Near-Native Protein Folding

Stefka Fidanova
Institute for Parallel Processing at Bulgarian Academy of Science, Sofia, Bulgaria

***Abstract**: The protein folding problem is a fundamental problem in computational molecular biology. The high resolution 3D structure of a protein is the key to the understanding and manipulating of its biochemical and cellular functions. Protein structure could be calculated from knowledge of its sequence and our understanding of the sequence-structure relationships. Various methods have been applied to solve protein folding problem. Our algorithm is based on reduced hydrophobic-polar (HP) model of protein structure. After that the folding problem is defined like optimization problem.*

***Keywords:** Protein folding, Hydrofobicity, 3D HP model.*

1. INTRODUCTION

Predicting the 3D structure of protein from their linear sequence is one of the major challenges in modern biology. Insights into the 3D structure of a protein are of great assistance when planning experiments aimed at the understanding of protein function and during the drug design process. The experimental elucidation of the 3D structure of proteins is however often hampered by difficulties in obtaining sufficient protein, diffracting crystals and many other technical aspects. Therefore the number of solved 3D structures increases only slowly. Proteins from different sources and sometimes diverse biological functions can have similar sequences and it is generally accepted the high sequence similarity is reflected by distinct structure similarity, but sometimes protein sequences with more than 30% identities have different structures and functions. However, in some cases proteins have similar functions and structures in the absence of high sequence identity.

The protein folding problem is a fundamental problem in molecular biology. Even under simplified lattice models the problem is hard and the standard computational approaches are not powerful enough to search for the correct structure in the huge conformation space.

Efforts to solve the protein folding problem have traditionally been rooted in two schools of thought. One is based on the principles of physics: that is, on the thermodynamic hypothesis, according to which the native structure of a protein corresponds to the global minimum of its free energy. The other school of thought is based on the principles of evolution. Thus methods have been developed to map the sequence of one protein (target) to the structure of another protein (template), to model the overall fold of the target based on that of the template and to infer how the target structure will be changed, related to the template, as a result of substitutions, insertions and detections [2].

According methods for protein-structure prediction has been divided into two classes: do novo modeling and comparative modeling. The de novo approach can be farther subdivided, those based exclusively on the physics of the interactions within the polypeptide chain and between the polypeptide and solvent, using heuristic methods [7, 9, 10], and knowledge-based methods that utilize statistical potential based on the analysis of recurrent patterns in known protein structures and sequences. The comparative modeling models structure by copying the coordinates of the templates in the

aligned core regions. The variable regions are modeled by taking fragments with similar sequences from a database [2, 5].

Due to the complexity of the protein folding problem, simplified models such as hydrophobic-polar (HP) model have become one of the major tools for studying protein structures. The HP model is based on the observation that the hydrophobic force is the main force determining the unique native conformation of globular proteins. The 3D HP model is generally based on 3-dimensional cubic lattice. The energy of a conformation is defined as the number of topological contacts between hydrophobic amino acids that are not neighbors in the given sequence. More specifically, a conformation with exactly *n* H-H contacts has energy $E = n \times (-1)$ for example. The HP protein folding problem is to find an energy-minimizing conformation for given HP sequence.

In this paper different approach is applied. Using HP model is explained the structures in protein conformation observed by biologists. It is do novo modeling first constructing secondary structure before competing it in tertiary structure.

2. HYDROPHOBIC-POLAR PROTEIN MODEL

Determining the functional conformation of a protein molecule from amino acid sequence remains a central problem in computational biology [12]. The experimental determination of these conformation is often difficult and time consuming. To solve this problem it is common practice to use simplified models [11,12].

The hydrophobic-hydrophilic (or hydrophobic-polar) model [6] describes the proteins, based on the fact that hydrophobic amino acids tend to be less explored to the aqueous solvent than the polar ones, thus resulting in the formation of a hydrophobic core in the spatial structure. Albert at all in [1] note that the hydrophobic effect among amino acids contributes so significant a portion of the total energy function that it is the most important force in determining a protein's structure. The hydrophobicity of an amino acid is a measure of the thermodynamic interaction between the side chain and water. The 20 amino acids are classified as hydrophobic (H) or polar (P) by degree of hydrophobicity. Then the HP model simplifies the protein folding problem by considering only two types of amino acids: H and P [4,8].

