Asian Journal of Biochemistry

ISSN 1815-9923

Academic
Journals Inc.

www.academicjournals.com

OPEN ACCESS Asian Journal of Biochemistry

ISSN 1815-9923
DOI: 10.3923/ajb.2016.149.155

Research Article
A Novel Conformation Generation Framework for *De novo* Protein Structure Prediction Using Hydrophobic-polar Model

[1]Sandhya P.N. Dubey, [1]N. Gopalakrishna Kini, [2]M. Sathish Kumar, [3]S. Balaji, [1]M.P. Sumana Bhat and [1]Harshad R. Kavathiyal

[1]Department of Computer Science and Engineering,
[2]Department of Electronics and Communication Engineering,
[3]Department of Biotechnology, Manipal Institute of Technology, Manipal University, Manipal, India

Abstract

Background and Objective: The Protein Structure Prediction (PSP) problem is one of the hardest problems in computational biology and technological research. To reduce the complexity of the problem, Dill proposed the Hydrophobic-Polar (HP) model which subsequently became a major tactic to the PSP problem. **Methodology:** In this study, a novel algorithm was proposed to solve the PSP problem using Dill's HP model. Here, protein conformation is modeled with square and triangular lattice. The proposed method is tested on a set of benchmark sequences and real protein sequence 4RXN taken from Protein Data Bank (PDB). **Results:** It is observed that, the proposed approach results in good quality conformation in term of hydrophobic contact with 89% of accuracy. Also, experimental results show that the structure modeled with triangular lattice is closer to the wet lab structure. **Conclusion:** Results indicate that the proposed approach is promising to generate good number HP conformation, with immense drop in time.

Key words: Dill's HP model, hydrophobic and hydrophilic amino acid, lattice model, NP-complete problem, protein folding, protein structure prediction

Received: December 14, 2015 Accepted: February 16, 2016 Published: April 15, 2016

Citation: Sandhya P.N. Dubey, N. Gopalakrishna Kini, M. Sathish Kumar, S. Balaji, M.P. Sumana Bhat and Harshad R. Kavathiyal, 2016. A novel conformation generation framework for *de novo* protein structure prediction using hydrophobic-polar model. Asian J. Biochem., 11: 149-155.

Corresponding Author: Sandhya P.N. Dubey, Department of Computer Sciences and Engineering, AB-5, Manipal Institute of Technology, Manipal-576104, Karnataka, India

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Protein is one of the most important macromolecule in all living organisms and more than half of dry weight of our body is made of protein[1]. Protein plays a diverse role within every organism on the planet earth and its functions are governed by the 3D structures. Basically each protein is made of 20 different kinds of amino acids connected together to form a linear chain. Further, based on the physiochemical properties of connected amino acids and its folding environment, it folds into a unique structure which leads it to perform a key function. However, the miss-folded protein causes many life threatening diseases such as Alzheimer, cystic fibrosis, mad cow, etc.[2]. Predicting the native structure of protein based on a given sequence of amino acids is referred as the Protein Structure Prediction (PSP) problem. Solutions to this problem will result in improvising the life on planet earth by combating many diseases, designing many new proteins, genetic engineering and increasing the yield of crops[3].

The limitation of wet lab structure prediction techniques has guided the PSP problem towards computational side for assisting in solving[4]. Computationally there are three approaches to deal with this problem viz., homology modeling, threading and *de novo* approach. First two approaches depend on known structures of protein. Based on available structures they attempt to solve the PSP problem and are called as template-based approach. However, till date only one percent of sequence's structure is known[5] and is coming up with many new sequences.

It is difficult to solve the PSP problem only by referring to the known structure. This limitation of template based modeling has given rise to a new area of research wherein PSP is computationally modeled based on physiochemical properties of amino acids. Attempt to solve the PSP with its primary sequence, properties of constituent amino acid and folding environment is referring as *de novo* PSP. In terms of pure computational approach, PSP is regarded as NP-hard problem.

