Motif Prediction in Amino Acid Interaction Networks

Omar GACI and Stefan BALEV *

Abstract—**In this paper we represent a protein as a graph where the vertices are amino acids and the edges are interactions between them. We propose a genetic algorithm of reconstructing the graph of interactions between secondary structure elements which describe the structural motifs. The performance of our algorithms is validated experimentally.**

Keywords: interaction network, protein structure, genetic algorithm

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.

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 a genetic algorithm of reconstructing the graph of interactions between secondary structure elements. 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].

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 $n+4$.

*Le Havre University, LITIS EA 4108, BP 540, 76058 Le Havre - France, email: {Omar.Gaci, Stefan.Balev}@univ-lehavre.fr

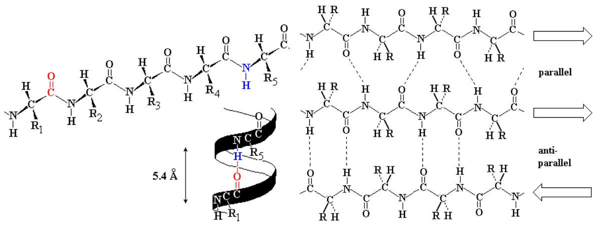


Figure 1: Left: an α-helix illustrated as ribbon diagram, there are 3.6 residues per turn corresponding to 5.4 Å. Right: A β-sheet composed by three strands.

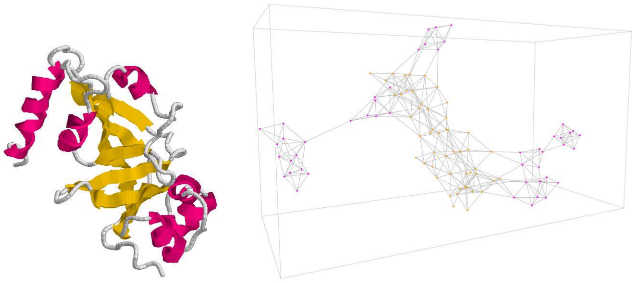


Figure 2: Protein 1DTP (left) and its SSE-IN (right).

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 [12]. 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.

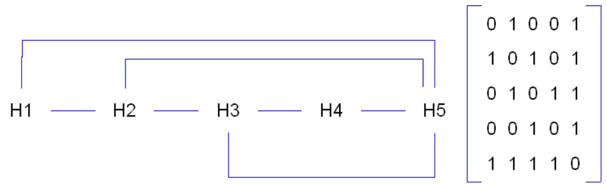
The reason of ignoring the amino acids not participating in SSE is simple. Evolution tends to preserve the structural core of proteins composed from SSE. In the other hand, the loops (regions between SSE) are not so important to the structure and hence, are subject to more mutations. That is why homologous 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.

In [14, 5, 2, 6] 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 Motif Prediction

In previous works [8, 9, 10], 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 properties of its SSE-IN, see [11]. 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 [13] 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.



$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

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.

To fold a SSE-IN, 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. In this paper, we use the same approach: to fold a SSE-IN we limit the toplogical space by associating a strucutral family to a sequence [11]. Since the strucutral motifs which describe a strucutral family are limited, we propose a genetic algorithm (GA) to enumerate all possibilities.

In this section, we present a method based on a GA to predict the graph whose vertices represent the SSE and edges represent spatial interactions between two amino acids involved in two different SSE, further this graph is called Secondary Structure Interaction Network (SS-IN), see Fig 3.

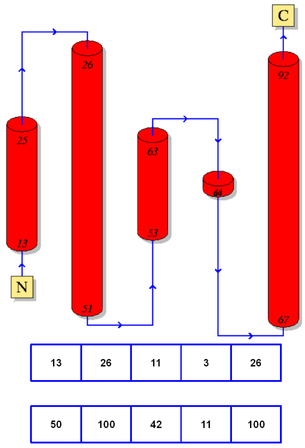
Thereafter, we use a dataset composed by proteins which have not fold families in the SCOP v1.73 classification and for which we have predicted a family in [11].

