International Journal of Bio-Science and Bio-Technology
Vol.8, No.1 (2016), pp.363-372
http://dx.doi.org/10.14257/ijbsbt.2016.8.1.32

A New Method for Refining3D Protein Predicted Models Based on 3D Random Walk

Shaheera Rashwan[1*] and Bayumy A. B. Youssef[2]

Informatics Research Institute, City of Scientific Research and Technological Applications, Borg Elarab, Alexandria, Egypt
[1]srashwan@mucsat.sci.eg, [2]bbayumy@mucsat.sci.eg

Abstract

Predicting protein 3D structures from the amino acid sequence is still a hard and unsolved task after five decades of efforts. High-resolution models can be built only if the target protein has a known homologue. If not, it must be built from scratch which yields in most cases to protein models of low-resolution, i.e. far from their native structure. In this paper, we present a new refinement method of 3D protein predicted models. The new method relies on the motion that the atoms in a protein take randomly in the 3D space leading to folding. Experimental results using the CASP benchmark show the assessment and the quality of the new method in comparison with the traditional methods that depend on molecular dynamics simulation.We prove that 55 % of cases were successfully enhanced by the new method.

***Keywords**: Protein Structure Prediction, Refinement, Molecular Dynamics, 3D Random Walk*

1. Introduction

Protein structure prediction is the prediction of the tertiary structure of a protein from its primary structure, which is the amino acid sequence. Solving such a problem is of great importance in the field of bioinformatics and especially in drug design.

The number of available protein sequences is increasing exponentially, about 5.3 million protein sequences were deposited in the UniProtKB database [1], with the great success of the genome sequence projects. However, due to the technical difficulties, the number of available protein structures is very far behind, the number of protein structures in the Protein Data Bank (PDB) [2] is only about 44,000, less than 1% of the proteinsequences. So, developingefficient computer-based algorithm to predicting 3D structures from sequences is the best way to fill up this gap.

We can divide the methods of protein structure prediction into two categories: template-based modeling methods (if similar proteins have been solved before) and free modeling methods (built from scratch). Low-resolution protein models often occur by free modeling methods (in other name: ab initio modeling). Here, we can see the need of developing a refinement method that enhances the protein structure models and brings them close to the native structure.

Previous developments in the area of refining 3D protein structure predicted models were based on the Molecular Dynamics (MD) Simulation. The molecular dynamics (MD) simulation is a technique by which one generates the atomic trajectories of a system of N particles by numerical integration of Newton's equation of motion, for a specific interatomic potential, with certain initial condition (IC) and boundary condition (BC). In structural biology, the MD method is frequently applied for ligand docking, simulations

[*] Corresponding Author

ISSN: 2233-7849 IJBSBT

of homology modeling and ab initio prediction of protein structure by simulating folding of the polypeptide chain from random coil.

Raval, *et al.,* [3] used Molecular Dynamics as a technique for homology model refinement using all-atom simulations, each at least 100 μs long and a physics-based force field shown to successfully fold fast-folding proteins. Dimaio, *et al.,* [4] developed an approach in which density maps generated from molecular replacement solutions for each set of starting protein structure models are used to guide energy optimization (minimization) by structure rebuilding, combinatorial sidechain packing, and torsion space minimization. New maps were generated using phase information from the energy-optimized models most consistent with the diffraction data. Zhang, *et al.,* [5] used MD simulations to refine protein structural models and checked in particular the possibility of reshaping the middle-range funnel of physics-based energy landscapes. They developed a Fragment-Guided Molecular Dynamics (FG-MD) algorithm, which combines the physical-based force field AMBER99 with knowledge-based H-bonding and repulsive potentials. Heo, *et al.,* [6] developed GalaxyRefine which first rebuilds all side-chain conformations and repeatedly relaxes the structure by short molecular dynamics simulations after side-chain repacking perturbations.

Bhattacharya and Cheng [7] proposed a two-step refinement protocol, called 3Drefine. The first step is based on optimization of hydrogen bonding (HB) network and the second step applies atomic-level energy minimization on the optimized model using a composite physics and knowledge-based force fields. Topf, *et al.,* [14] developed a new method for characterizing the structure of assemblycomponents by iterative comparative protein structure modeling andfitting into cryo-electron microscopy (cryoEM) density maps. They used a comparative model of a given component by consideringmany alignments between the target sequence and a relatedtemplate structure while optimizing the fit of a model into thecorresponding density map. Schroder, *et al.,* [15] developed a geometry-based algorithm that samples conformational space under constraints imposed by low-resolution density maps obtained from electron microscopy or X-ray crystallography experiments. A deformable elastic network (DEN) is used to restrain the sampling to prior knowledge of an approximate structure.

