Detecting Community Structure in Amino Acid Interaction Networks

Omar GACI, Stefan BALEV and Antoine DUTOT

Abstract—In this paper we introduce the notion of protein interaction network. This is a graph whose vertices are the protein's amino acids and whose edges are the interactions between them. Using a graph theory approach, we observe that according to their structural roles, the nodes interact differently. By leading a community structure detection, we confirm this specific behavior and describe the communities composition to finally propose a new approach to fold a protein interaction network.

Keywords—interaction network, protein structure, community structure detection.

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.

I. 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.

The rest of the paper is organized as follows. In section II we briefly present the main types of amino acid interactions which determine the protein structure. In section III we introduce our model of amino acid interaction networks. Section 4 presents the means to lead a community structure detection in interaction networks. In section 5 we study the amino acids interaction networks by topological measures and observe specific interactions. We lead also a community structure detection to confirm the behavior already described. Finally,

Le Havre University, LITIS EA 4108, BP 540, 76058 Le Havre - France, email: {Omar.Gaci, Stefan.Balev}@univ-lehavre.fr
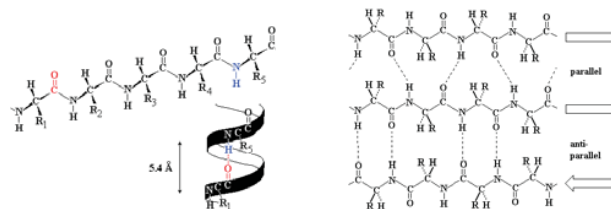
in section 6 we conclude and give some future research directions.

II. PROTEIN STRUCTURE

Unlike other biological macromolecules (e.g., DNA), proteins have complex, irregular structures. They are built up by amino acids that are linked by peptide bonds to form a polypeptide chain. We distinguish four levels of protein structure:

- The amino acid sequence of a protein's polypeptide chain is called its primary or one-dimensional (1D) structure. It can be considered as a word over the 20-letter amino acid alphabet.
- Different elements of the sequence form local regular secondary (2D) structures, such as α-helices or β-strands.
- The tertiary (3D) structure is formed by packing such structural elements into one or several compact globular units called domains.
- The final protein may contain several polypeptide chains arranged in a quaternary structure.

By formation of such tertiary and quaternary structure, amino acids far apart in the sequence are brought close together to form functional regions (active sites). The reader can find more on protein structure in ([5]).

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).

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).

III. AMINO ACID INTERACTION NETWORKS

The 3D structure of a protein is determined by the coordinates of its atoms. This information is available in Protein Data Bank (PDB) ([4]), which regroups all experimentally solved protein structures. Using the coordinates of two atoms, one can compute the distance between them. We define the distance between two amino acids as the distance between their C_α atoms. Considering the C_α atom as a "center" of the amino acid is an approximation, but it works well enough for our purposes. Let us denote by N the number of amino acids in the protein. A contact map matrix is a $N \times N$ 0-1 matrix, whose element (i, j) is one if there is a contact between amino acids i and j and zero otherwise. It provides useful information about the protein. For example, the secondary structure elements can be identified using this matrix. Indeed, α-helices spread along the main diagonal, while β-sheets appear as bands parallel or perpendicular to the main diagonal ([13]). There are different ways to define the contact between two amino acids. Our notion is based on spacial proximity, so that the contact map can consider non-covalent interactions. We say that two amino acids are in contact iff the distance between them is below a given threshold. A commonly used threshold is 7 Å and this is the value we use.

Consider a graph with N vertices (each vertex corresponds to an amino acid) and the contact map matrix as incidence matrix. It is called contact map graph. The contact map graph is an abstract description of the protein structure taking into account only the interactions between the amino acids. Now let us consider the subgraph induced by the set of amino acids participating in SSE. We call this graph SSE interaction network (SSE-IN) and this is the object we study in the present paper. The reason of ignoring the amino acids not participating in SSE is simple. Evolution tends to preserve the structural core of proteins composed from SSE. In the other hand, the loops (regions between SSE) are not so important to the structure and hence, are subject to more mutations. That is why homologous proteins tend to have relatively preserved structural cores and variable loop regions. 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. Here, we also consider the graph induced by the entire set of amino acids participating in folded proteins. We call this graph the three dimentional structure elements interaction network (3DSE-IN), see Fig. 2.

