Bioautomation, 2007, 7, 57 – 63 ISSN 1312 – 451X

Near-native Protein Structure Simulation

Stefka Fidanova

Institute for Parallel Processing – Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str.
1113 Sofia, Bulgaria
E-mail: stefka@parallel.bas.bg

Received: July 17, 2007 **Accepted: September 12, 2007**

Published: October 24, 2007

Abstract: *The protein folding problem is a fundamental problem in computational molecular biology and biochemical physics. The high resolution 3D structure of a protein is the key to the understanding and manipulating of its biochemical and cellular functions. All information necessary to fold a protein to its native structure is contained in its amino-acid sequence. Proteins structure could be calculated from knowledge of its sequence and our understanding of the sequence-structure relationships. Various optimization methods have been applied to formulation of the folding problem. There are two main approaches. The one is based on properties of homologous proteins. Other is based on reduced models of proteins structure like hydrophobic-polar (HP) protein model. After that, the folding problem is defined like optimization problem. It is a hard optimization problem and most of the authors apply Monte Carlo or metaheuristic methods to solve it. In this paper other approach will be used. By HP model is explained the structures of proteins conformation observed by biologists and is studied the correspondence between the primary and tertiary structures of the proteins.*

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

Introduction

The number of amino acids and their sequence give a protein its individua characteristics. The number of amino acids in each protein ranges approximately between 30 and 40000, although most proteins are around hundred amino acids in length. Each protein's sequence of amino acids determines how it folds into a unique three dimensional structure that is its minimum energy state. Knowledge of 3D structure of proteins is crucial to pharmacology and medical science for the following two important reasons. Most drugs work by attaching themselves to a protein so that they can either stabilize the normally folded structure or disrupt the folding pathway which lead to a harmful protein. Thus, knowing exact 3D shapes of proteins will help to design drugs.

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.

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. These models try to generally reflect different global characteristics of protein structures [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.

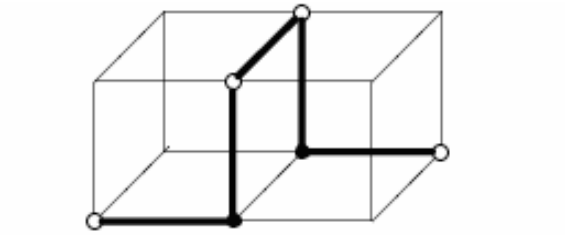

Fig. 1 HP protein representation on 3D cubic lattice, the black dots represent hydrophobic amino acids, the white dots represent polar

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.

Tertiary protein structure and folds

Tertiary structure describes the folding of the polypeptide chain to assemble the different secondary structure elements in a particular arrangement. As helices and sheets are units of secondary structure, so the domain is the unit of tertiary structure. Our aim is to explain structure elements and to predict the tertiary, using HP model on a cubic lattice. The main goal is to find an energy minimizing conformation. The main secondary structures are α-helix and β - sheets.

As is written in previous section, part of the amino acids are hydrophobic (H) and other are polar (P). Thus the polypeptide chain can be represented by binary chain which consists of H and P monomers. The problem of finding stead conformation becomes the problem to find a conformation with maximal number of non consecutive H-H contacts.

