BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **13**, No 4

Sofia • 2013

Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2013-0056

Application of HPD Model for Predicting Protein Mutations[1]

Stefka Fidanova

Institute of Information and Communication Technologies, 1113 Sofia
Email: stefka@parallel.bas.bg

Abstract: *The proteins are one of the most important part of the organisms. They are complex macromolecules that perform a vital function in all living beings. They are composed of a chain of amino acids. The biological function of a protein is determined by the way it is folded into a specific 3D structure, known as native conformation.*

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 realizations. Various methods have been applied to solve the protein folding problem. In this paper the protein is represented like a sequence over a 3-letter alphabet according to the specific functions of amino acids. After that the folding problem is defined as an optimization problem. Our protein model is multifunctional. It can be used to predict the 3D structure of the protein from its amino acid sequence. The model can predict the changes in the protein folding when several amino acids are mutated. A protein can be constructed by it with the needed 3D folding. In this paper we have concentrated on predicting protein folding changes when some amino acids are mutated.

Keywords: *Protein folding, hydrophobic and hydrophilic amino acids, destructor, HPD model, amino acids mutation.*

[1] This work has been partially supported by the European Commission Project ACOMIN.

1. Introduction

Predicting the structure of proteins 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 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 the solved 3D structures increases only slowly. Proteins from different sources and sometimes diverse biological functions can have similar sequences and it is generally accepted that high sequence similarity with more than 30 % identities have different structures and functions. However, in some cases proteins have functions and structures in the absence of high sequence identity.

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 the 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].

Accordingly, the methods for protein-structure prediction have been divided into two classes: de novo modeling and comparative modeling. The de novo approach can be further subdivided, those based exclusively on the physics of the interactions within the polypeptide chain and between the polypeptide and solvent, using heuristic methods [7, 11, 13] and knowledge-based methods that utilize the statistical potential based on the analysis of recurrent patterns in known 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 the Hydrophobic-Polar (HP) model has become one of the major tools for studying protein structures [6]. 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 a 3D cubic lattice. The energy of 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 $E = n\,(-1)$ for example. The HP protein folding problem is finding and energy-minimizing conformation for a given HP sequence.

In this paper a different approach is applied. In our previous work [8] we expand the HP model, adding a third letter D (HPD model) for Proline amino acid, because it has special biological functions. Using HPD model explains the structures in protein conformation observed by biologists. It is de novo modeling

first constructing secondary structures before completing them in a tertiary structure. In this work we concentrate on the application of HPD model for changes in protein folding when some amino acids mutate. This study is important because it can be used for the design of blockers and other drugs.

2. Extended hydrophobic-polar model

Determining the functional conformation of a protein molecule from an amino acid sequence remains a central problem in computational biology [14]. The experimental determination of these conformations is often difficult and time consuming. To solve this problem it is common practice to use simplified models [13, 14].

The hydrophobic-hydrophilic (or hydrophobic-polar) model 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. A l b e r t et al. [1] note that the hydrophobic effect among amino acids contributes to so significant 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 the measure of the thermodynamic interaction between the side chain and water. The 20 amino acids are classified as Hydrophobic (H) or Polar (P) by the degree of hydrophobicity. Then the HP model simplifies the protein folding problem by considering only two types of amino acids: H and P [4, 9, 16].

Polar amino acids are more ionic and bond well with water, while hydrophobic amino acids are less ionic and therefore do not bond so well with water. Therefore, folded proteins generally have polar amino acids on the outside of their folded structure and hydrophobic amino acids on the inside. In 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 (Fig. 1).

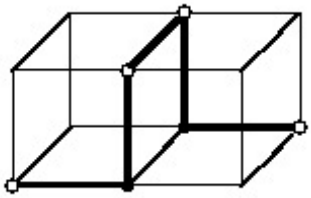

Fig. 1. HP protein representation on a 3D cubic lattice, the black dots represent hydrophobic amino acids, while the white dots represent polar amino acids

The free energy of conformations is defined as the negative number of nonconsecutive Hydrophobic-Hydrophobic (H-H) amino acids. The 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 a problem to find a conformation with the maximal number of H-H contacts. In spite of its apparent simplicity, the folding optimal structures of the HP model on a 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.

