Reconstructing Amino Acid Interaction Networks by an Ant Colony Approach

Omar Gaci, Stefan Balev

Reconstructing Amino Acid Interaction Networks by an Ant Colony Approach

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 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. We rely on a probability that two amino acids interact as a function of their physico-chemical properties coupled to an ant colony system to solve this problem.*

Key words: *ant algorithm, interaction network, protein structure*

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 chapter we consider the problem of reconstructing the network of amino acid interactions when only the protein sequence and its structural family are known.

In this study, we treat proteins as networks of interacting amino acid pairs [8]. 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 identify the interactions involved between the amino acids in the folded protein. First, we build an interaction probability between two amino acids as a function of their physico-chemical properties to predict the graph of the folded protein. Second, we develop an ant colony algorithm to fold or also to reconstruct a sequence SSE-IN.

1.1 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).

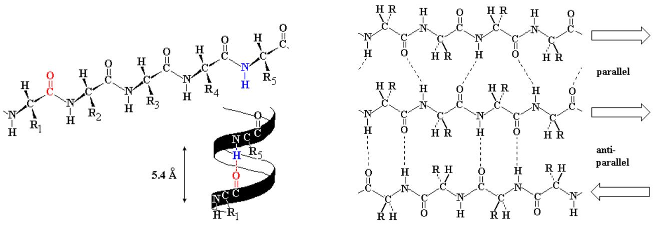


Fig. 1. Left: an α-helix illustrated as ribbon diagram, there are 3.6 residues per turn corresponding to 5.4 Å. Right: A β-sheet composed by three strands.

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

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

1.2 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 [14]. 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 [2] 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 [17, 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.

1.3 A Comparaison of Amino Acid Interactions Prediction

In previous works [10, 11, 13], we have studied the protein SSE-IN. We have identified notably some of their properties like the degree distribution or also the way

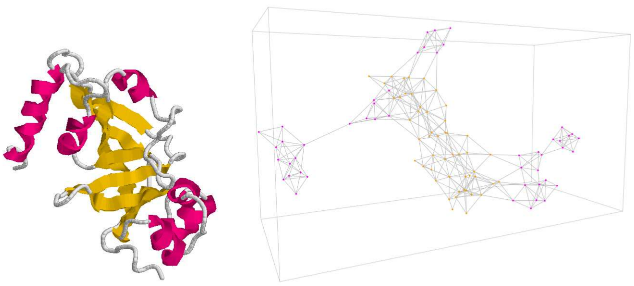
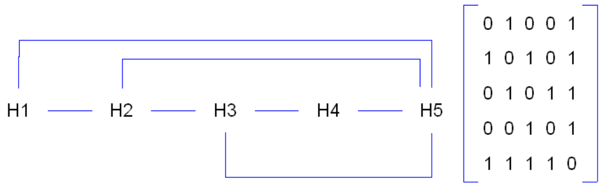


Fig. 2. Protein 1DTP (left) and its SSE-IN (right).

in which the amino acids interact. These works have allowed us to determine criteria discriminating the different structural families. We have established a parallel between structural families and topological metrics describing the protein SSE-IN.

Using these results, in [12] we have proposed a method of predicting the structural family of a protein. Then, a protein or an unknown sequence hasn't a family in the SCOP v1.73 classification. It is defined by its sequence in which the amino acids participating in the secondary structure. This preliminary step is usually ensured by threading methods [16] or also by hidden Markov models [1].

In [12], we consider that the motifs formed by the SSEs in an unknown sequence can be represented as an adjacency matrix where 1 means that the two SSEs are in contact. Then, we use a comparative model built with template proteins from the predicted family and we rely on a genetic algorithm (GA) to predict the unknown sequence matrix of motifs, see Fig.3.



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

Fig. 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 or to reconstruct it, 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. We use the same approach: to fold a SSE-IN we limit the topological space by associating a structural family to a sequence [12]. Since the structural motifs which describe a structural family are limited, we propose a GA to enumerate all possibilities.

Now, in this work, we want to identify the interactions involved between the amino acids in the folded protein, we proceed in two steps.

In section 2, we elaborate a method to predict the interactions between amino acids in the tertiary structure. We start by a disconnected graph from which we want to add edges between amino acids belonging to different SSEs. Relying on a comparative model we evaluate the quantity of edges to add and we build an interaction probability between amino acids considering their chemical properties. Our goal is to identify which parts are in interaction between two SSEs. To measure the performance of our method, we are interested in the contact zones correctly identified.

