A Study of Protein Structure using Amino Acid Interaction Networks

Omar Gaci, Stefan Balev

X

A study of protein structure using amino acid interaction networks

Omar Gaci and Stefan Balev
University of Le Havre
France

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 (Orengo, 1997) and SCOP (Murzin.1995). 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:

1. The amino acid sequence of a protein's polypeptide chain is called its primary or one- dimensional (1D) structure. It can be considered as a word over the 20-letter letter amino acid alphabet.

2. Different elements of the sequence form local regular secondary (2D) structures, such as α-helices or β-strands.
3. The tertiary (3D) structure is formed by packing such structural elements into one or several compact globular units called domains.
4. 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 (Branden & Tooze, 1999).

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

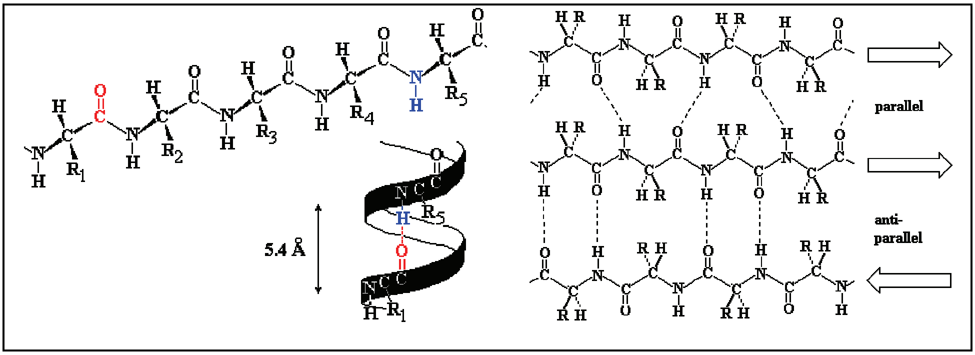

Fig. 1. Left: an α-helix illustrated as ribbon diagram, there are 3.6 residues per turn corresponding to 5.4 Å. Right: A β-sheet composed by three strands.

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

A β-sheet is build up from a combination of several regions of the polypeptide chain where hydrogen bonds can form between C'=O groups of one β strand and another NH group parallel to the first strand. There are two kinds of β-sheet formations, anti-parallel β-sheets (in which the two strands run in opposite directions) and parallel sheets (in which the two strands run in the same direction).

Based on the local organization of the secondary structure elements (SSE), proteins are divided in the following four classes (Levitt & Chothia, 1976):

1. All α, proteins have only α -helix secondary structure.
2. All β, proteins have only β -strand secondary structure.
3. α / β, proteins have mixed α -helix and β-strand secondary structure.
4. α + β, 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 (Murzin et al., 1995), and CATH, Class Architecture Topology Homology (Orengo et al., 1997). 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.

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) (Berman et al., 2000), 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 *Calpha* atoms. Considering the *Calpha* 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 x 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 (Ghosh et al., 2007). 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.

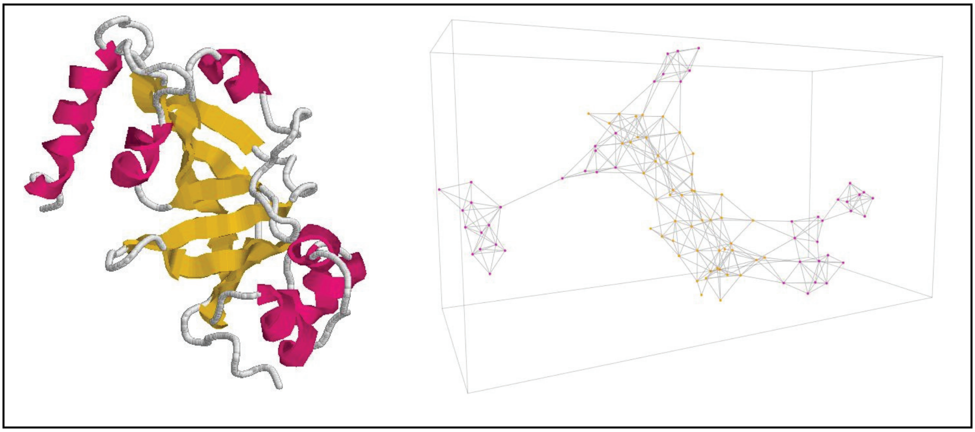

Fig. 2. Protein 1DTP (left) and its SSE-IN (right).

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 chapter. The reason of ignoring the amino acids not participating in SSE is simple. Evolution tends to preserve the structural core of proteins composed from SSE. In the other hand, the loops (regions between SSE) are not so important to the structure and hence, are subject to more mutations. That is why homologous proteins tend to have relatively preserved structural cores and variable loop regions. Thus, the structure determining interactions are those between amino acids belonging to the same SSE on local level and between different SSEs on global level. Fig. 2 gives an example of a protein and its SSE-IN.

In (Muppirala & Li, 2006) and (Brinda & Vishveshwara, 2005) the authors rely on similar models of amino acid interaction networks to study some of their properties, in particular concerning the role played by certain nodes or comparing the graph to general interaction networks models. Thanks to this point of view the protein folding problem can be tackled by graph theory approaches.

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

4. General Models of Interactions Networks

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

From physics and computer science to biology and social sciences, researchers have found that a broad variety of systems can be represented as networks, and that there is much to be learned by studying these networks. Indeed, the studies of the Web (Broder et al., 2000), of social networks (Wasserman & Faust, 1994) or of metabolic networks (Jeong et al., 2000) contribute to put in light common non-trivial properties of these networks which have *a priori* nothing in common. The ambition is to understand how the large networks are structured, how they evolve and what are the phenomena acting on their constitution and formation.

In this section we present the three main models of interaction networks by describing their specific properties. We also define several measures that we use in the next section in order to study SSE-IN empirically.

4.1 The Random Graph Model

The random graph models are one of the oldest network models, introduced in (Solomonoff & Rapoport, 1951) and further studied in (Erdõs & Rényi, 1959) and (Erdõs & Rényi, 1960). These works identify two different classes of random graphs, called Gn,ù and Gn,p and defined by the following connection rules:

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

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

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

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

The degree distribution is one of the important characteristics of this kind of networks because it affects their properties and behavior (Réka & Barabási, 2000). The random graph $G_{n,p}$ has a binomial degree distribution. The probability p_k that a randomly chosen vertex is connected to exactly k others is (Newman et al., 2001):

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

When n tends to infinity, this becomes:

$$p_k = \lim_{n \to \infty} \frac{n^k}{k!} (\frac{p}{1-p})^k (1-p)^n \approx \frac{z^k e^{-z}}{k!}$$

As we see in Fig. 3, Poisson distributions have different behavior for different mean degrees z. Each distribution has a clear peak close to k = z, followed by a tail that decays as 1/k! which is considerably quicker than exponential.

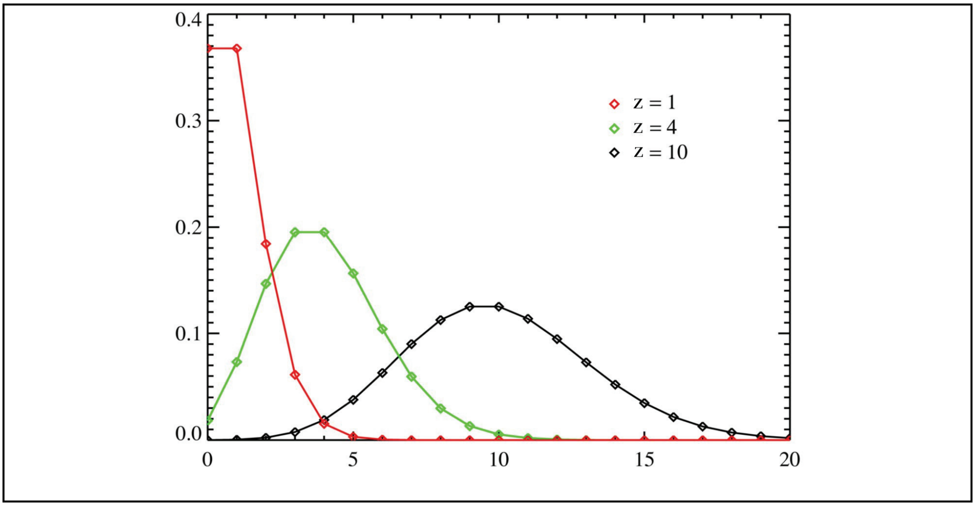

Fig. 3. Poisson distribution $p_k = \frac{z^k e^{-z}}{k!}$ with z = 1, 4 and 10.