To deal with NP-hardness, hierarchical approach has given a good result[6]. The first level of hierarchy, deals with the most simplified model and as it move towards the higher levels of hierarchy, more detailed approaches are included. Dill's HP model is one of the most successful modeling techniques to deal with PSP at first level of hierarchy. However, even with HP model, the PSP problem has been proven to be NP-complete problem[7]. There are many attempts to solve the PSP with HP model[8-11] using various approaches. In this study, *de novo* protein structure prediction problem with HP and functional model was addressed. Here, a new tactic was proposed, using memory based learning. Experimental results shows that proposed approach accelerate the convergence speed and the quality of obtained conformation also shows the agreement with the state-of-the-art results. This study is implemented on Visual Studio 2012 platform with C^{++} programming. Furthermore, predicted structures are visualized with OpenGL graphics.

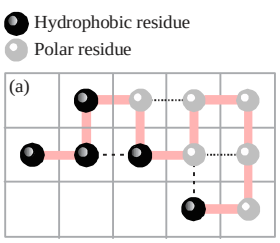
The remainder of the paper is organized as follows: section 2 give details on Dill's HP model, functional model and problem formulation. Section 3 provides detail on proposed algorithm. The results of performed experiment and discussion on obtained results are presented in section 4. Finally, section 5 provides the conclusion of the work and the future direction.

MATERIALS AND METHODS

Dill's HP model: Dill[12] proposed the Hydrophobic-Polar (HP) model by considering the behavior of amino acids in a folding environment perceived as aquatic. This model considers the amino acids as either Hydrophobic (H) or as Hydrophilic (polar P). Hydrophobic amino acids repel water and are buried deep inside the aquatic environment whereas, the hydrophilic amino acids act as strainer. The HP model is one of the most successful models to solve the PSP at coarse level[2,13-21]. Classification of amino acid is as follows: H = CFILMVWY and P = HATGPSQRNDEK amino acids are represented in their 1-letter code. However, this model has considered only the Hydrophobic-Hydrophobic (H-H) interaction by assigning the value of minus one for each such interaction, whereas, for other interaction (P-P, P-H or H-P) it assigns 0.

Functional model: Unlike Dill's HP model this model has considered the other interactions (H-P, P-H and P-P). This consideration has brought about one of the most important property of protein folding namely ligand binding site. Ligand binding site plays a key role in protein folding as it carries this feature for protein docking problem, helps in drug designing etc.[22]. Figure 1a depicts one possible conformation for sequence PPPHPHHHHHP in HP model. Figure 1b, c shows the energy table for HP and functional model respectively[23].

Problem formulation: The following constraints need to be satisfied when modeling the PSP on lattice:



(b)

	H	P
H	-1	0
P	0	0

(c)

	H	P
H	-2	1
P	1	1

Fig. 1(a-c): (a) One possible conformation in HP model for the sequence PPPHPHHHHHHP. Topological contact is shown with dotted lines, whereas consecutive elements of HP sequence are connected with solid red line. Free energy for this conformation is -2 in both HP and functional model, (b) HP energy matrix and (c) Functional energy matrix

Table 1: Benchmark sequence to test

Sequences	Sequence length	Max H-H
$PHP^2HPHP^4HP^5H$	18	4
$PHP^2HPHP^4HP^5HP^2H^2P$	23	6
$PHP^2HPHP^4HP^5HP^2H^2P^2$	28	7
$HPHP^2H^2PHP^2HPH^2P^2HPH$	20	9
$H^2P^2HP^2HP^2HP^2HP^2HP^2HP^2H^2$	24	9
$P^2HP^2H^2P^4H^2P^4H^2P^4H^2$	25	8

- **Self-Avoiding Walk (SAW):** Each amino acid must occupy only one lattice point, which no other amino acid can share
- Consecutive amino acids of protein primary sequence must occupy adjacent lattice point

