Ant Colony Approach to Predict Amino Acid Interaction Networks

Omar Gaci, Stefan Balev

Ant Colony Approach to Predict Amino Acid Interaction Networks

Omar GACI
Le Havre University
LITIS Laboratory
Le Havre, France
Email: omar.gaci@univ-lehavre.fr

Stefan BALEV
Le Havre University
LITIS Laboratory
Le Havre, France
Email: stefan.balev@univ-lehavre.fr

Abstract

In this paper we introduce the notion of protein interaction network. This is a graph whose vertices are the proteins amino acids and whose edges are the interactions between them. We consider the problem of reconstructing protein's interaction network from its amino acid sequence. An ant colony approach is used to solve this problem.

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 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 term this graph SSE interaction network (SSE-IN). We carry out a study to identify the interactions involved between the amino acids in the folded protein. To achieve this topological folding, we build an interaction probability between two amino acids as a function of their physico-chemical properties and we rely on this probability building an ant colony system to fold a SS-IN.

The rest of the paper is organized as follows. In section 2 we briefly present the main types of amino acid interactions which determine the protein structure. In section 3 we introduce our model of amino acid interaction networks. In section 4 we propose an ant colony approach to reconstruct the graph of interactions between amino acids involved in the secondary structure. Finally, in section 5 we conclude and give some future research directions.

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:

- 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 [4].

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.

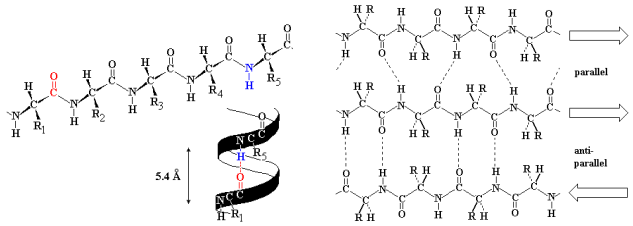

Figure 1. Left: an α-helix illustrated as ribbon diagram, there are 3.6 residues per turn corresponding to 5.4 Å. Right: A β-sheet composed by three strands.

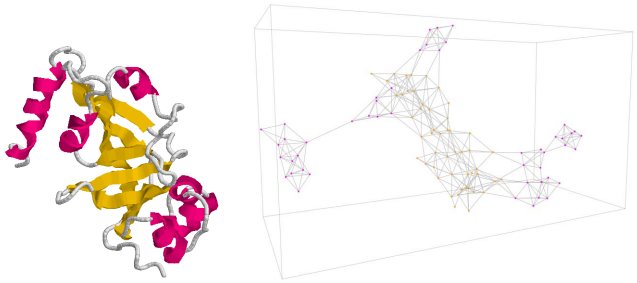

Figure 2. Protein 1DTP (left) and its SSE-IN (right).

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).

3. 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) [3], 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 [6] 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.

In [15], [5] 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.

4. Interaction Prediction

In previous works [8], [9], [11], we have studied the protein SSE-IN. We have identified notably some of their properties like the degree distribution or also the way in which the amino acids interact. These works have allowed us to determine criteria discrimminating the different strucutral families. We have established a parallel between structural families and topological metrics describing the protein SSE-IN.

Using these results, we have proposed a method to predict the family of an unclassified protein based on the topological

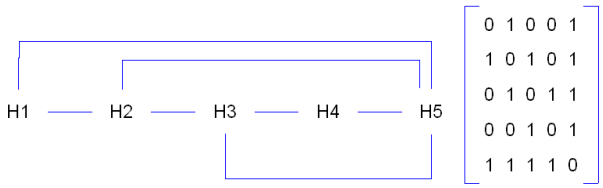

Figure 3. 2OUF SS-IN (left) and its associated incidence matrix (right). The vertices represent the different α-helices and an edge exists when two amino acids interact.

properties of its SSE-IN, see [12]. Thus, we consider a protein defined by its sequence in which the amino acids participating in the secondary strucutre are known. This preliminary step is usually ensured by threading methods [14] or also by hidden Markov models [1]. Then, we apply a method able to associate a family from which we rely to predict the fold shape of the protein. This work consists in predicting the family which is the most compatible to the unknown sequence. The following step, is to fold the unknown sequence SSE-IN relying on the family topological properties.