In ([16], [6], [2], [9]) 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.
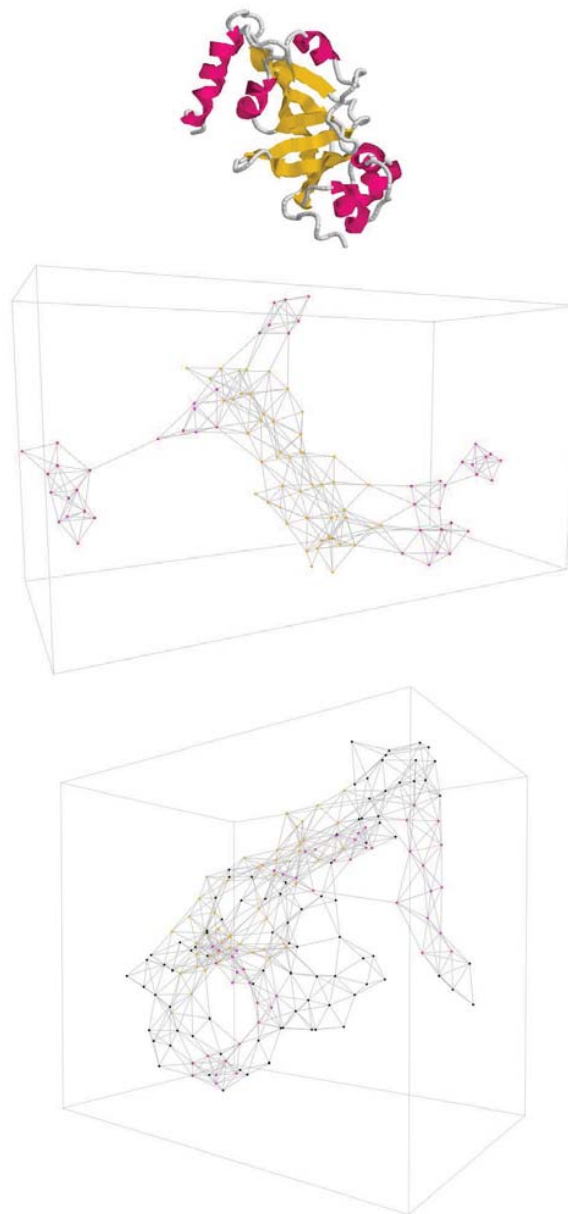

Fig. 2. Protein 1DTP (top), its SSE-IN and its 3DSE-IN (bottom).

IV. COMMUNITY STRUCTURE DETECTION

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. The purpose of these studies is to explain how elements interact inside the network and what are the general laws which govern the observed network properties.

From physics and computer science to biology and the social sciences, researchers have found that a broad variety of systems can be represented as networks, and that there is much
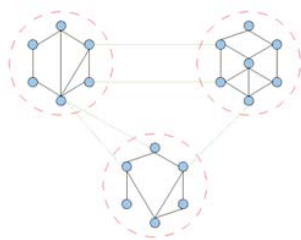
Fig. 3. An example of network with community structure as defined in this paper. There are three communities, denoted by the dashed circles, having dense internal links whereas there are connected by a lower density of external links.

to be learned by studing these networks ([7]). Indeed, the study of the Web ([1]), of social networks ([19]) or of metabolic networks ([15]) are contribute to put in light common non-trivial properties to these networks which have *a priori* nothing in common. The ambition is to understand how the large networks are structured, how they evolve and what are the phenomenom acting on their constitution and formation ([20]).

A. Definitions

Community structure is a network property which can be described as the gathering of vertices into groups such that there is a higher density of edges within groups than between them ([8]). The network topology let appear groups of nodes, called communities, within the density or also the number of edges, is higher than the density between these groups. The ability to identify these organisations can be helpful in understanding the structure of networks since it provide an interpretation of the topology by describing how vertices interact each other, see Fig. 3.

