictions imposed by the chemistry of proteins.

The current study has made evident the necessity for efficient algorithms for detecting topological cavities in residue networks as well as for better mathematical characterization of these structures. We hope that this work contributes to the search for other topological methods and algorithms for extracting more structural information from the topology of proteins represented as networks.

Acknowledgements

The author thanks Prof. C. Baysal for kindly supply the structures of the 595 protein residue networks used in this work and for continuous support. Dr. D. Turgut is also thanked for his valuable help with calculations related to the cutoff values. Constructive comments from anonymous referees and the Editor contribute to improve the presentation of this material. Partial financial support from the New Professor's Fund given by the Principal, University of Strathclyde is also acknowledged.

REFERENCES

1. Strogatz, S. H. 2001. Exploring complex networks. *Nature* 410:268-276.
2. Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Rev*. 45:167-256.
3. Boccaletti, S., V. Latora, Y. Moreno, M. Chávez, and D.-U. Hwang. 2006. Complex networks: Structure and dynamics. *Phys. Rep*. 424:175-308.
4. Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440-442.
5. Barabási A.-L., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286:509-512.
6. Song C., S. Havlin S., and H. A. Makse. 2005. Self-similarity of complex networks. *Nature* 433:392-395.
7. Milo R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: Simple building blocks of complex networks. *Science* 298:824-827.
8. Newman, M. J. E. 2003. Mixing patterns in networks. *Phys. Rev. E*. 67:026166.
9. Böde, C., I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, and P. Csermely. 2007. Network analysis of protein dynamics. *FEBS Lett*. 581:2776-2782.
10. Vendruscolo, M., N. V. Dokholyan, E. Paci, and M. Karplus. 2002. Small-world view of the amino acids that play a key role in a protein folding. *Phys. Rev. E*. 65:061910.
11. Atilgan, A. R., P. Akan, and C. Baysal. 2004. Small-world communication of residues and significance for protein dynamics. *Biophys. J*. 86:85–91.
12. Bagler, G., and S. Sinha. 2005. Network properties of protein structures. *Physica A*. 346:27–33.
13. Brinda, K. V., and S. Vishveshwara. 2005. A network representation of protein structures: Implications for protein stability. *Biophys. J*. 89:4159-4170.
14. Greene, L. H., and V. A. Higman. 2003. Uncovering network systems within protein structures. *J. Mol. Biol*. 334:781–791.
15. da Silveira, C. H., D. E. V. Pires, R. C. Minardi, C. Ribeiro, C. J. M. Veloso, J. C. D. Lopes, W. Meira Jr., G. Neshich, C. H. I. Ramos, R. Habesch, and M. M. Santoro. 2008. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins* 74:727-743.
16. Bartoli, L., P. Fariselli, and R. Casadio. 2007. The effect of backbone on the small-world properties of protein contact maps. *Phys. Biol*. 4:L1-L5.
17. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliand, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The protein data bank. *Nucleid Acids Res*. 28:235-242.
18. Fariselli, P., and R. Casadio. 1999. A neural network based predictor of residue contacts in proteins. *Protein Eng*. 12:15-21.
19. Biggs, N. 1993. Algebraic Graph Theory. Cambridge University Press, Cambridge, U.K.

20. Hoory, S., N. Lineal, and A. Wigderson. 2006. Expander graphs and their applications. *Bull. Am. Math. Soc*. 43:439-562.

21. Estrada, E. 2006. Spectral scaling and good expansion properties in complex networks. *Europhys. Lett.* 73:649-655.

22. Estrada, E. 2006. Network robustness. The interplay of expansibility and degree distribution. *Eur. Phys. J. B* 52:563-574.

23. Estrada, E., and J. A. Rodríguez-Velázquez. 2005. Subgraph centrality in complex networks. *Phys. Rev. E* 71:056103.

24. Estrada, E. 2007. Topological structural classes of complex networks. *Phys. Rev. E* 75:016103.

25. González-Diaz, H., Y. González-Diaz, L. Santana, F. M. Ubeira, and E. Uriarte. 2008. Proteomics, networks and connectivity indices. *Proteomics* 8:750-778.

26. Privalov, P. L. 1989. Thermodynamic problems of protein structure. *Annu. Rev. Biophys. Biophys. Chem.* 18:47-69.

27. Dengler, U., A. S. Siddiqui, and G. J. Barton. 2001. Protein structural domains: Analysis of the 3Dee domains database. *Proteins* 42:332-344.

28. Estrada, E. 2004. A protein folding degree measure and its dependence on crystal packing, protein size, secondary structure, and domain structural class. *J. Chem. Inf. Comp. Sci.* 44:1238-1250.

29. Chandrasekharan, N., V. S. Lakshmanan, and M. Medidi. 1993. Efficient parallel algorithms for finding chordless cycles in graphs. *Parall. Proc. Lett*. 3:165-170.

30. Lubiw, A. 1988/89. A note on odd/even cycles. *Discr. Appl. Math*. 22:87-92.

31. Hayward, R. 1987. Two classes of perfect graphs. Ph. D. Thesis. School of Computer Science, McGill University.

32. Spinard, J. P. 1991. Finding large holes. *Inform. Process. Lett*. 39:227-229.

33. Laskowski, R. A. 2001. PDBSum: Summary and analysis of PDB structures. *Nucleic Acids Res*. 29:221-222.

34. Liang, J., H. Edelsbrunner, and C. Woodward. 1998. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci*. 7:1884-1897.

