An Improvement of the Grid-based Hydrophobic-hydrophilic Model

Stefka Fidanova

Institute for Information and Communication Technologies, Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
E-mail: stefka@parallel.bas.bg

Received: September 10, 2009 **Accepted: December 15, 2009**

Published: July 30, 2010

Abstract: *Proteins are complex macromolecules that perform 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 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 protein folding problem. In this paper the protein is represented like a sequence over 3 letter alphabet according the specific functions of amino acids. After that the folding problem is defined like 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; by it can be constructed a protein with needed 3D folding.*

Keywords: *Protein folding, Hydrophobic and hydrophilic amino acids, Destructor.*

Introduction

Predicting the 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 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.

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].

According methods for protein-structure prediction has been divided into two classes: de novo modeling and comparative modeling. The de novo approach can be farther subdivided, those based exclusively on the physics of the interactions within the polipeptide chain and between the polypeptide and solvent, using heuristic methods [7, 10, 12], and knowledge-based

methods that utilize statistical potential based on the analysis of recurrent patterns in known structures and sequences. The comparative modelling 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 studding 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 3D 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 $E = n^{-1}$ for example. The HP protein folding problem is to find and energy-minimizing conformation for given HP sequence.

In this paper different approach is applied. We expand the HP model adding third letter D (HPD model) for Proline amino acid, because it has special biological functions. Using HPD model is explained the structures in protein conformation observed by biologists. It is de novo modeling first constructing secondary structures before completing them in tertiary structure.

Expanded hydrophobic-polar model

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

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. Albert et al. [1] note that the hydrophobic effect among amino acids contributes so significant 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, 15].

Polar amino acids are more ionic and bond well with water, while hydrophobic amino acids are less ionic and therefore do not bonds as 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 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.

The free energy of a conformations is defined as the negative number of nonconsecutive 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.

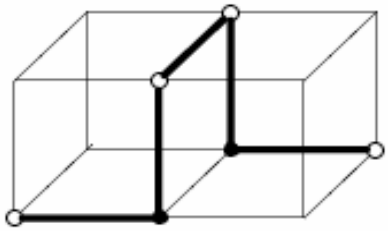

Fig. 1 HP protein representation on 3D cubic lattice, the black dots represent hydrophobic amino acids, the white dots represent polar

In spite of its apparent simplicity, folding optimal structures 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.

It is known that Proline amino acid has a special biological features [11]. In one side it is hydrophobic amino acid. In other side it acts as structural disruptor in the middle of secondary structure elements such as α-helices. However Proline is commonly found as a first residue of an α-helix. Therefore we expand HP model adding third letter D (disruptor) for Proline residue. So the problem to find the native folding of the protein becomes to find the folding with maximal number of H-H and H-D contacts, taking into account that D is at the beginning of the helix.

Protein folding

As is written in 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 three letters chain which consists of H, P and D monomers. The problem of finding stead conformation becomes the problem to find a conformation with 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 are 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, 9, 10, 12]. 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 to arise 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 is the folding, if the monomers chain has a special structure.

Let is consider polypeptide chain with only hydrophobic monomers or isolated polar monomers inside. As it is known it takes a form with 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 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 decrease. Let there is one D monomer inside a hydrophobic chain. Then the

hydrophobic helix is separated to two consecutive helices and the second helix starts with D monomer.

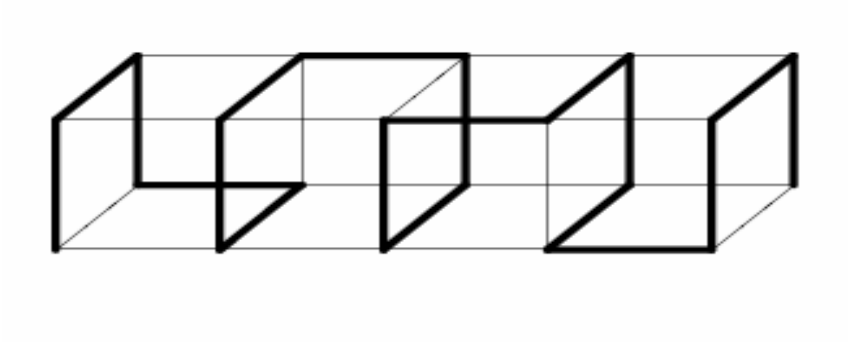

Fig. 2 Helix with 5 levels

Let the protein chain consists of long part of polar monomers and short part of one or two hydrophobic monomers at the ends. The hydrophobic monomers try to create a structure with 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 orthogonally packing of β-sheets.