Once a family has been predicted, we have been interested in the way that the different SSEs interact each others that is we want to predict the motifs involved in this sequence, see Fig. 3. In [10] we consider that the motifs formed by the SSEs in the unclassified sequence can be represented by an adjacency matrix where a 1 means that the two SSEs are in contact. Then, we use a comparative model built from the template proteins belonging to the predicted family and we rely on a genetic algorithm to predict the unclassified sequence matrix of motifs.

In this paper, we consider that the matrix of motif is correct. Our goal is to add inter-SSE edges to rebuild the sequence SSE-IN, see Fig. 4. We use a two-step approach following the hierarchical structure of the network. The first step consists in considering separately each pair of SSEs in interaction. We use an ant algorithm to identify the suitable interactions between amino acids belonging to these SSEs. When this step is done for all pairs of SSEs in interaction, we go back to global level and use another ant algorithm to select the final set of inter-SSE edges.

4.1. Prediction of the edge quantity

In [11] we have studied the quantity of edges between SSEs, also called shortcuts, and we have shown that it is bounded mainly due to the excluded volume effect. Nevertheless, even if the shortcut number can be estimated knowing its upper bound, the distribution of shortcuts between the different SSEs composing the sequence stays difficult to predict (actually it is too variable from one protein SSE-IN to another).

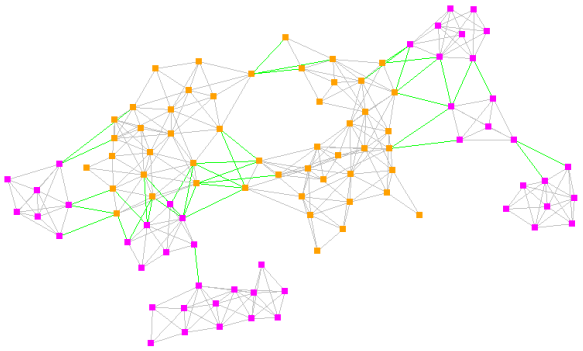

Figure 4. Protein 1DTP SSE-IN. Shortcut edges we want to predict are plotted in green.

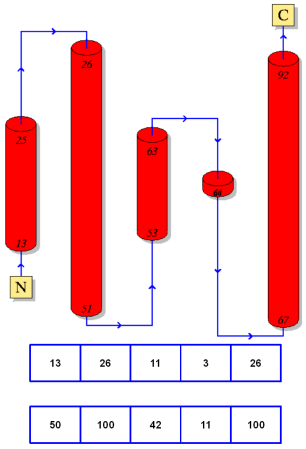

Figure 5. Building the array representation for the unclassified protein 2OUF.

To predict the shortcut edge rate of the sequence, we rely on the template proteins. We represent a protein as an array with one element by SSE in their order in the primary structure. Each cell is equals to the normalized size of the corresponding SSE, see Fig. 5.

From the predicted family, we choose the proteins with the same number of SSEs as the unclassified protein and we construct their arrays. Let $s = (s_1, \ldots, s_n)$ be the sequence array and $t = (t_1, \ldots, t_n)$ be the template protein array. We use the standard L_1 distance $d(s,t) = \sum_{i=1}^{n} |s_i - t_i|$. We admit 20% of difference, if $d(s,t) < 0.2 \sum_{i=1}^{n} s_i$, we consider that the template protein is near the sequence. Finally, we compute the average shortcut rate of the near proteins to predict the quantity of edges to add in the initial disconnected graph of the sequence. If no near proteins are found, we compute the edge quantity to add as the average of the template proteins shortcut rate.

