2006 IEEE Congress on Evolutionary Computation
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

A Guided Genetic Algorithm for Protein Folding Prediction Using 3D Hydrophobic-Hydrophilic Model

Md Tamjidul Hoque, Madhu Chetty and Laurence S. Dooley

***Abstract*—In this paper, a Guided Genetic Algorithm (GGA) has been presented for protein folding prediction (PFP) using 3D Hydrophobic-Hydrophilic (HP) model. Effective strategies have been formulated utilizing the core formation of the globular protein, which provides the guideline for the Genetic Algorithm (GA) while predicting protein folding. Building blocks containing *Hydrophobic* (H) – *Hydrophilic* (P or Polar) covalent bond are utilized such a way that it helps form a core that maximizes the fitness. A series of operators are developed including *Diagonal Move* and *Tilt Move* to assist in implementing the building blocks in three-dimensional space. The GGA outperformed Unger's GA in 3D HP model. The overall strategy incorporates a swing function that provides a mechanism to enable the GGA to test more potential solutions and also prevent it from developing a schema that may cause it to become trapped in local minima. Further, it helps the guidelines remain non-rigid. GGA provides improved and robust performance for PFP.**

I. Introduction

PROTEIN is the result of a three dimensional folding of a linear chain of amino acids. The chain called the *primary structure* is the sequential concatenation of amino acids taken from a set of 20 members only [1]. Two amino acids concatenate by releasing water and forming peptide bond. The folding of a protein called native state is unique for same sequence, which is generally the lowest free energy state. The protein folding prediction (PFP) is the problem of determining the native state of a protein from its primary structure. This prediction is of immense importance [2] because the 3D or tertiary structure determines the biological functions and its understanding is essential for drug designing [3].

The protein folding prediction (PFP) is a combinatorial optimization problem, which so far has evaded solution in most of the cases because of the astronomical number of potential solutions [4]. Systematic exhaustive search is infeasible especially for the long sequences. The complicated form of the energy function does not suggest any obvious search strategy. Most searches become trapped in one of the many local free energy minima characteristic of the energy landscape. In practice, X-ray Crystallography (XC) and Nuclear Magnetic Resonance (NMR) are used to determine the native conformation [5], but both the methods are time consuming and for some cases infeasible such as membrane protein.

The most successful approach in case of hard optimization problem like PFP so far, is based on hybrid evolutionary approach [4]. In this approach, the model does not deal with the full atomic description of the chain; only main chain atoms and sometimes an additional one-side chain atom representation are used. Moreover, a lattice model is used avoiding the continuous conformational space that simplifies many of the required calculations and enables some computation to be performed at the backend before the actual simulation begins. A simplified energy or fitness function is used. The HP model introduced by Dill [6] is such a model having these properties and mostly used. PFP in HP model has been proved to be NP-complete [7] [8]. Therefore deterministic approaches are not practical. Two ways of carrying out the search for low energy conformations are used [4]: enumeration for highly simplified cases and non-deterministic search technique such as Genetic Algorithm (GA), which is evolutionary in nature. Several other outstanding concepts such as a number of versions of Monte Carlo (MC), Evolutionary MC (EMC) [9] [10], Simulated Annealing (SA), and Tabu Search with GA (GTB) [11], Ant Colony Optimization [12] is mentionable. Statistical approaches such as Contact Interaction (CI) [13] and Chain Growth (CG) [14] have also been applied to PFP, however these techniques are all characterized by the fact that as the sequence length increases, and the accuracy reduces, except for enumeration [15] or exhaustive maneuvers such as in [16]. A prediction in 2D HP [17] helps to develop the strategies easily rather in 3D HP for obvious reasons. But the 3D extensions are equally important to make the prediction strategies mapping towards real PFP.

Our approach for PFP in 3D HP is using GA. GA reduces the need for highly accurate strategies, which would avoid requirement of redefining strategies separately for each individual sequence, i.e. a generic guideline does the purpose. Further, GA is driven by an implicit parallelism and generates significantly more successful descendants than random search. GA has been proved to outperform MC particularly for PFP in the HP model [18] [19]. Again PFP in HP model is one of the most challenging optimization problems that make the job of any search approach very difficult including GA. Our interest in this paper is to find out, how this problem pose difficulties for GA while

M. T. Hoque is with the Gippsland School of Information Technology, Monash University, Australia (corresponding author to provide phone: +61 3 5122 6778; fax: +61 3 9902 6842; e-mail: Tamjidul.Hoque@infotech.monash.edu.au).

M. Chetty is with the Gippsland School of Information Technology, Monash University, Australia (e-mail: madhu.chetty@infotech.monash.edu.au).

L. S. Dooley is with the Gippsland School of Information Technology, Monash University, Australia (e-mail: Laurence.Dooley@infotech.monash.edu.au).

0-7803-9487-9/06/$20.00/©2006 IEEE

searching optimum conformation and innovating new strategies to overcome them.