4.1 Overall description

Our GA works on a population of proteins which have the same number of SSE as the studied sequence. The initial population is composed of the proteins of the predicted family with the same number of SSE. Each individual of the population has associated SS-IN adjacency matrix. At each iteration we apply genetic operators in order to obtain new individuals with new adjacency matrices. Our fitness function does not use a measure based on the adjacency matrices, it is based only on the SSE sizes.

4.2 Genome structure

The genome structure of a protein SS-IN is an array of alleles. Each allele represents a SSE notably considering its size that is the number of amino acids which compose it. The size is normalized contributing to produce genomes whose alleles describe a value between 0 and 100. Obviously, the position of an allele corresponds to the SSE position it represents in the sequence, see Fig 4. In the same time, for each genome we associate its SS-IN incidence matrix.



13	26	11	3	26

50	100	42	11	100

Figure 4: Building the chromosome representation for the unclassified protein 2OUF.

50	100	42	11	100

18	8	100	35	42

Figure 5: The distance between two chromosomes is the sum of distances between the alleles, in this case (32+92+58+24+58).

4.3 Fitness function

We evaluate the performance of a chromosome by using the L_1 distance between this chromosome and the target sequence. An example is given in Fig 5.

4.4 Crossover operator

This operator uses two parents to produce two children. After generating a random cut position, we swap the both parts as shown in Fig 6. Nevertheless, this operator can produce incidence matrices which are not compatible with the fold family, we discuss this problem below.

4.5 Mutation operator

This operator is used for a small fraction (about 1%) of the generated children. It modifies the chromosome and the associated matrix. For the chromosomes, we define two operators: the two position swapping and the one position mutation. Concerning the associated matrix, we define four operators: the row translation, the column translation, the two position swapping and the one position mutation. The crossover and mutation operators may produce matrices which describe incoherent SS-IN compared to the predicted sequence fold family. To eliminate the wrong cases we develop a topological operator.

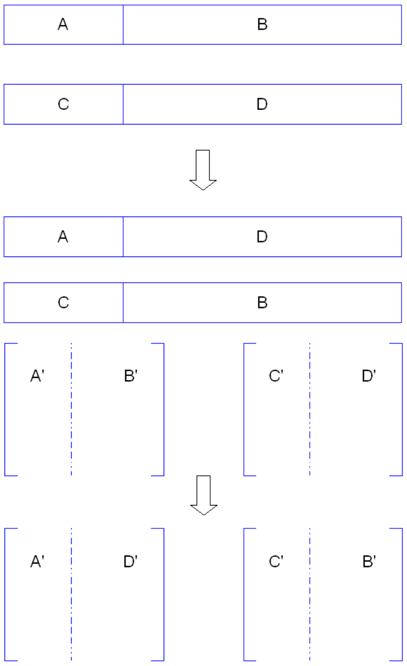


Figure 6: The crossover operators for the chromosomes (top) and for the associated matrix (bottom).

4.6 Topological operator

In [7, 8] we have shown that the protein SSE-IN can be described by their topological properties. We have shown that there exists a parallel between biological classification (notably the SCOP fold family level) and the SSE-IN topological properties. Here we exploit these properties to exclude the incompatible children generated by our GA. The principle is the following, we have predicted a fold family for the sequence from which we extract an initial population of chromosomes. Thus, we compute the diameter, the characteristic path length and the mean degree to evaluate the average topological properties of the family for the particular SSE number. Then, after the GA generates a new individual by crossover or mutation, we compare the associated SS-IN matrix with the properties of the initial population by admitting an error rate up to 20%. If the new individual is not compatible, it is rejected.

4.7 Algorithm implementation

Starting from an initial population of chromosomes from the predicted family, our algorithm modifies the population using the genetic operators defined above. The process is stopped when the fitness of the population stops increasing between two iterations, see Algorithm 1.

The genetic process is the following: after the initial population is built, we extract a fraction of parents according to their fitness and we reproduce them to produce children. Then, we select the new generation by including the chromosomes which are not among the parents plus a fraction of parents plus a fraction of children. It remains to compute the new generation fitness.