4.2 Small-World Networks

This network model was introduced in (Watts, 1999) as a model of social networks. It has been since adopted to treat phenomena in physics, computer science or social sciences. The model comes from the observation that many real-world networks have the following two properties:

- The small-world effect, meaning that most pairs of vertices are connected by a short path through the network. This phenomenon has two explanations. First, the concept of “shortcuts” through a network allows to join two distant vertices by a small number of edges (Watts, 1999).

. Second, the concept of “hubs”, vertices whose connectivity is higher than others provide bridges between distant vertices because most vertices are linked to them.

- High “clustering'”, meaning that there is a high probability that two vertices are connected one to another if they share the same neighbor.

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

Definition 2 The characteristic path length (Watts, 1999), denoted L, of a graph G is the median of the means of the shortest path lengths connecting each vertex *v* to all other vertices. More precisely, let *d(v, u)* be the length of the shortest path between two vertices *v* and *u* and let $\overline{d_v}$ be the average of *d(v, u)* over all $u \in V$. Then the characteristic path length is the median of { $\overline{d_v}$ }.

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

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

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

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

Definition 3 The local clustering coefficient (Watts, 1999), C_v, of a vertex *v* with k_v neighbors measures the density of the links in the neighborhood of *v*.

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

where the numerator is the number of edges in the neighborhood of *v* and the denominator is the number of all possible edges in this neighborhood. The clustering coefficient *C* of a graph is the average of the local clustering coefficients of all vertices:

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

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

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

Watts and Strogatz (Watts, 1999) defined a network to be a small-world if it shows both of the following properties:

1. Small world effect: $L \approx L_{RG}$
2. High clustering: $C >> C_{RG}$

4.3 Scale-Free Networks

The most important property of scale-free systems is their invariance to changes in scale. The term "scale-free" refers to a system defined by a functional form *f(x)* that remains unchanged within a multiplicative factor under rescaling of the independent variable *x*. Indeed, this means power-law forms, since these are the only solutions to *f(an) = b f(n)*, where *n* is the number of vertices (Newman, 2002). The scale-invariance property means that any part of the scale-free network is stochastically similar to the whole network and parameters are assumed to be independent of the system size (Jeong et al., 2000).
If n_k is the number of vertices having the degree *k*, we define p_k as the fraction of vertices that have degree *k* in the network:

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

The degree distribution can be expressed *via* the cumulative degree function (Newman, 2002), (Erdõs & Rényi, 1959):

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

which is the probabilty for a node to have a degree greater or equal to *k*.
By plotting the cumulative degree function one can observe how its tail evolves, following a power law or an exponential distribution.
The power law distribution is defined as following (Newman, 2002):

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

and the exponential distribution is defined by the next formula:

$$P_k \approx \sum_{k'=k}^{\infty} e^{-k'/\alpha} \approx e^{-k/\alpha}$$

Between this two distributions, there is a mixture of them where the distribution has a power law regime followed by a sharp cut-off, with an exponential decay of the tail, expressed by the next formula:

$$P_k \approx \sum_{k'=k}^{\infty} k'^{-\alpha} e^{-k'/\alpha} \approx k'^{(\alpha-1)}\, e^{-k/\alpha}$$

Like a power law distribution, it decreases polynomially, so that the number of vertices with weak degree is important while a reduced proportion of vertices having high degree exists. The last are called "hubs" that is sites with large connectivity through the network, see Fig. 4. The scale-free model depends mainly on the kind of degree distribution, thus a network is defined as a scale-free if:

- The degree distribution is a power law distribution P(k) ≈ $k^{-\alpha}$ over a part of its range.
- The distribution exposant satisfies $2 < \alpha \leq 3$ (Goh et al., 2001).

Amaral and al (Amaral et al., 2000) have studied networks whose cumulative degree distribution shape lets appear three kinds of networks. First, scale-free networks whose distribution decays as a power law with an exposant α satisfying bounds seen above. Second, see Fig. 4, broad-scale or truncated scale-free networks whose the degree distribution has a power law regime followed by a sharp cutoff. Third, single-scale networks whose degree distribution decays fast like an exponential.

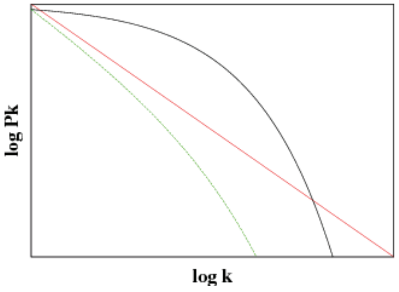

Fig. 4. Degree distribution described in (Amaral et al., 2000). The red line follows a power law, as for scale-free networks. The green line corresponds to truncated scale-free networks. The black curve corresponds to single-scale networks.

4.4 Topological Measures

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

Diameter and mean distance The distance in a graph *G = (V,E)* between two vertices *u,v* ∈ *V*, denoted by *d(u,v)*, is the length of the shortest path connecting *u* and *v* (Diestel, 2000). 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 (Diestel, 2000):

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

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

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

Density 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:

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

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 (Coleman & Moré, 1983).

Clustering coefficients Watts and Strogatz proposed a measure of clustering (Watts, 1999) and defined it as a measure of local vertices density, thus for each node *v*, the local clustering around its neighborhood 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 neighborhood. If we extend the previous defintion to the entire graph, the clustering is given by the expression:

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

The last defintion is mainly local because for each node, it involves only its neighborhood. The global clustering was studied by Newman et al. (Newman, 2001) and can be mesuared by the following formula:

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

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. A First Topological Description

In this section, we present a publication (Gaci & Balev, 2008b) where we study the protein SSE-IN behavior. We want to observe how proteins from a same structural family provide similar SSE-IN according to their topological properties. To do that, we propose topological measures which we apply on a sample of proteins to put in evidence the existence of equivalence between structural similarity and topological homogeneity in the resulting SSE-IN.
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
RossmannFold	CATH	α β	2576
TIM Barrel	CATH	α β	1051
Lysozyme	CATH	Mainly α	871
Globin-like	SCOP	All α	733
TIM β/α-barrel	SCOP	α/β	896
Lysozyme-like	SCOP	α+ β	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
RossmannFold	18.84
TIM Barrel	19.83
Lysozyme	12.81
Globin-like	15.65
TIM β/α-barrel	20.09
Lysozyme-like	12.85

Table 2. Average diameter for each family.

Figure 5 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 *z* 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* / *z* is about 2.5 for all the families. The last property is a characterization of all proteins' SSE-IN.

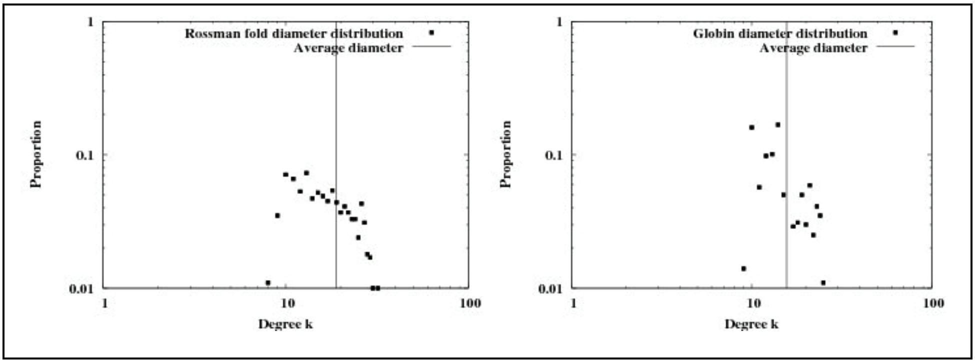

Fig. 5. Average diameter of Rossman fold, left, and beta/alpha-barrel, right.

Name	*z*
RossmannFold	7.26
TIM Barrel	7.79
Lysozyme	4.99
Globin-like	6.64
TIM β/α-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 2 and 3 one can see the inversely proportional relation between density in one hand, and diameter and average distance on the other.