Limitations of the MD method are related mostly to underlying molecular mechanics force fields. A single run of MD simulation optimizes potential energy, rather than free energy of the protein. Another limitation is its ignorance to the “randomization” of the trajectory taken by the atomic moves and interactions leading to folding.

In our work, we will replace the usage of the molecular dynamics method by the notion of Brownian motion (BM). Brownian motion is the random motion of particles suspended in a fluid resulting from their collision with the atoms or molecules in the gas or liquid. In physics, random walks are used as simple models of physical Brownian motion and diffusion such as the random moves of molecules in liquids and gases. In this paper, we develop a computer-based program that mimics the movements of the atoms in a molecule randomly in the 3D space in order to make the protein folding in its free minimum energy.

In this paper, we have proposed a method to refine computationally predicted protein models in order to bring them closer to the native state. The method is based on minimization of multi-body Van der Waals interaction potential by random walk. The paper is organized as follows: section 1 presents the introduction including some of the related works. Section 2 summarizes a background on the random walk in 3D and the energy function used in our work. Section 3 introduces the refinement method used for producing ab initio protein predicted models of higher resolution. Section 4 shows the experimental results. Section 6 concludes and discusses the new method and its results. Finally a list of references is given.

2. Background

2.1 Random Walk

Given an undirected, connected graph G(V,E) with vertices: |V| = n, and edges: |E| = m a random "step" in G is a move from some node u to a randomly selected neighbor v. A random walk is a sequence of these random steps starting from some initial node [12].

In our work, we will consider the atoms in the molecules of the protein as being the vertices and the edges are the paths that take the atoms leading to folding. There is no a real graph. This is just a grid search on protein coordinates. Only one atom is moved at a time, it is always moved in a coordinate direction not a random vector and it is always moved a fixed distance. The randomization here is about the direction of the step that the atom takes in the 3D space.

2.2 Energy Function

"The law of thermodynamics states that the natural or folded state of proteins is a global free energy minimum" [11]. Now the objective is to search the predicted model among a set of legal models that minimizes the energy. Van der Waals (vdW) interactions play a critical role in determining thestructure, stability, and function for a wide variety of systems.

Most of 3D protein structure refinement techniques [5, 7] depend on the Hydrogen bonding and ignore totally the van der Waal's forces. However, van der Waal's forces are important for a protein achieving its final shape. Although they are individually very weak, the sum of these interactions contributes substantial energy to the final three-dimensional shape of the protein [13].

In [8], DiStasio, *et al.,* introduce an efficientmethod that describes in an accurate manner the nonadditive many-bodyvdW energy contributions arising from interactions that cannot bemodeled by an effective pairwise approach.

According to [8], ThevdW-MB interaction energy for the full many-body system is computed as

$$W(r_{pq}) = (1 - \exp(-(r_{pq} / R_{pq}^{vdW})^{\beta})) / r_{pq} \tag{1}$$

Where r_{pq} is the inter atomic distance between atoms p and q, β is a range-separation parameter, and $R_{pq}^{vdW} = R_{p}^{vdW} + R_{q}^{vdW}$ is the vdW correlation length in terms of the individual vdW radii, also defined as functional of the density. The best values of the β parameter were found as 2.56 for the vdW-MB model.

The radii used for atoms are as follows: carbon, 1.7 Å; oxygen, 1.52 Å; nitrogen, 1.55 Å; Hydrogen, 1.2 Å; and sulfur, 1.8 Å.

3. Materials and Methods

3.1 Data Used

The CASP (critical assessment of techniques for protein structure prediction) is a community-wide experiment for protein structure prediction taking place every two years since 1994. The experiment provides users of structure prediction servers with an opportunity to assess the quality of the various methods and servers entirely blindly. It also provides the research community with an assessment of the state of the art in this field.

We extracted 20 protein sequences studied at the CASP 10 meetings from the public web page of the protein structure prediction center [9].

