A new off-lattice HP model with side-chains for protein folding problem

Ivan Todorin[1,*], *Nicola* Yanev [2], *Metodi* Traykov[3], *Borislav* Yurukov[4]

[1]Department of Communication and Computer Engineering, Faculty of Engineering, University Center for Advanced Bioinformatics Research, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., 2700 Blagoevgrad, Bulgaria
[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str., Block 8, 1113 Sofia, Bulgaria
[3]Department of Electrical Engineering, Electronics and Automatics, Faculty of Engineering, University Center for Advanced Bioinformatics Research, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., 2700 Blagoevgrad, Bulgaria
[4]Department of Informatics, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., 2700 Blagoevgrad, Bulgaria

Abstract. A HP like nonlinear programming model and a "brutforce" algorithm is proposed. The model takes into account highest degree of complexity – different size of radiuses of side-chains, mentioned below as radicals, and two stages of the algorithm, including random structure initially and subsequent purposeful folding process according to the physical forces and energies. The very preliminary computational runs favourably demonstrate the adequateness of the model and the efficiency of the algorithm.

1 Introduction

3D structure of proteins is the major factor that determines their biological activity. The synthesis of new proteins and the crystallographic analysis of their 3D structure is very slow and very expensive process. If we can predict the 3D structure of many proteins, than only proteins with expected properties have to be synthesized. That will increase the number of known structures in the databases for proteins, and they can be used for drug design. The prediction of the 3D structure of proteins, if we know only the primary structure – the amino-acid sequence, is a protein folding problem. The reason for this process of folding in water environment is the interaction between water molecules and between amino-acids and water molecules. As water molecule has higher polarity than amino-acids, there is a minimum of energy when the protein is folded, not to spoil water to water interconnections. The way of folding is determined by the polarity or the hydrophobicity of different amino-acids, so the 3D structure with minimum energy is the real case. [1,2,3] There is less energy when more hydrophobic (H) amino-acids (the hydrophobic type of amino acid depends on the middle nucleotide of the codon [4]) are in contact in the core of the folded 3D structure and more polar (P) amino-acids are in contact with water. As we know the amino-acid sequence and the hydrophobicity of every amino-acid, we can predict the 3D structure – this method is called HP folding [5]. The closest to our model type of HP model is the off-lattice one, following the approach know as HP folding.

2 Methodology

2.1 Description of the set of feasible solutions

Structure and contact defining constraints

Let $x_i, y_i, z_i \in R$ be the unknown coordinates of the alpha carbon atoms, $xr_i, yr_i, zr_i \in R$ be the coordinates of the centers of the radicals and r_i be the normalized radiuses of the radicals, for the i^{th}(j^{th}) amino acid in the peptide chain, where n is the number of amino acids:

$$0.9 \le \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2 + (z_i - z_{i+1})^2} \le 1.1$$

$$\sqrt{(x_i - xr_{i+1})^2 + (y_i - yr_{i+1})^2 + (z_i - zr_{i+1})^2} \le 1.1(r_i + 0.3)$$

$$0.6 \le \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

$$0.9(r_i + r_j) \le \sqrt{(xr_i - xr_j)^2 + (yr_i - yr_j)^2 + (zr_i - zr_j)^2}$$

$$0.9(r_j + 0.3) \le \sqrt{(x_i - xr_j)^2 + (y_i - yr_j)^2 + (z_i - zr_j)^2} \le 1.1(r\_j + 0.3)$$

The constraints below are on corresponding euclidean distances between i[th] and j[th] alpha carbon atoms and centers of radicals:

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \le 0.8 \quad (2)$$

$$\sqrt{(xr_i - xr_j)^2 + (yr_i - yr_j)^2 + (zr_i - zr_j)^2} \le 1.2(r_i + r_j)$$

The goal of the willpower folding is to maximize the contacts between the radicals of hydrophobic amino

* Corresponding author: itodorin@gmail.com

acids and the contacts between the alpha carbon atoms of all amino acids and the radicals with electrical charge not in contact – to find the maximum of the following objective function:

$$(3)\ F(fold) = \sum_{i\ contact\ j\ redicals} (h_i + h_j + wtw) + \sum_{i\ contact\ j\ \alpha\ carbons} wtw,$$

where h_i is the hydrophobicity value of the amino acid, *wtw* is a parameter for the influence of hydrogen bonds.