It is known that Prolline amino acid has a special biological feature [12]. On one side it is a hydrophobic amino acid. On the other side it acts as a structural disruptor in the middle of secondary structure elements, such as α helices. However, Proline is commonly found as the first residue of an α helix. Therefore we expand HP model adding a third letter D (Disruptor) for Proline residue. So the problem to find the native folding of the protein is to find the folding with the maximal number of H-H and H-D contacts, taking into account that D is at the beginning of the helix.

3. Protein folding

As written in the previous sections, some of the amino acids are hydrophobic (H), others are Polar (P) and Disruptors (D). Thus the polypeptide chain can be represented by a three-letters chain which consists of H, P and D monomers. The problem of finding a steady conformation becomes a problem to find a conformation with the maximal number of non-consecutive H-H and H-D contacts. Even under simplified lattice models the problem is hard and the standard computational approach is not powerful enough to search for the correct structure in the huge conformation space. Most of the authors use metaheuristic algorithms to solve the problem [7, 11, 10, 13]. The main disadvantage of metaheuristics is that they achieve close to the real folding for short proteins only. So our idea is to cut the monomers chain into shorter chains, to fold them and after that to connect the folded parts thus as to cause additional H-H and H-D contacts between the parts. The next question is how to cut the monomer chain. Therefore we try to understand what the folding is, if the monomers chain has a special structure.

Let us consider a polypeptide chain with only hydrophobic monomers or isolated polar monomers inside. As known, it takes a form with the minimal energy, i.e. with maximal H-H and H-D non-consecutive contacts. There are more possibilities for H-H and H-D contacts in helix than in sheets or other confirmation. On a 3D lattice the helix is represented with four monomers on every loop, see Fig. 2. If the diameter of the helix is larger, the number of H-H and H-D contacts decreases. Let there is one D monomer inside a hydrophobic chain. Then the hydrophobic helix is separated into two consecutive helices and the second helix starts with a D monomer.

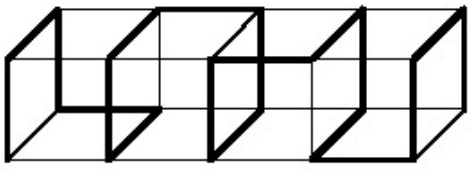

Fig. 2. Helix with five loops

Let the protein chain consists of a long part of polar monomers and a short part of one or two hydrophobic monomers at the ends. The hydrophobic monomers try

to create a structure with a greater number of H-H and H-D contacts. Every polar part forms a β sheet. Thus the chain is folded like parallel situated β sheets (hairpin). If there are several consecutive polar parts with one or two hydrophobic monomers between them, the fold is an orthogonal packing of β sheets.

The next configuration considered is two hydrophobic monomers followed by one polar monomer (PHHPHHPHH). Like in the previous cases the hydrophobic monomers create helix and the polar monomers situated on both sides of the hydrophobic. Thus the monomer chain creates a large helix consisting of four hydrophobic monomers and two polar monomers (Fig. 3).

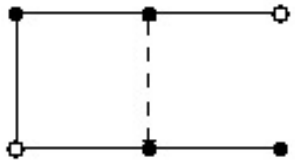

Fig. 3. A loop of a helix with four hydrophobic monomers and two polar. The black dots represent the hydrophobic monomers. The dash lines represent the H-H contacts

Let the protein chain consists of the repetition of one polar and one hydrophobic monomer (PHPHPH). This chain cannot create H-H contacts, but if there are two consecutive chains of this kind, with two polar or two hydrophobic monomers between them (PHPHPHPPHPHP or PHPHPHHPHPHPH), they fold like a hairpin. Other types of configurations we call unstructured and fold them using any metaheuristic method if they are large, or according to the other parts of the protein, thus creating the maximal number of H-H and H-D contacts.

4. Application of HPD model at a protein mutation

Some illnesses are provoked by viruses. Immunostimulators are used to treat them. Others are provoked by bacterias, then antibiotics are applied. There are illnesses provoked by wrong synthesis of proteins; these are autoimmune illnesses and they can be treated by blockers. In order to prepare a blocker, the protein provoking the illness is mutated in its inactive part. The mutation must be thus executed, that the new protein has the same or a similar folding. After that, from all candidates for a drug, the one is chosen without any or with less circumstantial effects. We apply HPD model on γ interferon.