3.2 Proposed Method

In our work, we propose a new method for refinement of 3D protein predicted models based on 3D Random walk and Van der Waals energy minimization.

The steps of the new method can be summarized as follows:

Step1: Input

Predict the initial 3D protein structure models from its amino acid sequence (read in its FASTA format) using one of the abintio methods which produces low resolution protein structure models. Here, we choose the HyperChem release 8.0 as the ab initio method used.

Step2: Initialize the energy

Calculate the initial total energy of the input protein structure using equation (1) and consider this energy value as being the least accepted energy for the protein to be folded.

Step3: Solve the problem

Do the atomic 3D random walk as follows:

- For all atoms in the protein
- Randomly select a vector from 6 coordinate vectors:{(+1,0,0),(-1,0,0),(0,+1,0),(0,-1,0),(0,0,+1),(0,0,-1)}
- Add the vector onto the coordinates of the atom
- Calculate the new energy of the protein according to the energy function equation (1)
- If energy improves, accept the new configuration, otherwise discard the configuration and try again until reaching the maximum number of walks defined by the user.

Step4: Output

The output predicted model is the model that has the lowest energy among all walks Clearly, the new method has two main advantages: first, the number of walks is user-defined which means that the user can stop the refinement procedure at the level of accuracy he defines and second, the model saves its initial state (walk no. 0) if there is no improvement in accuracy during the running of the method.

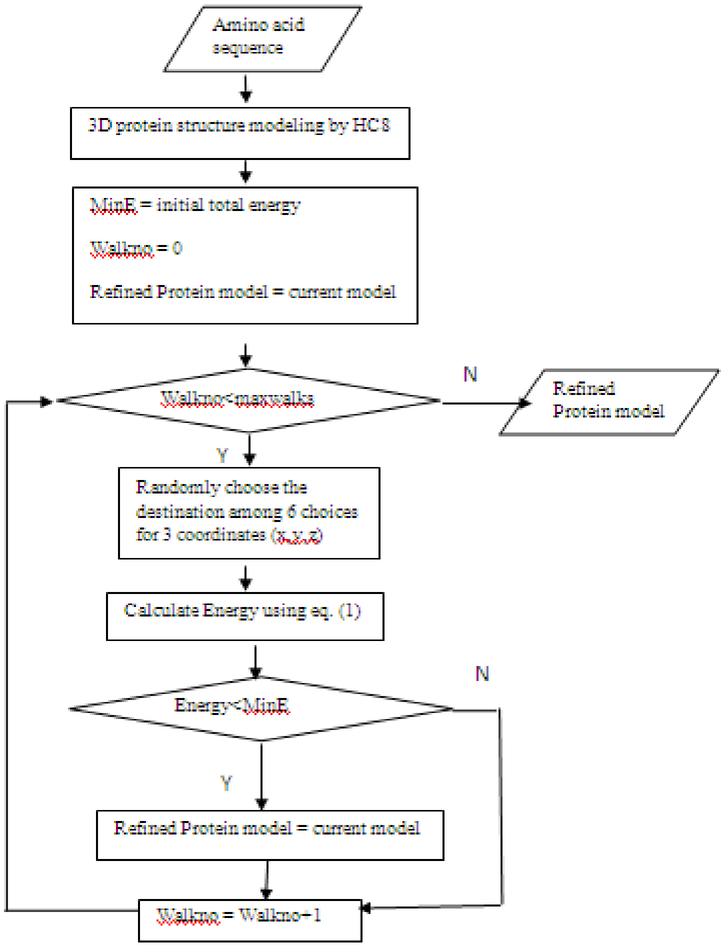
Amino acid sequence
3D protein structure modeling by HC8
MinE = initial total energy
Walkno = 0
Refined Protein model = current model
Walkno<maxwalks
N
Refined Protein model
Y
Randomly choose the destination among 6 choices for 3 coordinates (x,y,z)
Calculate Energy using eq. (1)
N
Energy<MinE
Y
Refined Protein model = current model
Walkno = Walkno+1

Figure 1. A Flowchart of the Suggested Method

4. Experimental Results

In this Section, we will present the result of applying our new procedure for protein structure prediction using ab initio modeling. We suggest a new refinement method for producing models of higher resolution. Implementation was done in Matlab version 7.10.