35. Weisel, M., E. Proschak, J. M. Kriegl, and G. Schneider. 2009. Form follows function: Shape analysis of protein cavities for receptor-based drug design. *Proteomics* 9:451-459.

36. Bohne-Lang, A., G. Wolf-Dieter, and R. Ranzinger. 2005. AISMIG: An Interactive Server-side Molecule Image Generator. *Nucleic Acids Res*. 33:705-709.

37. Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. 2002 Hierarchical Organization of Modularity in Metabolic Networks. *Science* 297:1551-1555.

38. Erdős, P., and A. Rényi, 1959. On random graphs. I. *Publ. Math.* 6:290–297.

39. Friedman, J. 1991. On the second eigenvalue and random walks in random *d*-regular graphs. *Combinatorica* 11:331–362.

40. Gkantsidis, C., A. Saberi, and M. Mihail. 2003. Conductance and congestion in power law graphs. In *Proceedings of the ACM Sigmetrics* June, 2003.

Figure captions

FIGURE 1 Construction of the residue network for protein with PDB code 1ash. The network is represented by using the 3D vector display program MAGE, (http://www.analytictech.com/downloaduc6.htm). The nodes are colored in the same way as the secondary structure elements given in the cartoon representation shown in this figure.

FIGURE 2 The four topological classes of networks which are theoretically possible according to the spectral scaling method. A model graph for each class (left) and the spectral scaling (right) are given for each model.

FIGURE 3 Classification of real-world networks in the topological classes illustrated in Fig. 2. Networks representing a variety of complex biological, ecological, informational, social and technological systems are represented by circles. Protein residues networks are represented by upper triangles.

FIGURE 4 A) Spectral scaling of 595 protein residue networks as a function of the number of residues in the protein. In the box there are all proteins which have $\xi^{+} \neq 0$ (5% of the total). B) Plot of the negative deviations from perfect scaling versus the number of residues in the proteins studied. The lower bound has been plotted by hand as an eye-guide.

FIGURE 5 Effect of the cutoff value on the topological structure of the residue network with PDB code 1ash. In the left part we illustrate the spectral scaling for networks obtained with cutoff values 5.0 Å (A), 7.0 Å (B), 10.0 Å (C) and 15.0 Å (D). In the right part of the figure we illustrate the three-dimensional structure of the protein in which the amino acids forming some topological cavities are represented by their surface. In the case of 5.0 Å the cavity illustrated contains 53 residues. The cavity illustrated for 7.0 Å (B) is the one reported in the main text. In the cases of cutoff values 10.0 Å (C) and 15.0 Å (D) we investigate only the effects of the cutoff value on the 15-nodes cavity found for 7.0 Å. The molecular structure and surface are drawn by using AISMIG (38).

FIGURE 6 A) Plot of the positive and negative deviations from perfect scaling for 595 protein residue networks together with the corresponding values for the random networks with 1000 nodes generated by using ER and BA models. B) Average shortest path length $L(p)$, clustering coefficient $C(p)$ and topological classes for the WS networks having 300 nodes.