Two H amino acids adjacent on lattice form an H-H interaction which causes reduction of free energy and is called as Topological Neighbor (TN)[24]. Each occurrence of a TN due to two hydrophobic residues lowers the total energy of the conformation by one in the HP model, whereas it lowers the energy by two in the functional model. The PSP was represented by a linear program as follows: For a given protein sequence $S = s_1, s_2,..., s_n$ of length n where each $s_i \in \{H, P\}$. Let l_{ijk} is the kth amino acid of the sequence at (I, j) position on the 2D lattice model where:

$$l_{ijk} = \begin{cases} 1 \text{ if kth amino acid at (i, j) is hydrophibic} \\ 0, \text{ if kth amino acid is polar} \end{cases}$$

In order to check the availability of lattice point and ascertain whether it is free or occupied and maintain the Boolean array where, k_{ij} represents the kth amino acid placed at (I, j) position in 2D array:

$$k_{ij} = \begin{cases} 1 \text{ if kth amino acid is placed at point (i, j)} \\ 0, \text{ otherwise} \end{cases}$$

Objective function is to maximize the H-H contact given with Eq. 1:

$$\text{Maximize } Z = \sum_{i,\, j:\, i+1<j} c_{ij} . e_{ij} \quad (1)$$

Where:

$$c_{ij} = \begin{cases} 1 \text{ if kth amino acid at i and j form TN} \\ 0, \text{ otherwise} \end{cases}$$

Energy values assign as follows:

$$\text{For HP model, } e_{ij} = \begin{cases} -1 \text{ if } s_i = s_j = H \\ 0, \text{ otherwise} \end{cases}$$

$$\text{For functional model, } e_{ij} = \begin{cases} -2 \text{ if } s_i = s_j = H \\ 1, \text{ otherwise} \end{cases}$$

With the above energy function, the PSP problem was modeled on square and triangular lattice for both the energy definitions. Further, the comparative analysis of both the cases was performed with standard benchmark sequences as listed in Table 1.

Proposed algorithm: In order to reduce the search time and to improve the quality of conformations, a novel structure prediction algorithm was developed. Proposed algorithm is grounded on fact that while folding, protein follows some specific folding pathways called as folding rule. These folding rules are defined as follows: (1) Each amino acid computes its corresponding neighbors with which it forms the hydrophobic contact and (2) Each amino acid maintains its memory from previous interactions to find the hydrophobic neighbor, thus resulting in hydrophobic core construction.

Table 2: Comparison table on the basis of time and H-H contact

Yoon[26]				Sayantan and Nanda[25]		Proposed algorithm	
IP1	H-H count	IP2	H-H count	Time	H-H count	Time	H-H count
92 min	4	>10 h	4	9 sec	4	5 sec	2
>10 h	6	>10 h	6	23 min	6	798 sec	4
>10 h	7	>10 h	7	12 h	7	6980 sec	3
N/A	N/A	N/A	N/A	36 sec	9	8 sec	6
N/A	N/A	N/A	N/A	50 min	9	1278 sec	6
N/A	N/A	N/A	N/A	123 min	8	3491sec	6

N/A: Not considered in the study

Algorithm 1:

Procedure: Conformation generation
Input: S_i = {H, P}$^+$, Search Size//i = 0 to n, where n is the length of input sequence
Output: Optimal_conf, Energy

```
index←0
while index<Search Size do
    startPos = (0, 0)
    posList.add (startPos)
    secPos = eastneighbour (startPos)
    posList.add (secPos)
    for I = 2 to n-1
if Si = = H then
        Memory = 0
        for j = i-2 to 0
        if Sj = = H then
        Memory++
        else
        Memory - -
    endif
  endfor
  Choose the position Pos next to maximum memory
  else
  select the move randomly Pos for P residue
  endif
  endfor
  posList.add (Pos)
  return posList as conf
  Calculate the Energy of conf
  Store conf, Energy
  Index + +
end while
Return conf with maximum Energy
```

The aim of this study is to come up with an appropriate method to generate good quality conformation for HP model by satisfying all the constraints mentioned in the previous section. The efficiency of the proposed algorithm is validated using the test beds listed in Table 1. Table 2 presents a comparison of our work with the approaches proposed by Sayantan and Nanda[25] and Yoon[26] which deals with 2D square lattices. Proposed framework generates good quality conformation with inordinate drop in time.

