*2012 International Congress on Informatics, Environment, Energy and Applications-IEEA 2012*
*IPCSIT vol.38 (2012) © (2012) IACSIT Press, Singapore*

# Protein Structure Prediction using 2D HP Lattice Model Based on Integer Programming Approach

Sayantan Mandal[a], Nanda Dulal Jana[b]

[a,b]Department of Information Technology
National Institute of Technology
National Institute of Technology, Durgapur West Bengal – 713209, INDIA
[a]pikusayantan@gmail.com, [b]nanda.jana@gmail.com

**Abstract.** To predict structure of protein from primary amino acid sequence is computationally difficult. Protein folds on a lattice called conformation predict a native confirmation which has maximum topological hydrophobic contact. In this paper we address Integer Programming approach to predict protein structure using 2D square HP lattice. We proposed method that bound the conformational space where protein sequence are folded and discard the symmetric conformation. The proposed method applied on some benchmark sequence of proteins. The experimental result shows the efficiency and effectiveness of the proposed method.

**Keywords:** Protein structure prediction, Integer programming, HP lattice model, Bioinformatics, Hydrophilic hydrophobic amino acid, Protein

## 1. Introduction

A protein is a sequence formed by combination 20 amino acids. Amino acids are characterized by polar (hydrophilic) and non polar (hydrophobic) based on its residue. These amino acids are connected to each other with peptide bond to form a protein sequence [5].

Protein sequences are folded on lattice with non overlapping amino acid chain. These self avoiding conformations produced native structures which have minimum energy configuration [2]. The native structure is the 3D structure of the protein sequence. We need to know the native protein structure so that we can find out the function of that protein structure.

There are many optimization algorithms used to predict the native confirmation of a protein. Again taking deterministic approach will take long time as the search space is huge, and searching in each and every folding in that search space takes lots of time.

In this work protein structure prediction is formulated by integer programming method. Confirmation space contains huge number of confirmation for a protein. Finding the native confirmation from this space takes lots of time. Therefore it is necessary to bound the search space and discard the symmetric confirmation from that space. Here we used backtracking method to avoid non-overlapping amino acids. This concept helped a lot in reducing time to find the native confirmation.

## 2. HP lattice model

HP lattice model was introduced by Dill [1] in 1987. The different types of 20 amino acids are categorized in hydrophobic (H) and hydrophilic (P) depending on affinity towards water. The HP lattice model has some advantage for PSP, these are:

(i) Two amino acids are said to be consecutive if peptide bond (covalent bond) occur between two amino acids. When there is a contact between two non consecutive hydrophobic (H) amino acids takes place then it causes reduction of free energy of the resultant molecule. So, protein reaches ground state, when number of non consecutive hydrophobic amino acids contact is maximize. (ii) Two hydrophobic (H) amino acids adjacent on lattice form an H-H contact. (iii) Hydrophobic (H) amino acid cluster in middle and hydrophilic amino acid surrounding the cluster [4]. (iv) Each confirmation is a graph where at each vertex resides amino acid and edges denotes peptide bonds. (v) In this model even amino acid is always contact with odd amino acid as adjacent on-lattice and vice versa. (vi) In 2D HP square lattice model any amino acid except first and last amino acid, has two peptide bonds (adjacent on sequence) and maximum of two topological contacts (adjacent on lattice).

HP lattice model of the optimized protein sequence 'HPHPPHHPHPPHPHHPPHPH'. Black circles (white circles) are hydrophobic (hydrophilic) amino acids. Continuous line represents polypeptide bonds and the dotted line denotes H-H contacts. Here maximum H-H contact is 9.