Fig. 1

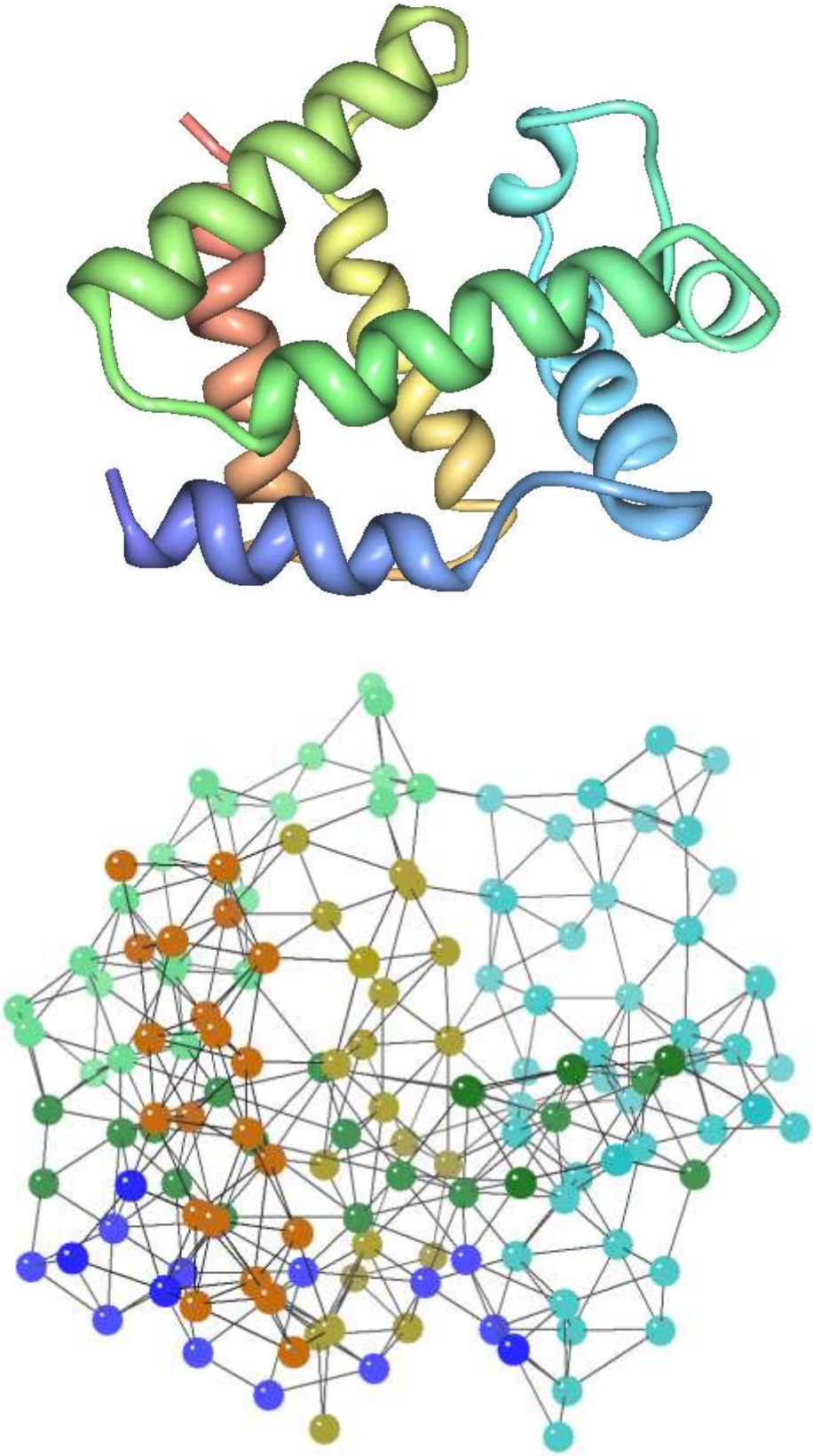
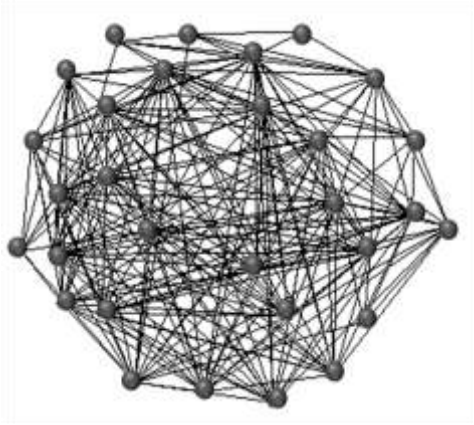
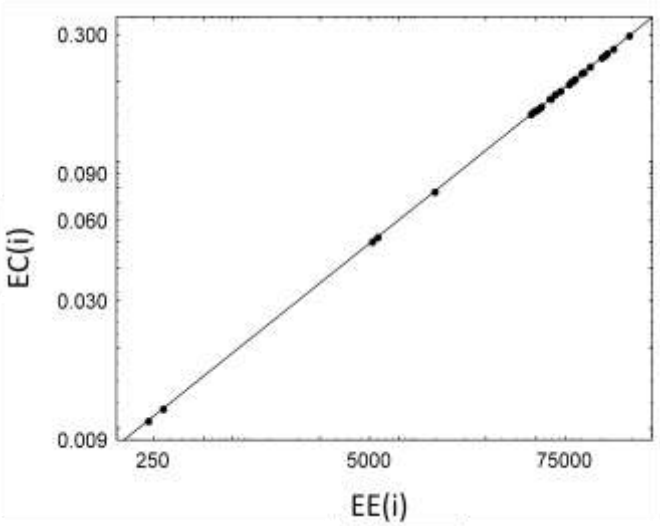
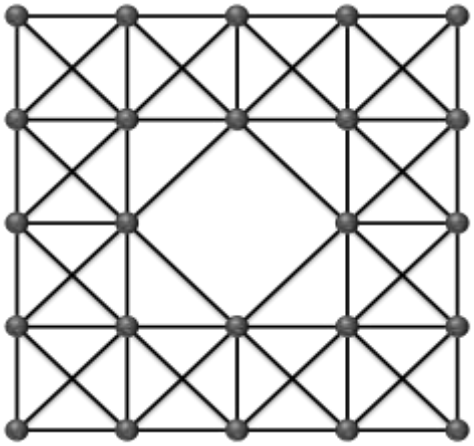
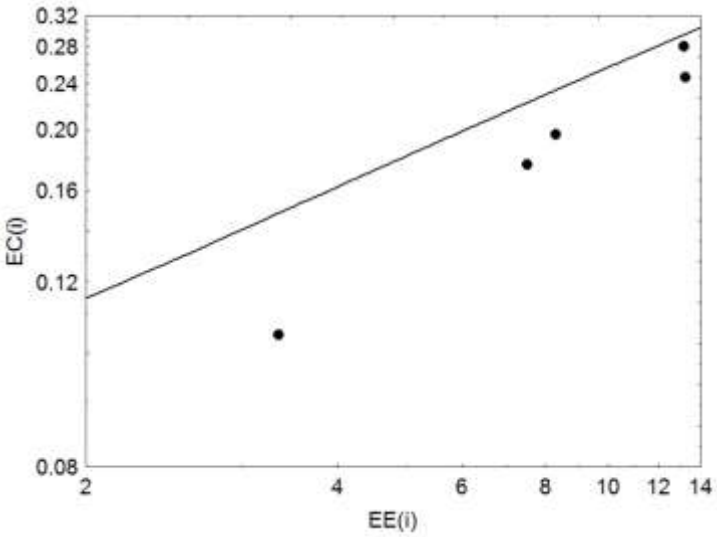
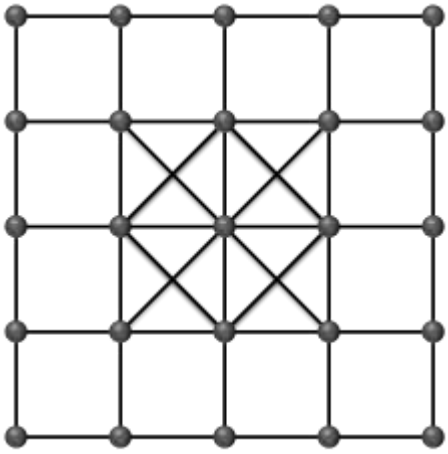


