Chapter 45
A Study of the Protein Folding Problem by a Simulation Model

Omar Gaci

Abstract In this paper, we propose a simulation model to study the protein folding problem. We describe the main properties of proteins and describe the protein folding problem according to the existing approaches. Then, we propose to simulate the folding process when a protein is represented by an amino acid interaction network. This is a graph whose vertices are the proteins amino acids and whose edges are the 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.

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.

O. Gaci
Le Havre University, 25 rue Phillipe Lebon, 76600 Le Havre, France
e-mail: omar.gaci@gmail.com

S.-I. Ao et al. (eds.), *Machine Learning and Systems Engineering*,
Lecture Notes in Electrical Engineering 68,
DOI 10.1007/978-90-481-9419-3_45,

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 propose to study the protein folding problem. We describe this biological process through the historical approaches to solve this problem. Then, we treat proteins as networks of interacting amino acid pairs [1]. 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 begin by recapitulating relative works about this kind of study model. Then, we present a genetic algorithm able to reconstruct the graph whose vertices represent the SSE and edges represent spatial interactions between them. In other words, this graph is another way to describe the motifs involved in the protein secondary structures.

2 The Protein Folding Problem

Several tens of thousands of protein sequences are encoded in the human genome. A protein is comparable to an amino acid chain which folds to adopt its tertiary structure. Thus, this 3D structure enables a protein to achieve its biological function. In vivo, each protein must quickly find its native structure, functional, among innumerable alternative conformations.

The protein 3D structure prediction is one of the most important problems of bioinformatics and remains however still irresolute in the majority of cases. The problem is summarized by the following question: being given a protein defined by its sequence of amino acids, which is its native structure? In other words, we want to determine the structure whose amino acids are correctly organized in three dimensions in order to this protein can achieve correctly its biological function.

Unfortunately, the exact answer is not always possible that is why the researchers have developed study models to provide a feasible solution for any unknown sequences. However, models to fold proteins bring back to NP-Hard optimization problems [2]. Those kinds of models consider a conformational space where the modeled protein tries to reach its minimum energy level which corresponds to its native structure.

Therefore, any algorithm of resolution seems improbable and ineffective; the fact is that in the absolute no study model is yet able to entirely define the general principles of the protein folding.

2.1 The Levinthal Paradox

The first observation of spontaneous and reversible folding in vitro was carried out by Anfinsen [3]. He deduced that the native structure of a protein corresponds

to a conformation with a minimal free energy, at least under suitable environmental conditions. But if the protein folding is indeed under thermodynamic control, a judicious question is to know how a protein can find, in a reasonable time, its structure of lower energy among an astronomical number of possible conformations.

As example, a protein of 100 residues can adopt 2^{100} ($\approx 10^{30}$) distinct conformations when we suppose that only two possibilities are accessible to each residue. If the passage from a conformation to another is carried out in 10^{-13} s (which corresponds to time necessary for a rotation around a connection), this protein would need at least 10^{17} s, i.e. approximately three billion years, "to test" all possible conformations. The proteins however manage to find their native structures in a lapse of time which is about the millisecond at the second. The apparent incompatibility between these facts, raised initially by Levinthal [4], was quickly set up in paradox and made run enormously ink since.

Levinthal gives the solution of its paradox: proteins do not explore the integrality of their conformational space, and their folding needs to be "guided", for example, via the fast formation of certain interactions which would be determining for the continuation of the process.

2.2 Motivations

To be able to understand how a protein accomplishes its biological function, and to be able to act on the cellular processes in which the protein intervenes, it is essential to know its structure. Many protein native structures were determined experimentally – primarily by crystallography with X-rays or by Nuclear Magnetic Resonance (NMR) – and indexed in a database accessible to all, Protein Data Bank (PDB) [5].

However, the application of these experimental techniques consumes a considerable time [6, 7]. Indeed, the number of protein sequences known [8] is much more important than the number of solved structures [5], this gap continues to grow quickly.

The design of methods making it possible to predict the protein structure from its sequence is a problem whose stakes are major, and which fascine many of scientists for several decades. Various tracks were followed with an aim of solving this problem, elementary in theory but extremely complex in practice.

3 Approaches to Study the Protein Folding Problem

The existing models for the protein folding problem study depend, amongst other things, on the way that the native structure is supposed be reached. Either, a protein folds following a preferential folding path [9], or a protein folds by searching the native state among an energetic landscape organized as a funnel.