Technics to succeed in detecting community structure falls into two groups, bisection and hierarchical which itself is divided into two broad classes, agglomerative and divisive ([18]).

The agglomerative method is based on an empty network (n vertices with no edges) from which edges are added between vertex pairs according to a particular similarity measure. One can stop the process at any step and the resulting components are considered as the network communities. In a divisive method the network is considered in its entirety with all its vertices and edges. Edges are removed when they link the weakest connected pairs of vertices. After some iterations, the network is divided into small components which are the network communities.

Both method can be halted after any edge removal, the two process can be illustrated as a tree or dendrogram where the ith horizontal cuts represent the network community structure at the ith step, see Fig. 4.

B. Algorithm

In ([21]) the authors suggest using a force-based graph layout to detect communities.

A graph layout is a spacial arrangement of nodes so that strongly connected nodes are close together and lightly or unconnected nodes are far form each others. Most often such layout is used to better visualize a graph. One of the most successful layout method is based on repulsive and attractive forces. Intuitively the idea is that all vertices repulse each others, but vertices connected by an edge attract. The attraction forces counteract the repulsion forces for vertices connected to each others whereas other vertices are pushed far away. This is very close to the community definition given by ([8]).

There are several force-based layout algorithms. The one that where used here is based on ([10]). Such algorithms tend to slow a lot when the number of nodes grow due to their $O(n^2)$ complexity. However, it is possible to use recursive space decompositions techniques like the ones used in fast mutlipole methods ([14], [3]) to speed up things. This has been done in our reference implementation.

Once a graph layout has been found communities are detected using a hierarchical divisive technique. Edges are cut using a value that may take into account several factors, but at least the edge spacial length computed by the force-based algorithm. After the edge is cut, the layout algorithm is run anew. Each time an edge is cut, a *modularity*, see below, measure is made. Edges are cut as long as the modularity grows. This procedure ends when it is no more possible to grow the modularity. Communities correspond then to the connected components of the graph.

It is possible to improve this algorithm by changing the way an edge is cut. Instead of only considering the edges spacial length, it is possible to use a length modified by several factors. One possible factor we used in our implementation is called "proximity". We define intuitively the proximity of two nodes A and B as the number of common neighbors for A and B :

$$proximity = \frac{\text{number of common neighbors}}{\text{combined degree}}$$

With :

$$combineddegree = \frac{\text{degree of } A + \text{ degree of } B}{2}$$

Then we can divide the spacial length of an edge by this proximity measure to ensure nodes that share a lot of neighbors are closer.

The algorithm relies on an evaluation function namely the *modularity* which measure the quality of a particular division of a network to perform community detection. The modularity is defined by the next formula ([17]):

$$Q = \sum_i (e_{ii} - a_i^2)$$

If the network is divided into k communities, e is a (k,k) symmetric matrix whose element e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j. The row (or column) sums $a_i = \sum_j e_{ij}$ represent the fraction of edges that connect to vertices in community i ([17]). Because the algorithm need to remove the total m edges and since each iteration takes $O(mn)$ time, the worst-case running time of the algorithm is $O(m^2 n)$ or $O(n^3)$ on a sparse network.
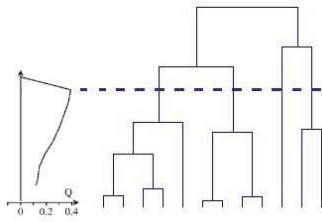
Fig. 4. A hierarchical tree or dendrogram computed by hierarchical detection of community structure algorithm. Vertices of the network are at the bottom, they belong to communities when they are linked by a horizontal line. The dotted-line indicates the cut level choosen when the modularity is maximum which guaranteed the strongest community structure.

TABLE I
STRUCTURAL FAMILIES STUDIED. WE CHOOSE ONLY FAMILIES WHICH CONTAIN MORE THAN 100 PROTEINS, FOR A TOTAL OF 18294 PROTEINS. WE HAVE WORKED WITH THE SCOP 1.73 CLASSIFICATION.