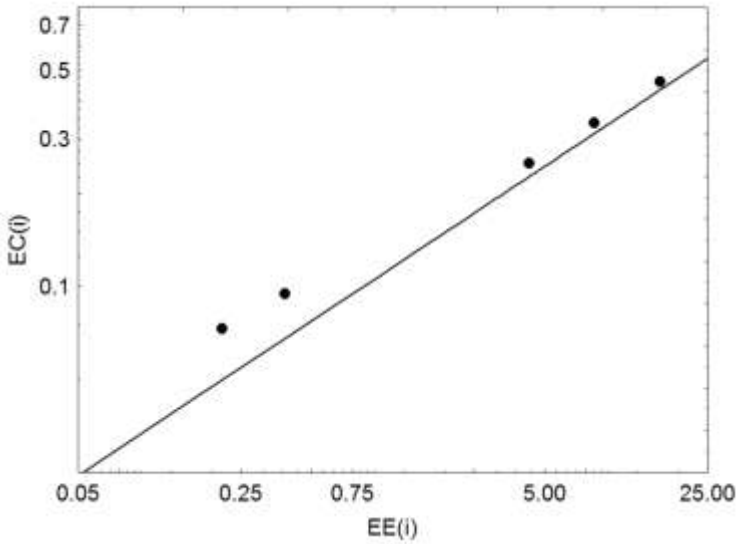
Fig. 2

Network Model	Spectral Scaling
A) 	$\Delta \log EC(i) \cong 0, \forall i \in V \Rightarrow [EC(i)]^2 \sinh(\lambda_1) \cong EE_{odd}(i).$ 
B) 	$\Delta \log EC(i) \leq 0 \Rightarrow [EC(i)]^2 \sinh(\lambda_1) \leq EE_{odd}(i), \forall i \in V.$ 

C)

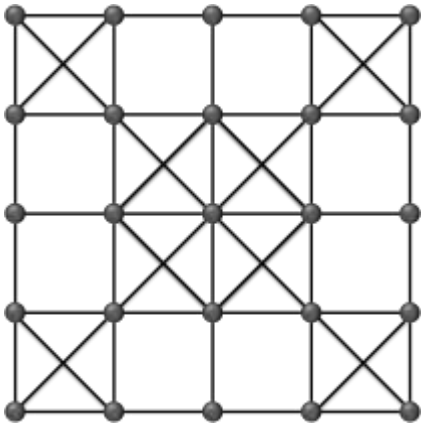
$$\Delta \log EC(i) \geq 0 \Rightarrow [EC(i)]^2 \sinh(\lambda_1) \geq EE_{odd}(i), \forall i \in V.$$





D)

$$\Delta \log EC(p) \leq 0,\ p \in V \text{ and } \Delta \log EC(q) > 0, q \in V.$$