The first hypothesis implies the existence of preferential paths for the folding process. In the simplest case, the folding mechanism is comparable to a linear reaction. Thus, when the steps are enough specifics, only a local region of the conformational space will be explored. This concept is nowadays obsolete since we know the existence of parallel folding paths.

The second hypothesis defines the folding by the following way (Dill, 1997):

> a parallel flow process of an ensemble of chain molecules; folding is seen as more like trickle of water down mountainsides of complex shapes, and less like flow through a single gallery.

In other words, the folding can be described as a set of transitions between structures whose energies become weaker. It allows guiding the protein by a funnel effect toward the conformational state whose energy level is the minimum that is the native conformation. Then, the polypeptide chain explores only a fraction of the accessible states.

This last hypothesis is the one accepted in this chapter, the protein folding is a process by which a large number of conformations are accessible and which leads a sequence into its native structure with the lowest energy level (see Fig. 1).

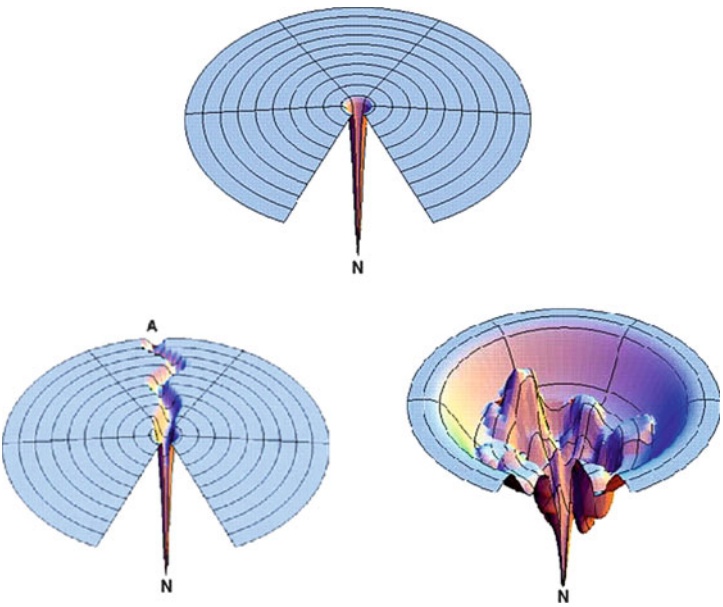

Fig. 1 Evolution in the description of folding paths in an energy landscape. Top, the protein folding according to the Anfinsen theory: a protein can adopt a large number of conformations. Bottom-left, a protein folds by following only one specific path in the conformational space. Bottom-right, from a denatured conformation, a protein searches its native structure whose energy level is minimum in a minimum time

3.1 Latest Approach

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.

These studies are lead with the aim to explain how elements interact with each other inside the network and what the general laws which govern the observed network properties are.

From physics and computer science to biology and social sciences, researchers have found that a broad variety of systems can be represented as networks, and that there is much to be learned by studying these networks. Indeed, the studies of the Web [10], of social networks [11] or of metabolic networks [12] contribute to put in light common non-trivial properties of 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 phenomena acting on their constitution and formation.

In [13], the authors propose to consider a protein as an interaction network whose vertices represent the amino acids and an edge describes a specific type of interaction (which is not the same according to the object of study). Thus, a protein, molecule composed of atoms becomes a set constituted by individuals (the amino acids), by interactions (to be defined according to the study) which evolves in a particular environment (describing the experimental conditions).

The vocabulary evolves but the aim remains the same, we want to better understand the protein folding process by the way of the modeling. The interaction network of a protein is initially the one built from the primary structure. The goal is to predict the graph of the tertiary structure through a discrete simulation process.

3.2 The Amino Acid Interaction Network

The 3D structure of a protein is represented by the coordinates of its atoms. This information is available in Protein Data Bank (PDB), 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 $N\,x\,N$ *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 spatial proximity, so that the contact map can consider non-covalent interactions. We say that two amino acids are in contact if and only if the distance between them is below a given threshold. A commonly used threshold is 7 Å [1] 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.

First, we consider the graph induced by the entire set of amino acids participating in folded proteins. We call this graph the three dimensional structure elements interaction network (3DSE-IN), see Fig. 2.