Name	$\delta(G)$
RossmannFold	0.033
TIM Barrel	0.03
Lysozyme	0.038
Globin-like	0.034
TIM β/α-barrel	0.029
Lysozyme-like	0.042

Table 4. Average density for each family.

The mean degree, *z*, 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	z
RossmannFold	7.2
TIM Barrel	7.17
Lysozyme	6.82
Globin-like	7.69
TIM β/α-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 6. We can remark that the curves follow a power law distribution and can be approximated by the following power-law function:

$$p_k = 141.29k^{-\alpha}, where \alpha \approx 2.99$$

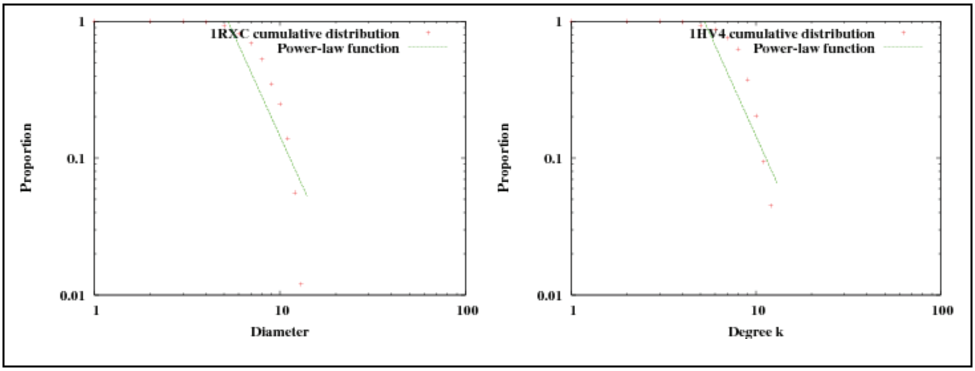

Fig. 6. Cumulative degree distribution for 1RXC from Rossman fold, left, and 1HV4 from TIM beta/alpha-barrel, right.

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 7, follows a Poisson distribution whose peak is reached for a degree near *z*. 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.

5.4 Clustering Coefficients

The local clustering Clocal measures the fraction of pairs of a vertex's neighbors and the global clustering Cglobal 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.

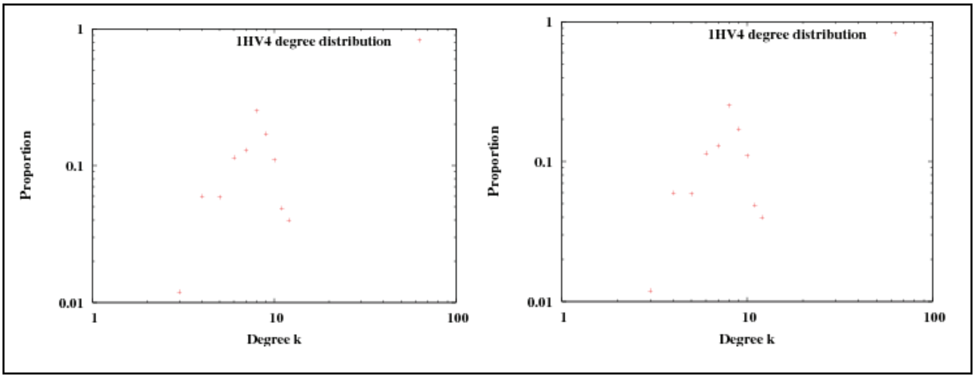

Fig. 7. Degree distribution for 1RXC from Rossman fold, left, and 1HV4 from TIM beta/alpha-barrel, right.

Name	*z*
RossmannFold	7.2
TIM Barrel	7.17
Lysozyme	6.82
Globin-like	7.69
TIM β/α-barrel	7.15
Lysozyme-like	6.81

Table 6. Clustering coefficients for each family.

5.5 Consequences

In this first part 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 discriminating 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.

6. Comparisons with the Scale-Free Networks

We have worked on the entire PDB file available the 18th October 2007, that is 46679 files. We lead our simulations by considering all proteins SSE-IN without limiting us on a particular protein family or basing us on a classification. Indeed, our goal is to identify a general model for proteins SSE-IN and so that we don't discriminate proteins due to their biological function. Thus, despite the heterogeneity we can expect among folded proteins, we assume that a general model of Interaction Networks emerge and can be identified from protein SSE-IN which came from combinaison of the 20 amino acids.

6.1 General Behavior

We compute the cumulative degree distribution for all proteins SSE-IN, a sample of our results is presented on Fig 8. We can remark that the curves describe a power law regime followed by the sharp cut-off. The power law function is expressed as following:

$$p_k = 213.413 k^{-\alpha}, where \alpha \approx 3.2$$

while the distribution is approximated by the next function:

$$P_k = 1.48347 k^{0.962514} \exp^{-k/2.12615}$$

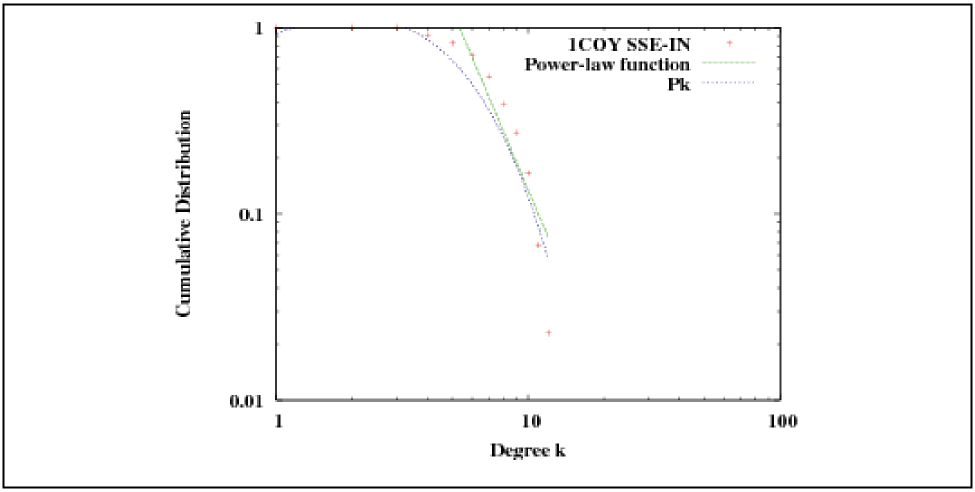

Fig. 8. Cumulative degree distribution for protein 1COY SSE-IN.

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

First, we are interested in the degree distribution and mainly its shape, see Fig 9. We can see that degree distribution follows a Poisson distribution whose peak is reached for a degree near z. This result provides precision about how the vertices are connected within SSE-IN. It implies that the degree of the vertices is homogenous. In other words, a major part of them has a connectivity enough close to the mean degree. Consequently, the cumulative distribution depends on the mean degree value which acts as a threshold beyond which it decreases as an exponential since it's approximed via P_k.

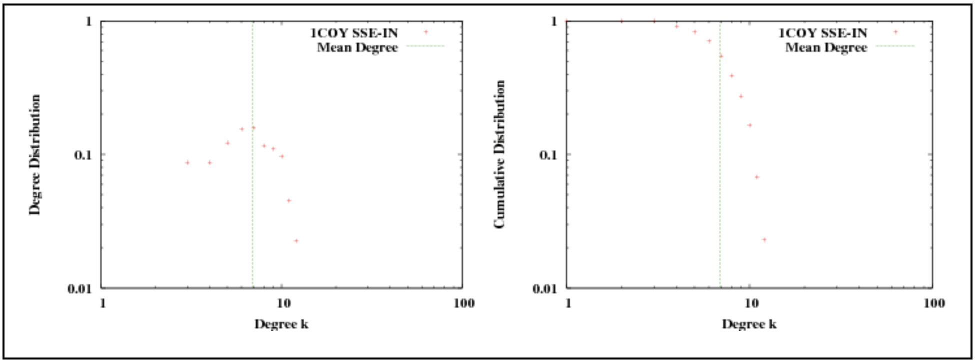

Fig. 9. Degree and cumulative distribution for 1COY SSE-IN. They decrease for degree values greater than the mean degree.