2.2 Model description

The main purpose of the developed algorithm is to create a structure with low potential energy, starting from randomly folded form, instead of looking for the best structure among wholly randomly generated forms. Each amino acid is represented as the position of the alpha carbon atom and the position of the center of the radical with three-dimensional coordinates. The first stage is to generate a three-dimensional shape by randomly turning the peptide chain in 90 degrees and a distance 1 between the alpha carbon atoms in the peptide chain and between each alpha carbon and the radical center of the same amino-acid. For prevention of making rather spread or extreamly tight, there is used variable constrain of spreading [6]. The second stage, which is called "Willpower folding" [7], is to purposefully modify the 3D in order to minimize the energy while using now non-integer coordinates.

2.3 Willpower folding

The steps of the algorithm for maximizing (3) subject to (1) and (2) are given below:

1. Locate the conditional center as the arithmetic mean of the three directions:
$$x_c = \sum_{i=1}^{n} \frac{xr_i}{n};\ y_c = \sum_{i=1}^{n} \frac{yr_i}{n}; z_c = \sum_{i=1\ldots n} \frac{zr_i}{n}$$
2. Displace the coordinates of the radicals in proportion to their hydrophobicity value h_i, if the $h_i > 0$ direction is to the center of the molecule, and if $h_i < 0$ is opposite: $x_{ni} = x_i + 0.02h_i$, $x_c > x_i$; $x_{i\_new} = x_i - 0.02h_i$, $x_c < x_i$; $y_{i\_new} = y_i + 0.02h_i$, $y_c > y_i$; $y_{i\_new} = y_i - 0.02h_i$, $y_c < y_i$; $z_{i\_new} = z_i + 0.02h_i$, $z_c > z_i$; $z_{i\_new} = x_i - 0.02h_i$, $z_c < z_i$;
3. Displace the coordinates of the alpha carbon atoms to the center of the molecule: $x_{i\_new} = x_i + 0.01$, $x_c > x_i$; $x_{i\_new} = x_i - 0.01$, $x_c < x_i$; $y_{i\_new} = y_i + 0.01$, $y_c > y_i$; $y_{i\_new} = y_i - 0.01$, $y_c < y_i$; $z_{i\_new} = z_i + 0.01$, $z_c > z_i$; $z_{i\_new} = x_i - 0.01$, $z_c < z_i$;
4. Correct the positions of all alpha carbon atoms and radicals in order to preserve the peptide chain, such as alpha-carbon i approaching $i + 1$ and radical i approaching the alpha carbon atom i if $\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2 + (z_i - z_{i+1})^2} \le 1.1$, $x_{i\_new} = x_i + (x_{i+1} - x_i)/10$; $y_{i\_new} = y_i + (y_{i+1} - y_i)/10$; $z_{i\_new} = z_i + (z_{i+1} - z_i)/10$; $xr_{i\_new} = xr_i + (x_i - xr_i)/10$; $yr_{i\_new} = yr_i + (y_i - yr_i)/10$; $zr_{i\_new} = zr_i + (z_i - zr_i)/10$;
5. Separate the coordinates of the radicals of electrically charged amino acids from one another if $\sqrt{(xr_i - xr_j)^2 + (yr_i - yr_j)^2 + (zr_i - zr_j)^2} \le 2$, $xr_{i\_new} = xr_i - (xr_j - xr_i)/10$; $yr_{i\_new} = yr_i - (yr_j - yr_i)/10$; $zr_{i\_new} = zr_i - (zr_j - zr_i)/10$; $xr_{j\_new} = xr_j - (xr_i - xr_j)/10$; $yr_{j\_new} = yr_j - (yr_i - yr_j)/10$; $zr_{j\_new} = zr_j - (zr_i - zr_j)/10$;