As well, we consider the subgraph induced by the set of amino acids participating in SSE. We call this graph SSE interaction network (SSE-IN) (see Fig. 2).

In [15] the authors rely on amino acid interaction networks (more precisely they use SSE-IN) to study some of their properties, in particular concerning the role played by certain nodes or comparing the graph to general interaction networks models. Thus, thanks to this point of view the Protein Folding Problem can be tackle by the graph theory.

To manipulate a SSE-IN or a 3DSE-IN, we need a PDB file which is transformed by a parser we have developed. This parser generates a new file which is read by the GraphStream library [16] to display the SSE-IN in two or three dimensions.

4 Folding a Protein in a Topological Space by Bio-Inspired Methods

In this section, we treat proteins as amino acid interaction networks (see Fig. 2). We describe a bio-inspired method we use to fold amino acid interaction networks. In particular, we want to fold a SSE-IN to predict the motifs which describe the secondary structure.

4.1 Genetic Algorithms

The concept of genetic algorithms has been proposed by John Holland [17] to describe adaptive systems according to biological process.

The genetic algorithms are inspired from the concept of natural selection proposed by Charles Darwin. The vocabulary employed here is the one relative to the evolution theory and the genetic. We speak about individuals (potential solutions), populations, genes (which are the variables), chromosomes, parents, descendants, reproductions, etc.

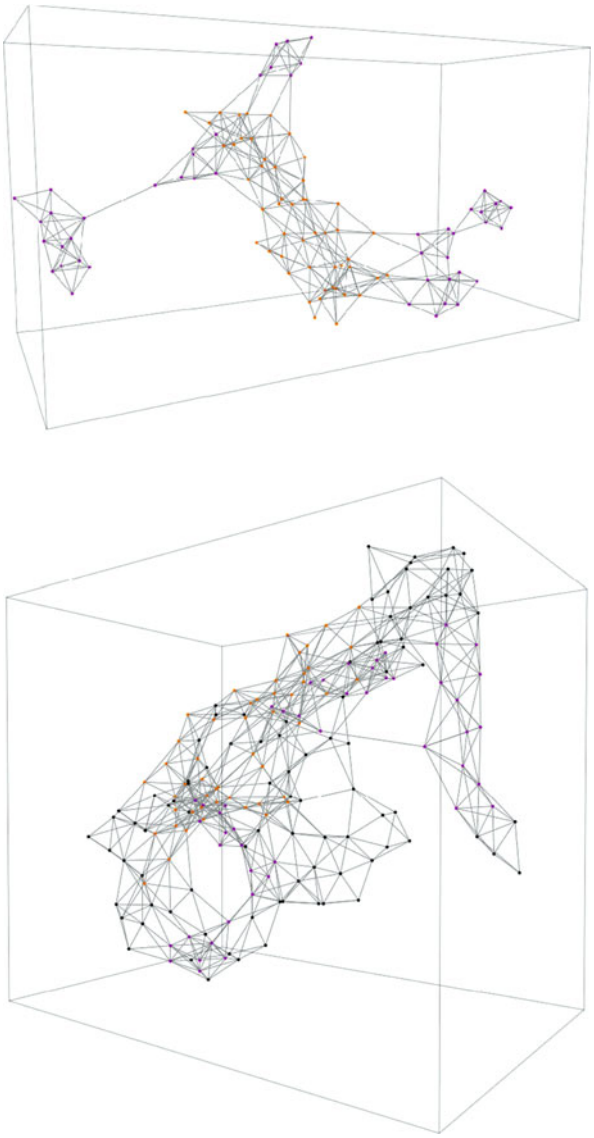

Fig. 2 Protein 1DTP SSE-IN (*top*) and the 1DTP 3DSE-IN (*bottom*). From a pdb file a parser we have developed produces a new file which corresponds to the SSE-IN graph displayed by the GraphStream library [16]

At the beginning, we conserve a population among which it exists a solution which is not yet optimal. Then, the genetic algorithm make evolves this population by an iterative process. Certain individuals reproduce themselves, others mute or disappear and only the well adapted individuals are supposed to survive. The

genetic heritage between generations must help to produce individuals which are better and better adapted to correspond to the optimal solution.