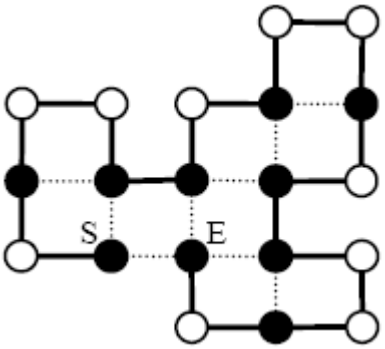


Fig 1: optimized 2D HP lattice model

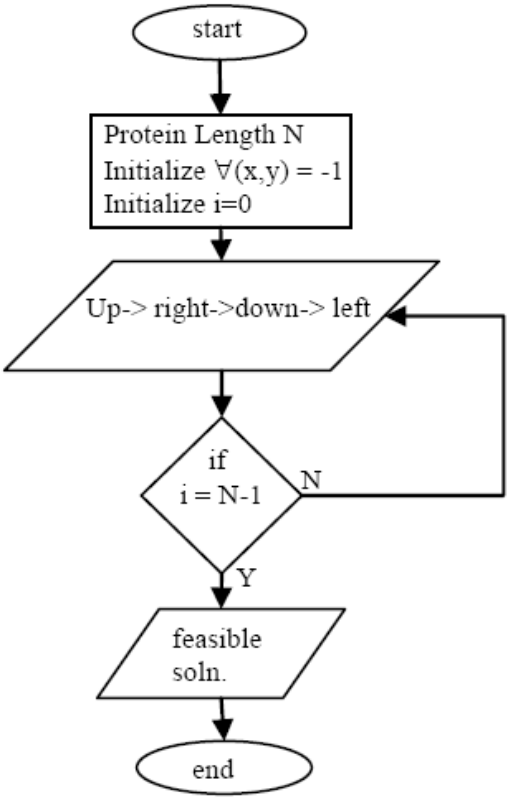


Fig 2: Basic flowchart for searching feasible folding

## 3. Previous Works

Protein structure prediction using Integer programming in [6], it takes much longer time for small sequence. It searched many invalid folding makes the program run for much more time. It also checked folding which are valid but those folding don't give optimum results at any cost. In [6] some algorithm place a particular amino acid by defining distance or range in lattice from centre coordinates of their lattice model. In that way it reduces placing of an amino acid in impossible location. But still there remains many invalid folding.

## 4. Methodology

### 4.1. Integer programming formulation of PSP

In this section we represent a mathematical model of protein structure in 2D HP lattice model. The mathematical formulation can be describe in the following manner

Objective: Maximize the number of H-H contacts.

Subject to:

Assignment: Each amino acid must occupy one lattice point.

Non-overlapping: No two amino acids may share the same lattice point.

Connectivity: Every two amino acids that are consecutive in the protein sequence must occupy adjacent lattice point.

Therefore, the objective is to maximize the H-H contacts and obtain the feasible solution, satisfying the constraints, the assignment, non-overlapping and the connectivity.

We consider a protein sequence of n amino acids

$X_{ijk}$ is the $K^{th}$ amino acid of the sequence at $(i, j)$ position on the 2D HP lattice model

$$\text{Let Xijk} \begin{cases} 1 \text{ if Kth amino acid at } (i,j) \text{ is hydrophobic} \\ 0 \text{ if Kth amino acid is polar} \end{cases}$$

$K_{ij}$ is Kth amino acid placed at (i,j)

$$\text{Let } K_{ij} \begin{cases} 1 \text{ if } K^{th} \text{ amino acid is placed at point}(i,j) \\ 0 \text{ otherwise} \end{cases}$$

$Y_{ijku} = 1$ if both $K_{ij}$ and $K'_{i(j+1)}$ both are hydrophobic, i.e. check between $k^{th}$ node and node above it

$Y_{ijkr} = 1$ if both $K_{ij}$ and $K'_{(i+1)j}$ both are hydrophobic, i.e. check between $k^{th}$ node and node right side of it.