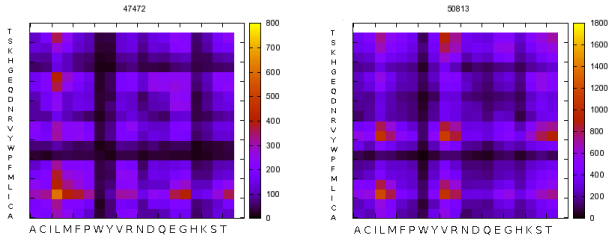

Figure 6. Occurrence matrix of edges inter-SSE for the family 47472, *All alpha* class (left) and 50813, *All beta* class (right).

4.2. Prediction of the nodes to be connected

To add an edge between two SSEs we use a probability built from the predicted family. First, we choose to add the same quantity of edges between SSEs in contact, then the number of edges between two SSEs is equal to the number of shortcut predicted previously divided by the number of SSEs in contact. Second, to add an edge between two SSEs we use a probability built from the predicted family. Indeed, we consider the occurrence of a shortcut as a function of the amino acids it links. Fig. 6 shows the shortcut occurrence matrix of the families *All alpha* 47472 and *All beta* 50813. We remark that shortcuts between amino acids Leucine are the most frequent.

The problem of deciding how to add edges between two SSEs can be stated as follows. Let $A = a_1 \dots a_s$ and $B = b_1 \dots b_1 \dots b_t$ be two SSEs.

We want to add k edges among the $s \times t$ possibilities. For $i \in [1, s]$ and $j \in [1, t]$ the probability to link the amino acids a_i and b_j will be proportional to the number of occurrences of the shortcut (a_i, b_j) denoted $S(a_i, b_j)$:

$$w_{ij} \sim S(a_i, b_j)$$

In order to add approximately k edges, we need

$$\sum_{i=1}^{s} \sum_{j=1}^{t} w_{ij} = k$$

and hence

$$w_{ij} = \frac{kS(a_i, b_j)}{\sum_{i=1}^{s} \sum_{j=1}^{t} S(a_i, b_j)}$$

Actually, for each pair of SSEs in contact we add the $s \times t$ possible edges whose weight is w_{ij}.

4.3. Local algorithm

This algorithm is used to identify the suitable shortcuts between a pair of SSEs in interaction. We start with a graph in which each node form the first SSE is connected to each node from the second SSE. The edges have weights w_{ij} corresponding to the relative frequency of each kind of shortcut in the family, computed as explained previously. Each ant walks on the graph dropping pheromone on its way. This ant system has to reinforce the suitable shortcuts. When the emergence of specific shortcuts is done, we keep these edges for the next global research. An overall description is given as algorithm 1.

Algorithm 1: Algorithm to the emergence of shortcut edges between two SSEs using an ant colony.

```
begin
    create n ants
    while stop condition do
        for each ant a do
            moveAnt(a)
        updatePheromone()
    selectShortcut(phero_min)
end
```

The local research process by ant colony is the following. First, we create n ants which is equals is equal to the total number of nodes involved in the two SSEs which are respectively composed of s and t amino acids. The initial positions of the ants are chosen randomly, so at the beginning more ants may be at the same node. If an ant is on the node i, the probability to choose the next node j, denoted p_{ij}, is defined as follows:

$$p_{ij} = \frac{[\tau_{ij}]^{\alpha} \cdot [w_{ij}]^{\beta}}{\sum_{k \in V(i)} [\tau_{ij}]^{\alpha} \cdot [w_{ij}]^{\beta}}$$

The heuristic vector or also the weight, w_{ij}, associated to the shortcut edges has been already computed:

$$w_{ij} = \frac{kS(a_i, b_j)}{\sum_{i=1}^{s} \sum_{j=1}^{t} S(a_i, b_j)}$$

For the edges connecting nodes from the same SSE, we take as weight the average of the weights of inter-SSE edges:

$$\overline{w} = \frac{1}{st} \sum_{i=1}^{s} \sum j = 1^t w_{ij}$$

After each move we update the pheromone quantity on the inter-SSE edges using the formula

$$x\tau_{ij} \leftarrow (1 - \rho)\tau_{ij} + a_{ij}\Delta\tau$$

where a_{ij} is the number of ants that moved on the edge (i, j) and $\Delta\tau$ is the quantity of pheromone dropped by each ant. As far as the edges belonging to the same SSE are concerned, we keep the pheromone rate on them equal to the average pheromone rate on the inter-SSE edges

$$\overline{\tau} = \frac{1}{st} \sum_{i=1}^{s} \sum j = 1^t \tau_{ij}$$

In this way ants move inside an SSE completely randomly, while if they decide to change the SSE they are guided by the edge weights ant the pheromone rates.