6. Reduce the distance radical coordinates of closely spaced hydrophobic amino acids if $\sqrt{(xr_i - xr_j)^2 + (yr_i - yr_j)^2 + (zr_i - zr_j)^2} \le 2$, $xr_{i\_new} = xr_i + (xr_j - xr_i)/20$; $yr_{i\_new} = yr_i + (yr_j - yr_i)/20$; $zr_{i\_new} = zr_i + (zr_j - zr_i)/20$; $xr_{j\_new} = xr_j + (xr_i - xr_j)/20$; $yr_{j\_new} = yr_j + (yr_i - yr_j)/20$; $zr_{j\_new} = zr_j + (zr_i - zr_j)/20$;
7. Reduce the distance the coordinates of the alpha carbon atoms of closely spaced amino acids if $\sqrt{(xr_i - xr_j)^2 + (yr_i - yr_j)^2 + (zr_i - zr_j)^2} \le 2$, $xr_{i\_new} = xr_i + (xr_j - xr_i)/20$; $yr_{i\_new} = yr_i + (yr_j - yr_i)/20$; $zr_{i\_new} = zr_i + (zr_j - zr_i)/20$; $xr_{j\_new} = xr_j + (xr_i - xr_j)/20$; $yr_{j\_new} = yr_j + (yr_i - yr_j)/20$; $zr_{j\_new} = zr_j + (zr_i - zr_j)/20$;
8. Adjust the positions of all alpha carbon atoms to avoid overlapping if $0.6 > \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$ as $j \ne i+1$, and $0.9 > \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$ as $j = i+1$, $x_{i\_new} = x_i - 0.1(x_j - x_i)/((x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2)$; $y_{i\_new} = y_i - 0.1(y_j - y_i)/((x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2)$; $z_{i\_new} = z_i - 0.1(z_j - z_i)/((x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2)$;
9. Correct the positions of all radicals to avoid overlapping if $0.9(r_i + r_j) > \sqrt{(xr_i - xr_j)^2 + (yr_i - yr_j)^2 + (zr_i - zr_j)^2}$, $xr_{i\_new} = xr_i - 0.1(xr_j - xr_i)/((xr_j - xr_i)^2 + (yr_j - yr_i)^2 + (zr_j - zr_i)^2)$; $yr_{i\_new} = yr_i - 0.1(yr_j - yr_i)/((xr_j - xr_i)^2 + (yr_j - yr_i)^2 + (\mathrm{zr_j} - \mathrm{zr_i})^2)$; $zr_{i\_new} = zr_i - 0.1(zr_j - zr_i)/((xr_j - xr_i)^2 + (yr_j - yr_i)^2 + (zr_j - zr_i)^2)$;
10. Adjust positions to avoid overlapping between the alpha carbon atom and the radical, if $0.9(r_j + 0.3) > \sqrt{(xr_i - xr_j)^2 + (yr_i - yr_j)^2 + (zr_i - zr_j)^2}$, $xr_{i\_new} = xr_i - 0.1(xr_j - x_i)/((xr_j - r_i)^2 + (yr_j - y_i)^2 + (zr_j - z_i)^2)$; $yr_{i\_new} = yr_i - 0.1(yr_j - y_i)/((xr_j - x_i)^2 + (yr_j - y_i)^2 + (\mathrm{zr_j} - \mathrm{z_i})^2)$; $zr_{i\_new} = zr_i - 0.1(zr_j - z_i)/((xr_j - r_i)^2 + (yr_j - r_i)^2 + (zr_j - z_i)^2)$;

2.4 Thermo effect

What we call "Thermo effect" is to make a small random move in every step of Willpower folding, which corresponds to the real environment. This prevents building equal structures if the initial random fold is the same and gives more chance to find better structure.