In section 3, we want to identify which are the edges which link two SSEs, see Fig. 4. considering that the predicted matrix of motif is correct, see Fig. 3. To do that, we start by isolating two SSEs which are in contact and we link each node from the first SSE with each node from the second SSE. Then, it remains to select the suitable edges by using an ant colony approach. When the edges have been collected for each couple of SSE, we consider the disconnected SSE-IN (that is there are none edges inter-SSE yet) in which we add the edges identified during the previous approach. Finally, we select again the global suitable edges using an global ant colony approach.

Further, 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 [12].

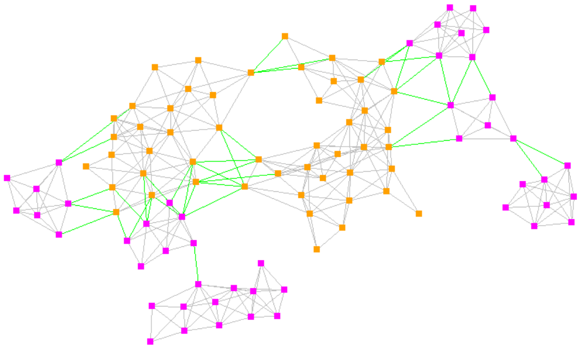


Fig. 4. Protein 1DTP SSE-IN. Interaction we want to predict by an ant system are plotted in green.

2 A Stochastic Interaction Prediction

Here, we want to identify which zones are in contact between two SSEs assuming that the predicted matrix of motif is correct. We start by considering that the sequence SSE-IN is a disconnected graph with no edges between the amino acids. Our goal is to add edges between amino acids belonging to different SSE to rebuild the sequence

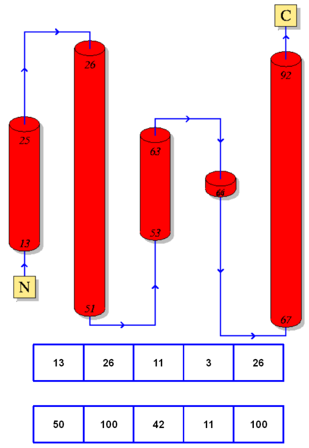
SSE-IN. Thus, to predict the sequence SSE-IN we have to predict the quantity of edges to add and we have to identify the nodes which will be connected.

2.1 Prediction of the edge quantity

In [13] 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).

To predict the shortcut edge rate of the sequence, we rely on the template proteins. We present a protein as an array with elements corresponding to the SSEs, Each cell represents a SSE notably considering its size that is the number of amino acids which compose it. The size is normalized contributing to produce arrays whose cells describe a value between 0 and 100, 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.



13	26	11	3	26

50	100	42	11	100

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

2.2 Contact zones between the nodes to be connected

Our goal is to predict the motifs involved in the unknown sequence, we want to identify which parts are in interaction between two SSEs, that is why we are interested in predicting the zones of contact. Thus, when we add an edge in the disconnected sequence SSE-IN, what is important is not the two nodes linked but rather the zones which the nodes can cover. When measuring the performance of our method, we will be interested in the contact zones correctly identified. We define the notion of a zone for a helix or a sheet as follows.

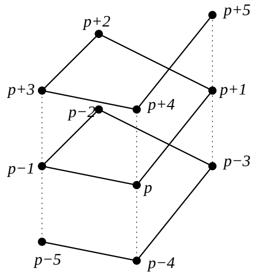


Fig. 6. Contact zones for a node p belonging to a helix

For a node belonging to a helix in position p the contact zones in which this node participates are the quadrilaterals we can form with its spatial neighbours, namely $(p, p+1, p+5, p+4)$, $(p, p+4, p+3, p-1)$, $(p, p-1, p-5, p-4)$ and $(p, p-4, p-3, p+1)$, see Fig. 6.

For a node belonging to a strand in position p, the contact zone contains the node and its two neighbours in the chain, $(p-1, p, p+1)$.