The algorithm stops when a maximum number of iterations is reached. It can also stop earlier if the maximum pheromone rate is k times bigger than the average pheromone rate on the edges. In this case we consider that the reinforcement is sufficient. At the end, we keep only edges with pheromone quantity exceeding a threshold $phero_{min}$.

4.4. Global algorithm

Once the local algorithm has provided inter-SSE edges, we build the SSE-IN composed by these edges. Now our goal is to keep only the most appropriate among them. The number of edges to keep, m_P, we choose as in section 4.1. Global algorithm works similarly to the local one, but this time the ants move on the entire SSE-IN graph. Their number is equal to the total number of nodes. After the end of this algorithm we sort the edges by decreasing pheromone rate and keep only the first m_P of them.

Finally, we compare the topological properties of the obtained graph to the average properties of the corresponding family and accept or reject, we discuss this step in the next section

4.5. Measure and Reliability

In [7], [8] we have shown that the protein SSE-IN can be described by their topological properties. We have shown that it exists a parallel between biological classification (notably the SCOP fold family level) and the SSE-IN topological properties. Here we exploit these properties to excluded the incompatible SSE-IN built by our ant colony approach. The principle is the following, we have predicted a fold family for the sequence from which we extract the template proteins. Thus, we compute the diameter, the characteristic path length and the mean degree to evaluate the average topological properties of the family for a particular SSE number. Then, after we have built the sequence SSE-IN, we compare its topological properties with the themplate ones. We admit an error up to 20% to accept the built sequence SSE-IN. If the built SSE-IN is not compatible, it is rejected.

As well, we evaluate the prediction of edge inter-SSE added. We compare the predicted value, denoted m_P, with the real value, denoted m_R:

$$accuracy = 1 - \frac{|m_R - m_P|}{m_P}$$

Table 1. Shortcut edge research by an ant colony approach. The score measures the intersection between the effective and predicted shortcut edges. The algorithm parameter values are: $\alpha = 25$, $\beta = 12$, $\rho = 0.7$, $\Delta\tau = 4000$, $k = 2$, $phero_{min} = 0.8$.

SCOP Family	Protein Number and Size	Score	Standard Deviation
46688 (*All* α)	17 - [27 ; 46]	79.136	8.11
47472	10 - [98 ; 125]	70.69	15.53
46457	25 - [129 ; 135]	74.56	9.04
48112	11 - [194 ; 200]	68.3	14.94
48507	18 - [203 ; 214]	64.48	7.84
46457	16 - [241 ; 281]	63.37	16.93
48507	20 - [387 ; 422]	61.94	9.42
50629 (*All* β)	6 - [54 ; 66]	76.13	6.39
50813	11 - [90 ; 111]	72.56	5.86
48725	24 - [120 ; 124]	76.69	11.96
50629	13 - [124 ; 128]	71.3	14.44
50875	14 - [133 ; 224]	74.48	14.14

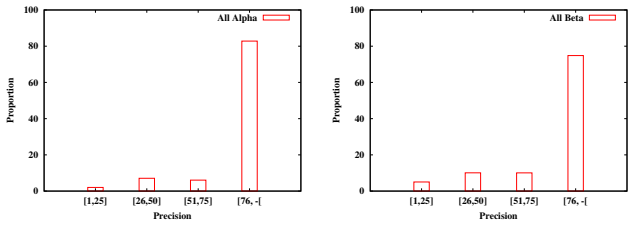

Figure 7. Precision of the edge quantity to add, the most prediction are close to the real values.