Algorithm 1: Genetic algorithm for SS-IN adjacency matrix determination.

```
Data:
pop: Current chromosome population
parents: Set of parents
children: Set of children

begin
    pop ⟵ setInitialPopulation();
    while fitness(pop) is increasing do
        parents ⟵ parentExtraction(pop);
        children ⟵ parentCrossing(parents);
        children ⟵ childrenMutation(children) ;
        children ⟵ exclusionByTopology(children);
        pop ⟵ selection(pop, children);
end
```

4.8 Algorithm performance

At the end of our GA, the final population contains individuals close to the target protein in terms of SSE length distribution because of the choice of our fitness function. As a side effect, their associated matrices are supposed to be close to the adjacency matrix of the studied protein.

In order to test the performance of our GA, we pick randomly three chromosomes from the final popualtion and we compare their associated matrices to the sequence SS-IN adjacency matrix. To evaluate the difference between two matrices, we use an error rate defined as the number of wrong elements divided by the size of the matrix. The dataset we use is composed of 698 proteins belonging to the *All alpha* class and 413 proteins belonging to the *All beta* class, see Fig 7.

The average error rate for *All alpha* class is 16.7% and for *All beta* class it is 14.3%. The maximum error rate is 25%. As shown in Fig 8, the error rate strongly depends on the initial population size. Indeed, when the initial population contains sufficient number of individuals, the genetic diversity ensures better SS-IN prediction. When we have sufficient number of sample proteins from the predicted family, we expect more reliable results. Note for example that when the initial population contains at least 10 individuals, the error rate is always less than 15%.

5 Conclusions and Future Work

In this paper, we are interested in the way the SSEs interact. We propose a genetic algorithm trying to construct the interaction network of SSEs (SS-IN). The GA starts

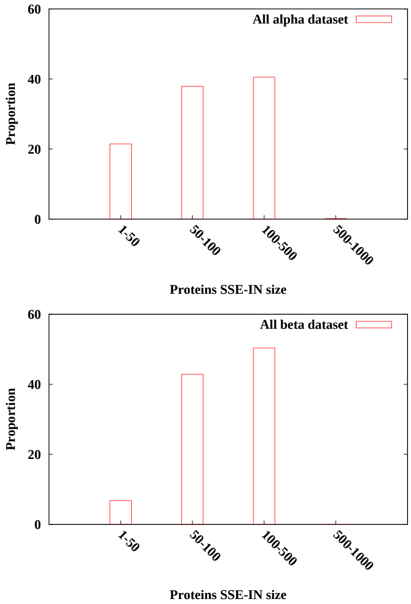


Figure 7: The dataset we use is composed by 698 proteins from the *All alpha* class (top) and by 413 proteins from the *All beta* class (bottom).

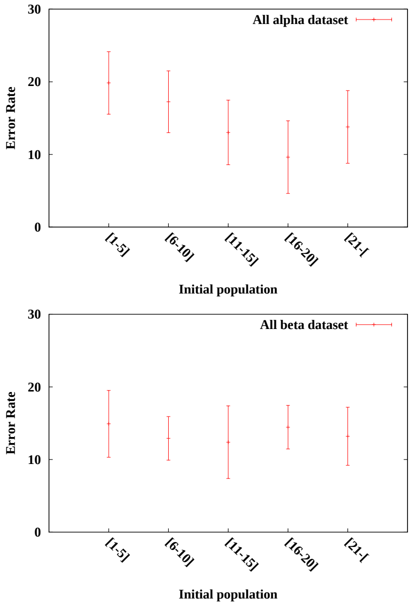


Figure 8: Error rate as a function of the initial population size. When the initial size is more than 10, the error rate becomes less than 15%.

with a population of real proteins from the predicted family. To complete the standard crossover and mutation operators, we introduce a topological operator which excludes the individuals incompatible with the fold family. The GA produces SS-IN with maximum error rate about 25% in the general case. The performance depends on the number of available sample proteins from the predicted family, when this number is greater than 10, the error rate is below 15%.

The next step would be to use the SS-IN prediction in order to build the interaction graph of the target sequence in amino acid level (SSE-IN).

The characterization we propose constitutes a new approach to the protein folding problem. The properties identified here, but also other properties we studied [9, 10], 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] 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. 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.

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

[12] 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.

[13] 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.

[14] 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.