For a hard optimization problem like PFP, GA faces two type of difficulty, explicit and implicit to GA operations. As the search proceeds using simple GA, the similarity is found to be increasing among the population and diversity reduces. As a result, GA gets stuck to sub-optimal solution. On the other hand, by crossover or mutation operations GA may not produce valid conformation since the conformation needs to be a self-avoiding walk (SAW), which is implicit to the PFP. Further, as the optimum conformation is highly likely be a compact one, crossover and mutation become ineffectual by producing increasingly non-SAW conformation and the search rarely get any progress. Our strategies to overcome this difficulty are based on core formation concept. For the formation of proper core boundary additional constraint, which is generic in nature, has been combined with the existing fitness function. Their strategic combination helps predict the optimum conformation effectively and in a complementary manner.

The remainder of the paper is organized as follows. In Section II, the HP model and metaphorical view of the protein core has been described, while Section III provides a proof of the optimal shape of the H-Core in three dimensions. Section IV describes the fine set of sub-conformations used as building block for HP mixed layer. Section V discusses detail implementation of search procedure including result, while Section VI discusses the theoretical aspect of the overall approach using GGA. Finally, Section VII draws the conclusions.

II. The HP Lattice model

The HP model has been introduced by Dill [6] based on the observation that the hydrophobic forces are dominating the protein folding. In the model, amino acids are represented as a reduced set of ‘H’ (Hydrophobic or Non-Polar) and ‘P’ (Hydrophilic or Polar) only. The protein conformations of the sequence are placed as a self-avoiding walk (SAW) on a 2D square or 3D cube lattice. The energy of a given conformation is defined as a number of topological neighboring (TN) contacts between those Hs which are not sequential with respect to the sequence. The PFP is formally defined as a given amino-acid sequence, $s = s_1, s_2, s_s, \cdots, s_m$, ($m$ = total amino acids in the sequence) a conformation c needs to be formed where, $c^* \in C(s)$, energy $E^* = E(C) = \min\{E(c) \mid c \in C\}$ [12]. Here, $C(s)$ is the set of all valid (i.e. SAW) conformations of s. If the number of TNs in a conformation c is q then the value of $E(c)$ is defined as $E(c) = -q$. In a 2D HP model (Fig. 1, (a)) a non-terminal and a terminal residue both having 4 neighbour can have maximum of 2 TNs and 3 TNs respectively. In case of 3D, maximum possible neighbours are 6 in numbers and the maximum TNs are 4 and 5 respectively for a non-terminal and terminal residue of the sequence.

It is well known [20] that the Hs form the protein core freeing energy. The Ps, having affinity with the solvent tend to remain in the outer surface. This paper visualizes the folded protein through the 3D HP model as a three-layered kernel (Fig. 1(b)). The inner kernel, called the H-Core [20] [21], is compact and mainly formed of Hs while the outer kernel consists mostly of Ps. The H-Core Centre is called HCC (defined in Section III, B). The composite thin layer between the two kernels consists of those Hs that are covalent bonded with Ps, which for the purpose of this paper is referred to as the *HP mixed layer*.

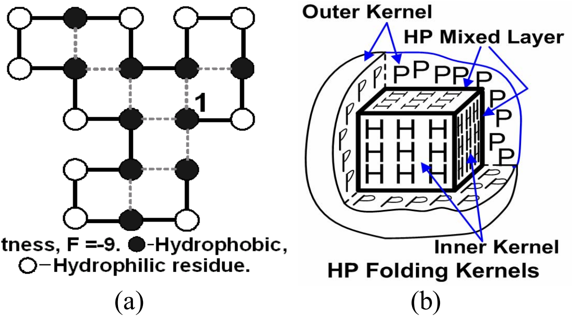


Fig. 1. (a) Conformation in a 2D HP Model shown by solid line. Dotted line indicates TN. (b) A 3D metaphoric HP folding kernels.

III. Optimum Shape of 3D H-Core

This section develops a proof for the optimum shape of the H-Core, under the assumption that the segment is a sequence of Hs only and it is a variation of the proof presented in [20]. In 3D HP Model, every H can have a maximum of 6 neighboring (Forward, Backward, Left, Right, Up and Down) residues, therefore H has 6 sides. The positioning of H inside the core (assuming a rectangular box) can be categorized based upon the number of its position within the core (Table I), such as H at corner, edge, plane and interior will respectively have 3, 4, 5 and 6 (sides) inside core. The objective is now to determine the shape of the H-Core that will maximize the total number of sides inside the core. A