4.6. Simulations and Results

We have tested our method according to the predicted family because the probability of adding edge is determined by the family occurrence matrix. We have used the same dataset of sequences whose family has been predicted. For each protein, we have launched 180 simulations and we accept the built SSE-IN only when its topological properties are compatible to the template properties. The results are presented in Tab. 1. The score is the percentage of correctly predicted shortcut edges between the sequence SSE-IN and the SSE-IN we have reconstructed.

Since the edge quantity to add is in the most cases accurate, we understand that the global intersection score depends on the local researches lead for each pair of SSEs in contact. The plot, see Fig. 8, confirms this tendancy, if the local researches select at least 75% of the correct shortcuts edge, the global intersection score stays superior to 75% and evolves around 78% for the *All alpha* class and 76% for the *All beta* class.

To recapitulate, we show that the global score evolves around 70% of precision. The average score decreases for big proteins (of size more than 200). For these proteins there

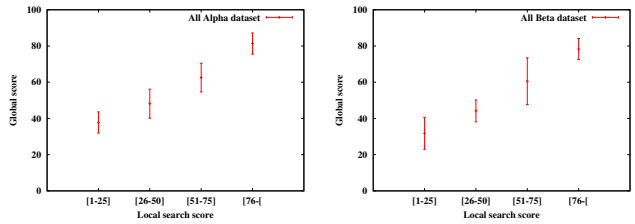

Figure 8. The global score depends on the local researches. When the local researches provide the correct edges, the global score stays close to 78% for the *All alpha* class and 76% for the *All beta* class.

are many pairs of SSE in contact and the errors made by the local algorithm accumulate to global level.

5. Conclusion

We propose in this paper a means to fold a protein SSE-IN. To do that, we rely on the Levinthal hypothesis also called the kinetic hypothesis. Thus, the folding process is oriented and the proteins don't explore their entire conformational space. Here, we use the same approach: to fold a SSE-IN we limit the toplogical space by associating to a sequence a strucutral family [12].

Since the strucutral motifs which are described in a strucutral family are limited, we predict by this way the SSE-IN motifs [10].

Then, we use an ant colony system to predict the nodes involved when two SSEs are in contact. We predict the inter-SSE edges by an ant colony approach and we repeat the same process at the global level. By measuring the resulting SSE-IN, we remark that their topology depends on the local research. If, during the local folding, we collect the right edges (more than 75%) then the global score stays close to 76% of the real edges inter-SSE.

The characterization we propose constitutes a new approach to the protein folding problem. Here we propose to fold a protein SSE-IN relying on topological properties. We use these properties to guide a folding simulation in the topological pathway from unfolded to folded state.

References

[1] K. Asai, S. Hayamizu, and K. Handa. Prediction of protein secondary structure by the hidden markov model. *Comput. Appl. Biosci.*, 9(2), 1993.

[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] 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.

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

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

[6] 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.

[7] O. Gaci and S. Balev. Characterization of amino acid interaction networks in proteins. In *JOBIM 2008*, pages 59–60, 2008.

[8] O. Gaci and S. Balev. Proteins: From structural classification to amino acid interaction networks. In *Proceedings of BIOCOMP'08*, volume II, pages 728–734. CSREA Press, 2008.

[9] 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.

[10] O. Gaci and S. Balev. Motif prediction in amino acid interaction networks. In *Proceedings of the International Conference on Computational Biology*. IAENG, 2009. 6 pages.

[11] 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.

[12] O. Gaci and S. Balev. Prediction of protein families by topological inference. In *Proceedings of the 1st International Conference on Bioinformatics*, 2010. 7 pages, in submission, http://www-lih.univ-lehavre.fr/~gaci/bioinformaticsGACI-BALEV.pdf.

[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] B. Mirny and L. Shakhnovich. Protein structure prediction by threading: Why it works and why it does not. *J. Mol. Biol.*, 283(2):507–526, 1998.

[15] 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.