So, we need to maximize H-H bond

$$Maximise\ Z = \sum_{K=0}^{N-1} (Yijku + Yijkr) \qquad \forall k, i, j$$

Non-Overlapping

$$\sum_{K=0}^{N-1} K_{ij} \leq 1 \qquad \forall i, j$$

Connectivity

Let $K^{th}$ amino acid is placed at $(i, j)$, so $K_{ij} = 1$, then,

$$(K+1)_{(i+1)j} + (K+1)_{i(j+1)} + (K+1)_{(i-1)j} + (K+1)_{i(j-1)} = 1$$

## 4.2. Bound area for search

While searching all the feasible folding makes algorithm to run for long time. To reduce runtime we will bound the searching area, so that folding are compact and searching is done in defined bounded area. We know that hydrophobic amino acid forms the inner core and polar amino acid surrounds that inner core of that amino acid and hence forms a compact structure. For example many structures like which form straight graph or double line graph (for big length series) it is obvious that it will not give maximum HH contacts, so we can discard them. Structure with many amino acids reside without any neighbour amino topologically connected are also discarded as it doesn't form compact structure.

We have bounded the region using the formula $P = 2(\sqrt{n} + 3)$ where n is the length of amino acid. Length and breadth of the conformational space is P.

## 4.3. Proposed Methodology

We will set 1st amino acid at center of the lattice(N,N) and 2nd one just right of it (N+1,N). Fixing them will avoid rotational symmetry. We follow a sequence for traversal like this Up→Right→Down→Left, let call this rule URDL rule. Each amino acid will have this set of rule one after another. If one amino acid is moved to new node on lattice then this rule will apply form start for that amino acid. There will be a pointer to track current working node that is to be placed. Now according to the rule 3rd amino acid will move up

direction, and so like that next amino acid will also be placed in up direction till the last amino acid is placed and pointer will point to N-1$^{th}$ node. Whenever last amino acid is successfully placed it means we got one complete folding and calculates its HH contact. Now last amino acid will move right according to rule and completes another folding. Now, last node will try to place in the down direction with respect to n-1$^{th}$ node, but there is already N-2$^{th}$ node is placed so it will not place nth amino acid in down direction and move to left direction according to rule and completes another folding. Now all the combination is completed, so it will backtrack one node back to N-2$^{th}$ amino acid and pointer points this node. Now the N-1$^{th}$ node is placed to right to N-2$^{th}$ node and pointer points to this node. Now as it got new coordinate the rule will be applicable from start. Hence, N$^{th}$ amino acid will be placed in the up direction with respect to N-1$^{th}$ amino acid and completes another fold. In this way it keeps on placing amino acid on lattice and checks its HH bonds. While placing we took many consideration about boundary. Flowchart shown in fig. 2 shows the movement algorithm.

### 4.4. Evaluating HH contact

When we are getting valid folding on square lattice model, we are checking for number of amino acid present in the sequence. We start checking from first amino acid and gradually move to last amino acid. We first check if that amino acid is hydrophobic or polar

$$X_{ijk}= 1$$

If amino acid is polar then we check nodes in up direction and node in left direction.

$Y_{ijku}= 1$ if both $K_{ij}$ and $K'_{i(j+1)}$ both are hydrophobic, i.e. check between k$^{th}$ node and node above it.

$Y_{ijkr}= 1$ if both $K_{ij}$ and $K'_{(i+1)j}$ both are hydrophobic, i.e. check between k$^{th}$ node and node right side of it.

Let number of HH bond in amino acid side chain= S

Then number of amino HH bond in folding but not in chain:

$$\text{Total HH bond}= \sum_{K=0}^{N-1}\left(Y_{ijku} + Y_{ijkr}\right) - S \quad \forall i, j$$

When checking surrounding for even amino acid we will check its adjacency with odd amino acid discarding all even amino acid and vice versa.