2.3 Prediction of the nodes to be connected

The matrix of motifs determined by the GA (in [12]) indicates which SSEs are in contact and it remains to predict which zones will be in contact. 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. 7 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 $S(a_i, b_j)$:

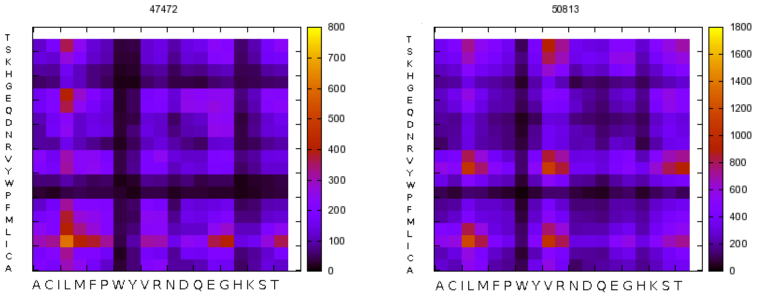


Fig. 7. Occurrence matrix of inter-SSE edges in the family *All alpha* 47472 (left) and *All beta* 50813 (right).

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

In order to add approximately k edges, we need

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

and hence

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

2.4 Measures and Reliability

In [9, 10] we have shown that the proteins SSE-IN can be described by their topological properties. We have shown that there exists a parallel between structural biological classification (notably the SCOP fold family level) and the SSE-IN topological properties. Here we use these properties to exclude the incompatible SSE-IN built by adding of edges using the probabilities p_{ij}, defined above. The principle is the following, we have predicted a fold family for the sequence from which we extract the template proteins. Then, we compute the diameter, the characteristic path length and the mean degree of the extracted proteins to evaluate the average topological properties of the family for the particular SSE number. Then, after we have built the sequence SSE-IN, we compare its topological properties to the average template properties. We admit an error rate up to 20% to accept the built sequence SSE-IN.

We also compare of the edge quantity added the real number of edges. Let m_P be the predicted number of edges and m_R be the real one. Then

$$\text{precision} = 1 - \frac{|m_R - m_P|}{m_P}$$

Table 1. Experimental results by family. The score measures the intersection between the effective and the predicted contact zones.

Class	SCOP Family	Protein Number	Protein Size	Score	Standard Deviation
All α	46688	17	27-46	68.65	14.89
	47472	10	98-125	64.25	6.45
	46457	25	129-135	64.75	13.37
	48112	11	194-200	76.87	18.80
	48507	18	203-214	60.08	10.19
	46457	16	241-281	69.40	5.53
	48507	20	387-422	65.27	16.19
All β	50629	6	54-66	75.71	7.54
	50813	11	90-111	69.13	8.28
	48725	24	120-124	63.17	16.87
	50629	13	124-128	65.22	12.65
	50875	14	133-224	66.36	14.97

2.5 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 (in [12]). For each protein, we have launched 60 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 contact zones.

The scores evolve around 65% meaning that our probabilistic method is relatively reliable. Moreover, we measure the accuracy of the shortcut quantity prediction to see that our prediction stays close to the reality, see Fig. 8. Consequently, our results validate the prediction done in [12].

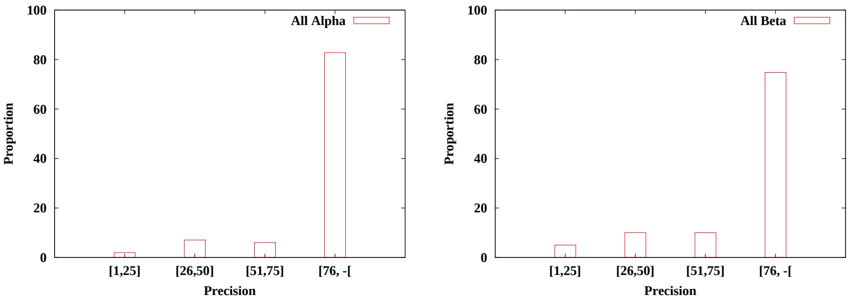


Fig. 8. Precision of the edge quantity added. The most of the predictions are close the real value.

Nevertheless, the presented results are average and there are fluctuations as indicated by the standard deviation values. One can easily understand that the intersection score depends before all on the number of edges we decided to add. If the prediction of the edge quantity to add is precise, then the score is significantly improved. To show this correlation, we have plotted the scores as a function of the shortcut prediction error rate, see Fig. 9. It appears that when the shortcut prediction error rate is less than 25%, the score of contact zone reaches 70%.