We choose the step size in the random walk to be 1Å and the number of walks used in experiments was 20 walks. This choice of the step size and the small number of walks can be justified by the saying that even we do very slight movements of the atoms in random directions, the refined 3D protein structure become closer to its native structure. This proves the efficiency of the suggested method.

We assess the quality of the predicted models with refinement using ProQServer[10], which is a neural network-based method to predict the quality of a protein model that extracts structural features, such as frequency of atom–atom contacts, and predicts the quality of a model, as measured either by LGscore or MaxSub.

We compare our results with the results of the refinement method GalaxyRefine[6]. We choose this method of refinement as it is an example of refinement methods that rely on the molecular dynamics simulation and depend on only Hydrogen bonding with total ignorance of the effect of the van der Waals forces in the protein folding process.

Figure 2 shows the 3D predicted protein models for the first three protein sequences data using HyperChem 8.0 without refinement (in column a), HyperChem 8.0 with refinement using GalaxyRefine (in column b) and HyperChem 8.0 with refinement using our method described in section 3 (in column c). The GalaxyRefine web server produces five refined predicted models. We choose, for comparison, the model with the lowest RMSD (Root Mean Square Deviation) as being the best model.

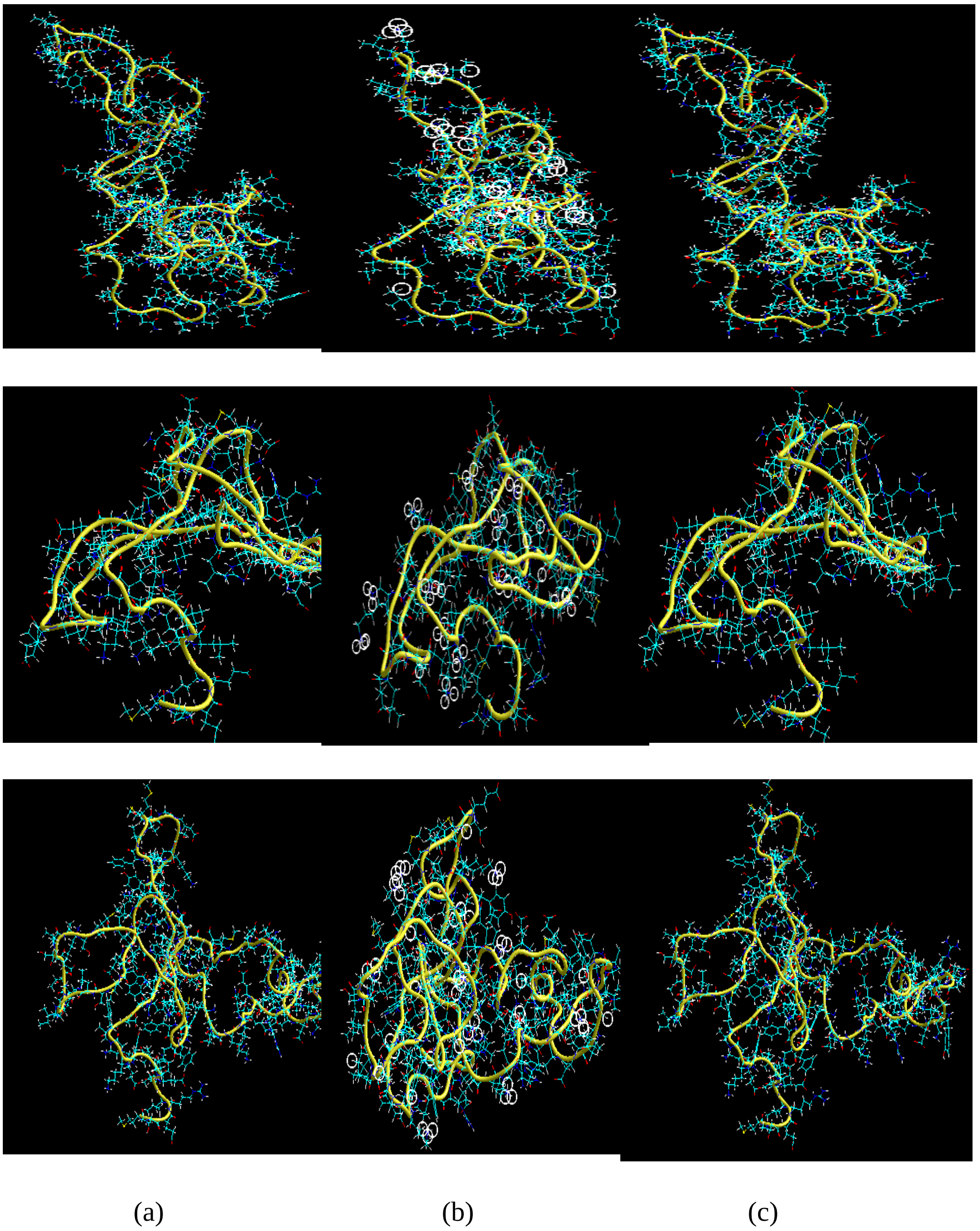

Figure 2. The 3D Predicted Model of the First Three Protein Sequences using (a) HyperChem Release 8.0 without Refinement, (b) Refinement using GalaxyRefine and (c) Our Suggested Refinement Method

Table 1 shows the LG_score of the initial model produced by HyperChem 8.0(HC8), compared to the LG_score produced by HC8 followed by GalaxyRefine (HC8-GR) and to that produced by HC8 followed by our suggested method (HC8-RW). In the table, we present also the number of walks done for the initial model to reach its free minimum energy among all walks. Maximum number of walks is 20 walks and of course the walk 0 represents the initial model.