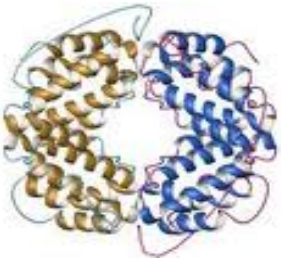

Fig. 4. γ interferon

The mutation is made by replacing amino acids in positions 86, 87, 88 by other amino acids. The replacements used and their HPD representations can be seen in

Table 1. These positions are chosen because they are far from the active part of the protein.

Table 1. Mutations in positions 86, 87, 88

No	Mutation	HPD representation	No	Mutation	HPD representation
1	Pro Tyr Leu	HPH	53	Arg Pro Ser	PDP
2	Pro Asn Tyr	HPP	54	Arg Ser Cys	PPH
3	Trp Ser Ser	HPP	55	Pro Phe Leu	DHH
4	Val Ser Arg	HPP	56	Leu Tyr Pro	HPD
5	Pro Leu Ser	HHP	57	Pro Val Phe	DHH
6	His Val Cys	PPH	58	Pro Met Phe	DHH
7	Pro Tyr Val	HPH	59	Ser Phe Phe	PHH
8	Arg Ser Ser	PPP	60	His Ala Ala	PHH
9	Phe Ser Arg	HPP	61	Pro Phe Ser	DHP
10	Pro Cys Cys	HHH	62	Ala Thr Ala	HPH
11	Pro Ser Val	HPP	63	Leu Phe Ser	HHP
12	Thr Phe Trp	PHH	64	Leu Val Ser	HHP
13	Leu Pro Phe	HDH	65	Phe Leu Val	HHH
14	Asp Leu Leu	PHH	66	Phe Leu Val	HHH
15	Ala His Leu	HPH	67	Pro Arg Ser	DPP
16	Thr Val Leu	PHH	68	Pro Arg Ser	DPP
17	Cys Phe Pro	HHD	69	Pro Arg Ser	DPP
18	Ser Thr Phe	PPH	70	Phe Ser Arg	HPP
19	Pro Ser Pro	DPD	71	Leu Tyr Phe	HPH
20	Ser Ser Leu	PPH	72	Arg Ser Ala	PPH
21	Val Ser Gly	HPH	73	Gln Phe His	PHP
22	Thr Pro Thr	PDP	74	Val Leu Leu	HHH
23	Cys His Phe	HPH	75	Val Leu Pro	HHD
24	Ser Val Ser	PHP	76	Val Ser Ala	HPH
25	Glu Met Pro	PHD	77	Thr Leu Val	PHH
26	Leu Thr Pro	HPD	78	Gln Ala Gly	PHH
27	Leu Pro Pro	HDD	79	Leu Ser Val	HPH
28	Pro Pro Thr	DDP	80	Ser Leu Phe	PHH
29	Phe Ser Leu	HPH	81	Tyr Ala Phe	PHH
30	Phe Phe Pro	HHD	82	His Tyr Pro	PPD
31	Leu Cys Pro	HHD	83	Ala Ser Leu	HPH
32	Pro Ser Ala	DPH	84	Phe Pro Leu	HDH
33	Asp Leu Leu	PHH	85	Pro Pro Ser	HHP
34	Ala Phe Phe	HHH	86	Thr Asn Gly	PPH
35	Leu Leu His	HHP	87	Val Ser Pro	HPD
36	Thr Leu Leu	PHH	88	Ser Pro Pro	PDD
37	Phe Thr Ala	HPH	89	Phe Pro Ser	HDP
38	His Pro Leu	PDH	90	Cys Ser Pro	HPD
39	Phe Thr Arg	HPP	91	Cys Ala Pro	HHD
40	Arg Leu Arg	PHP	92	Ser Phe Cys	PHH
41	Pro Leu Ala	DHH	93	Leu Phe Glu	HHP
42	Phe Cys Arg	HHP	94	Phe Thr Pro	HPD
43	His Ser Arg	PPP	95	His Gln Arg	PPP
44	Pro Tyr Pro	DPD	96	Leu Ser Ser	HPP
45	Ser Leu Leu	PHH	97	Trp Leu Ser	HHP
46	Trp Ser Ala	HPH	98	Leu Thr Ala	HPH
47	Trp Ser Ala	HPH	99	Ser Phe Cys	PHH
48	Ala Ile Pro	HHD	100	Ile Ser Asp	HPP
49	Arg Pro Val	PDH	101	Phe Tyr Thr	HPP
50	Phe Cys Arg	HHP	102	Pro Leu Phe	DHH
51	Pro Phe Ala	DHH	103	LysLysGln	PPP
52	Arg Arg Ser	PPP			

