Bioinformatics research of 3D structure of the proteins

Radoslav Stoev, Kiril Gashteovski, Ivan Todorin, Ivan Trenchev[1], Nadezhda Borisova, Metodi Traykov
South-West University "Neofit Rilski", Blagoevgrad, Bulgaria

***Abstract**: In this article we will represent a HP model for protein folding using their properties in water environment. This model is based on combinatorial algorithms. Some of the examples will be solved by Branch and Bound. The calculations will be performed via CLPEX and MATLAB. Graphical software for presentation of the results will be demonstrated.*
***Keywords:** Bioinformatics, 3D structure, HP folding.*

1. INTRODUCTION

The HP folding in lattice models, which are used for prediction of the 3D structure of proteins, are based on the fact, that in the cell environment, peptides fold in a way, in which more hydrophobic amino-acids are closed together in a kernel, in contact between them, and more polar amino-acids are oriented outside in contact with the water in the cell environment [5]. The first HP model of this kind has been presented by Ken Dill in 1985. Such a form of the folded peptide is with minimum of the potential energy, because of the hydrogen bonds between water molecules and polar amino-acids and it is more stable, so we can expect that a 3D structure, which is folded like this, will be the real case. In this way we can predict the 3D structure of a protein from its primary structure [3]. Today many primary structures of proteins are known and less their 3D structures, which determine the properties of the proteins. Databases with 3D structures are used in Drug Design, for simulation of ligand-target interactions, for pocket detection of targets, for virtual screening, etc. Enlarging such databases is in great interest of pharmacology, as it helps for cheaper and faster Drug discovery [1, 2, 4].

These models use combinatorial algorithms to find out a conformation of the folded protein, which has so many contacts between hydrophobic amino-acids, as possible. These contacts have to be counted for many possible conformations, which is very hard computational problem, because the number of possible folds is extremely large. The problem to be found out the structure with most contacts, which seems to be with minimum of

the potential energy, is NP-complete [6, 7]. We have a possibility to obtain a fold, which is close to the optimal, but we might miss to check the best one. Also we assume that the amino-acids are only two types – hydrophobic (H) and polar or hydrophilic (P), with equal size and ordered in cubic lattice of two types – face-to-face cubic lattice (FFC) and face-to-center cubic lattice (FCC) [8]. Last is necessary to improve the time for solving the problem by the contemporary computational resources. Different models may be improved for accuracy or for speed by choosing the lattice – FCC is more accurate, using backbone chain [9], including only the carbon atoms of the peptide connection or using side-chain structure, in which the other part of the amino-acid is oriented in one of the possible sides in its own position in the lattice, etc.

2. METHODOLOGY

In the following model, an input sequence of numbered amino-acids is used, in which every member is defined as H – in case of a hydrophobic amino-acid, and as P – in case of a hydrophilic (polar) amino-acid [6]:

INPUT PPHP...H

All amino-acids have to be put in a cubic lattice with m rows and m columns, which lattice is transformed in a row:

m x m – lattice (i,j) i,j $\xrightarrow{(i-1)m+j}$ i

Every position in the lattice and its occupation of any member of the input sequence is assigned as a point, so if member k is put in the position i – this variable has value 1, otherwise it has value 0:

m^2n – points $x_{ik}\{1,0\}$

Every member k has to be put just in one position:

$$\sum_{i=1}^{n} x_{ik} = 1 \text{, for every k}$$

Every position i might be occupied only by one member or it might be free:

$$\sum_{k=1}^{N} x_{ik} = 1 \text{, for every i}$$

If a position is occupied by a member, we must have at least one occupied position neighbor to it, or if it is free, we might have zero or more occupied neighbor positions:

$$x_{ik} \leq \sum_{j \in G(i)} x_{jk+1} \text{ , G(i) is a set of all neighbors of } x_{ik}$$

We define additional variable y, which has value 0 or 1 if two neighbor cells in the lattice are occupied by hydrophobic amino-acids:

$$x_{ik} \leq \sum_{j \in G(i)} y_{ikjl}$$

where G= {l-k>2∩Sk=H∩Sl=H}

$$x_{jl} \leq \sum_{i \in G(j)} y_{ikjl}$$

G(j) – set of points, neighbor to j

The final goal is the maximum number of contacts between hydrophobic amino-acids, which is represented by our variable y:

$$\max \sum y_{ikjl}$$

We will present a table with classification of amino-acids according to their hydrophobic properties [9].