The next configuration considered is two hydrophobic monomers followed by one polar monomer (PHHPHHPHH). 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 and two polar monomers, see Fig. 3.

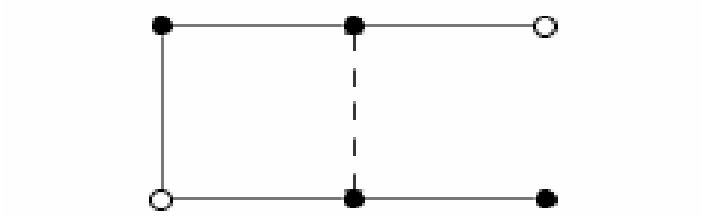

Fig. 3 A loop of a helix with fore hydrophobic monomers and two polar. Black dots represent the hydrophobic monomers. Dash-line represents the H-H contacts.

Let the protein chain consists of repetition of one polar and one hydrophobic monomers (PHPHPH). This chain can not 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 hairpin. Other types of configurations we call unstructured and we fold them using some metaheuristic method if they are large or according to other parts of the protein, thus to create maximal number of H-H and H-D contacts.

Experimental results

We test our ideas on proteins with known folding. Like tests we chose the following proteins: Glycoprotein, Leucocin A, ATP Syntease and Bacteriorhodopsin.

Glycoprotein

The amino acid sequence of the Glycoprotein is: GAHWGVLAGIAYFSMVGDWAK. Its HPD representation is: HHPHHHHHHHHPHPHHHPHHP. The HPD chain consists of 21 predominantly hydrophobic monomers and isolated polar monomers. So the folding with maximal number H-H contacts is helix with 5 loops (Fig. 4).

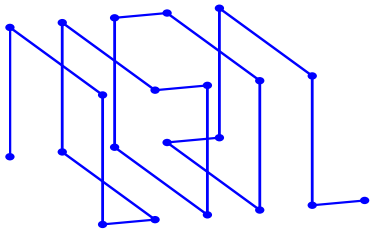

Fig. 4 Glycoprotein folded by our algorithm

Let us consider the real folding of Glycoprotein (Fig. 5). It is observed that it consists of 5 loops helix. Thus we can conclude that there is very high similarity between real folding and our folding for Glycoprotein.

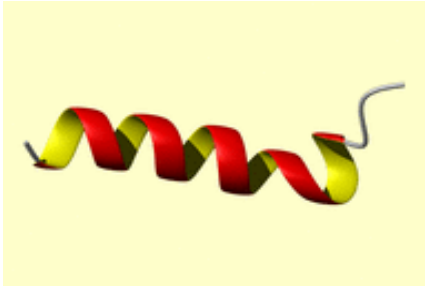

Fig. 5 Glycoprotein folding

Leucocin A

The amino acid sequence of Leucocin A is: KYYGNGVHCTKSGCSVAWGQAFSAGVHRL ANGGNGFW. Its HPD representation is: PPPHPHHPHPPPHHPHHHHPHHPHHHPPHHPH HPHHH. We cut the HPD chain of Leucocin A of 3 parts as follows: the first part consists of 15 monomers; the second part consists of 12 monomers; the third part consists of 10 monomers. The first part consists of 3 polar amino acids followed by HP, HH, PH, 3 polar amino acids and HHP. So the chain flex on the first HH monomers, thus H-H contact arise between the first and the third H monomers. After then it flex again, thus that arise H-H contact between the forth and the last H monomers from the first part. This is hairpins like folding. The hydrophobic amino acids predominate in the second part. Therefore it folds like helix with 3 loops. The third part consists of repetition of PHH monomers. So it folds like large helix with 6 monomers on the loop. When we assemble the protein we try to create additional H-H and H-D contacts between the parts. We put the first part (hairpin) to be perpendicular to the axis of the α-helix. Thus there are two additional H-H contacts between the first and the second part. We put the third part (large helix) to have the same axis as the second part, because thus there are three additional H-H contacts between them. After assembling the three parts we achieve the folding, represented on Fig. 6.