The amino acid chain of γ interferon is as follows:

Gln Asp Pro Tyr Val Lys Glu Ala Glu Asn^{10} Leu Lys Lys Tyr Phe Asn Ala Gly His Ser^{20}

Asp Val Ala Asp Asn Gly Thr Leu Phe Leu^{30} Gly Ile Leu Lys Asn Trp Lys Glu Glu Ser^{40}

Asp Arg Lys Ile Met Gln Ser Glu Ile Val^{50} Ser Phe Tyr Phe Lys Leu Phe Lys Asn Phe^{60}

Lys Asp Asp Gln Ser Ile Gln Lys Ser Val^{70} Glu Thr Ile Lys Glu Asp Met Asn Val Lys^{80}

Phe Phe Asn Ser Asn Lys Lys Lys Arg Asp^{90} Asp Phe Glu Lys Leu Thr Asn Tyr Ser Val^{100}

Thr Asp Leu Asn Val Gln Arg Lys Ala Ile^{110} His Glu Leu Ile Gln Val Met Ala Glu Leu^{120}

Ser Pro Ala Ala Lys Thr Gly Lys Arg Lys^{130} Arg Ser Gln Met Leu Phe Arg Gly Arg Arg^{140}

Ala Ser Gln^{143}

The amino acids in positions 86, 87 and 88 are Lys, Lys, Lys and their HPD representation is PPP because they are polar. In Fig. 5 the helix structure of γ interferon is represented. Position 86 is in helix D and positions 87 and 88 are in helix E.

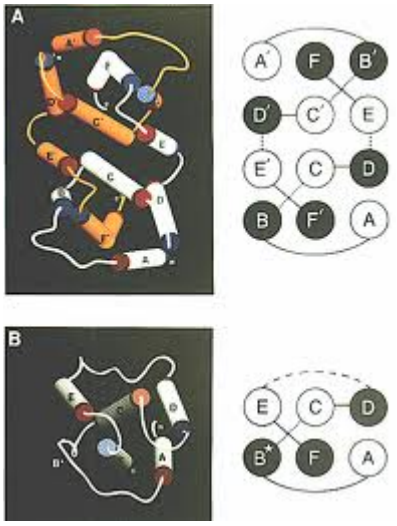

Fig. 5. Helices in γ interferon

After HPD representation of the mutations, there are 24 different kinds of mutations instead of 103. All mutations which have HPD representation PPP will

not change the protein structure, because the HPD representation of the original amino acids at positions 86, 87, 88 have the same representation. The Pro amino acid is a destructor and cuts the helixes, but if it is at the beginning or at the end of the helix it will not change it, therefore the mutations with HPD representation PDP, PDH, PDD, HDP, HDH , HDD will not change the protein structure. If the HPD representation of the mutation is PPD, then the helix E becomes shorter because of the Pro (D) at position 88. If the HPD representation of the mutation is PHP, PHH, PPH, then the structure of the protein will be unchanged, because the number of H-H contacts increases without changing the number of polar amino acids in the unstructured part. If the HPD representation of the mutation is HPD, PHD, DPD, HHD, then the helix E becomes shorter because of the Pro (D) at position 88. If the HPD representation of the mutation is HHP, HHH, HPP, HPH, the number of the hydrophobic amino acids increases and the structure remains unchanged. If the HPD representation of the mutation is DHP, DHH, DPP, DPH, DDP, the helix E will become larger, because there is Pro (D) before helix E.

