Building a Parallel Between Structural and Topological Properties

Omar GACI and Stefan BALEV

Le Havre University
LITIS EA 4108, BP 540, 76058 Le Havre - France
omar.gaci@univ-lehavre.fr

Summary. *In this chapter, we study the amino acid interaction networks. This is a graph whose vertices are the protein's amino acids and whose edges are the interactions between them. Using a graph theory approach, we identify a number of properties of these networks. Some of them are common to all proteins, while others depend on the structure arrangement. We rely on this last group of properties to illustrate the correlation between structural and topological properties. Then, we propose a topological space where proteins from a same family tend to be grouped.*

1 Introduction

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 study, we treat proteins as networks of interacting amino acid pairs [2]. In particular, we consider the subgraph induced by the set of amino acids participating in the secondary structure also called Secondary Structure Elements (SSE). We call this graph SSE interaction network (SSE-IN). We carry out a study to describe the structural families of proteins when they are represented as interaction networks. We show how the properties of these networks are related to the structure of the corresponding protein. Thus, we propose a topological space where proteins from the same family tend to be grouped. By this way, we draw a parallel between structural and topological properties.

2 A Topological Study

The purpose of our work is to offer a graph theory interpretation of the hierarchical protein classifications. Indeed, when a protein belongs to a hierarchical level according to its structural properties then one can say also that the protein SSE-IN belongs to the same level. Thus, the topological properties of a SSE-IN are a consequence of the protein structural family. It implies that a SSE-IN is described by specific topological properties relative to the protein structural classification.

The first step before studying the proteins SSE-IN is to select them according to their SSE arrangements. Then, a protein belongs to a CATH [4] topology level or a SCOP [3] fold level if all its domains are the same. We have worked with the CATH v3.1.0 and SCOP 1.73 classifications. We have computed topological measures for three families of each hierarchical classification, namely SCOP and CATH (see Tab. 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 thousand amino acids in SSE. Among these proteins we limit the redundancy, the families contain at the maximum 20% of homologous proteins.

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

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

Table 2. Average of metric values for each family [1]. The column l regroups the average mean distances of SSE-IN. The column D represents the average diameter, δ is the average density and z the average mean degree for each studied family.

Name	l	D	δ	z
Rossmann fold	7.26	18.84	0.033	7.20
TIM Barrel	7.79	19.83	0.030	7.17
Lysozyme	4.99	12.81	0.038	6.82
Globin-like	6.64	15.65	0.034	7.69
TIM β/α-barrel	7.86	20.09	0.029	7.15
Lysozyme-like	5.03	12.85	0.042	6.81

2.1 Diameter and mean distance

Table 2 (column D) 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.

The diameter being an upper bound of distances in interaction networks, we expect that the mean distance l will be lower than D. Table 2 (column l) 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/l is about 2.5 for all the families. The last property is a characterization of all proteins' SSE-IN.

2.2 Density and mean degree

The density measures the ratio between the number of available edges and the number of all possible edges. Results presented in Table 2 (column δ) 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 Table 2 (columns l, D and δ) one can see the inversely proportional relation between density in one hand, and diameter and average distance on the other.

The mean degree is presented in Table 2 (column z). 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.

2.3 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 1. We can remark that the curves follow a power law distribution which can be approximated by the following power-law function:

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

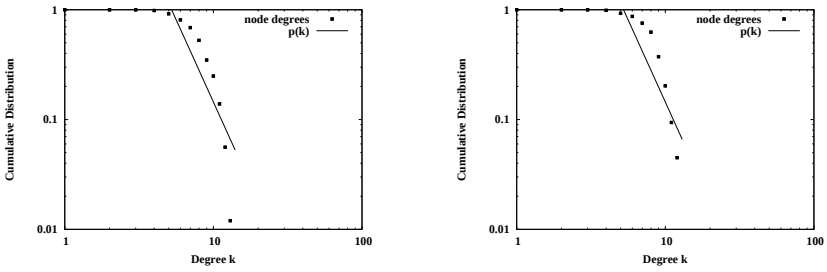

Fig. 1. Cumulative degree distribution for 1RXC from Rossman fold, left, and 1HV4 from TIM beta/alpha-barrel, right.