Table 1. The LG_score of HC8, HC8-GR and HC8-RW

LG_score				
Target No.	**HC8**	**HC8-GR**	**HC8-RW**	**No. of Walks**
1	-0.278	-0.378	-0.293	20
2	-0.217	-0.143	-0.217	0
3	-0.267	0.865	-0.209	5
4	-0.159	-0.563	-0.159	0
5	-0.073	-0.109	-0.066	20
6	-0.34	-0.623	-0.357	3
7	-0.293	-0.044	-0.396	19
8	-0.14	-0.347	-0.188	20
9	-0.27	-0.07	-0.231	18
10	-0.222	-0.197	-0.28	7
11	0.034	-0.414	-0.055	11
12	-0.146	-0.434	-0.181	17
13	0.009	1.371	0.036	17
14	-0.094	0.67	-0.082	11
15	-0.308	-0.682	-0.375	19
16	-0.205	0.805	-0.185	11
17	-0.303	0.32	-0.28	1
18	-0.253	1.375	-0.313	16
19	-0.142	-0.381	-0.132	17
20	-0.191	-0.717	-0.165	17

Table 2 shows the MaxSub of the initial model produced by HyperChem 8.0(HC8), compared to the MaxSub produced by HC8 followed by GalaxyRefine (HC8-GR) and to that produced by HC8 followed by our suggested method (HC8-RW).

Table 2. The MaxSub of HC8, HC8-GR and HC8-RW

MaxSub				
Target No.	**HC8**	**HC8-GR**	**HC8-RW**	**No. of Walks**
1	-0.069	-0.048	-0.069	20
2	-0.077	-0.12	-0.077	0
3	-0.073	-0.006	-0.063	5
4	-0.057	-0.074	-0.057	0
5	-0.051	-0.05	-0.048	20
6	-0.08	-0.088	-0.083	3
7	-0.071	-0.023	-0.085	19
8	-0.063	-0.037	-0.065	20
9	-0.056	-0.105	-0.051	18
10	-0.047	0.024	-0.51	7
11	-0.039	-0.052	-0.046	11
12	-0.054	-0.077	-0.061	17
13	-0.058	0.05	-0.057	17
14	-0.048	-1.097	-0.047	11
15	-0.047	-0.014	-0.6	19
16	-0.091	-0.154	-0.082	11
17	-0.067	-0.011	-0.065	1
18	-0.056	0.01	-0.068	16
19	-0.073	-0.053	-0.068	17
20	-0.046	-0.056	-0.045	17

The shaded cells in Tables 1 and 2 represent the success that was achieved by our new method in terms of LG_score and MaxSub respectively. Notice that in Table 1, the number of sequences where the new method accomplished good performance was 11 sequences over 20 *i.e.*, 55% of the total cases. We consider the case where the optimum number of walks is 0 as a successful case and we argue that by the fact that the new method either enhance the model or at least preserve the initial good state of the model. Also notice that in Table 2, the number of sequences where the new method accomplished good performance was 12 sequences over 20 i.e. 60% of the total cases.