Class	Family Number	Protein Number
All α	12	2968
All β	17	6372
α/β	18	5197
$\alpha + \beta$	16	3757

V. EXPERIMENTAL RESULTS

In this paper we propose to identify the community structure of amino acid interaction networks built from folded protein. This work, lead to divide up a graph into connected components representing communities which regroup the nodes the most connected. We study the protein 3DSE-IN and in particular their community structures. We want to to put in evidence how amino acids interact each other within the tertiary structure and what are amino acids which tend to group togheter. Thus, our goal is to describe the composition of communities and identify nodes which interact highly.

Based on the SCOP classification and more precisely on the fold families (see Tab. 1), we have selected a total of 18294 proteins and studied their SSE-IN and also their 3DSE-IN to finally describe their community structure composition. Thus, 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 the results only for two studied classes. Moreover, we choose to not plot the average deviation when it disturbs the clarity of the figures.

Now, we describe the topological properties of the protein 3DSE-IN and further we will return to the community structure detection using the algorithm presented in section 4.2.

A. Topological Properties

First, we plot the distribution of the two interaction networks, see Fig. 5, to see that our dataset describes proteins whose size, their amino acid number, is most frequently between 100 and 500 residues.
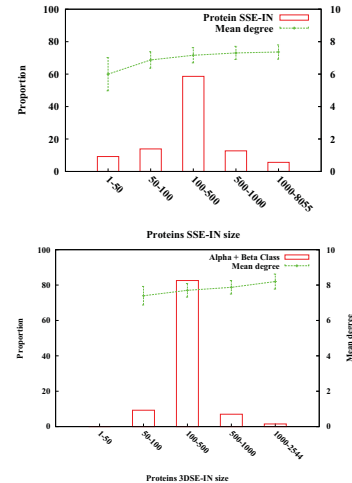

Fig. 5. Mean degree distribution according to protein SSE-IN size. It evolves with values close, between 6 and 8. Most of the SSE-IN or 3DSE-IN have a size 100 and 500 amino acids.

The mean degree, denoted z, provides a relation between the edge number m as a function of the node number n, it is defined by the next formula:

$$z = \frac{2m}{n}$$

From the plot, see Fig. 5, we deduct bounds for the mean degree which involves according to the following relation: For $n \geq 10$,

$$6 < z_{\text{SSE-IN}} < 9 \quad \text{and} \quad 6 < z_{\text{3DSE-IN}} < 9$$

These results help to understand how amino acids interact within a protein interaction network since it gives an estimation of the quantity of edges getting evolved as a function of the IN size. The reader can find a more detailed mean degree study in ([11]).

Second, we continue our description of the interaction networks. Thus, the nature of an edge depends on nodes which it allows to link. In a protein SSE-IN the edges can join two nodes from the same SSE or not. In a 3DSE-IN, the edges can link nodes whose structural role is different. Then, we distinguish the nodes which participate in the secondary structure and the others (from the loop regions) and we obtain:

$$\begin{cases} m_{\text{SSE-IN}} = m_H + m_S + m_{H-S} = m_{2D} \\ m_{\text{3DSE-IN}} = m_{2D} + m_{3D} + m_{2D-3D} \end{cases}$$

The fraction m_H or m_S designates the edges which link only nodes which are from the same SSE while m_{H-S} link two nodes belonging to different SSE. As well, m_{3D} edges join only nodes which intervene in the loop regions. In a protein 3DSE a node belongs to the secondary structure or not, thus: $n_{\text{3DSE-IN}} = n_{2D} + n_{3D}$.