3 Computational Results

First, using our definition of contact (2), we find the contacts in the real structure of protein 1UUB, using the coordinates of alpha carbons in PDB file of the protein data bank (Figure 1).

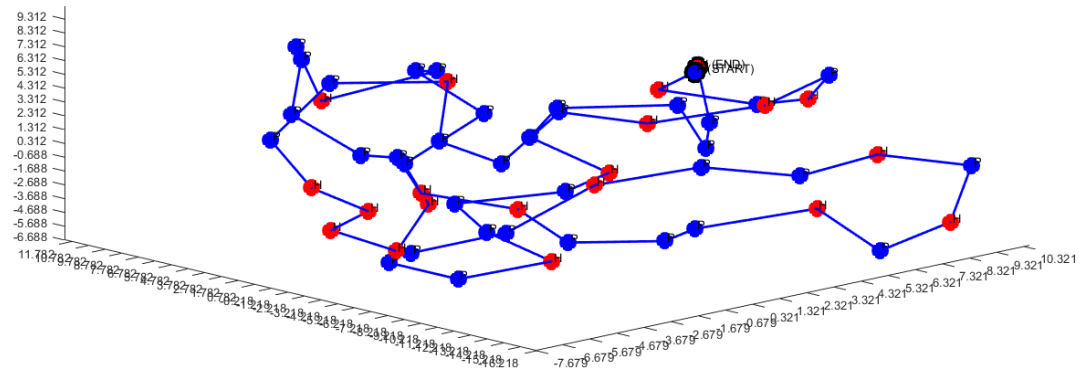

Fig. 1. Real structure of protein 1UUB

There are 125 contacts according to the constraints (2) above.

The test run of the program, realizing the algorithm on 1UUB obtains the following results: among only 100000 randomly generated structures, the value of the evaluation function is 1279.2 and 206 contacts – 59 matches with the real one (Figure 2).. The time needed was 70 min on PC, Intel i5, 2.26GHz.

Remark: for the source code in C++, the list of values coordinates and the list of contacts' couples, mail to itodorin@gmail.com.

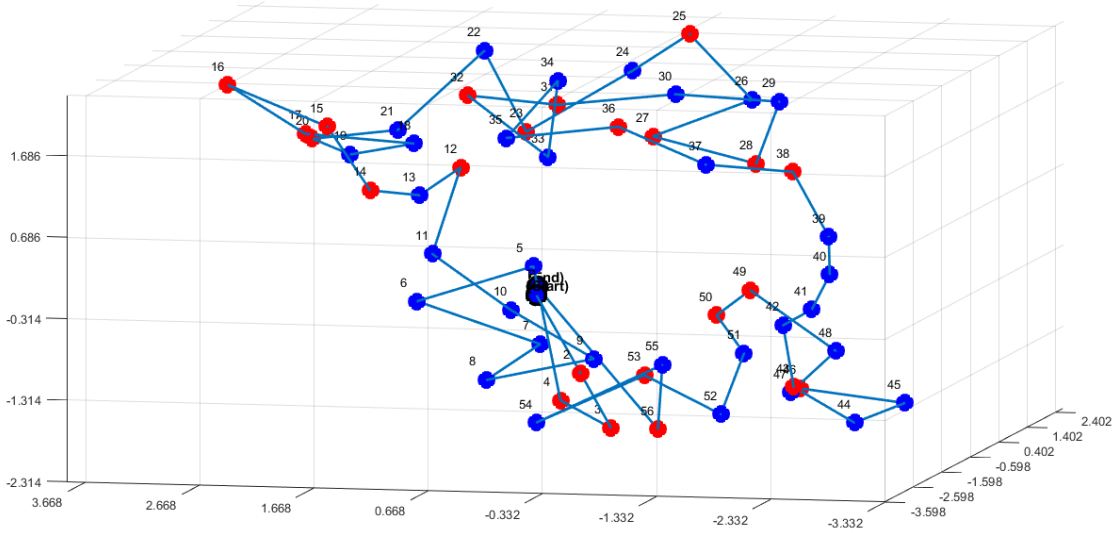

Fig. 2. The structure with 59 matches

The other experiment we done, is to find out how much better structure could be built, classifying fold not by evaluation function, but by ratio contacts matching ratio. In this way we obtain the following structure with 62 contacts matches out of 120 (Figure 3).