In order to make a comparison, the mutations were tested on GROMACS (GROningen MAchine for Chemical Simulations). Every one of the mutations was run for 10 hours. There is 80 % coincidence between our methodology and GROMACS. The other 20 % are when 10 hours are not enough to finish the calculations. Thus we can conclude that our methodology gives very close to the real results in protein mutation and that the method is very fast.

5. Conclusion

We have proposed a methodology for protein folding prediction. We have tested our ideas on 103 mutations of γ interferon. We compare the results achieved by GROMACS simulation and the coincidence is 80 %. We can conclude that our methodology is very fast and gives a result close to the real one, applying protein mutations.

References

1. Albert, B., D. Bray, S. A. Jonson, J. Lewis, M. Raff, K. Roberts, P. Walter. Essential Cell Biology: An Introduction to the Molecular Biology of the Cell. Garland Publishing, Inc., 1998.
2. Balev, S. Solving the Protein Threading Problem by Lagrangian Relaxation, Algorithms in Bioinformatics. S. Istrail, P. Pevzner, M. Waterman, Eds. – In: Lecture Notes in Computer Sciences. No 3240. Springer, 2004, 182-193.
3. Berger, B., T. Leighton. Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-Complete. – Computational Biology, Vol. **5**, 1998, 27-40.
4. Chandru, V., A. Dattasharma, V. S. A. Kumar. The Algorithm of Folding Protein on Lattice. – Discrete Applied Mathematics, Vol. **127**, 2003, No 1, 145-161.
5. Chotia, C. One Thousand Families for the Molecular Biologist. – Nature Biotechnology, Vol. **22**, 2004, 1317-1321.
6. Dill, K., K. M. Fiebig, H. S. Chan. Cooperativity in Protein-Folding Kinetics. Nat. Acad. Sci., USA, 1993, 1942-1946.

7. F i d a n o v a, S. 3D HP Protein Folding Problem Using Ant Algorithm– In: Proc. of BioPS Int. Conf., Sofia, Bulgaria, III. 2006, 19-26.
8. F i d a n o v a, S. HPD Model for Protein Structure Simulation. – In: Proc. of 5th International Conference Computer Science'2009, Sofia, Bulgaria, 2010, 2006, 336-341. ISBN 978-954-438-853-9.
9. H e u n, V. Approximate Protein Folding in the HP Side Chain Model on Extended Cubic Lattices. – Discrete Applied Mathematics, Vol. **127**, 2003, No 1, 63-177.
10. H o q u e, T., M. C h e t t y, A. S a t t a r. Extended HP Model for Protein Structure Prediction. – Computational Biology, Vol. **16**, 2009, No 1, 85-103.
11. K r a s n o g o r, N., D. P e l t a, P. M. L o p e z, P. M o c c i o l a, P. de la C a n a. Genetic Algorithms for the Protein Folding Problem: A Critical View. Engineering of Intelligent Systems. C. Alpaydin, Ed. ICSC Academic Press, 1998, 353-360.
12. L e v i t, M. Effect of Proline Residues on Protein Folding. – Molecular Biology, Vol. **145**, 1981, 251-263.
13. L i a n g, F., W. H. W o n g. Evolutionary Monte Carlo for Protein Folding Simulations. – Chemical Physics, Vol. **115**, 2001, No7, 444-451.
14. L y n g s o, R. B., C. N. S. P e d e r s e n. Protein Folding in the 2D HP Model. – In: Proc. of 1st Journees Ouverst: Biology, Informatique et Mathematiques, Montpellier, France, 2000.
15. P e d e r s e n, J. T., J. M o u l t. Genetic Algorithm for Protein Structure Prediction. – Curr. Opin. Struct. Biol., Vol. **6**, 1996, 227-231.
16. K h o d a b a k h s h i, A. M., J. M a n u c h, A. R a f i e y, A. G u p t a. Stable Structure Approximating Inverse Problem Folding on 2D Hydrophilic-Polar-Cysteine (HPC) Model. – Computational Biology, Vol. **16**, 2009, No 1, 19-30.