By interesting us in protein SSE-IN, we describe the way that their constituents interact each other. We evaluate the interaction level between the α-helices and the β-sheets by
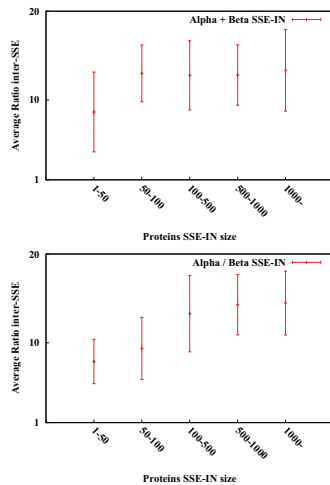
Fig. 6. Ratio of inter-SSE edges (r) as a function of the network size for the two classes of studied proteins.It's bounded and doesn't exceed 20% of the total edge number.

Fig. 7. Fraction of nodes in protein 3DSE-IN for the classes $\alpha + \beta$, top, and α/β, bottom, according to their structural role. Nodes from the secondary structure constitute on average 59.84% of the total node number for $\alpha + \beta$ class and 64.8% for α/β class.

computing the ratio of edges, denoted $r = m_{H-S}/m_{\text{SSE-IN}}$, linking two different SSEs, see Fig. 6. By this way, we measure the interaction level inter-SSE.

Fig. 6 shows that the number of inter-SSE edges is quite variable, but is bounded and does not exceed 20%, its the consequence of the excluded volume effect, since the number of residues that can physically reside within a given radius is limited.

Now, we study only the 3DSE-IN topological properties to find a general behavior which can influence the community structure detection.

We describe the 3DSE-IN according to the nature of their nodes and their edges, see Fig. 7-8.

These plots confirm clearly the assumption about which the secondary structure elements are considered as the structural fragments most conserved through the folded proteins. Thus, these elements composed in the most majority the protein 3DSE-IN around 61% of the total node number. Their interaction is measured by the fraction of edges which link them, denoted m_{2D}, which represents around 54.9% of the total edge number. As well, the fraction m_{2D} and m_{2D-3D} are equivalent through all protein 3DSE-IN when $n_{3\text{DSE-IN}} > 100$. This means that a node from the loop regions, n_{3D}, has an equal probability to interact with a n_{2D} node than with an another n_{3D} node.

It is interesting to understand the way that the nodes from the loop regions, n_{3D}, interact in the protein 3DSE-IN. Indeed, see Fig. 9, they tend to group together and consequently we can except to encounter local clouds of them interacting with n_{2D} nodes.

Once we have described the general behavior of protein 3DSE-IN, we can refine this study by searching correlations between previous measures. We observe, see Fig. 7-8, that the fraction of node n_{2D} involved in the protein 3DSE-IN determines the way that the nodes interact each other. To validate this relation of cause and effect, we describe the nature
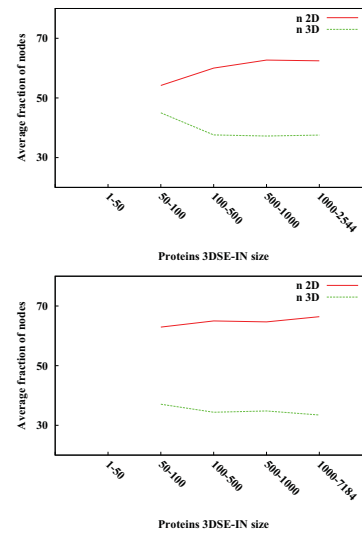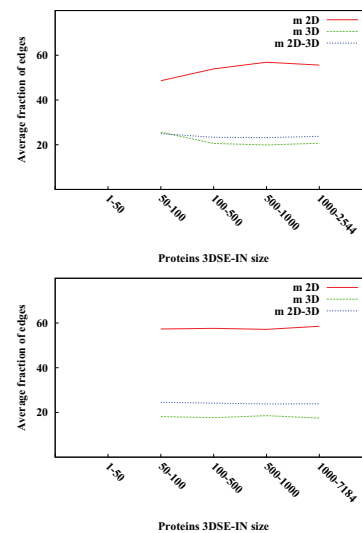

Fig. 8. Fraction of edges in protein 3DSE-IN for the classes $\alpha + \beta$, top, and α/β, bottom, according to the nodes they link.

of the edges as a function of the $n_{2D}/n_{3\text{DSE-IN}}$ rate.