Name	Symbol	Classification	Name	Symbol	Classification
Alanine	A	Hydrophobic	Leucine	L	Hydrophobic
Arginine	R	Polar	Lysine	K	Polar
Asparagine	N	Polar	Methionine	M	Hydrophobic
Aspartic Acid	D	Polar	Phenylalanine	F	Hydrophobic
Cysteine	C	Polar	Proline	P	Hydrophobic
Glutamic Acid	E	Polar	Serine	S	Polar
Glutamine	Q	Polar	Threonine	T	Polar
Glycine	G	Polar	Tryptophan	W	Hydrophobic
Histidine	H	Polar	Tyrosine	Y	Polar
Isoleucine	I	Hydrophobic	Valine	V	Hydrophobic

TABLE 1: Hydrophobic / Polar classification of the 20 a-amino-acids.

We have developed a software, written on C++, which transform protein sequence in FASTA format into a sequence of 0 and 1, where 0 is for a polar amino-acid and 1 is for a hydrophobic amino-acid.

3. RESULTS

We will present three examples. The first and the third example are calculated on a 3D lattice and the second example on a 2D lattice [5, 7, 8].

All calculations will be done by using the software package CPLEX, for which we have a license for academic use from the IBM [8]. The package has been installed on an IBM server.

For every example we have developed a software, which calculates an input LP file for CPLEX. The time for calculation of the three examples is generally 36,5 hours.

We have also developed a software for visualization in 3Ds Max.
The following calculations are done for the protein 1101010101.

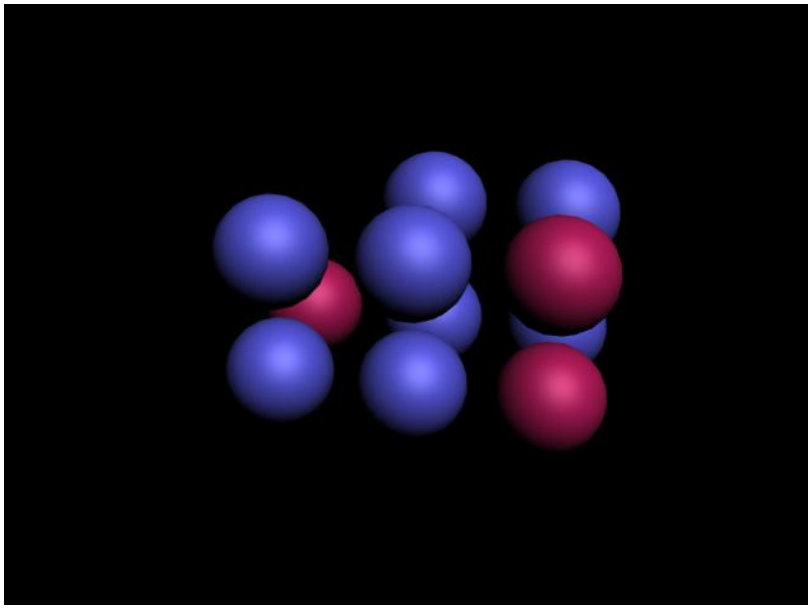

Fig. 1: Solution of the protein 1101010101.

The next figure represents the result for the described protein in 2D lattice.