RESULTS

Effect of neighbor selection strategy: This has increase the number of close conformation (compact conformation in term of space occupied) and reduced the number of conformation which is far related with native structure.

Impact of neighbor selection w.r.t. fitness function: Neighbor selection strategy has improved the convergence of folding toward the native conformation as in early folding each amino acid try to form the cluster of same residue type and distance from the other. This has results in better fitness value in the early folding (Table 2).

Impact of neighbor selection w.r.t. time: Neighbor selection strategy has reduced the search spaces size (as it removes the far related conformation) has resulted in intense drop for computation time (Fig. 2).

Effect of memory based learning: This has cent percent cut the folding of sequence into non-SAW conformation, results into better search space with all SAW conformation.

Impact of memory based learning w.r.t. fitness function: Memory based learning has lessened the time to attain a better fitness function and results into early convergence to native conformation.

Impact of memory based learning w.r.t. time: Memory based learning has removed the chances of any two residues to occupy the same location on lattice. Hence, it helps in generating only SAW conformation. Moreover, it also helps forming conformation with higher number of H-H count as every amino acid checks it nearest neighbor and try to take a lattice point closest to similar residue. This has improved the quality of possible conformation space to be search and reduced the search space size.

Statistical analysis test: To relate the nearness of the predicted conformation with real structure, Rubredoxin protein (4RXN) was considered. In wet lab, Rubredoxin structure is determined through x-ray crystallography. The 4RXN is one of the earliest benchmark proteins which serve as the test case[27]. The structure of 4RXN was modeled with proposed approach on both the square and triangular lattices and visualized with ACD/ChemSketch (ver 12.0) software as shown in Fig. 3.

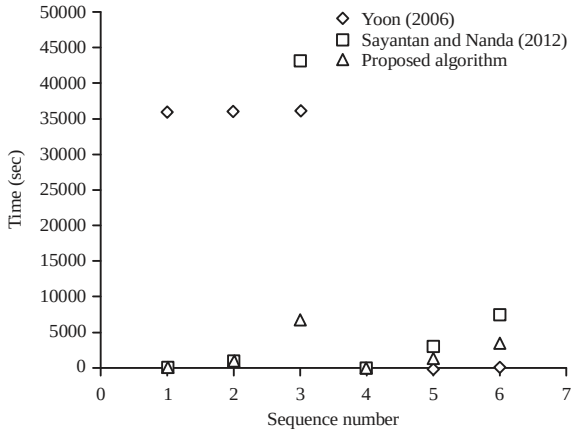


Fig. 2: Comparision of time (sec) to generate initial conformation for benchmark sequences

However, the 4RXN obtained conformation was evaluated (Fig. 3d) with wet lab conformation (Fig. 3c) it's not guarantee the cent percent correct modelling as wet lab conformation is presented over the 3D space whereas conformation obtained with the proposed approach is represented in the 2D space.

The evaluation of predicted structure is done in term of statistical analysis unit, sensitivity, specificity and accuracy, presented in Table 3.

Proposed approach has shown a good agreement between predicted conformation and experimental determined structure. Accuracy of predicted conformation is obtained to 89 and 76% for 4RXN, respectively on triangular and square lattice. Triangular lattice based implementation has given the good utility for proposed algorithm with 88, 89 and 89% sensitivity, specificity and accuracy respectively as shown in Fig. 4. Over both square and triangular lattice, proposed approach gives high values of specificity.