4.2 Motif Prediction

In previous works [18], 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 discriminating the different structural 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 deduce the family of an unclassified protein based on the topological properties of its SSE-IN, see [19]. Thus, we consider a protein defined by its sequence in which the amino acids participating in the secondary structure are known. 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 associating 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.

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 topological space by associating a structural family to a sequence [19]. Since the structural motifs which describe a structural 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).

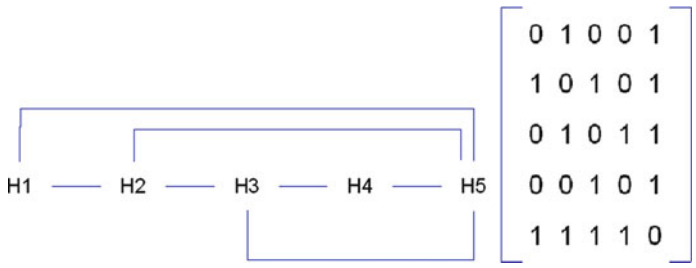

Fig. 3 2OUF SS-IN (*left*) and its associated incidence matrix (*right*). The vertices represent the different α-helices and an edge exists when two amino acids interact

4.3 Dataset

Thereafter, we use a dataset composed by proteins which have not fold families in the SCOP v1.73 classification and for which we have associated a family in [19].

4.4 Overall Description

The GA has to predict the adjacency matrix of an unknown sequence when it is represented by a chromosome. Then, the initial population is composed of proteins of the associated family with the same number of SSEs. During the genetic process, genetic operators are applied to create new individuals with new adjacency matrices. We want to predict the studied protein adjacency matrix when only its chromosome is known.

Here, we represent a protein by 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. In the same time, for each genome we associate its SS-IN incidence matrix.

The fitness function we use to evaluate the performance of a chromosome is the L_1 distance between this chromosome and the target sequence.

4.5 Genetic Operators

Our GA uses the common genetic operators and also a specific topological operator.

The crossover operator uses two parents to produce two children. It produces two new chromosomes and matrices. After generating two random cut positions, (one applied on chromosomes and another on matrices), we swap respectively the both chromosome parts and the both matrices parts. This operator can produce incidence matrices which are not compatible with the structural family, the topological operator solve this problem.

The mutation 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.

These common operators may produce matrices which describe incoherent SS-IN compared to the associated sequence fold family. To eliminate the wrong cases we develop a topological operator.

The topological operator is used to exclude the incompatible children generated by our GA. The principle is the following; we have deduced 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.

4.6 Algorithm

Starting from an initial population of chromosomes from the associated family, the population evolves according to the genetic operators. When the global population fitness cannot increase between two generations, the process is stopped, see Algorithm 1.