Polar amino acids are more ionic and bond well with water, while hydrophobic amino acids are less ionic and therefore do not bond as well with water. Therefore folded proteins generally have polar amino acids on the outside of their folded structures and hydrophobic amino acids on the inside. In the HP model the amino acid sequence is abstracted to a binary sequence of monomers that are either hydrophobic or polar. The structure is a chain whose monomers are on the nodes of a three-dimensional cubic lattice, see Fig. 1.

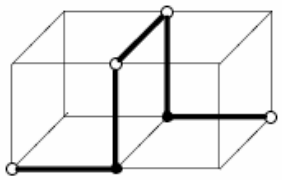

Fig.1 HP protein representation on 3D cubic lattice, the black dots represent hydrophobic amino acids, the white dots represent polar.

The free energy of a conformations is defined as the negative number of non-consecutive hydrophobic-hydrophobic (H-H) contacts. A contact is defined as two non-

consecutive monomers in the chain occupying adjacent sites in the lattice. Thus the problem to find a conformation with less energy, becomes the problem to find a conformation with maximal number of H-H contacts.

In spite of its apparent simplicity finding optimal strictures of the HP model on cubic lattice has been classified as a NP-complete problem [3]. The 3D HP protein folding problem can be formally defined as follows: Given an amino acid sequence $s = s_1, s_2, \ldots, s_n$, find an energy minimizing conformation of *s*, i.e. find $c^s \in C(s)$ such that $E^s = E(c^s) = \min\{E(c) \mid c \in C\}$, where *C*(*s*) is the set of all valid conformations for *s*, and *E* is the energy of the conformation.

3. PROTEIN FOLDING

The problem of finding stead conformation becomes the problem to find a conformation with maximal number of non consecutive H-H contacts. Let is considered polypeptide chain with only hydrophobic monomers or isolated polar monomers inside. As is known it will take a form with minimal energy, i.e. with maximal H-H non consecutive contacts. There are more possibilities for H-H contacts in helix than in sheet. On 3D lattice the helix is represented with four monomers on a level, see Fig. 2.

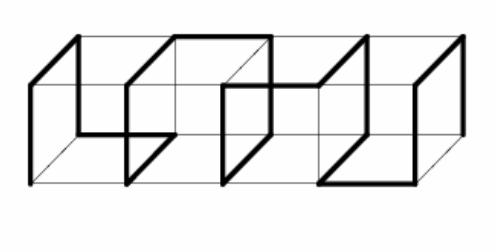

Fig. 2 Helix with 5 levels.

If the diameter of the helix is larger, the number of H-H contacts decrease.

Let the protein chain consists of long part of polar monomers and short part of one or two hydrophobic monomers. The hydrophobic monomers try to create a structure with greater number of H-H contacts. Every polar part forms a β-sheet. Thus the chain is folded like parallel situated β-sheets (hairpin) if there are one H-H contact at every of the ends of the chain or orthogonally packing of β-sheets in other case.

The next configuration considered is two hydrophobic monomers followed by one polar monomers. Like in previous cases the hydrophobic monomers create helix and the polar monomers are situated in the both sides of the hydrophobic. Thus the monomer chain creates larger helix consisting four hydrophobic monomersl and two polar monomers, see Fig.3.

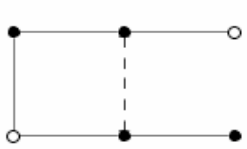

Fig. 3. A level of helix with four hydrophobic monomers inside and two polar. Black dots represent the hydrophobic monomers. Dash-lines represent the H-H contacts.