The plot, see Fig. 10, shows that the average interacting rate m_{2D-3D} between n_{2D} nodes and n_{3D} nodes stays constant independently of the proportion of n_{2D} nodes presents in the protein 3DSE-IN. It implies that the n_{3D} nodes have a bounded interaction level. Thus, the n_{3D} nodes will form edges with n_{2D} nodes for a maximum of 25% of the total edge number. To this limit be reached, n_{3D} nodes will group together to form clouds whose only the "surface" will interact with n_{2D} nodes.

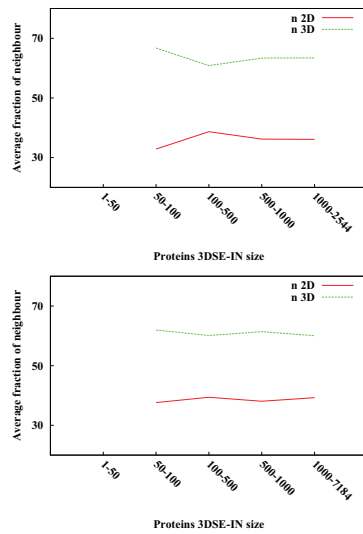This behavior illustrates perfectly the general folded protein

Fig. 9. Fraction of neighbour of nodes n_{3D} in protein 3DSE-IN. The n_{3D} nodes aggregate each others since near than 64% of their neighbourhood is composed by others n_{3D} nodes.
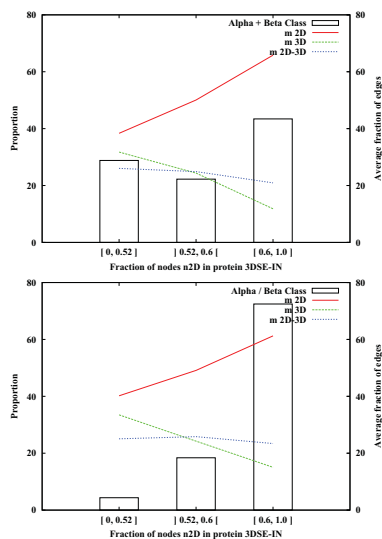

Fig. 10. Variation of the edge natures as a function of the $n_{2D}/n_{\text{3DSE-IN}}$ fraction in protein 3DSE-IN.

shape, that is the hydrophobic side chains are packed into the interior of the protein creating a hydrophobic core and a hydrophilic surface. The fact that the SSE are in the interior of the folded protein implies therefore that interaction between them and the others amino acids are bounded, because their contact surface is limited. As well, when the fraction of node $n_{2D}/n_{\text{3DSE-IN}}$ represents less than 60% of the total node number, n_{3D} nodes have tendency to interact each other in proportion superior or equal than they interact with n_{2D} nodes.

The consequence is to see appear clouds of n_{3D} nodes able to disturb the community structure already determined in the protein SSE-IN.
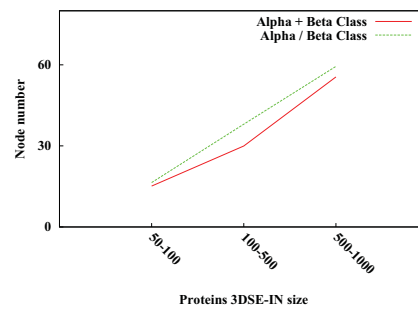

Fig. 11. Evolution of communities size as function of protein 3DSE-IN size

B. Community Structure Detection

Now we have described the behavior of the protein 3DSE-IN, we are interested in the community structure composition. The algorithm we use to perform the community detection is given here:

Algorithm 1: Algorithm to detect community structure in protein 3DSE-IN.

Input:

m : Edge set

Data:

Q : modularity in the network.

r_e : Fraction length/proximity for the edge e.