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

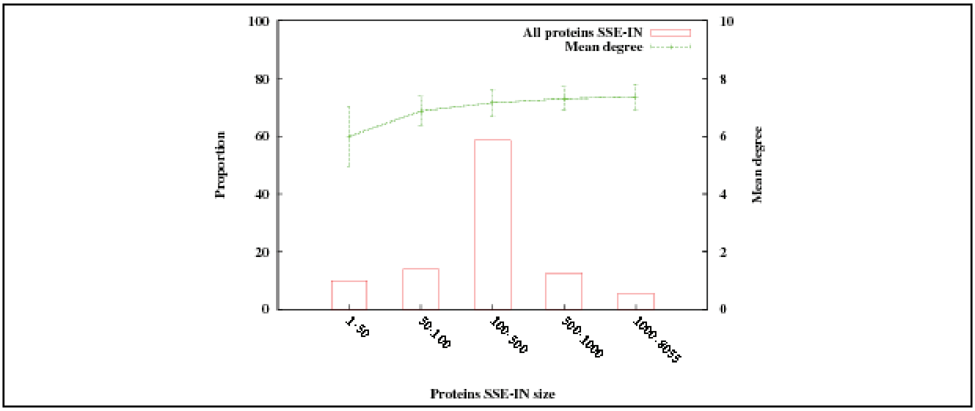

Fig. 10. Mean degree distribution according to protein SSE-IN size. It evolves with values enough close, between 6 and 8.

To illustrate the mean degree homogeneity we choose two proteins, namely 1SE9 and 1AON with size respectively 50 and 4988. Their size ratio is approximately 100. Even if the mean degree are slightly different, the distributions are very similar, see Fig 11.

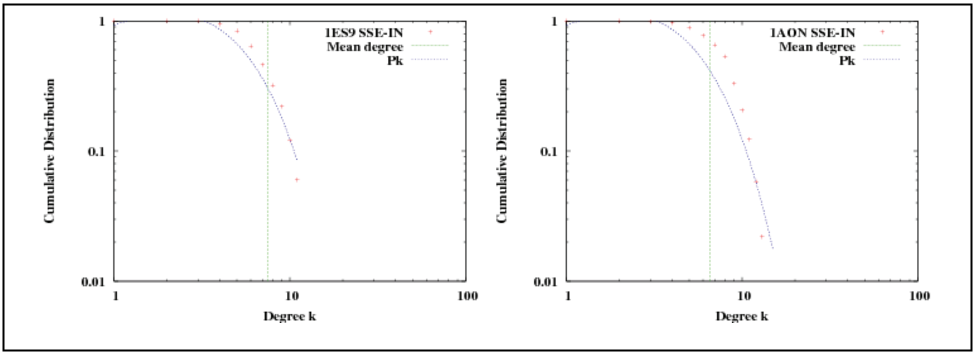

Fig. 11. Cumulative degree distribution of 1SE9 and 1AON SSE-IN whose size equal 50 and and 4988. Despite their important size difference, their mean degree stay close and worth respectively 6.6 and 7.5

To recapitulate, we show that the mean degree values constitute a threshold for protein SSE-IN cumulative degree distribution. For degrees lower than the mean degree it decreases slowly and after this threshold its decrease is fast compared to an exponential one, as shown Fig 8, 9 and 10.

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

Since the degree distribution depends on the mean degree value, we compare for each node its degree as function of z, see Fig 12, to illustrate how nodes interact and in particular to highlight the weak fraction of highly connected nodes, also called hubs, less than 5 % of the total node number.

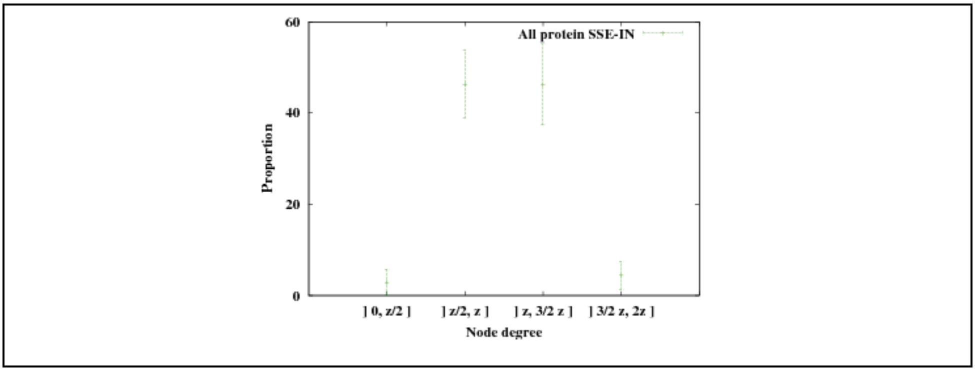

Fig. 12. Degree of nodes in protein SSE-IN as function of z. Less than 5 % of nodes are hubs that is having a 'high' connectivity.

An interesting study is to put in evidence the biological properties of nodes whose connectivity is marginal. Thus, we search to identify nodes which highly interact with their neighbors in the folded proteins. To do that, we have proceeded by grouping the proteins according to their secondary structure. Indeed, we have already shown (Gaci & Balev, 2008b) that the protein SSE-IN topologies from structural classifications are homogeneous and established a parallel between structural and topological properties. Based on the SCOP classification and more precisely on the fold families, we have selected a total of 18294 proteins, see Table 7, and studied their SSE-IN to describe the specificities of the hubs.

Class	Number of Families	Number of Proteins
All Alpha	12	2968
All Beta	17	6372
Alpha / Beta	18	5197
Alpha + Beta	16	3757

Table 7. Structural families studied for the Scale-Free properties. We choose only families which count more than 100 proteins, for a total of 18294 proteins.We have worked with the SCOP 1.7.3 classification, a protein belongs to a SCOP fold level if all its domains are the same.

For each protein SSE-IN from a SCOP fold level we identify the nodes whose degree is high, see Fig 12. By grouping nodes as a function of amino acids they represent, we finally obtain the amino acid connectivity score by fold families. By repeating this process at the SCOP class level, we calculate and normalize the cumulated connectivity level of amino acids playing the role of hubs, see Fig 13.

If we consider that the role played by an amino acid inside a folded protein is equivalent to its interaction degree, then these plots show that despite a functional diversity between the 4 SCOP classes, there are globally the same amino acids which interact most, namely the Ala, Cys, Gly, Leu, Val. Therefore, the amino acids having a high connectivity interact independently of the protein biological function, we will explain this behavior further. Thus, the figure shows clearly that the amino acid Ala plays a central role in the studied proteins independently of the classification of their secondary structure.

We also compute the occurrence level of hubs, that is their number of appearances on each protein SSE-IN. We cumulated this score at the SCOP class level to obtain the probability for each amino acid to be a hub having a high connected occurrence in a protein SSE-IN, see Fig 14.

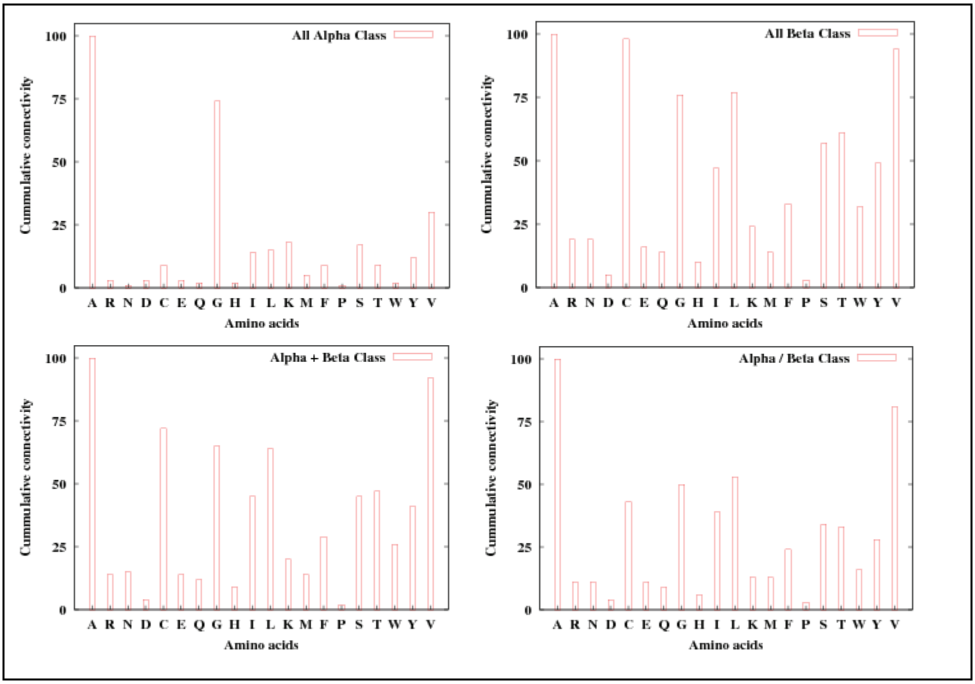

Fig. 13. Normalisation of the amino acid cumulative connectivity when they act as hubs for each SCOP class level. The amino acid Ala interacts higher than the others independently of proteins classification.