Let the protein chain consists of repetition of one hydrophobic and two polar monomers. This case is very similar to previous one, but because there is only one hydrophobic monomer between two polar and the hydrophobic monomers can not

create their own helix, they create two parallel columns. Thus the monomer's chain creates a helix consisting two hydrophobic monomers in the middle of every level and four polar monomers, two in both sides. Let the protein chain consists of repetition of two hydrophobic and one polar monomers. Like in previous case the monomer chain creates helix consisting two hydrophobic monomers in the middle of the every level and alternated polar and hydrophobic monomers in two sides. Other types of configurations fold according to other parts of the protein, thus to create maximal number of H-H contacts.

4. EXPERIMENTAL RESULTS

We test our ideas on proteins with known folding. Like tests we choose Pheromone Er22 and Bacteriocin Leucocin A proteins. Both consist of 37 monomers. The amino acid chain of Pheromone Er22 is: DICDIAIAQCSLTLCQDCENTPICELAVKGSCP-PPWS. Its HP representation is: PHPPHHHHPPPHPHPPPPPPPHHPHHHHPHPPHH-HHP. We cut the HP chain in fife parts as follows: (1) PHPPHHHH; (2) PP; (3) PHPH; (4) PPPPPPP; (5) HHPHHHHPHPPHHHHP. Hydrophobic amino acids predominate in the first and in the fifth parts. Thus they form helixes, the first part forms helix with two levels and the fifth part forms helix with four levels (loops). The third part is folded thus to create a maximal number H-H contacts with the first part. Thus it creates something similar to one loop helix. The forth part consists of only hydrophobic amino acids and it forms tight structures which connect the fifth part. We put the helixes, formed by first and fifth parts, parallel each of other. Thus there are additional H-H contacts between them. The achieved folding can be seen at Fig 4.

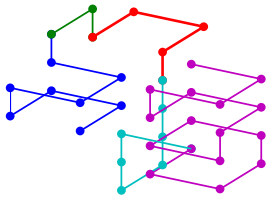

Fig. 4. Pheromone Er22 folding achieved by our algorithm.

Let us consider the real folding of Pheromone Er22, see Fig. 5. We observe that it consists of three parallel situated helixes. The first helix consists of two loops like in our folding. The second helix consists of one loop, like the third part of our folding. The third helix consists of three loops and there is another unstructured loop after it. The fifth part of our folding consists of four loops. Thus we can conclude that there is high similarity between real folding and our folding for Pheromone Er22.

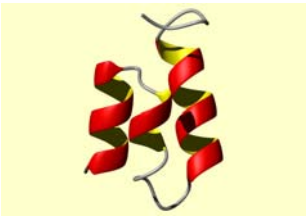

Fig.5. Real Pheromone Er22 folding.

The amino acid chain of Leucocin A is: LYYGNGVHCTKSGCSVAWGQAF-SAGVHRLANGGNGFW. It HP representation is: PPPHPHHPPPPPHPPHHHHPHHP HHHPPHHPHHPHHH. We cut the HP chain of the Leucocin A of three parts as follows: the first part consists of 15 amino acids; the second part consists of 12 amino acids; the third part consists of 10 amino acids. The polar amino acids predominate in the first part. There are several hydrophobic amino acids inside the first part, thus it folds like parallel situated hairpin. The hydrophobic amino acids predominate in the second part. Therefore it folds like helix with three loops. We put the helix orthogonally to the hairpin, because thus there are additional H-H contacts between the hydrophobic ends of the hairpin and the amino acids of the helix. The third part consists of repetitions of one polar and two hydrophobic amino acids. Thus it folds like large helix. After assembling the three parts, we achieve the folding represented on Fig. 6.

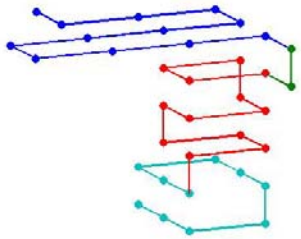

Fig.6 Leucocin A folded by our algorithm

On the Fig. 7 is real folding of Leucocin A. We observe unfolded part and hairpin at the beginning, followed by orthogonally situated helix with three loops. The folding ends with unstructured part, which looks like large loop, exactly like the third part of our folding. We conclude that there is vary high similarity between original and our folding.