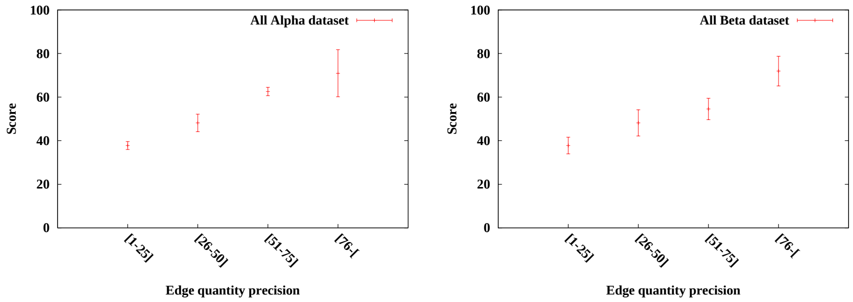


Fig. 9. Score as a function of the precision of the predicted edge quantity

3 Interaction Prediction by Ant Colony Algorithm

Here, we continue the work done in the previous section since we want to predict the nodes which are involved when two SSEs are in contact. We propose an ant colony algorithm to perform this last step of folding a SSE-IN.

In this section we consider that the matrix of motifs is already known. Our goal is to add inter-SSE edges and to reconstruct the 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.

3.1 Ant systems

The studies of social insects have shown that their behavior is self-organized [7] notably concerning the food search. The self-organisation is the process during which structures emerge at collective level resulting from the multiple but simple interactions between the individuals of a population.

The ants have the particularity to communicate using volatile chemical substances called pheromones. The ants drop them to form pheromone trails which lead the other ants and allow for example to find the shortest path between their nest and a food source [15, 3] although the ants do not have a global vision of their environment.

The general concept of the ant systems relies on agents collaborating to search a solution using indirect communication mechanism inspired by the pheromone trails. The artificial ants have the following characteristics:

- an ant is able to perceive locally its environment.
- it has a memory to save the solution it is building.
- it can adapt the pheromone quantity to drop as a function of the solution it is building.

The local perception given to ants is also called heuristic value. Together with the pheromone quantity, it influences their choices.

Formally, at each point where a choice is needed, the artificial ants use probabilistic rules to determine their moves. Let τ_{ij} be the quantity of pheromone present on a path between the points i and j and w_{ij} be the heuristic value associated to this path. Then, the probability p_{ij} that an ant reaches the point j from the point i is given by the next formula:

$$p_{ij} = \frac{[\tau_{ij}]^{\alpha} \ . \ [w_{ij}]^{\beta}}{\sum_{k \in V(i)} [\tau_{ik}]^{\alpha} \ . \ [\eta_{ik}]^{\beta}} \tag{1}$$

where $V(i)$ is the neighborhood of point i. The parameters α and β control the relative influence of the pheromone and the heuristic values.

In nature, as in computer models, the pheromone evaporates. Let ρ be the pheromone persistence factor. Then $1-\rho$ is the evaporation rate. The pheromone quantity is updated according to this rate:

$$\tau_{ij} = (1-\rho)\ \tau_{ij} \ + \ \Delta\tau_{ij}$$

where $\Delta\tau_{ij}$ is the pheromone dropped by the ants on the path between i and j.

The local choices made by the ants will influence the global behavior of the colony. A global behavior emerges from the interactions between the ants without being specified or planned at insect level. In this way, the ant systems allow to find solutions of different problems relying on the population local interactions.

3.2 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. The edge weights are computed as explained in section 2.3. Each ant walks on the graph dropping pheromone on its way. At the end, we will keep the edges with highest pheromone rate and use them for the next phase. An overall description is given as algorithm 1.

The number of ants in our colony is equal to the total number of nodes. The initial positions of the ants are chosen randomly, so at the beginning more ants may be at the same node. Each ant chooses the next node to move randomly, with probabilities defined by eq. 1. The weights of the inter-SSE edges are computed as explained in section 2.3:

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
```

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

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

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

where a_{ij} is the number of ants that moved on the edge (i, j) and $\varDelta\tau$ 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 $phero_{min}$.

3.3 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 2.1. 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 it as explained in section 2.4.

3.4 Results

The experimental settings used in this section are the same as in section 2.5. Tab. 2 summarizes the results. In most of the cases the score is greater than 70% and better than the score obtained by the stochastic approach (compare to Tab. 1). The average score decreases for big proteins (of size more than 200). For these proteins there are many pairs of SSE in contact and the errors made by the local algorithm accumulate to global level.

Table 2. Results of the colony approach. The score measures the intersection between the effective and predicted shortcut edges. The algorithm parameter values are: $\alpha = 25$, $\beta = 12$, $\rho = 0.7$, $\Delta\tau = 4000$, $k = 2$, $phero_{threshold} = 0.8$.

Class	**SCOP Family**	**Protein Number**	**Protein Size**	**Score**	**Standard Deviation**
All α	46688	17	27-46	79.136	8.114
	47472	10	98-125	70.692	15.53
	46457	25	129-135	74.567	9.047
	48112	11	194-200	68.3	14.942
	48507	18	203-214	64.488	7.842
	46457	16	241-281	63.375	16.931
	48507	20	387-422	61.947	9.429
All β	50629	6	54-66	76.136	6.391
	50813	11	90-111	72.567	5.867
	48725	24	120-124	76.692	11.964
	50629	13	124-128	71.3	14.44
	50875	14	133-224	74.488	14.141

4 Conclusion

We propose in this paper a means to fold a protein SSE-IN. Thus, we rely on the Levinthal hypothesis, since the folding process is oriented, we observe four steps to fold a SSE-IN.

First, we limit the topological space by associating to a sequence a structural family [12].

Second, we propose a genetic algorithm trying to construct the interaction network of SSEs. It allows enumerating the possible structural motifs.

Third, we develop a method to identify the motifs involved in a sequence SSE-IN. Then, we build a probability of node interaction between different SSEs and we use a comparative model to predict the quantity of inter-SSE edges to add. The score obtained measure the percentage of correctly predicted contact zones. When the inter-SSE edges prediction error rate is less than 25%, the score of contact zone reaches 70%.

Fourth, we use an ant colony system to predict the nodes involved when two SSEs are in contact. 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. Indeed, 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. Asai, K., Hayamizu, S., and Handa, K., 1993, "Prediction of protein secondary structure by the hidden markov model," Comput. Appl. Biosci., 9(2).
2. Atilgan, A. R., Akan, P., and Baysal, C., 2004, "Small-world communication of residues and significance for protein dynamics," Biophys J., 86(1 Pt 1), pp. 85–91.
3. Beckers, R., Deneubourg, J. L., and Goss, S., 1992, "Trails and u-turns in the selection of a path by the ant lasius niger," J. Theor. Biol., 159, pp.397–415.
4. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne., P. E., 2000, "The protein data bank," Nucleic Acids Research, 28, pp. 235–242.
5. Branden, C., and Tooze, J., 1999, Introduction to protein structure, Garland Publishing, London, UK.
6. Brinda, K. V., and Vishveshwara, S., 2005, "A network representation of protein structures: implications for protein stability," Biophys J., 89(6), pp. 4159–4170.
7. Deneubourg, J. L., Pasteels, J. M., and Verhaeghe, J. C., 1983, "Probabilistic behaviour in ants : a strategy of errors ?" Journal of Theoretical Biology., 105, pp. 259–271.
8. Dokholyan, N. V., Li, L., Ding, F., and Shakhnovich, E. I., 2002, "Topological determinants of protein folding," Proc Natl Acad Sci USA, 99(13), pp. 8637–8641.
9. Gaci, O., and Balev, S., 2008, "Characterization of amino acid interaction networks in proteins," Proc. Journées Ouvertes en Biologie, Informatique et Mathématiques, Lille, France, pp. 59–60.
10. Gaci, O., and Balev, S., 2008, "Proteins: From structural classification to amino acid interaction networks," Proc. the 2008 International Conference on Bioinformatics and Computational Biology, CSREA Press, Las Vegas, pp. 728–734.
11. Gaci, O., and Balev, S., 2009, "Hubs identification in amino acids interaction networks," Proc. 7th ACS/IEEE International Conference on Computer Systems and Applications, Rabat, 7 pages.
12. Gaci, O., and Balev, S., 2009, "Prediction of protein families and motifs by topological inference," Advances in Bioinformatics., in submission.

13. Gaci, O., and Balev, S., 2009, "The small-world model for amino acid interaction networks," Proc. the IEEE AINA 2009, workshop on Bioinformatics and Life Science Modeling and Computing, Bradford, 6 pages.
14. Ghosh, A., Brinda, K. V., and Vishveshwara, S., 2007, "Dynamics of lysozyme structure network: probing the process of unfolding," Biophys J., 92(7), pp. 2523–2535.
15. Goss, S., Aron, S., Deneubourg, J. L., and Pasteels, J. M., 1989, "Self-organized shortcuts in the argentine ant," Naturwissenchaften, 76(12), pp. 579–581.
16. Mirny, B., and Shakhnovich, L., 1998, "Protein structure prediction by threading: Why it works and why it does not," J. Mol. Biol., 283(2), pp. 507–526.
17. Muppirala, U. K. and Li., Z., 2006, "A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues," Protein Eng. Des. Sel., 19(6), pp. 265–275.