We can conclude that almost 55% of cases were successfully enhanced by the new method. Also, we can see that in the rest of the cases the decrease in the model quality was very small in comparison with the Galaxy Refine method. Why? The answer is that we do not destruct the initial model and resolve it from the beginning as in the Galaxy Refine method… We just do very small steps in a random way and calculate the total energy of each model. The refined model is not so far from the initial one.

5. Conclusions and Future Work

In this paper, we presented a new method for refining 3D protein structure models. The new method is based on 3D random walk on a graph where nodes are the atoms and the edges are the inter-molecular distances between those atoms that form the residues of the protein sequence. At each walk, we calculated the total energy of the protein to be minimized. This method has enhanced the protein models in 55% of the cases done in the experiments.

As a future work, we suggest the usage of the Markov chain algorithm instead of the random walk. Markov chain also, like random walk, identifies the random processes but add a transition probability matrix where we can define the steps of the atoms. The atoms in a protein can move in any direction but with preserving the protein structure topology, *i.e.*, there are states where the atoms cannot move to and those states can have the probability 0 in the transition probability matrix. The use of a Markov chain presumes a discrete set of states. The main challenge in this work is to be able to discretize the protein state space. Such enhancement can decrease the cases to be examined in the traditional random walk and also may lead to more accurate results.

References

[1] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, *et al.*, “The universal protein resource (UniProt)”, Nucleic acids research, vol. 33, no. suppl. 1, **(2005)**, D154-D159.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, “The protein data bank”, Nucleic acids research, vol. 28, no. 1, **(2000)**, pp. 235-242.

[3] A. Raval, S. Piana, M. P. Eastwood, R. O. Dror and D. E. Shaw, “Refinement of protein structure homology models via long, all-atom molecular dynamics simulations”, Proteins: Structure, Function, and Bioinformatics, vol. 80, no. 8, **(2012)**, pp. 2071-2079.

[4] F. Di Maio, T. C. Terwilliger, R. J. Read, A. Wlodawer, G. Oberdorfer, U. Wagner, E. Valkov, *et al.*, “Improved molecular replacement by density-and energy-guided protein structure optimization”, Nature, vol. 473, no. 7348, **(2011)**, pp. 540-543.

[5] J. Zhang, Y. Liang and Y. Zhang, “Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling”, Structure, vol. 19, no. 12, **(2011)**, pp. 1784-1795.

[6] L. Heo, H. Park and C. Seok, “GalaxyRefine: protein structure refinement driven by side-chain repacking”, Nucleic acids research, vol. 41, no. W1, **(2013)**, pp. W384-W388.

[7] D. Bhattacharya and J. Cheng, “3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization”, Proteins: Structure, Function, and Bioinformatics, vol. 81.1, **(2013)**, pp. 119-131.

[8] R. A. Di Stasio, O. A. von Lilienfeld and A. Tkatchenko, “Collective many-body van der Waals interactions in molecular systems”, Proceedings of the National Academy of Sciences, vol. 109, no. 37, **(2012)**, pp. 14791-14795.

[9] “Protein Structure Prediction Center (CASP)”, http://predictioncenter.org/.

[10] B. Wallner and A. Elofsson, “Can correct protein models be identified?”, Protein Sci., vol. 12, no. 5, **(2003)**, pp. 1073-1086.

[11] T. Jiang, Q. Cui, G. Shi and S. Ma, "Protein folding simulations of the hydrophobic–hydrophilic model by combining tabu search with genetic algorithms", The Journal of chemical physics, vol. 119, no. 8, **(2003)**, pp. 4592-4596.

[12] L. Lovász, "Random walks on graphs: A survey", Combinatorics, Paul erdos is eighty, vol. 2, no. 1, **(1993)**, pp. 1-46.

[13] "Cliffsnotes: Tertiary Structure", http://www.cliffsnotes.com/sciences/biology/biochemistry-i/protein-structure/tertiary-structure.

[14] M. Topf, M. L. Baker, M. A. Marti-Renom, W. Chiu and A. Sali, "Refinement of protein structures by iterative comparative modeling and CryoEM density fitting", Journal of molecular biology, vol. 357, no. 5, **(2006)**, pp. 1655-1668.

[15] G. F. Schröder, A. T. Brunger and M. Levitt, "Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution", Structure, vol. 15, no. 12, **(2007)**, pp. 1630-1641.