## 5. Results and Discussion

In this section, we explain the result obtained by our proposed method on various benchmark sequences [7] and compare this with [6]. In table 1 shows the benchmark sequence with length and maximum H-H contact known to date.

| Seq No | Sequence | Seq. Len | Max H-H |
|---|---|---|---|
| 1 | $PHP^2HPHP^4HP^5H$ | 18 | 4 |
| 2 | $PHP^2HPHP^4HP^5HP^2H^2P$ | 23 | 6 |
| 3 | $PHP^2HPHP^4HP^5HP^2H^2P^2HP^3$ | 28 | 7 |
| 4 | $HPHP^2H^2PHP^2HPH^2P^2HPH$ | 20 | 9 |
| 5 | $H^2P^2HP^2HP^2HP^2HP^2HP^2HP^2H^2$ | 24 | 9 |
| 6 | $P^2HP^2H^2P^4H^2P^4H^2P^4H^2$ | 25 | 8 |

Table 1: benchmark sequence of PSP

| Seq No. | Max HH bond | Hyun-suk Yoon [6] | | | | Proposed algorithm | |
|---|---|---|---|---|---|---|---|
| | | IP1 | Max | IP2 | Max | time | Max |
| 1 | 4 | 92 min | 4 | >10hrs | 4 | 9 sec | 4 |
| 2 | 6 | >10 hrs | 6 | >10hrs | 6 | 23min | 6 |
| 3 | 7 | >10 hrs | 7 | >10hrs | 7 | 12hrs | 7 |
| 4 | 9 | N/A | N/A | N/A | N/A | 36sec | 9 |
| 5 | 9 | N/A | N/A | N/A | N/A | 50min | 9 |
| 6 | 8 | N/A | N/A | N/A | N/A | 123min | 8 |

Table 2: comparison

In table 2 shows the comparison of the results with our proposed methods. 'N/A' represent they are not considered this sequence with their experiment. The maximum H-H contacts obtained by [6] and proposed method are same, but time taken to evaluate the structure by other method is less than the previous approach. They did not mention what was the exact time taken to evaluate, just mentioned greater than 10 hrs. In this paper we give the exact time taken to evaluate structure of the corresponding sequence.

It is clearly shown that time taken by this algorithm is less than the algorithm defined in IP1 or IP2. On 28 length sequence the time taken shown is greater than 10hrs, but we took approx 12hrs which is also greater than 10hrs but on length 23 it also shown time taken is greater than 10hrs, so we can say that time taken for length 28 is more than 12hrs.

## 6. Conclusion and future work

In this work, protein structure prediction problem is formulated by Integer programming method and proposed algorithm with bounded conformational search space predicts the structure of a protein. Many symmetric structures are also been excluded from conformational search space. From the experimental results it has been shown that our proposed algorithm is very effective and efficient in PSP. We would like to apply proposed algorithm to cubic lattice by applying more parameters on this algorithm.

## 7. References


[1] Lau, K. and Dill, K. A., "*A lattice statistical mechanics model of the conformation and sequence spaces of proteins*" Macromolecules, vol. 22, pp. 3986–3997, 1989

[2] Charles J. Epstein, Robert F. Goldberger, and Christian B. Anfinsen. "*The genetic control of tertiary protein structure: Studies with model systems*" In Cold Spring Harbor Symposium on Quantitative Biology, pages 439–449, 1963. Vol. 28.

[3] William E. Hart and Alantha Newman. "*Protein Structure Prediction with Lattice Models*". 2001 by CRC Press.

[4] Kit Fun Lau and Ken A. Dill. "*A lattice statistical mechanics model of the conformation and sequence spaces of proteins*" Macromolecules, 22:3986–3997, 1989.

[5] V Chandru, A DattaSharma, and V S A Kumar. "*The algorithmics of folding proteins on lattices. Discrete Applied Mathematics*", 127(1):145–161, Apr 2003.

[6] Hyun-suk Yoon. "*Optimization Approaches to Protein Folding*",Phd. Thesis, School of Industrial and System Engineering,Georgia Institute of Technology, December 2006.

[7] Thang N. Bui and Gnanasekaran Sundarraj. "*An Efficient Genetic Algorithm for Predicting Protein Tertiary Structures in the 2D HP Model*", GECCO'05, June 25–29, 2005, Washington, DC, USA.

[8] Wayne L. Winston, "*Operation Research: Applications and Algorithms*" , Fourth Edition.