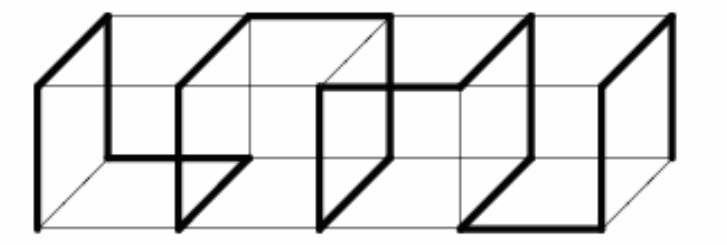

Fig. 2 Helix with 5 levels

Let is considered polypeptide chain with only hydrophobic monomers. 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 cubic lattice the helix will be represented with four monomers on a level, see Fig. 2. If the diameter of the helix is larger the number of H-H contacts decrease. Let the helix be divided in to two parallel situated helices with almost equal number of levels. Let h is the number of levels of the initial helix. Thus 3 H-H contacts are destroyed and $2\lfloor h/2 \rfloor - 1$ contacts are created. To increase the number of H-H contacts is needed $2\lfloor h/2 \rfloor - 1 > 3$, i.e. $2\lfloor h/2 \rfloor > 4$ hence $h \geq 6$ or if a hydrophobic string consists more than 24 monomers it will create more then one helices. Let us divide every of two helices in two almost equal parts and put the four new helices like four-helix bundle. In this case the destroyed H-H contacts are 7 and new contacts are $4\lfloor h/2 \rfloor - 2$. Thus to increase the number of H-H contacts $4\lfloor h/2 \rfloor - 2 > 7$ or $h \geq 6$. If the topology of the four helices are other the number of H-H contacts are less. Thus can be concluded that if there are more than 24 consecutive hydrophobic monomers, they will create more than one helix or the maximal length of the hydrophobic helix is 24 monomers.

Let us consider chain with only polar monomers at the beginning or at the end. The polar monomers do not create H-H contacts, thus this protein will have unstructured part at the beginning or at the end respectively.

Let considered cases with polar monomers inside the protein chain. If the configuration consists of more than two polar monomers and minimum 4 hydrophobic monomers in every of two sides, the conformation with maximal number of H-H contacts is two parallel situated helices and β - sheet. If the configuration consists of one polar monomer and more than 3 hydrophobic monomers in every of two sides, the conformation with a maximal number of H-H contacts is two parallel situated helices.

Let the protein chain consists of long part of polar monomers and short parts of one or two hydrophobic monomers. The hydrophobic monomers will try to create a structure with greater number of H-H contacts. As is written above this form is helix. Every polar part will form a β - sheet. Thus the chain is folded like orthogonal packing of β - sheets.

Let the protein chain consists of repetition of the following group of monomers: HHHPPHHHPPPHHPPP. The hydrophobic monomers try to create H-H contacts. Like in upper case this form is helix. The difference with upper case is that the polar parts are too short to form β - sheets. Thus the protein conformation is two parallel helices and the hydrophobic monomers are in the interior part between the helices with the polar monomers in the exterior part. Let this helix consists of h levels. If it is divided on two, maximum 4 H-H contacts will be destroyed and $2\lfloor h/2 \rfloor$ new contacts will be created. Thus like in a case of

hydrophobic helix the optimal number of helices levels is 6, i.e. 24 monomers. The helices will be situated on two parallel lines with hydrophobic part inside and polar part outside, see Fig. 3.

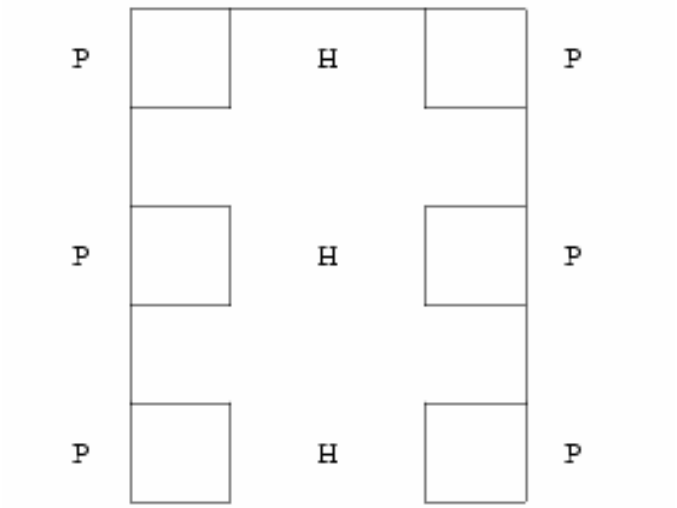

Fig. 3 Parallel helices with hydrophobic part inside and polar part outside

Let the protein chain consists of repetition of one hydrophobic and one polar monomers. Thus the distances between hydrophobic monomers are too short to create there own folding. Thus this part will be folded according other already folded parts.

The next configuration considered is two hydrophobic monomers followed by two 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 large helix consisting four hydrophobic monomers in the middle of every level and four polar monomers, two in every side, see Fig. 4. Let this helix consists of h levels. If it is split on two, 4 H-H contacts will be destroyed and $\lfloor h/2 \rfloor$ new H-H contacts will be created. Thus the optimal number of helix level is 6, i.e. 48 monomers.