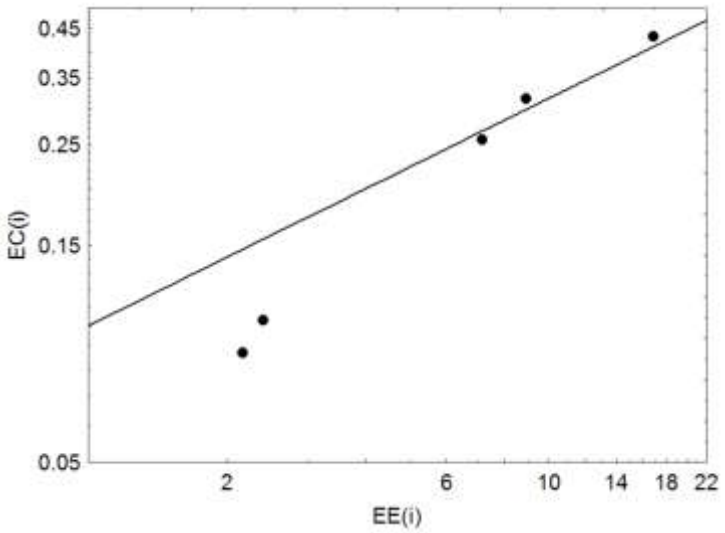


Fig. 3

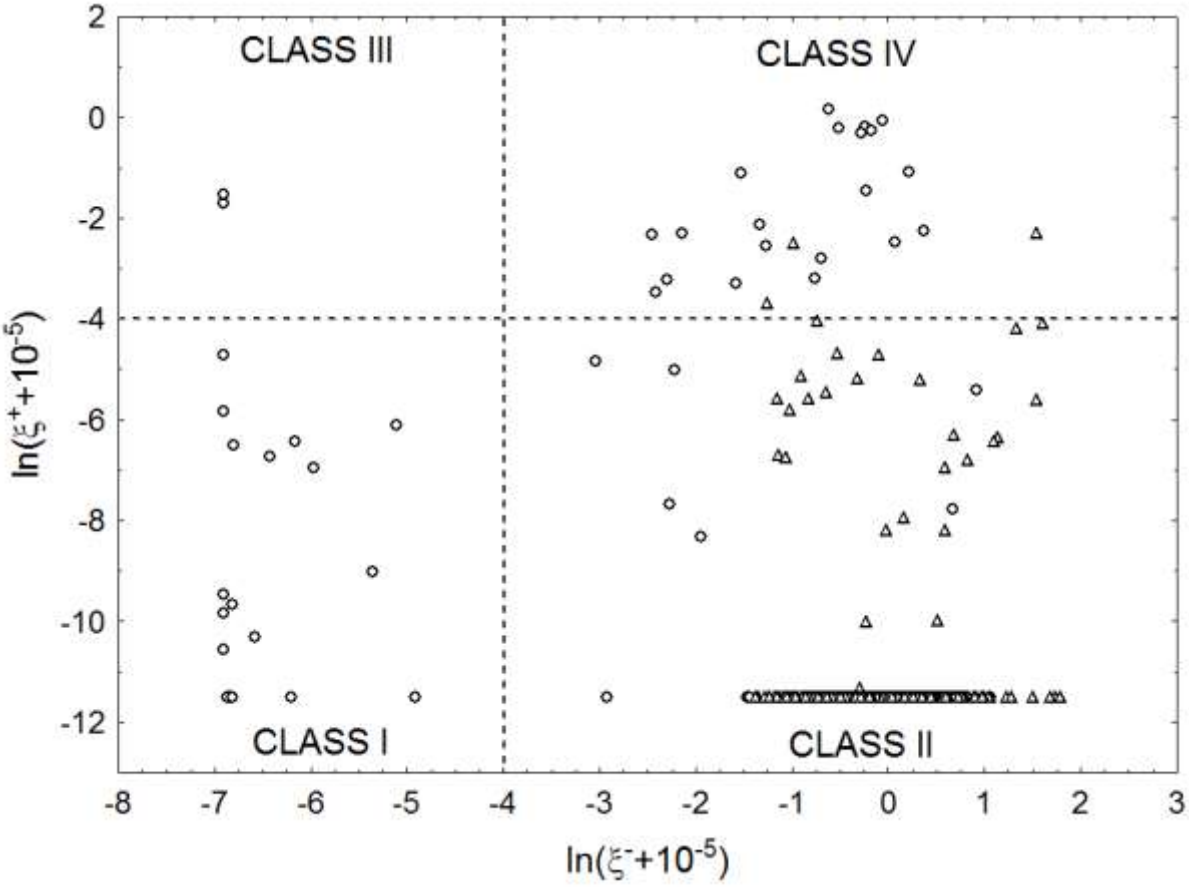
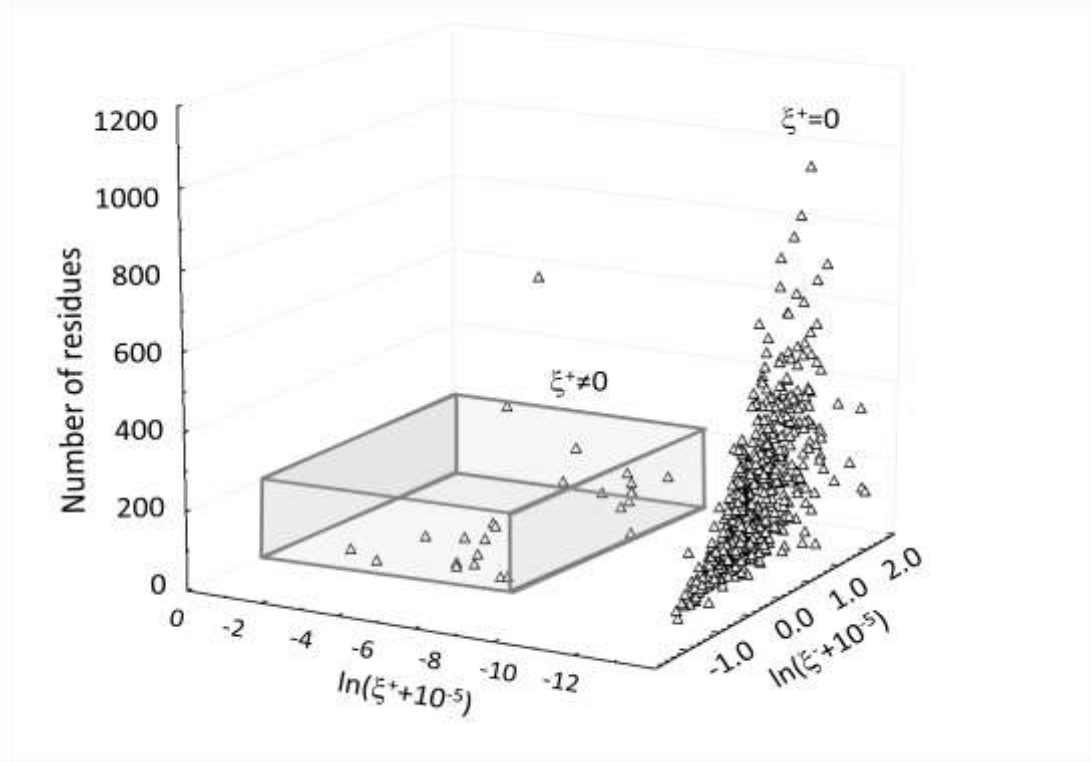