Table 3: Statistical analysis result for protein 4RXN

	Sensitivity (%)	Specificity (%)	Accuracy (%)
Square lattice	60	88	76
Triangular lattice	88	89	89

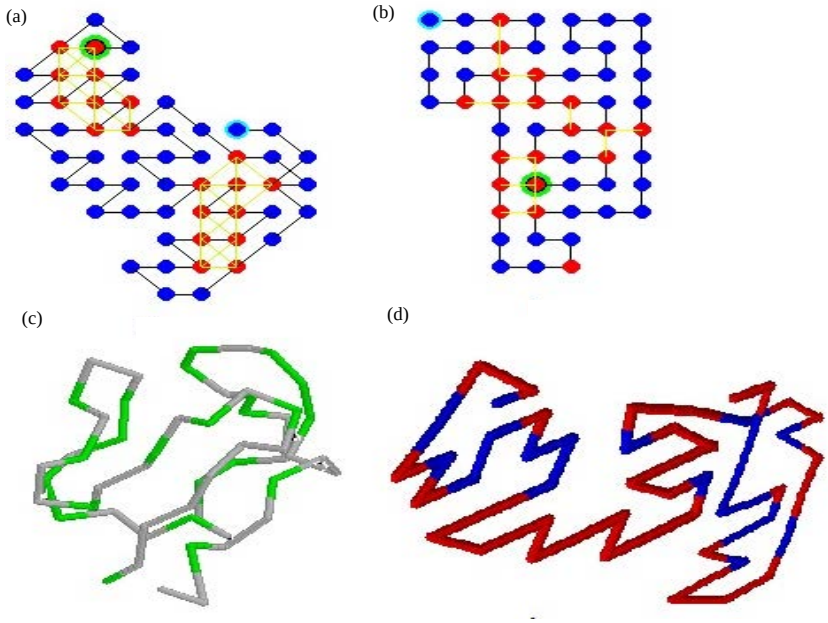


Fig. 3(a-d): Conformation for PDB ID: 4RXN: AMKKYTCTVCGYIYDPEDGDPDDGVNPGTDFKDIPDDWVCPLCGKDEFEEVEE. The HP conversion: HPPHPHPHHPHHHPPPPPPPPPHPPPPPHPPHPPPHHHPHHPHPPPPHPPHPP. (a-b) Structure on triangular and square lattices, respectively, (c) Backbone strand of 4RXN visualized with RasMol (v.2.7.5) and (d) Backbone strand with proposed approach

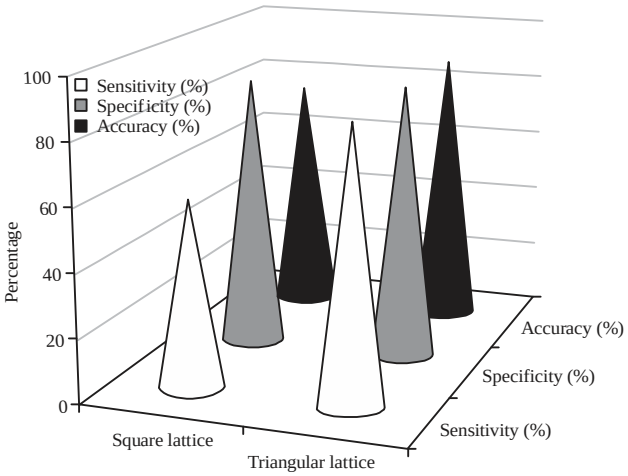


Fig. 4: Graph for sensitivity, specificity and accuracy obtained for proposed approach

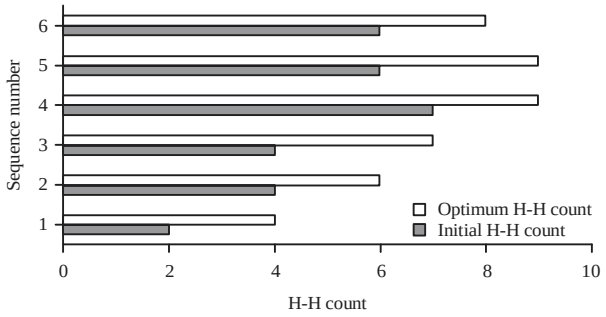


Fig. 5: Relative analysis of proposed apporach (initial H-H count) w.r.t state-of-art-the-art (optimum H-H count) objective function values