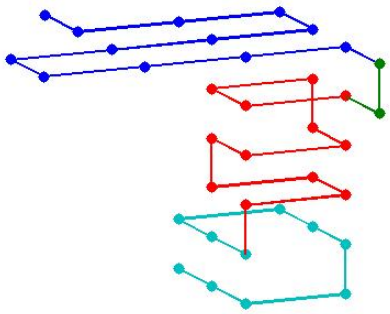

Fig. 6 Leucocin A folded by our algorithm

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

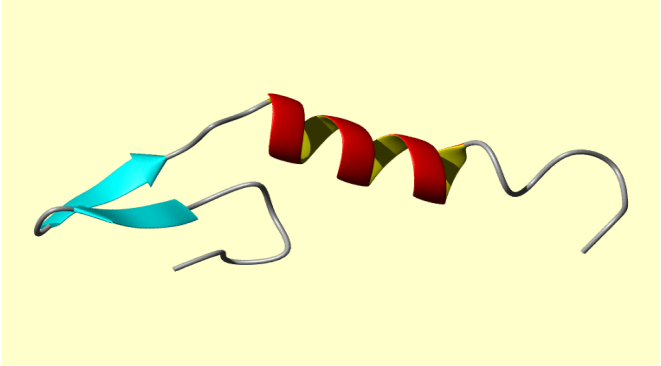

Fig. 7 Leucocin A folding

ATP Synthase

The amino acid chain of ATP Synthase is: MNLNATILGQAIAFVLFVLFCMKYVWPPLM AAI. The HPD representation is: HPHPHPHHHPHHHHHHHHHHHHPPHHDDHHHHH. We cut the HPD chain of the ATP Synthase of three parts as follows: the first part consists of 6 monomers, the second part consists of 20 monomers and the third part consists of 7 monomers. The first part is unstructured and it can not do their own folding. So we will fold it according the second part, thus to create maximal number of H-H contacts between them. The second and the third part consists of predominated hydrophobic amino acids and Proline (which is a hydrophobic too), thus if we follow only the hydrophobic-polar model this two parts will create a helix with 7 loops. But as we mention in the previous section, the Proline amino acid acts as a structural disruptor and it is commonly found as a first residue of an α-helix. Therefore in our HPD model, which takes in to account Proline residues, the second part ends before the Proline monomers and the third part starts with them. Thus the second part folds like 5 loops helix followed by two loops helix (the third part). Thus there are three H-D and H-D contacts between them. If the helixes are parallel each of other the H-H and H-D contacts will be only two. The first part is folded as it is shown on Fig. 8 and thus it creates 3 additional H-H contacts.

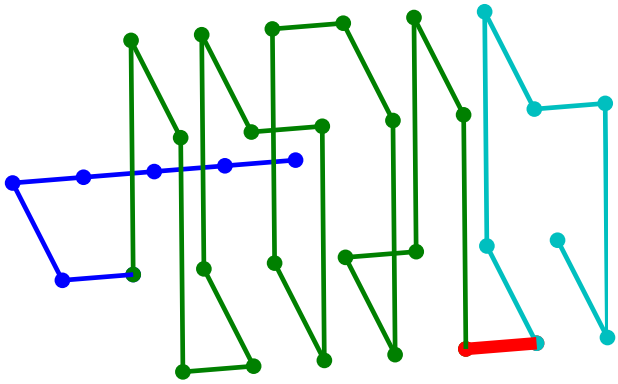

Fig. 8 ATP Synthase folded by our algorithm

Let us consider the real folding of ATP Synthase (see Fig. 9). It is observed short unfolded part, followed by two consecutive helices. The first helix consists of 5 loops and the second consists of 2 loops. So we can conclude that there is very high similarity between our and real folding for ATP Synthase.