CLASS III
CLASS IV
CLASS I
CLASS II
2
0
-2
-4
-6
-8
-10
-12
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
ln(ξ+10-5)
ln(ξ-+10-5)

Fig. 4

A)


1200
1000
800
600
400
200
0
Number of residues
ξ+=0
ξ+≠0
0
-2
-4
-6
-8
-10
-12
ln(ξ+10-5)
-1.0
0.0
1.0
2.0
ln(ξ-+10-5)

B)

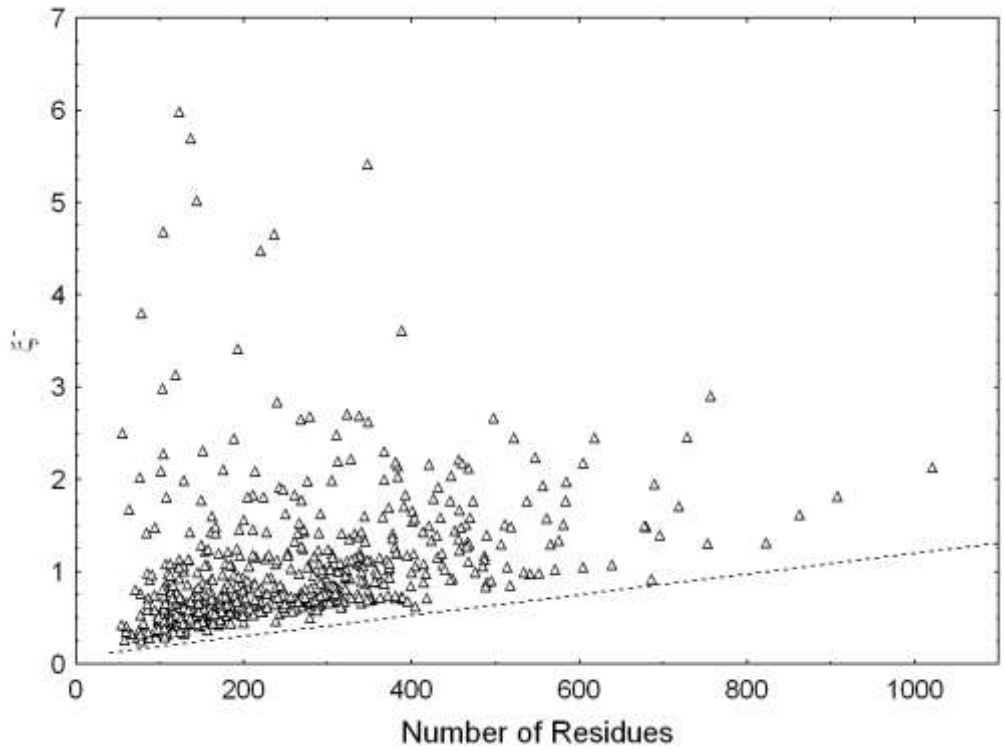
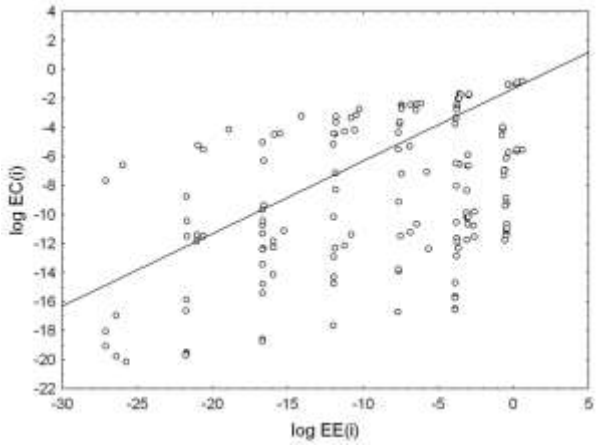
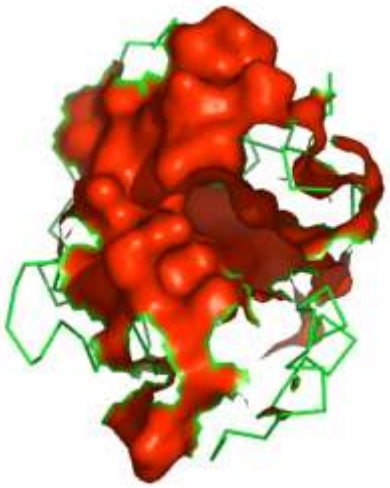