We can remark the existence of peaks which show clearly a higly tendency of amino acids Ala, Cys, Gly, Leu and Val to have a high interactivy within the folded protein. Thus, there are the same amino acids which play the role of hubs independently of the structural family, we provide below the cause of this result.

Now, we want to describe the way in which the hubs appear in the folded protein. Indeed, we study the distribution of the hubs position as a function of the structural class of the SSE-IN to identify variations dependent or not on the biological function of proteins. To make this study, we attribute an incremental position so that the H extremity has a position 100. Then, each time a hub exists in a SSE-IN, we increment its position occurence number and finally normalize by the maximum to obtain the occurence rate of hubs according to their positions in the SSE-IN, see Fig. 15.

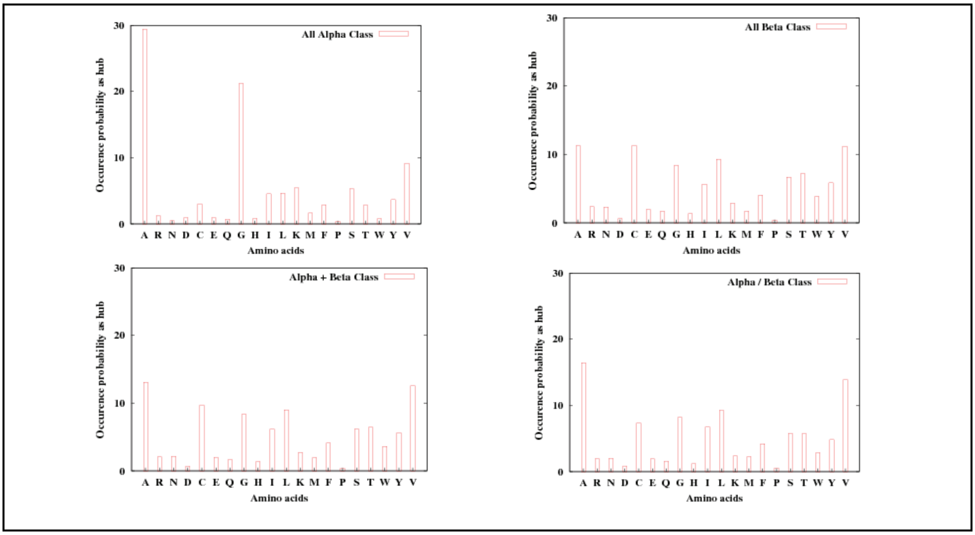

Fig. 14. Occurrence probability of hub amino acids for each SCOP Class level. The probability that an amino acid ALA has a high interaction level is superior than the others independently of proteins classification.

The results show the existence of favorable region in which the hub apparition is higher than somewhere else. This favorable localization is strongly visible for the All Alpha class where the hubs have a tendency to interact around the positions 20, 40 or 80. The distribution of the hub positions is the most homogenous for the alpha/beta class which involve a dependence to the SSE-IN topology since it not possible to find more than one strong favorable area which appear around the position 60.

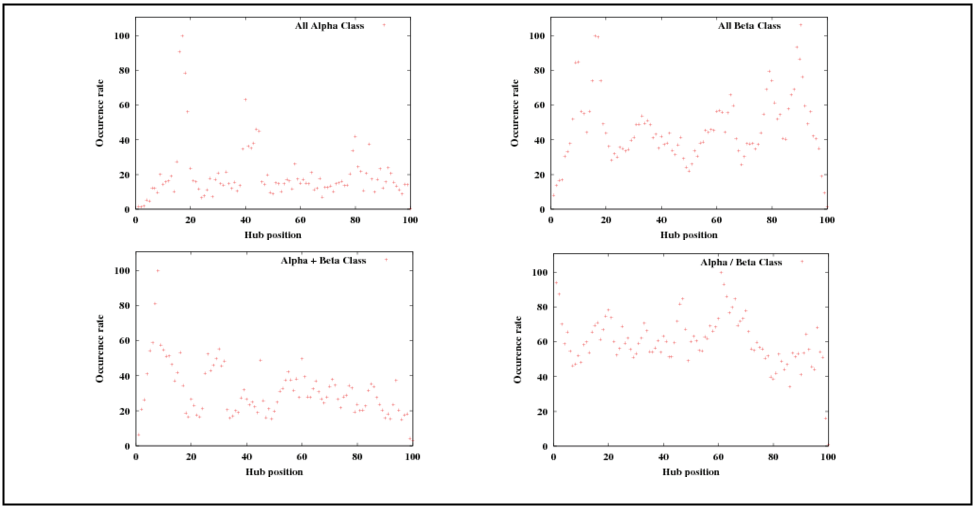

Fig. 15. Occurrence rate of hubs according to their position for each SCOP Class level.

To illustrate the existence of hubs favorable localization in the SSE-IN, we rely on the rich-club phenomenon (Colizza et al., 2006) according to which highly connected nodes have tendancy to be connected to one another. We compute the rich-club connectivity of a hub as the ratio of the number of links to the maximum number of links between nodes belonging to the rich-club.

It appears, see Fig 16, that certain hubs are isolated mainly when the rich-club connectivity is low (position 0, 30, 60 for All Alpha SSE-IN) where as the favorable hub localizations correspond to a high coefficient.

The main observations about hubs behaviour are the following. First, there are four amino acids which have higher probability to have a high connectivity. Second, hubs have tendency to act in particular region in the SSE-IN. These two observations lead us to compute only the occurrence rate of the most frequently encountered hub Ala according to its positions, see Fig 17.

By comparing the figure 15 and 17, it appears clearly that the most highly occurrence rate for the 4 classes correspond to the position of the amino acid Ala. Therefore, we can establish a relation between amino acid position in the primary structure and hub apparition. Then, certain amino acid that is the Ala, Cys, Gly, Leu and Val act as hub because they appear in the protein sequence in favourable region which involve a high interaction within the folded structure.

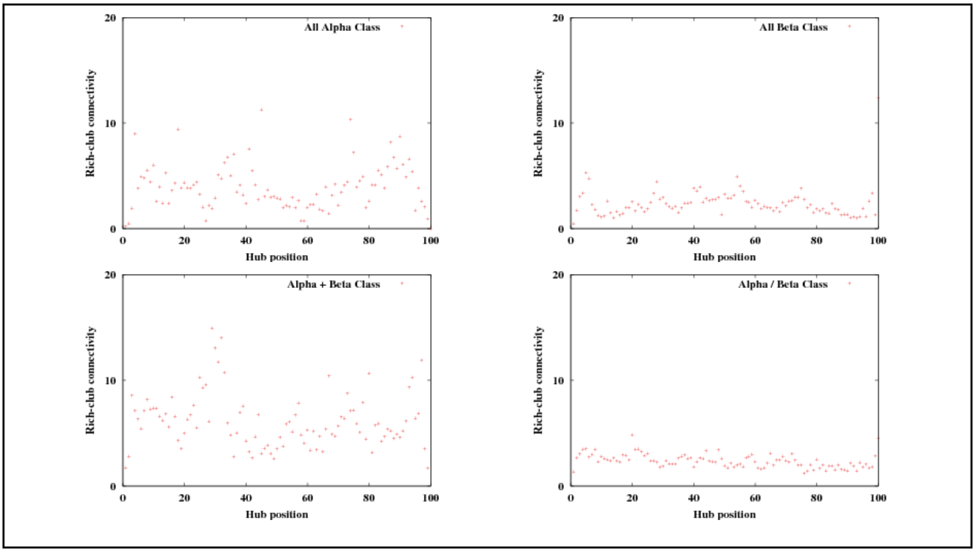

Fig. 16. Hub rich-club connectivity as function of hub position for each SCOP class level.