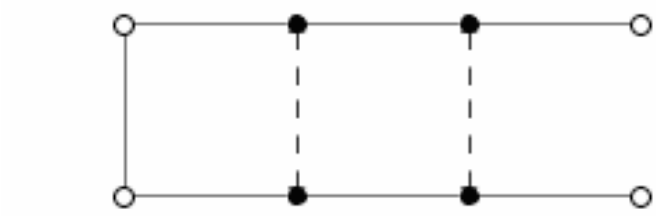

Fig. 4 A level of helix with four hydrophobic monomers inside and two polar in the both out side. Black dots represent the hydrophobic monomers. Dash-lines represent the H-H contacts

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, see Fig. 5. Like in upper case the optimal number of helix level is 6, i.e. 36 monomers.

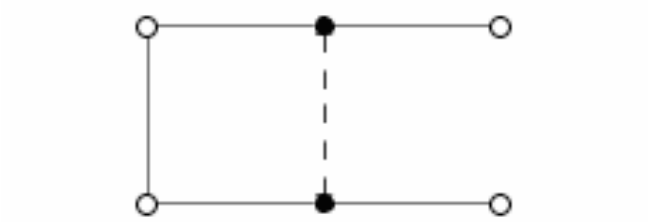

Fig. 5 A level of helix with two hydrophobic monomers inside and two polar in the both out side. Black dots represent the hydrophobic monomers. Dash-lines represent the H-H contacts.

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, see Fig. 6. If it is split on two, 2 H-H contacts will be destroyed and $2\lfloor h/2 \rfloor$ new H-H contacts will be created. Thus the optimal number of helix level is 3, i.e. 18 monomers. In all other configurations the number of helix levels is 6 only in this case it is less.

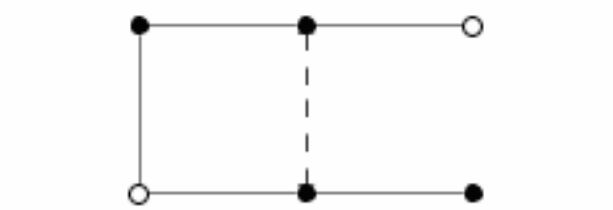

Fig. 6 A level of helix with repetition of two hydrophobic and one polar monomers. Black dots represent the hydrophobic monomers. Dash-lines represent the H-H contacts

The last considered configuration is a protein chain with repetition of three hydrophobic monomers followed by tree polar monomers. This configuration can not be structured part like helix and sheets, thus it forms unstructured part which folds according to other parts of the protein.

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 length of the helices and unstructured parts. It can be a basis for more precise folding prediction algorithm.

Acknowledgements

Stefka Fidanova was supported by the European Community grand RISCK 21 and by the Bulgarian Ministry of Education by the grand "Virtual screening and computer modeling for drug design".

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, 3240, 182-193.
3. Berger B., T. Leighton (1998). Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-complete, J. Comput. Biology, 5, 27-40.
4. Chandru V., A. Dattasharm, V. S. A. Kumar (2003). The Algorithmic of Folding Protein on Lattices, J. Discrete Applied Mathematics, 27(1), 145-161.
5. Chotia C. (2004). One Thousand Families for the Molecular Biologist, J. Nature Biotechnology, 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, 22, 3986-3997.
7. Fidanova S. (2006). 3D HP Protein Folding Problem using Ant Algorithm, - Proc. of Int. Conf. "Bioproces Systems 2006 - BioPS'06", Sofia, Bulgaria, III.19-III.26.
8. Heun V. (2003). Approximate Protein Folding in the HP side Chain Model on Extended Cubic Lattices, J. Discrete Applied Mathematics, 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, 115(7), 444-451.
11. Lyngso R. B., C. N. S. Pedersen (2000). Protein Folding in the 2D HP Model, - Proc. of the 1st Journees Ouvert: Biologie, Informatique et Mathematiques, JOBIM, Montpellier, (in French).
12. Pedersen J. T., J. Moult (1996). Genetic Algorithms for Protein Structure Prediction, Curr. Opin. Struct. Biol., 6, 227-231.