DISCUSSION

Proposed approach results in fast convergence to generate good number of valid conformation. Obtained conformations shows the good agreement with state-of-the-art results in term of H-H count. Further, computed approximations will be served as initial solutions for the PSP problem in hierarchical approach[14]. In on lattice modelling of PSP with HP model H-H count is only parameter for quantative assessment of predicted structure[11]. Although, proposed approach has not reached the optimal H-H count value as claimed in Islam and Chetty[24], Shatabda *et al.*[10], Yoon[26] and Sayantan and Nanda[25], results are significant as it has been generated in single run with inordinate drop in time (Fig. 2). In terms of population based strategies, it is a result of first generation whereas reported the-state-of-results are final optimum outcome after thousands of run[24]. Moreover, proposed approach obtains the objective value near to optimal value (>50%) in the first round of evaluation (Fig. 5).

The comparative analyses of square and triangular lattice model indicate that the triangular lattice[16] is presenting the nearer results (89% accuracy) with the experimental structure compared to square lattice. Furthermore we had observed that even though functional model[9] has considered the ligand binding concept still there is a need to use more physiochemical properties of amino acids which may in turn result in higher accuracy.

Objective of this proposed framework is to generate good initial conformation which can be directly used in any metaheuristics technique to approach the PSP problem. Proposed framework can be incorporated on any platform which supports the graphic file to visualize the conformation. Also as it is developed with object oriented programming concept there are many option to adept the framework and directly make use of this.

CONCLUSION

This study presented the exact method to generate the preliminary conformation to address the PSP problem with HP model. Our long term interest is to develop an efficient complete solution to deal with the PSP problem which can predicts the 3D structure of protein from its primary sequences. Further, results in improving the life on planet earth by combating many diseases, designing new proteins, genetic engineering, increasing the yield of crop.

REFERENCES

1. Cooper, G.M. and R.E. Hausman, 2006. The Cell: A Molecular Approach. 4th Edn., Sinauer Associates, Inc., Sunderland, ISBN-13: 978-0878932191, Pages: 745.
2. Samudrala, R., Y. Xia, E. Huang and M. Levitt, 1999. Ab initio protein structure prediction using a combined hierarchical approach. Proteins: Struct. Funct. Bioinform., 37: 194-198.
3. Helles, G., 2008. A comparative study of the reported performance of *ab initio* protein structure prediction algorithms. J. R. Soc. Interface, 5: 387-396.
4. Shmygelska, A. and H.H. Hoos, 2003. An Improved Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem. In: Advances in Artificial Intelligence, Xiang, Y. and B. Chaib-Draa (Eds.). Springer, Berlin, ISBN: 978-3-540-40300-5, pp: 400-417.
5. Marks, D.S., T.A. Hopf and C. Sander, 2012. Protein structure prediction from sequence variation. Nature Biotechnol., 30: 1072-1080.