6.2 Mean Degree Evolution

Since the mean degree plays the role of a threshold beyond which the cumulative degree distribution decreases exponentially, it is interesting to study its evolution with the size of the network, see Fig. 10. It appears that the mean degree increases very slightly with the size of the network. Even for networks with size ratio of 100, the mean degree ratio is only 1.15, see Fig. 11.

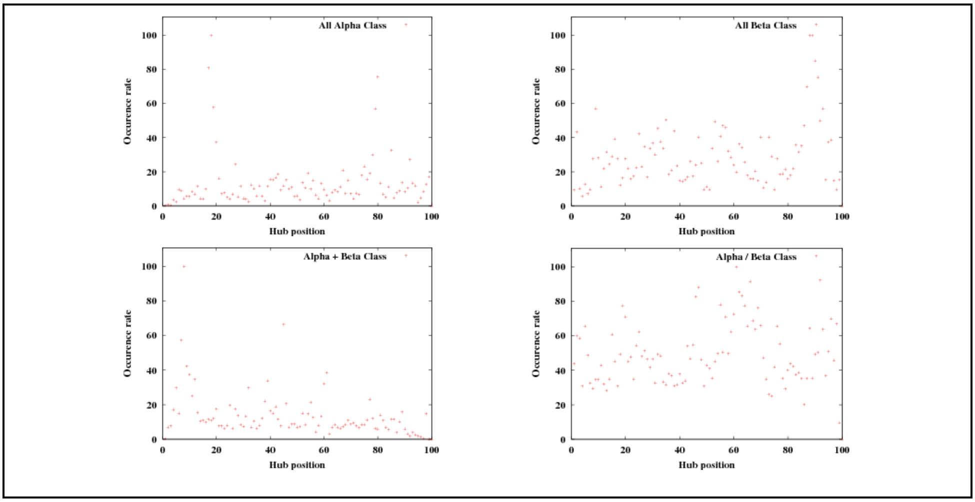

Fig. 17. Occurrence rate of the hub amino acid Ala for each SCOP Class level.

Whatever the size of the network is, we observe that the mean degree is always between 5 and 8. This mean degree interval is a common property characterizing all SSE-IN. In order to explain this property, let us consider the structure of our networks. They are composed of densely connected subgraphs corresponding to secondary structure elements (see Fig. 18. The number of edges connecting different subgraphs is relatively small, but these edges are the most important, since they correspond to interactions determining the tertiary structure.
We start by computing the mean degree in each SSE subgraph. The results are shown on Fig. 19. We can see that the mean degree evolution at microscopic level is almost the same as at macroscopic level (compare to Fig. 10). Independently of the SSE size and type, the mean degree of each SSE subgraph, z_{SSE} , is always bounded:

$$z_{\min} < z_{SSE} < z_{\max} \tag{1}$$

when the size of the network is more than 10. In the general case $z_{min} = 5$ and $z_{max} = 8$, but when we consider a specific SSE size and type, finer bounds can be found (see Fig. 19).

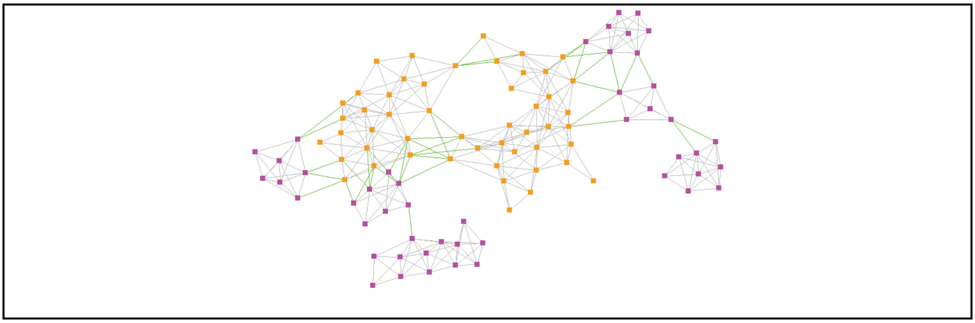

Fig. 18. SSE-IN of 1DTP protein. The edges connecting different SSE are green.

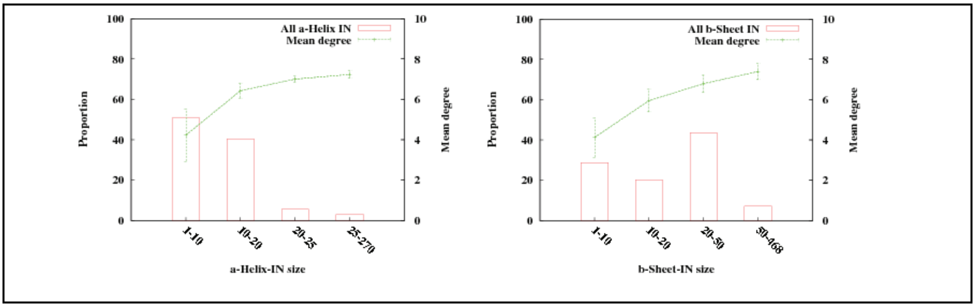

Fig. 19. SSE subgraphs size distribution and mean degree as a function of the size.

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

$$\frac{m}{n} = \frac{\sum_{i=1}^{s} m_i + m_{\text{int}\,er}}{\sum_{i=1}^{s} n_i} = \frac{\sum_{i=1}^{s} m_i}{\sum_{i=1}^{s} n_i} + r\frac{m}{n} \tag{2}$$

and hence for the mean degree *z* we have

$$z = \frac{2m}{n} = \frac{2}{1-r}\frac{\sum_{i=1}^{s} m_i}{\sum_{i=1}^{s} n_i} \tag{3}$$

On the other hand, from (1) it follows that

$$\frac{z_{\min}}{2} n_i < m_i < \frac{z_{\max}}{2} n_i, i = 1,\ldots,s \tag{4}$$

By summing up the last equation we obtain

$$\frac{z_{\min}}{2} < \frac{\sum_{i=1}^{s} m_i}{\sum_{i=1}^{s} n_i} < \frac{z_{\max}}{2} \tag{5}$$

which together with (3) gives

$$\frac{z_{\min}}{1-r} < z < \frac{z_{\max}}{1-r} \tag{6}$$

The last equation gives finer bounds on the mean degree. It shows that the bounds on z depend not only on the bounds on z_{SSE}, but also on the ratio of inter-SSE edges. A higher proportion of inter-SSE edges shifts up the bounds. Proteins with bigger size have more SSEs and hence more links between different SSEs. This explains the increase of the mean degree with the size of the networks. Fig. 20 shows that the number of inter-SSE edges is quite variable, but is bounded and does not exceed *20%*, it's the consequence of the excluded volume effect, since the number of residues that can physically reside within a given radius is limited.

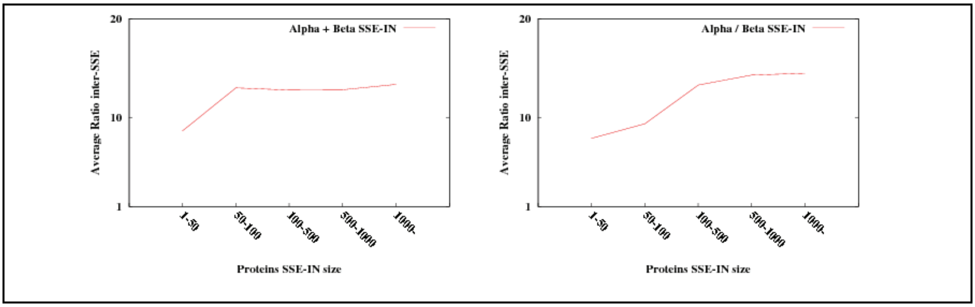

Fig. 20. Ratio of inter-SSE edges (*r*) as a function of the network size for the four classes of studied proteins.

6.3 Consequences

During our study, we have identified a means to approximate the cumulative degree distribution of the SSE-IN. This function describes a power law regime followed by a sharp cut-off. The SSE-In studied are consequently truncated scale-free networks. The study of the hubs shows that certain amino acid play a central role independently of the protein classification. The nature of these hubs depends on their positions in the primary structure.