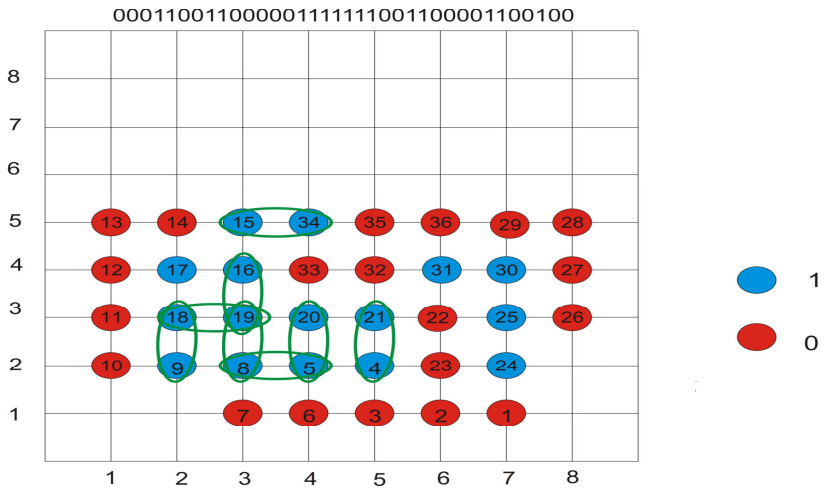

Fig. 2: Solution of the protein 0001100110000011111111001100001100100.

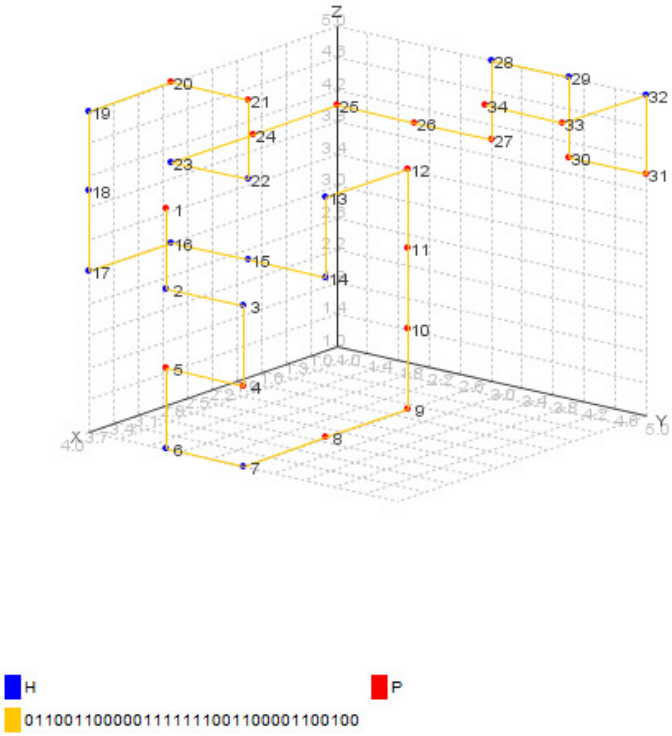

Fig. 3: Solution of the protein 10100110100101100101

4. CONCLUSIONS

Our mathematical model can be used in the gene engineering for modelling of the 3D structure of the RNA. RNA has enzyme activity and we can predict some structures and properties of it.

5. ACKNOWLEDGEMENTS

This study was supported by Grant DVU 01/0197 of Ministry of Education and Science.

6. REFERENCES

[1] Duan Y., Kollman P. A. (2001) Computational protein folding: From lattice to all-atom. IBM Systems Journal, 40, 297–309.
[2] Dyson F.. (2008) Birds and frogs. Notics of the American Mathematical Society, 56:212–223 .
[3] Guo Y.Z., Feng E., Wang Y. (2007) Optimal HP configurations of proteins by combining local search with elastic net algorithm. J. Biochem. Biophys. Methods 70, 335–340
[4] Hengyun L., Yang G. (2009) Extremal Optimization for protein folding simulations on the lattice Computers and Mathematics with Applications 57 1855-1861
[5] Istrail S., Hurd A., Lippert R., Walenz B., Batzoglou S., Conway J. H., Peyerl F. W. (2000). Prediction of self-assembly of energetic tiles and dominoes: Experiments, mathematics, and software. Sandia National Labs Technical Report, April.
[6] Jiang M., Zhu B. (2005) Protein folding in the hexagonal lattice in the hp model. J. of Bioinformatics and Computational Biology, 3 19–34.
[7] Namsu A. Park S., (2010) Finding an Upper Bound for the Number of Contacts in Hydrophobic-Hydrophilic Protein Structure Prediction Model Journal Of Computational Biology 17(4), 647–656
[8] Neumann J. V. (1947) The mathematician. in ”The Works of the Mind”, University of Chicago Press, 180–196
[9] Shakhnovich E. (1996) Modeling protein folding: the beauty and power of simplicity. Folding and Design, 1, 50–54.