6. Xia, Y., E.S. Huang, M. Levitt and R. Samudrala, 2000. *Ab initio* construction of protein tertiary structures using a hierarchical approach. J. Mol. Biol., 300: 171-185.
7. Wroe, R., E. Bornberg-Bauer and H.S. Chan, 2005. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: Robustness of the superfunnel paradigm. Biophys. J., 88: 118-131.
8. Hart, W.E. and S.C. Istrail, 1996. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. J. Comput. Biol., 3: 53-96.
9. Cutello, V., G. Nicosia, M. Pavone and J. Timmis, 2007. An immune algorithm for protein structure prediction on lattice models. IEEE Trans. Evol. Comput., 11: 101-117.
10. Shatabda, S., M.A.H. Newton, M.A. Rashid, D.N. Pham and A. Sattar, 2013. The road not taken: Retreat and diverge in local search for simplified protein structure prediction. BMC Bioinform. 10.1186/1471-2105-14-S2-S19
11. Dubey, S.P., S. Balaji, N.G. Kini and S. Kumar, 2015. A comparative study of various meta-heuristic algorithms for ab initio protein structure prediction on 2D hydrophobic-polar model. Proceedings of the Advances in Intelligent Systems and Computing, (AISC'15), IIT Roorkee, India.
12. Dill, K.A., 1985. Theory for the folding and stability of globular proteins. Biochemistry, 24: 1501-1509.
13. Unger, R. and J. Moult, 1993. Genetic algorithms for protein folding simulations. J. Mol. Biol., 231: 75-81.
14. Hoque, M.T., M. Chetty and L.S. Dooley, 2006. A Hybrid Genetic Algorithm for 2D FCC Hydrophobic-Hydrophilic Lattice Model to Predict Protein Folding. In: AI 2006: Advances in Artificial Intelligence, Sattar, A. and B.H. Kang (Eds.). Springer, Berlin, ISBN: 978-3-540-49787-5, pp: 867-876.
15. Islam, M.K. and M. Chetty, 2010. Clustered memetic algorithm for protein structure prediction. Proceedings of the IEEE Congress on Evolutionary Computation, July 18-23, 2010, IEEE, Barcelona, Spain, pp: 1-8.
16. Su, S.C., C.J. Lin and C.K. Ting, 2011. An effective hybrid of hill climbing and genetic algorithm for 2D triangular protein structure prediction. Proteome Sci., Vol. 9. 10.1186/1477-5956-9-S1-S19
17. Rashid, M.A., S. Shatabda, M.A.H. Newton, M.T. Hoque and A. Sattar, 2014. A parallel framework for multipoint spiral search in *ab initio* protein structure prediction. Adv. Bioinform. 10.1155/2014/985968
18. Shatabda, S., M.A.H. Newton, M.A. Rashid, D.N. Pham and A. Sattar, 2014. How good are simplified models for protein structure prediction? Adv. Bioinform. 10.1155/2014/867179
19. Shaw, D.L., A.S.M.S. Islam, M.S. Rahman and M. Hasan, 2014. Protein folding in HP model on hexagonal lattices with diagonals. BMC Bioinform., Vol. 14. 10.1186/1471-2105-15-S2-S7
20. Shehu, A. and K.D. Jong, 2015. Evolutionary algorithms for protein structure modeling. Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation July 11-15, 2015, Madrid, Spain, pp: 533-545.
21. Tsay, J.J., S.C. Su and C.S. Yu, 2015. A multi-objective approach for protein structure prediction based on an energy model and backbone angle preferences. Int. J. Mol. Sci., 16: 15136-15149.
22. Bonneau, R. and D. Baker, 2001. Ab initio protein structure prediction: Progress and prospects. Annu. Rev. Biophys. Biomol. Struct., 30: 173-189.
23. Blackburne, B.P. and J.D. Hirst, 2001. Evolution of functional model proteins. J. Chem. Phys., 115: 1935-1942.
24. Islam, K. and M. Chetty, 2013. Clustered Memetic algorithm with local heuristics for ab initio protein structure prediction. IEEE Trans. Evol. Comput., 17: 558-576.
25. Sayantan, M. and D.J. Nanda, 2012. Protein structure prediction using 2D HP lattice model based on integer programming approach. Proceedings of the International Congress on Informatics, Environment, Energy and Application, March 17-18, 2012, Singapore, pp: 1-5.
26. Yoon, H.S., 2006. Optimization approaches to protein folding. Ph.D. Thesis, School of Industrial and System Engineering, Georgia Institute of Technology.
27. Adman, E.T., L.C. Sieker and L.H. Jensen, 1991. Structure of rubredoxin from *Desulfovibrio vulgaris* at 1.5 A resolution. J. Mol. Biol., 217: 337-351.