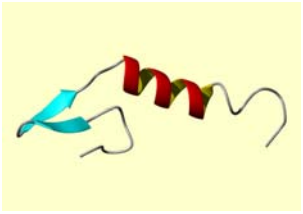

Fig. 7. Real folding of Leucocin A.

5. CONCLUSION

Protein folding is one of the main problems that occur in bio-informatics. It requires knowledge from different disciplines like biology, physical-chemistry. Most of the scientists develop comparison methods, but there are too inaccurate and slow. Other apply metaheuristics but they do not give good results for long proteins yet. Most successful so far approach is fragment assembly. Its relatively low computational cost makes it very useful for large-scale analyses. However, all template-based methods suffer from the fundamental limitation of being able to recognize only folds that have already been observed. Our idea is hybrid between do novo modeling and fragmentation assembly. The HP protein model on 3D lattice is used to model different fragments arising in protein folding. Thus shortcomings of other methods are avoided: the limitations of comparative methods to being already observed and the limitations of constructive methods to can fold well only short proteins. This paper is more theoretical. It explaines

the structures which arise in a tertiary protein form, like helices and β-sheets, maximal and unstructured parts. It can be a basis for more precise folding prediction algorithm.

Acknowledgement: Stefka Fidanova was supported by the Bulgarian Ministry of Education by the grand ``Virtual screening and computer modeling for drug design''.

6. REFERENCES

[1] Albert B., D. Bray, A. Jonson, J. Lewis, M. Raff, K. Roberts, P. Walter (1998) *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell,* Garland Publishing Inc.

[2] Balev S. (2004) Solving the Protein Threading Problem by Lagrangian Relaxation, 4th Int. Workshop on *Algorithms in Bioinformatics*, Bergen, Noeway, LNCS No 3240, 182-193.

[3] Berger B., T. Leighton (1998) Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-complete, *J. Comput. Biology*, Vol 5, 27-40.

[4] Chandru V., A. Dattasharma, V. S. A. Kumar (2003) The Algorithmic of Folding Protein on Lattices, *J. Discrete Applied Mathematics*, Vol 127(1), 145-161.

[5] Chotia C. (2004) One Thousand Families for the Molecular Biologist, *J. Nature Biotechnology*, Vol 22, 1317-1321.

[6] Dill K. A., K. F. Lau (1989) A Lattice Statistical mechanics Model of the Conformational Sequence Spaces of Proteins, *J. Macromolecules*, Vol 22, 3986-3997.

[7] Fidanova S. (2006) 3D HP Protein Folding Problem Using Ant Algorithm, In Proc. of *BioPS Int. Conf*., Sofia, Bulgaria, III.19-26.

[8] Heun V. (2003) Approximate Protein Folding in the HP side Chain Model on Extended Cubic Lattices, *J. Discrete Applied Mathematics*, Vol 127(1), 163-177.

[9] Krasnogor N., D. Petta, P. M. Lopez, P. Mocciola, E. de la Cana (1998) Genetic Algorithm for the Protein Folding Problem: A Critical View, *Engineering of Intelligent Systems*, Alpaydin C. editor, ICSC Academic press., 353-360.

[10] Liang F., W. H. Wang (2001) Evolutionary Monte Carlo for Protein Folding Simulations, *J. Chemical Physics*, Vol 115(7), 444-451.

[11] Lyngso R.B., Pedersen C.N.S. (2000) Protein Folding in the 2D HP Model, In Proceedings of the 1st Journees Ouvert: *Biologie, Informatique et Mathematiques*, JOBIM, Montpellier. (in French)

[12] Pedersen J.T., Moult J. (1996). Genetic Algorithms for Protein Structure Prediction, *Curr. Opin. Struct. Biol*. 6, 227-231.