$r_{\text{3DSE-IN}}$: Average rate length/proximity for the entire 3DSE-IN.

begin
    $Q \longleftarrow 0$
    while Q *is increasing* **do**
        foreach *Edge* $e \in m$ **do**
            computeProximity(e);
            computeLength(e);
        $m \longleftarrow m - \{\text{Edges with } r_e > r_{\text{3DSE-IN}}\};$
        $Q \longleftarrow computeModularity();$
end

We have realized simulations according to the protein SCOP class level, see Tab. 1. A first observation concerns the average size evolution of detected organisations, see Fig. 11. Its evolution is linear and the organisation sizes double when the interaction network size increases by doubling. This means that one can describes the entire 3DSE-IN, at a macroscopic level, as the association of microscopic communities whose elements are highly interacting. Thus, the local interactions contribute to the formation of strong communities with an average size depending on the interaction network size. Thus, these interactions between organisations finally determine the global protein 3DSE-IN.

By continuing our description, we are interested in the organisations constitution. At first sight, one can expect to encounter communities composed exclusively by nodes participating in the secondary structure because they form effectively
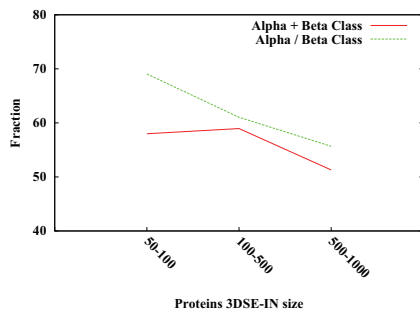
Fig. 12. Average fraction of nodes n_{2D} as function of protein 3DSE-IN in communities
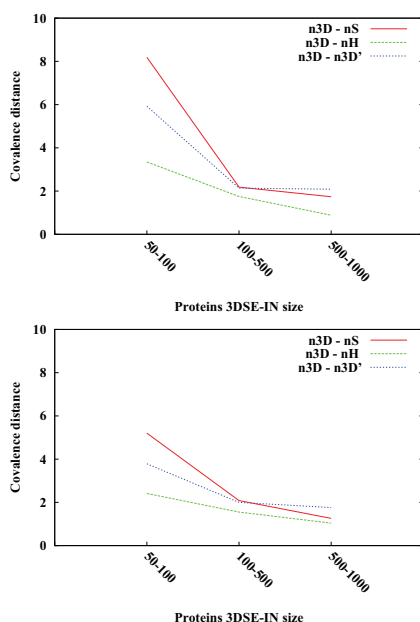

Fig. 13. Average covalence distance between n_{3D} nodes and their neighbors in communities for $\alpha + \beta$ class, top, and α/β class, bottom.

dense subgraph within 3DSE-IN.

We measure the average fraction of node n_{2D} in organisations, see Fig. 12, to evaluate in what proportion nodes from the secondary structure group together and interact. The plot shows that the communities tend to be composed predominantly by n_{2D} nodes. It implies that it exists communities with a weak fraction of nodes from the loop regions which interact with others n_{2D} nodes.

To identify suitably what are the n_{3D} nodes which interact highly within communities, we rely on the covalence distance, denoted cd, defined as follows. The covalence distance between two nodes, u and v corresponds to the number of covalence bonds between amino acids they represent in the primary structure. By this way, in each organisation we are interested in n_{3D} nodes and we evaluate their average covalence distance computed with their whole neighbors, we denote by n_H and n_S a node from a helix or sheet subgraph.

The plot provides interpretations about how nodes from the

tertiary structure interact with the others. First, n_{3D} nodes which interact highly with n_{2D} nodes from helixes are near of them in the primary structure meaning that helices have a local interaction within a folded protein. Second, n_{2D} from sheets are able to interact with distant n_{3D} nodes. This behavior is the consequence of the specific shape of sheets composed by strands which interact each other despite of a high covalence distance due to loop regions. To illustrate this phenomenon, we can consider the following example. Let a folded protein with 100 amino acids having a sheet composed by three strands s_1, s_2, s_3 defined by intervals $[10, 15]$, $[25, 30]$ and $[45, 50]$. Then, a community structure detection will provide an organisation grouping the three strands and others n_{3D} nodes. Here we are interested in a particular node, for example one representing the amino acid number 51. This node is connected with the node 50 of s_3 and also with the node 30 of s_2 and consequently its normalised average covalence distance will be high. Nevertheless, this phenomenon is eased for large 3DSE-IN notably because the strands are close each others.