7. Comparisons with the Small-World Networks

In this section we present an empirical characterization of SSE-IN properties. In order to choose our data sample, we have used the SCOP classification. We have worked with the SCOP 1.7.3 files. We have computed the measures from the section (4.2.) for the four mains classes of SCOP (see Table 7). Each class provides a broad sample guarantying more general results and avoiding fluctuations. Moreover, these four classes contain proteins of very different sizes, varying from several dozens to several thousands amino acids in SSE. The results obtained for the different classes are very similar, that is why in the rest of this section we show only the results for

In a previous work (Gaci & Balev, 2008a), we have studied the two properties of SSE-IN related to the small-world model, namely their L/L_{RG} and C/C_{RG} ratios. Our results show (see Fig. 21) that nearly 60% of the proteins have SSE-IN consisting of between 100 and 500 amino acids. The small-world properties are satisfied mainly when the size of the network does not exceed 500 amino acids and there are about 15.3% small-world networks among all SSE-IN. Fig. 4 explains the reason for this low rate. One can see that although highly clustered, most SSE-IN do not satisfy the first small-world property.

To explain the results presented on Fig. 22, note that the mean degree z is not very different from one SSE-IN to another and is generally independent from the size. When the mean degree is fixed, the characteristic path length of a random graph grows like *log n* and its clustering coefficient has n^{-1} behavior. We can see that there is no clear relationship between the characteristic path length and the SSE-IN size. There are proteins with close sizes but very different path lengths. However, in general it grows faster than logarithmically with the size. The figure also shows that SSE-IN are very highly clustered. The C/C_{RG} has clearly linear behavior, hence the clustering coefficient (like the mean degree) is independent from the size.

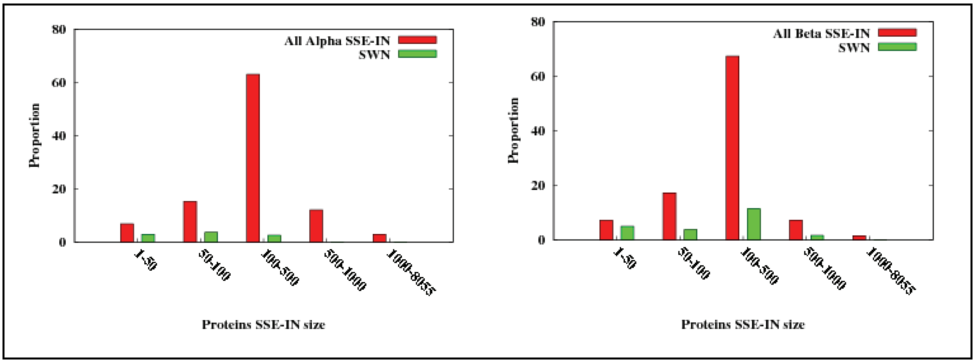

Fig. 21. Size distribution of proteins SSE-IN and small-world networks ratio.

On the other hand, the subnetworks corresponding to single SSEs are almost all small world networks. The last observation brings us to consider the edges whose extremities belong to different SSEs (see Fig. 18). These "shortcuts" represent the interactions between different SSEs and they determine the tertiary protein structure. They provide short paths between different network regions and bigger number of shortcuts implies smaller characteristic path length.

Computing the average connected component size, we observe that this size does not exceed 100 for the smallworld networks (see Fig. 23). Consequently, we identify a necessary condition for SSE-INs to be small-world, their average connected component size has to be less than 100.

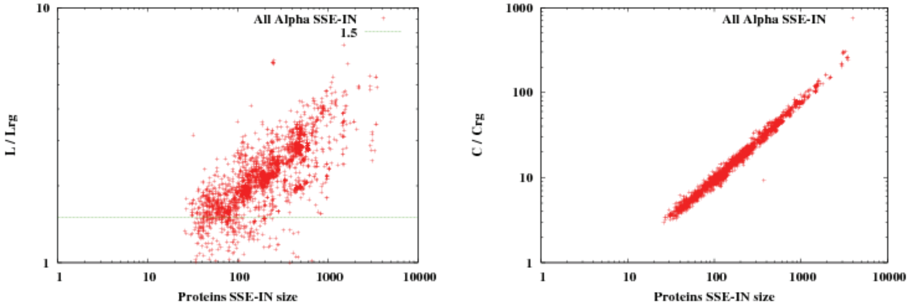

Fig. 22. L/L_{RG} (left) and C/C_{RG} (right) ratios as a function of SSE-IN size.

In (Gaci & Balev, 2008a) we show that the protein SSE-IN components interact with each other in the same way, no matter if they are small-world or not, see Fig 23. Then, it is interesting to study not the quantity of interaction inter-SSE but rather their quality, that is we want to put in evidence the role of nodes allowing interaction between different SSEs.
In graph theory, the betweenness of a node is defined as the total number of shortest paths between pairs of nodes that pass through this node. It measures the influence of a node in a network. The betweenness of a node t, denoted B(t) is defined as follows:

$$B(t) = \sum_{u \neq v, u \neq t, v \neq t} \frac{\sigma_{uv}(t)}{\sigma_{uv}}$$

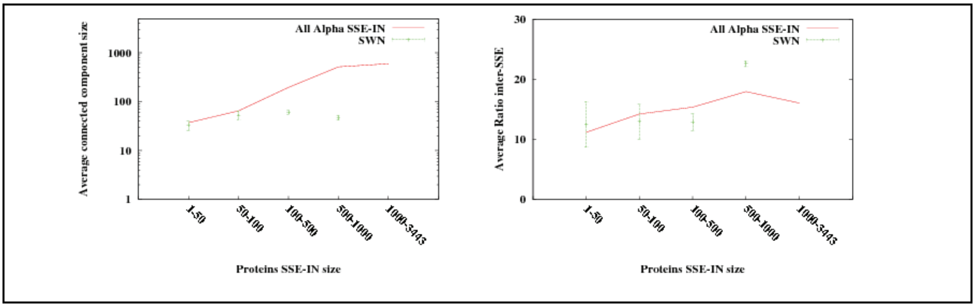

Fig. 23. Average connected component size (left) and shortcut ratio (right) as a function of protein SSE-In size.

where σ_{uv} is the number of shortest paths between the nodes u and v, and σ_{uv} *(t)* is the number of shortest paths between u to v that pass through t.
The average betweenness of nodes interacting with other secondary structure motifs is shown on Fig 24. We can see that the average betweenness of the small-world networks is similar to the average betweenness of the other networks. This means that neither the quantity nor the quality of nodes that link different SSEs guarantees a short characteristic path length.
To better understand the properties of SSE-IN which are small-world, we introduce the secondary structure components interaction network, SSC-IN. This is a network in which each SSE is contracted to a single node. There is an edge between two SSEs s_1 and s_2 if theres is at least one interaction between an amino acid belonging to s_1 and an amino acid

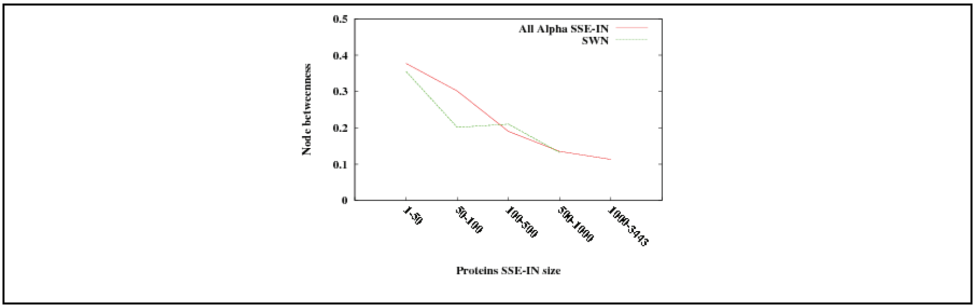

Fig. 24. Betweenness of nodes which link different SSEs.

is at least one interaction between an amino acid belonging to s_1 and an amino acid belonging to s_2. A weight is associated to each edge. The weight of the edge between s_1 and s_2 is the average distance (measured in the corresponding SSE-IN) between all pairs of amino acids belonging to s_1 and s_2. SSC-IN is an abstraction of SSE-IN which puts in evidence the interactions between SSEs. Fig. 25 shows an example of SSC-IN. We computed the characteristic path lengths and the clustering coefficients for all protein SSC-IN, the results show once again that the small world effect is the discriminant property in the SSC-IN. Indeed, for a small-world SSC-IN the characteristic path length grows linearly with the size of the underlying SSE-IN and it never exceeds *4*. The last observation can be correlated with the necessary condition for a SSE-IN to satisfy the small world model. Indeed, the average connected component size is bounded by *100* nodes in small world SSEIN. This has for consequence that the corresponding SSC-IN has bounded average connected component size and therefore a bounded characteristic path length. Then, a protein SSE-IN with size $n > 25$ is small world if and only if its SSC-IN characteristic path length is on the line shown in Fig 26

7.1 Consequences

In this study, we put in evidence in which proportion the SSE-IN satisfy the small-world model. Their constituent that is the subgraphs representing the SSE are in the most majority small-world.