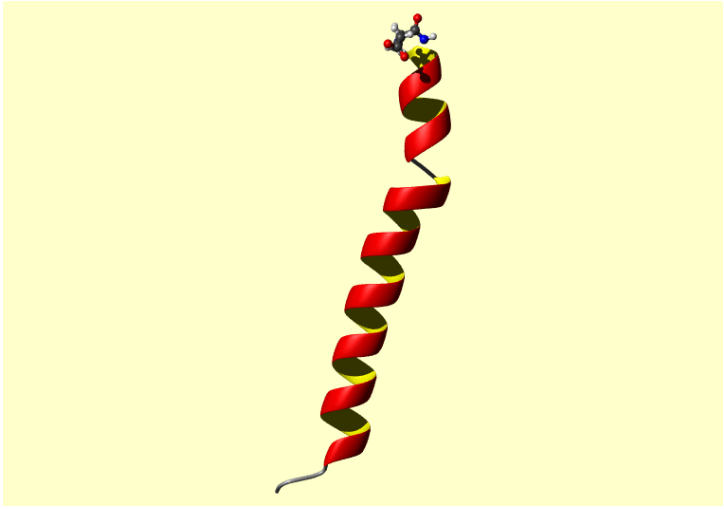

Fig. 9 ATP Synthase folding

Bacteriorhodopsin

The amino acid sequence of Bacteriorhodopsin is: QAQITGRPEWIWLALGTALMGLGTLY FLVKGMGVSDPDAKKFYAITTLVPAIAFTMYLSMLLGYGLTMVPF. Its HPD representation is: PHPHPHPDPHHHHHHHPHHHHHHPHPHHHPHHHHPPDPHPPHPHHPPHHDH HHHPHPHPHHHHPHHPHHDH. We cut the HPD chain of Bacteriorhodopsin on 5 parts as follows: the first part consists of 8 monomers, the second part consists of 29 monomers, the third part consists of 13 monomers, the forth part consists of 14 monomers and the fifth part consists of 8 monomers. When we cut the chain we use the D (Proline) monomer for beginning of the next part except for the fifth part. The first part can not be folded separately and we fold it according the second part. The hydrophobic amino acids predominate the second part, so it folds like 7 loop helix. The hydrophobic amino acids predominate the third and the forth parts. There is Proline (D) monomer at the beginning of the forth part. Therefore this two parts fold like consecutive helices, one with 3 loops and other with 3 and half loops. We put this two helices parallel to the first one (folding of the second part), because thus arise

7 additional H-H and H-D contacts between the helices. The fifth part is a repetition of PHH monomers, so it folds like large helix with 6 monomers on the loop (see Fig. 10).

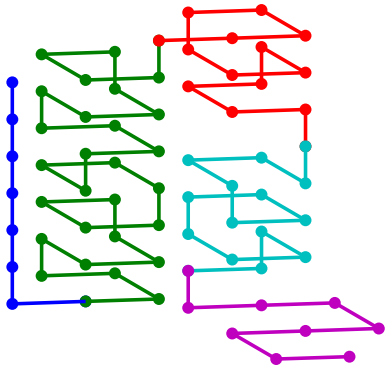

Fig. 10 Bacteriorhodopsin folded by our algorithm

Let us consider the real folding of Bacteriorhodopsin, Fig 11. We observe short unfolded part at the beginning followed by helix with 7 loops. Parallel to it there are two helices, first with 3 and second with 3 and half loops. At the end there is unstructured part like large loop. So we can conclude that there is very high similarity between our and real folding.