We observe the same results for all studied proteins. To explain this behavior, we have to rely on two facts. First, the mean degree of all proteins SSE-IN evolves

weakly (see Table 2, column z). Second, the degree distribution, see Figure 2, 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.

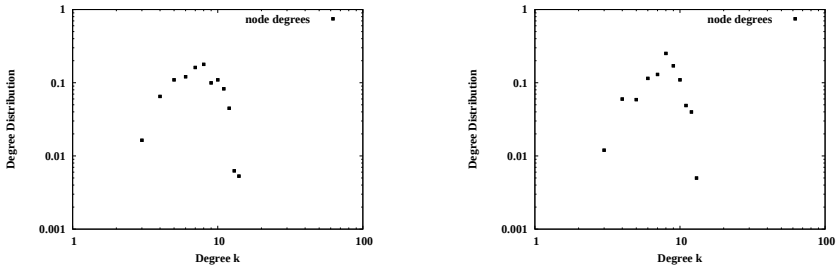

Fig. 2. Degree distribution for 1RXC from Rossman fold, left, and 1HV4 from TIM beta/alpha-barrel, right.

3 A Topological Space

In the previous section, we give different means to describe a protein structural family characterizing their SSE-IN. Some of properties, like diameter and density, allow discriminating two distinct families, while others, like mean degree and 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.

Here, we exploit this hypothesis proposing a topological space where a protein is described by its SSE-IN topology. Then, we want to project the structural families into this topological space to put in evidence that the proteins from a same family have SSE-INs which are grouped in this topological space. Consequently, we have to determine the dimensions of this topological space, that is, we want to identify which are the topological criterions able to discriminate the SSE-IN according to their families.

To build our topological space, we rely on the study done in the previous section and we apply it on another dataset, see Table 3. This new dataset is composed only of structural families from the *All Alpha* class in SCOP v1.73 classification.

First, we know that the mean distances and also the density are the discriminant metrics between the SSE-IN from different structural families. We plot a 3D topological space, see Fig 3, where the x axe represents the SSE-IN size, denoted N, the y axe represents the densities, denoted G, and the z axe represents the means distances, denoted L. The plots confirm that the study done in the previous section is reliable since the dimensions we use provide a topological space where the

Table 3. SCOP Fold families from *All alpha* class used to build our topological space.

SCOP ID	Family Name	Protein Number
46457	Globin-like	817
46688	DNA/RNA-binding 3-helical bundle	370
47472	EF Hand-like	313
48507	Nuclear receptor ligand-binding domain	223
48112	Heme-dependent peroxidases	207
48618	Phospholipase A2, PLA2	186
47112	Histone-fold	156
46625	Cytochrome c	148
48263	Cytochrome P450	146

proteins SSE-IN from the same structural family are grouped. Consequently, the parallel between structural and topological properties can be illustrated through the topological space we propose.

Second, we remark that among the protein SSE-IN belonging to a same structural family, there are some of them which have a size very close. Then, the proteins are grouped around a particular value n to form clusters. Thus, we can also describe the structural families' topological space describing the cluster properties that they form.

To characterize the clusters observed in the topological space, we have to define them. A cluster, denoted $c_{n=i}$, defined in the neighbourhood of a specific SSE-IN size equals to i satisfy:

$$p_{cluster} \in i \pm radius \geq r \times p_{family}$$

where $p_{cluster}$ designates the number of proteins in the cluster and p_{family} is the total number of proteins considered in the structural family. The parameter r is a threshold and we use the value of 25%. With this definition, we except that some clusters overlap each others. In this case, we merge them and consider the cluster center equals to i and the cluster radius is the average length between the minimum and the maximum SSE-IN size involved in the cluster.