7
6
5
4
3
2
1
0
0
200
400
600
800
1000
Number of Residues

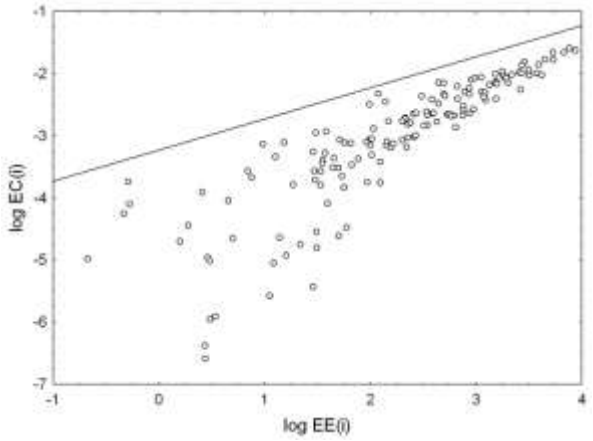
Fig. 5

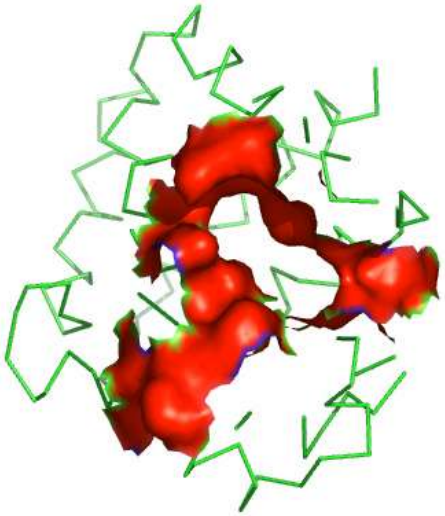
$r_C = 5.0\,\text{Å}$



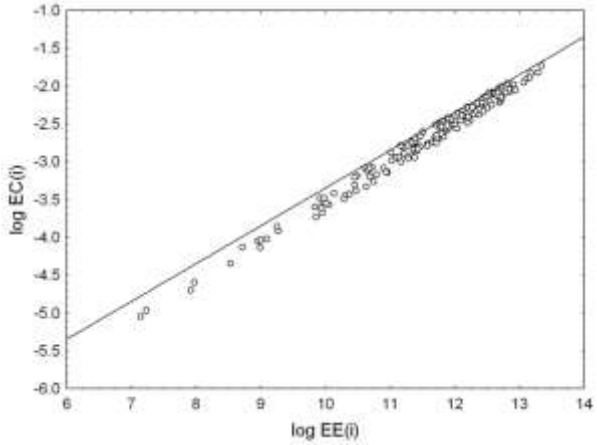


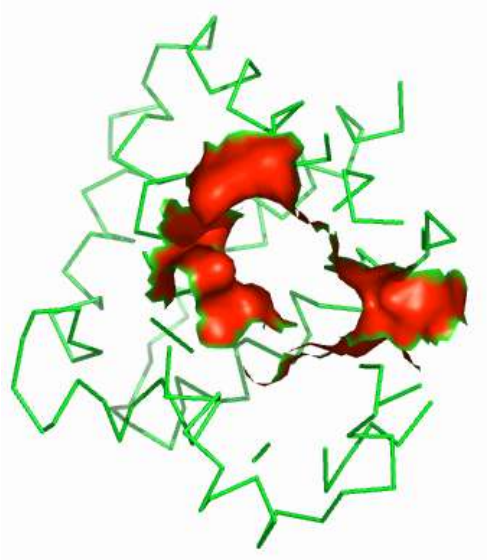
$r_C = 7.0\,\text{Å}$



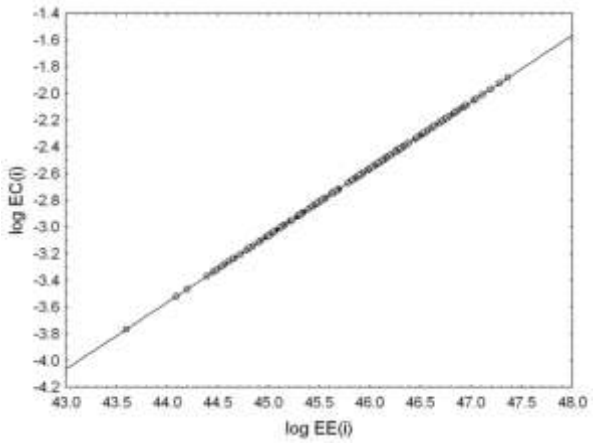


$r_C = 10.0\,\text{Å}$





$r_C = 15.0\,\text{Å}$


log EC(i)
log EE(i)
-1.4
-1.6
-1.8
-2.0
-2.2
-2.4
-2.6
-2.8
-3.0
-3.2
-3.4
-3.6
-3.8
-4.0
-4.2
43.0
43.5
44.0
44.5
45.0
45.5
46.0
46.5
47.0
47.5
48.0