The interactions between these SSE being limiting, it not allows to conserve a short distance when the SSE-IN size increases. Thus, to a SSE-IN belongs to the small-world model, it necessary that the average connected component size doesn't exceed 100 residues. Further, we consider a SSE-IN abstraction, called SSC-IN in which each node represents a SSE. Then, their study provides an alternative way to verify the SSE-IN behavior. Indeed, the characteristic path length in the SSC-IN is enough to infer to deduct if the SSE-IN belongs or not to the small-world model.

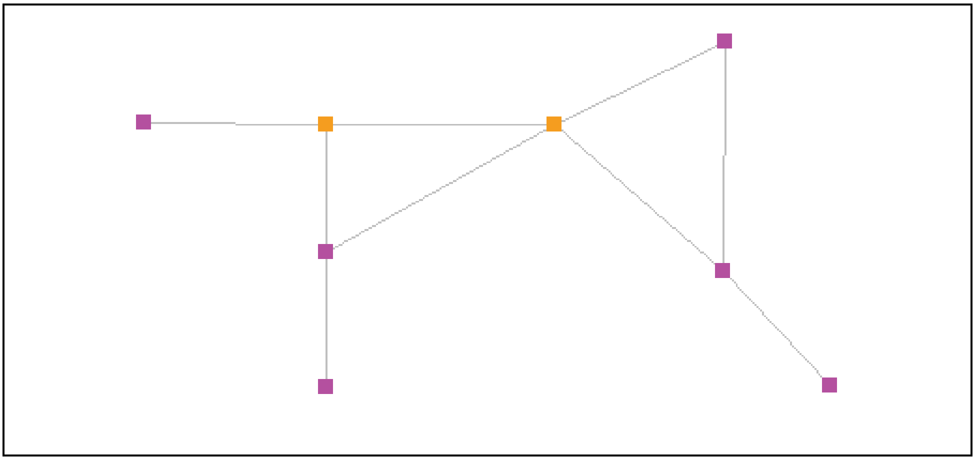

Fig. 25. Protein 1DTP SSC-IN. The edges weight corresponds to the mean distance between the two motifs in the SSE-IN.

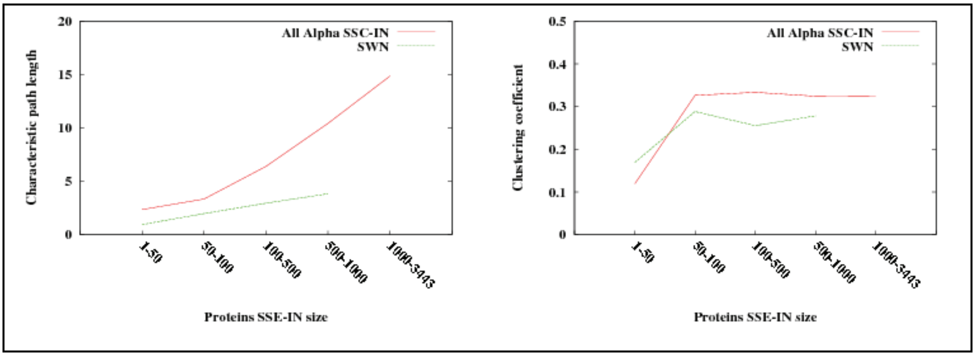

Fig. 26. Characteristic path length in SSC-IN. The distance in SW SSC-IN grows linearly and is bounded by 4.

8. Conclusion

In this chapter, we exploit the possibilities offer by a modelisation using interaction networks. The main hypothesis we use is the conservation from structural toward topological properties. The comparison of the SSE-IN with the general model of interaction networks allows us to describe the consequence of the folding dynamic.

The characterization we propose constitutes a first step of a new approach to the protein folding problem. The properties identified here, but also other properties we plan to study, can give us an insight on the folding process. They can be used to guide a folding simulation in the topological pathway from unfolded to folded state.

9. References

Amaral, L. A. N.; Scala, A.; Barthélémy, M. & Stanley, H. E. (2000). Classes of small-world networks. *Proc. Natl. Acad. Sci USA*, Vol. 97, No. 21, pp. 11149-11152

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. & Bourne., P. E. (2000). The protein data bank. *Nucleic Acids Research*, Vol. 28, pp. 235--242

Branden, C. & Tooze, J. (1999). *Introduction to protein structure*, Garland Publishing.

Brinda, K. V. & Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophys J*, Vol.89, No.6, pp. 4159-4170

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A. & Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, Vol.33, pp.309-320

Coleman, T. F. & Moré, J. J. (1983). Estimation of sparse jacobian matrices and graph coloring problems. *SIAM Journal on Numerical Analysis*, Vol.20, pp. 187-209

Colizza, V.; Flammini, A.; Serrano, M. A. & Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, Vol. 2, pp. 110

Diestel, R. (2000). *Graph Theory*. Springer Verlag, Princeton, New Jersey

Erdõs, P. & Rényi, A. (1959). On random graphs {I}. *Publicationes Mathematicae*, Vol. 6, pp. 290-297

Erdõs, P. & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, Vol. 7, pp. 17

Gaci, O. & Balev, S. (2008a). A general model for amino acid interaction networks. *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 34, pp. 401-405

Gaci, O. & Balev, S. (2008b). Proteins: From structural classification to amino acid interaction networks. *Proceedings of BIOCOMP'08*, pp. 728-734, Las Vegas, CSREA Press

Ghosh, A.; Brinda, K. V. & Vishveshwara, S. (2007). Dynamics of lysozyme structure network: probing the process of unfolding. *Biophys J*, Vol. 92, No. 7, pp. 2523-2535

Goh, K-.I.; Kahng, B. & Kim, D. (2001). Universal behavior of load distribution in scale-free networks. *Phys. Rev.*, Vol. 87

Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N. & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, Vol. 407, No. 6804, pp. 651-654

Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, Vol. 261, pp. 552-558

Muppirala, U. K. & Li, Z. (2006). A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues. *Protein Eng Des Sel*, Vol. 19, No. 6, pp. 265-275

Murzin, A. G.; Brenner, S. E.; Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of the protein database for the investigation of sequence and structures. *J. Mol. Biol.*, Vol. 247, pp. 536-540

Newman, M .E. J. (2001). The strucuture of scientific collaboration networks. *Proc. Natl. Acad. Sci USA.*, Vol. 98, pp. 404-409

Newman, M. E. J. (2002). The structure and function of networks. *Computer Physics Communications.*, Vol. 147, pp. 40-45

Newman, M. E. J.; Strogatz, S. H. & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, Vol. 64

Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T. ; Swindells, M. B. & Thornton, J.~M. (1997). CATH - a hierarchic classification of protein domain structures. *Structure*, Vol. 5, pp. 1093-1108

Réka, A. & Barabási, A.-L. (2000). Topology of evolving networks: local events and universality. *Physical Review Letters*, Vol. 85, No. 24

Solomonoff, R. & Rapoport, A. (1951). Connectivity of random nets. *Bull. Math. Biophys.*, Vol. 13, pp. 107–111

Wasserman, S. & Faust, K. (1994). Social network analysis : methods and applications, In Structural analysis in the social sciences, Cambridge University Press, Cambridge, 1994

Watts, D. J. (1999). *Small Worlds*, Princeton University Press, Princeton, New Jersey, 1999