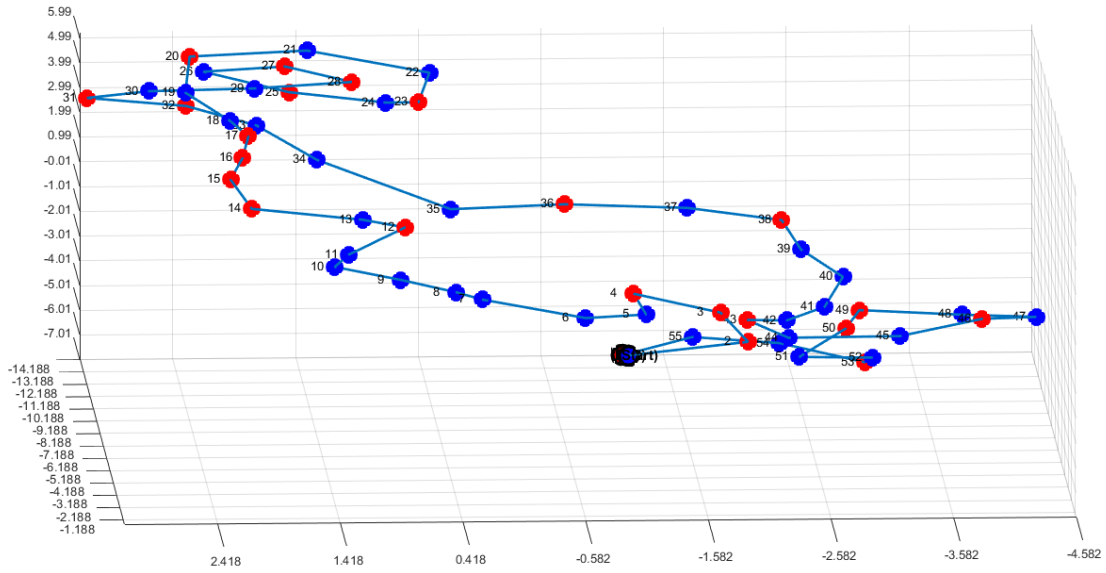

Fig. 3. The structure with 62 contacts matches

4 Conclusions

This approach for protein folding prediction, using "Willpower folding" with "Thermo effect", appear to be faster than models for finding best structure among randomly built structures. Therefore it is possible rather more structures of proteins to be processed for the same time. Even more, the model is new (age under year) and might be further developed in direction of granularity – placing every atom, and the evaluation function could take into account the real impact of the distribution of electronic density, such degree of granularity is hard to be execute considering computational time for models for finding best structure among randomly built structures.

5. Acknowledgements

This work is partially supported by the project of the Bulgarian National Science Fund, entitled: "Bioinformatics research: protein folding, docking and prediction of biological activity", code NSF I02/16, 12.12.14.

References

1. A. Kolinski and J. Skolnick, Proteins **18**, 338-352 (1994).
2. B. Berger, and T. Leighton, Journal of Computational Biology **5**, 27-40 (1998).
3. H. Greenberg, W. Hart, G. Lancia, INFORMS Journal on Computing **16**, 211-231 (2004).
4. P. Milanov, I. Trenchev, N. Pencheva, Mathematica Balkanica **17**, 157-165 (2004).
5. S. Istrail, and F. Lam, Commun. Inf. Syst. **9**, 303-346 (2009).
6. I. Todorin, Proceedings of the Fifth International Scientific Conference – FMNS2013 **1**, 188-192 (2013).
7. I. Todorin, SDNS **1**, 1-6 (2017) (in printing).