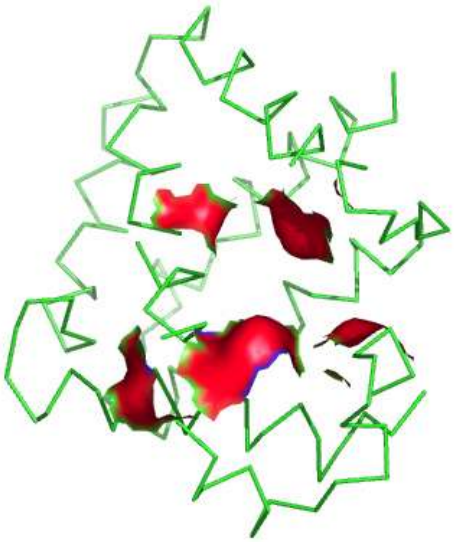
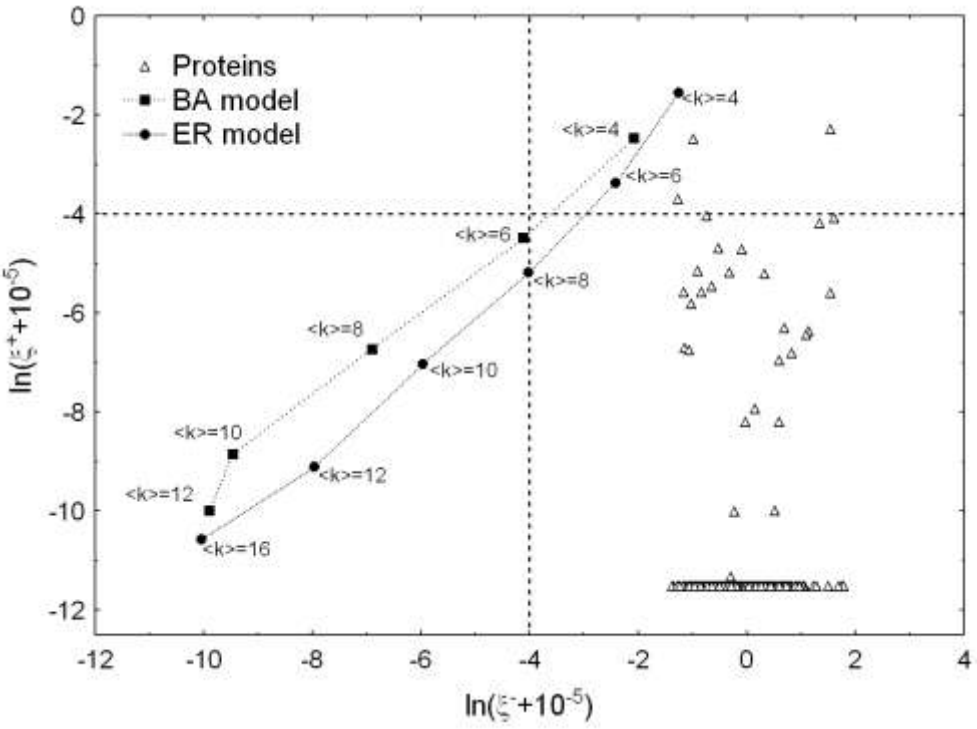
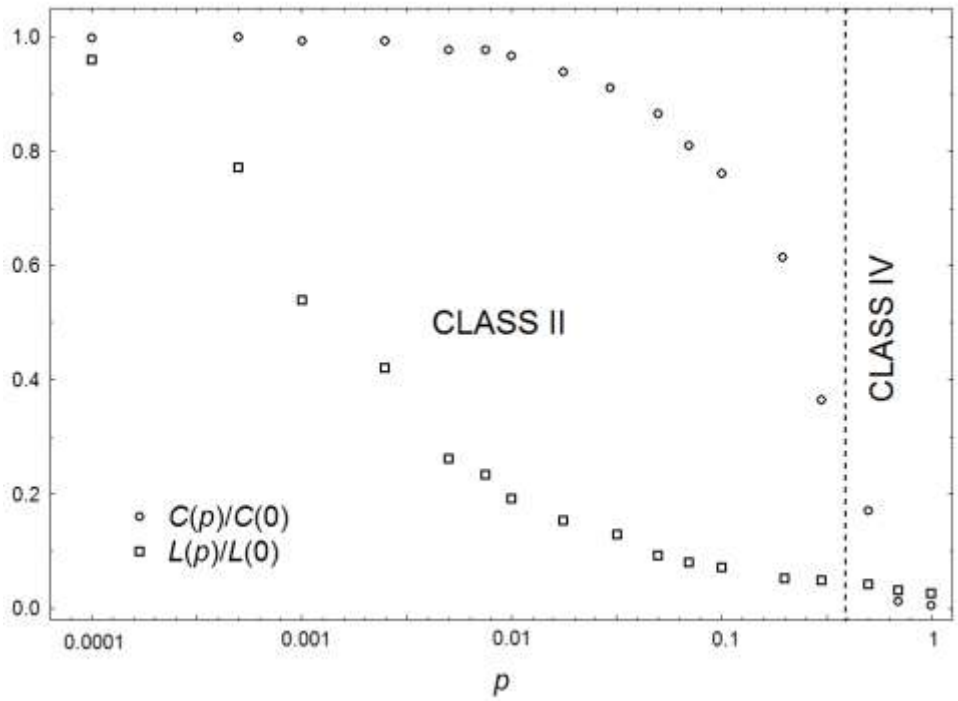


Fig. 6

A)


Proteins
BA model
ER model
<k>=4
<k>=6
<k>=8
<k>=10
<k>=12
<k>=16
ln(ξ⁺+10⁻⁵)
ln(ξ⁻+10⁻⁵)

B)


CLASS II
CLASS IV
C(p)/C(0)
L(p)/L(0)
p