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
```

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.

When the algorithm stops, the final population is composed of 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 that we want to predict.

In order to test the performance of our GA, we pick randomly three chromosomes from the final population 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. A structural family has been associated to this dataset in [19].

The average error rate for the *All alpha* class is 16.7% and for the *All beta* class it is 14.3%. The maximum error rate is 25%. As shown in Fig. 4, 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 associated family, we expect more reliable results. Note for example that when

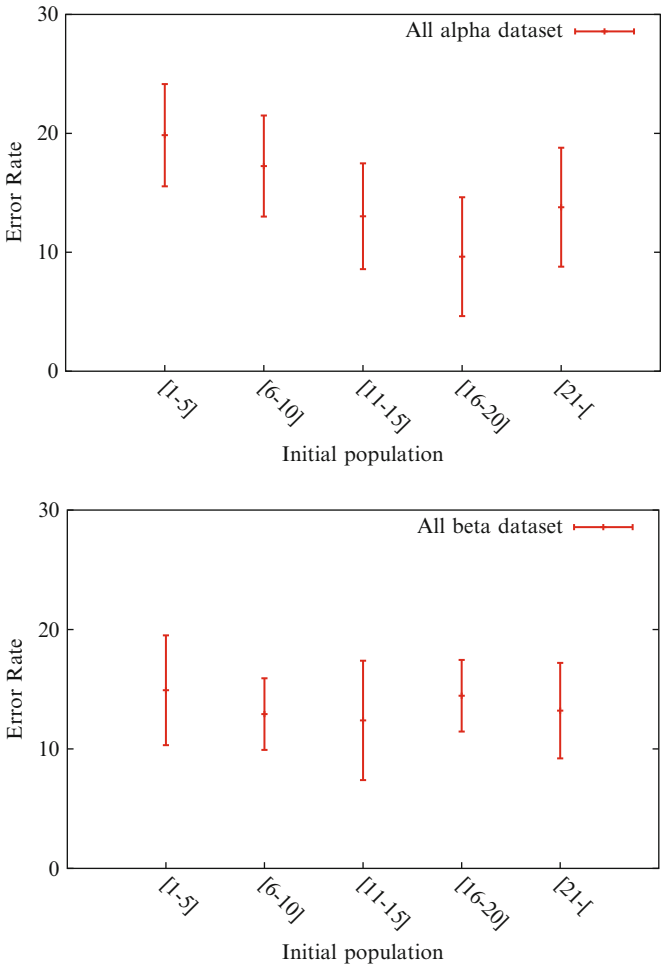

Fig. 4 Error rate as a function of the initial population size. When the initial population size is more than 10, the error rate becomes less than 15%

the initial population contains at least ten individuals, the error rate is always less than 15%.

5 Conclusions

In this paper, we present a simulation model to study the protein folding problem. We describe the reasons for which this biological process is a problem not yet solved.

We summarize relative works about how to fold an amino acid interaction networks. We need to limit the topological space so that the folding predictions become more accurate. We propose a genetic algorithm trying to construct the interaction network of SSEs (SS-IN). The GA starts 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 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. A.R. Atilgan, P. Akan, C. Baysal. Small-world communication of residues and significance for protein dynamics. Biophys. J. **86**(1 Pt 1), 85–91 (2004)
2. K.A. Dill, S. Bromberg, K.Z. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, H.S. Chan, Principles of protein folding: a perspective from simple exact models. Protein Sci. **4**(4), 561–602 (1995)
3. C.B. Anfinsen, Principles that govern the folding of protein chains. Science **181**, 223–230 (1973)
4. C. Levinthal. Are there pathways for protein folding? J. Chim. Phys. **65**, 44–45 (1968)
5. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, P.E. Bourne, I.N. Shindyalov, The protein data bank. Nucleic Acids Res. **28**, 235–242 (2000)
6. M. Kataoka, Y. Goto, X-ray solution scattering studies of protein folding. Folding Des. **1**, 107–114 (1996)
7. K.W. Plaxco, C.M. Dobson, Time-relaxed biophysical methods in the study of protein folding. Curr. Opin. Struct. Biol. **6**, 630–636 (1996)
8. A. Bairoch, R. Apweiler, The swiss-prot protein sequence database and its supplement trembl. Nucleic Acids Res. **28**, 45–48 (2000)
9. R.L. Baldwin, Why is protein folding so fast? Proc. Natl. Acad. Sci. USA **93**, 2627–2628 (1996)
10. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. Graph structure in the Web. Comput. Netw. **33**(1–6), 309–320, (2000)

11. S. Wasserman, K. Faust, Social network analysis: methods and applications. *Structural Analysis in the Social Sciences*, vol. 8 (Cambridge University Press, Cambridge, 1994)
12. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.-L. Barabàsi, The large-scale organization of metabolic networks. Nature **407**(6804), 651–654 (2000)
13. N.V. Dokholyan, L. Li, F. Ding, E.I. Shakhnovich, Topological determinants of protein folding. Proc. Natl. Acad. Sci. USA **99**(13), 8637–8641 (2002)
14. A. Ghosh, K.V. Brinda, S. Vishveshwara, Dynamics of lysozyme structure network: probing the process of unfolding. Biophys. J. **92**(7), 2523–2535, (2007)
15. U.K. Muppirala, 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 (2006)
16. A. Dutot, F. Guinand, D. Olivier, Y. Pigné, GraphStream: A Tool for bridging the gap between Complex Systems and Dynamic Graphs, Proc of EPNACS: Emergent Properties in Natural and Artificial Complex Systems, Dresden, Germany, 137–143 (2007)
17. J.H. Holland, *Adaptation in Natural and Artificial System* (MIT Press, Cambridge, MA, 1992)
18. O. Gaci, Building a parallel between structural and topological properties. In Advances in Computational Biology (Springer, 2010)
19. O. Gaci, Building a topological inference exploiting qualitative criteria. Evol. Boinformatics (2010)