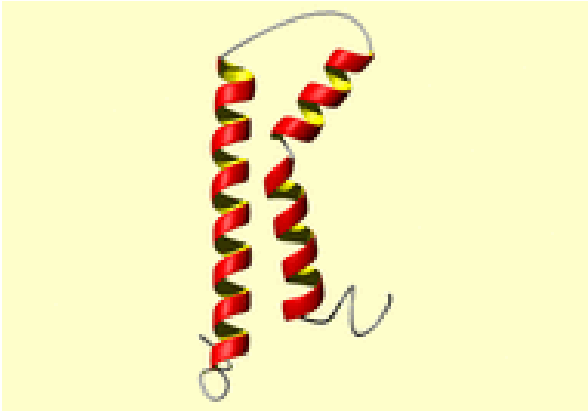

Fig. 11 Bacteriorhodopsin folding

Conclusion

Protein folding is one of the main problem that occur in bio-informatics. It requires knowledge from different disciplines like biology, physical-chemistry, mathematics. Most of the scientists develop comparison methods, but there are too inaccurate and slow. Others, apply metaheuristic methods but they do not achieve good results for long proteins. Most successful so far approach is fragment assembling. Its relatively low computational cost makes it very useful for large-scale analysis. 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 de novo modelling and fragmentation assembling. The HPD protein model on 3D lattice is used to model different fragments arising in protein folding. Thus shortcoming 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. Moreover, the known methods can be used only to predict the protein folding when the amino acid sequence is known. Our method can be used to construct a new protein with needed folding, which is a very important in pharmacology for new drug

design. It can be used also to predict changes in folding when some of the monomers are mutated, which is vary important for producing blockers. This paper explains the structures which arise in a tertiary protein form, like helices and sheets. We compare folding, achieved by our algorithm, with real folding and observe very high similarity. We can conclude that our idea gives encouraging results and it can be a basis for more precise folding prediction and protein construction algorithm.

Acknowledgment

Stefka Fidanova is supported by the Bulgarian Ministry of Education by the grant "Virtual screening and computer modeling for drug design".

References

1. Albert B., D. Bray, S. 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, Algorithms in Bioinformatics, Springer, Lecture Notes in Computer Sciences, 3240, 182-193.
3. Berger B., T. Leighton (1998). Protein Folding in the Hydrophobic-hydrophilic (HP) Model is NP-complete, Computational Biology, 5, 27-40.
4. Chandru V., A. Dattasharma, V. S. A Kumar (2003). The Algorithm of Folding Protein on Lattice, Discrete Applied Mathematics, 127(1), 145-161.
5. Chotia C. (2004). One Thousand Families for the Molecular Biologist, Nature Biotechnology, 22, 1317-1321.
6. Dill K., K. M. Fiebig, H. S. Chan (1993). Cooperativity in Protein-folding Kinetics, Nat. Acad. Sci., USA, 1942-1946.
7. Fidanova S. (2006). 3D HP Protein Folding Problem using Ant Algorithm, Proc. of Int. Symp. "Bioprocess Systems – BioPS'2006", Sofia, Bulgaria, III.19-III.26.
8. Heun V. (2003) Approximate Protein Folding in the HP Side Chain Model on Extended Cubic Lattices, Discrete Applied Mathematics, 127(1), 163-177.
9. Hoque T., M. Chetty, A. Sattar (2009). Extended HP Model for Protein Structure Prediction, Computational Biology, 16(1), 85-103.
10. Krasnogor N., D. Pelta, P. M. Lopez, P. Mocciola, P. de la Cana (1998). Genetic Algorithms for the Protein Folding Problem: A Critical View, Engineering of Intelligent Systems (Ed. Alpaydin C.), ICSC Academic Press, 353-360.
11. Levitt M. (1981). Effect of Proline Residues on Protein Folding, Molecular Biology, 145, 251-263.
12. Liang F., W. H.Wong (2001). Evolutionary Monte Carlo for Protein Folding Simulations, Chemical Physics, 115(7), 444-451.
13. Lyngso R. B., C. N. S. Pedersen (2000). Protein Folding in the 2D HP Model, Proc. of the 1st Journees Ouverst: Biology, Informatique et Mathematiques, Montpellier, France.
14. Pedersen J. T., J. Moult (1996). Genetic Algorithm for Protein Structure Prediction, Curr. Opin. Struct. Biol., 6, 227-231.
15. Khodabakhshi A. M., J. Manuch, A. Rafiey, A. Gupta (2009). Stable Structure Approximating Inverse Problem Folding on 2D Hydrophilic-Polar-Cysteine (HPC) Model, Computational Biology, 16(1), 19-30.

Assoc. Prof. Stefka Fidanova, Ph.D.
E-mail: stefka@parallel.bas.bg

Assoc. Prof. Stefka Fidanova prepares M.Sc. degree in Applied Mathematics and Ph.D. in Computer Science at Sofia University. Now she works in Institute for Information and Communication Technologies at Bulgarian Academy of Sciences. Her main scientific interests are in the field of combinatorial optimization, parallel algorithms and applications.