Table 4 shows how the clusters appear in our topological space. We remark that each family has a specific cluster distribution meaning that the topological space we built is reliable to regroup the SSE-IN according to their structural families. The radius is in the most case around 50 meaning that the clusters regroup proteins whose size is comparable. The cluster sizes show how the proteins from a family are grouped around a particular neighbourhood.

This cluster description is actually a consequence of the family composition. Indeed, the families regroup proteins having a size enough close notably because their secondary structures are similar.

4 Conclusion

In this chapter, we consider a protein as an interaction network of amino acids of a protein (SSE-IN) and study some of the properties of these networks. It appears

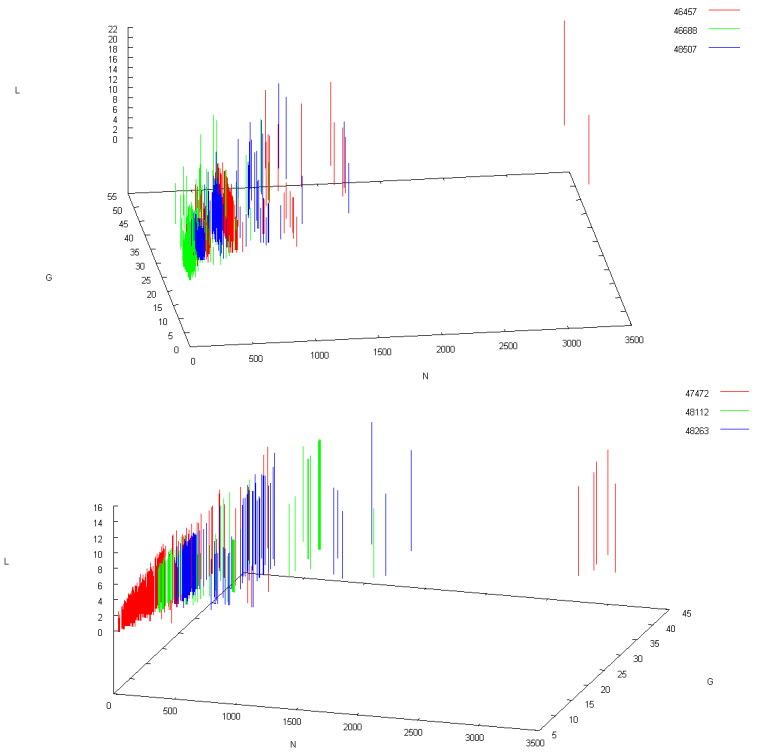

Fig. 3. A 3D topological space. The x axe represents the SSE-IN size, y the density and z the average distances. The proteins from a same family tend to be grouped.

Table 4. Cluster description for each family. The cluster size is expressed as the percentage of the total protein number in families.

SCOP ID	Cluster center	Cluster radius	Cluster size
46457	125	45	33.5
	485	45	39.4
46688	60	60	67.1
47472	90	70	78.6
48507	195	45	51.6
48112	190	50	77.3
48618	75	45	67.2
47112	595	45	76.9
46625	60	60	76.4
48263	275	45	50.7

that specific properties, like diameter and density, allow discriminating two distinct families, whereas others are common to all SSE-IN. Thus, proteins whose structural properties are similar will also have similar SSE-IN properties. In this way our model allows us to draw a parallel between biology and graph theory.

To illustrate the parallel between structural and topological properties, we propose a topological space whose dimensions are metrics enough discriminant between SSE-INs from different structural families. Then, the topological space let appears some clusters where the proteins from a same family are grouped. The description of these clusters contributes to distinguish by a new means the structural families relying on topological criterion. Through our topological space, we propose a means to describe a structural family by topological measures.

References

1. Diestel R. 2000. *Graph Theory.* Princeton: Springer Verlag.
2. Dokholyan NV, Li L, Ding F, Shakhnovich EI (2002) Topological determinants of protein folding. Proc Natl Acad Sci USA. 99(13):8637-8641
3. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of the protein database for the investigation of sequence and structures. J. Mol. Biol. 247:536-540
4. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH - a hierarchic classification of protein domain structures. Structure. 5:1093-1108