C. Synthesis

We study how loop nodes, n_{3D}, interact in the 3DSE-IN by plotting their node neighborhood distribution. Thus, they tend to group together to aggregate. Consequently we exccept to encounter local clouds which interact with n_{2D} nodes. We describe the nature of edges to observe that the average interaction level m_{2D-3D} stays constant independently of the $n_{2D}/n_{3DSE-IN}$ rate. This implies that n_{3D} nodes have a bounded interaction level. They form edges with the secondary structure nodes for a maximum of 25% of the total edge number. Once the limit is reached, n_{3D} nodes group together to form clouds. Only the *surface* of the clouds interacts with n_{2D} nodes.

The previous result implies a new approach to fold the amino acid graph which can be decomposed into three successive steps:

- we predict the graph composed only by the secondary structure
- we add the loops which fold independently
- we predict the global graph where the two parts are already folded

Our study allows bettering describing the collective behaviour of amino acids during the folding process. This behaviour is a consequence of the general protein folding principle, according to which amino acids from the secondary structure are compacted inside the tertiary structure limiting the contact surface with amino acids from the loops.

VI. CONCLUSION

In this article we show that the community structure detection leading in protein 3DSE-IN is a means to isolate the more interacting amino acid in the folded protein. Inside organisations, nodes participating in the secondary structure are majority and interact with nodes from loop regions. To identify these last, we measure their average covalence distance with their neighbourhood, results show that each communities have

local interaction because of their constituents are close in the primary structure.

A continuation of our work will be to consider a protein 3DSE-IN as a complex system where it exists local organisations whose mutual interactions allow the emergence of the global interaction network. In this case, each amino acid will be considered as a local agent able to interact according to its structure level and its environment. The emergence of a global topology would by the shape of the folded protein.

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 studied ([11], [12]), 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.

REFERENCES

[1] R. Albert, H. Jeong, and A.-L. Barabási. The diameter of the world wide web. *Nature*, 401:130–131, 1999.

[2] A. R. Atilgan, P. Akan, and C. Baysal. Small-world communication of residues and significance for protein dynamics. *Biophys J*, 86(1 Pt 1):85–91, January 2004.

[3] R. Beatson and L. Greengard. A short course on fast multipole methods.

[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[5] C. Branden and J. Tooze. *Introduction to protein structure*. Garland Publishing, 1999.

[6] K. V. Brinda and S. Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophys J*, 89(6):4159–4170, December 2005.

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

[8] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[9] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich. Topological determinants of protein folding. *Proc Natl Acad Sci U S A*, 99(13):8637–8641, June 2002.

[10] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software - Practice and Experience*, 21(11):1129–1164, 1991.

[11] O. Gaci and S. Balev. Hubs identification in amino acids interaction networks. In *Proceedings of the 7th ACS/IEEE International Conference on Computer Systems and Applications*, 2009. 7 pages.

[12] O. Gaci and S. Balev. The small-world model for amino acid interaction networks. In *Proceedings of the IEEE AINA 2009, workshop on Bioinformatics and Life Science Modeling and Computing*, 2009. 6 pages.

[13] A. Ghosh, K. V. Brinda, and S. Vishveshwara. Dynamics of lysozyme structure network: probing the process of unfolding. *Biophys J*, 92(7):2523–2535, April 2007.

[14] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.

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

[16] U. K. Muppirala and Z. Li. A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues. *Protein Eng Des Sel*, 19(6):265–275, June 2006.

[17] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.

[18] John P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, January 2000.

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

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

[21] Bo Yang and Da-You Liu. Incremental algorithm for detecting community structure in dynamic networks. In *Proceedings of The 4th International Conference on Machine Learning and Cybernetics*, pages 2284–2290, August 2005.