TABLE I
Total H sides inside the core

Position	Count	H Sides	Total H Sides
Corner	*8*	*3*	*24*
Edge (1)	4 *(l-2)*	*4*	*16l-32*
Edge (2)	4*(w-2)*	*4*	*16w-32*
Edge (3)	4*(h-2)*	*4*	*16h-32*
Plane (1)	*2(l-2) (w-2)*	*5*	*10(l-2) (w-2)*
Plane (2)	*2(w-2) (h-2)*	*5*	*10(w-2) (h-2)*
Plane (3)	*2(l-2) (h-2)*	*5*	*10(l-2) (h-2)*
Interior	*(l-2)(w-2)(h-2)*	*6*	*6(l-2)(w-2)(h-2)*
Total inside bonding sides, *B*=			*6lw-2(lw+lh+wh)*

shape with fewer corners is preferable otherwise a number of H sides may be outside the core. Among all possible 3D shapes drawn in lattice model, rectangular box will have fewer corners, i.e. least loss of the H sides going out of the core. Let us assume the *length*, *width* and *height* of a rectangular box are l, w and h respectively. The total number of Hs inside the core equals the core volume (V_{box}), that is,

$$V_{\text{box}} = lwh \tag{1}$$

From Table I, we get

$$B = 6lw - 2(lw + lh + wh) \tag{2}$$

From [17], we find that in 2D the optimum core is a square, therefore,

$$l = w \tag{3}$$

Using (1) and (3) in (2), we get,

$$B = 6V_{\text{box}} - 2\left(l^2 + 2\frac{V_{\text{box}}}{l}\right) \tag{4}$$

To maximize B, (4) is differentiated with respect to l and with V_{box} = constant. Using $\frac{dB}{dl} = 0$, we obtain,

$$l = \sqrt[3]{V_{\text{box}}} = w \tag{5}$$

Equation (4) will maximize B since, $\frac{d^2B}{dl^2} \leq 0$. Using (5) in (1) we get,

$$h = \sqrt[3]{V_{\text{box}}} \tag{6}$$

Thus, a cubic volume can form the best core cavity. Now, let n_H be the number of total H in a sequence. The number of those H immediately covalent bonded with P and being on surface (Plane, Edge and Corner) in a cubic core can be expressed as, $\left(6n_H^{2/3} - 12n_H^{1/3} + 8\right)$. Therefore, the probability of H being on non-corner is,

$$\Pr_{\text{non-corner}} = \frac{\left(3\left\lfloor n_H^{2/3}\right\rfloor - 6\left\lfloor n_H^{1/3}\right\rfloor\right)}{\left(3\left\lfloor n_H^{2/3}\right\rfloor - 6\left\lfloor n_H^{1/3}\right\rfloor + 4\right)} \tag{7}$$

The probability of H being in the corner is,

$$\Pr_{\text{corner}} = \frac{4}{\left(3\left\lfloor n_H^{2/3}\right\rfloor - 6\left\lfloor n_H^{1/3}\right\rfloor + 4\right)} \tag{8}$$

These measures given above are referred in the following sections.

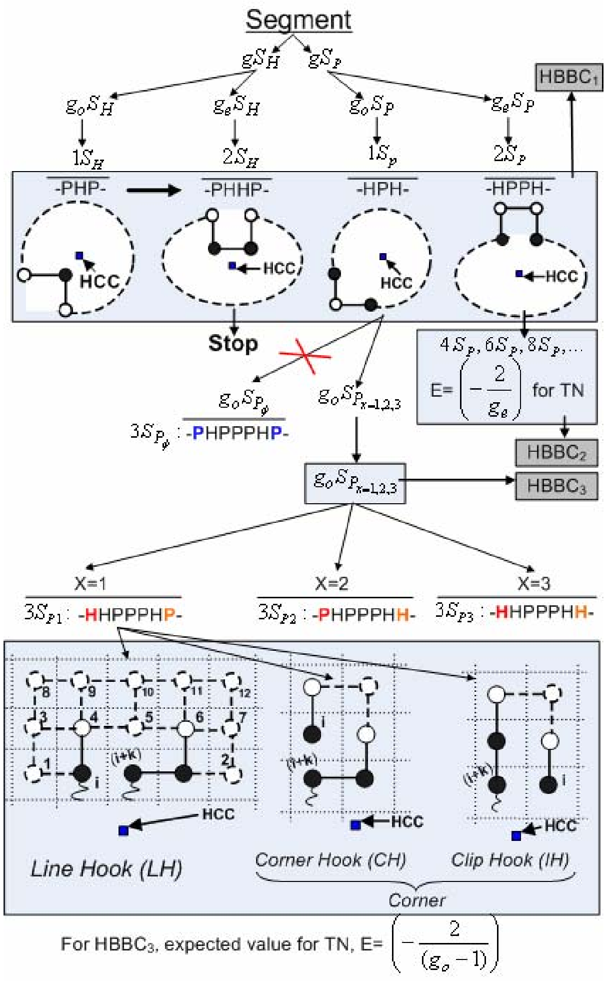


Fig. 2. Highly probable sub-conformations of corresponding sub-sequences for the HP mixed layer.

IV. Sub-Confirmations for HP- mixed Layer

To form the cavity, it is straightforward to think of placing the P of a *-HP-* segment on the opposite side of H with respect to the developing HCC, while searching for the desired conformation. With this placement, the cavity would tend to form a spherical shape, which is not the desired cubic one. To address these problems, sub-conformations that are highly probable corresponding to sub-sequences is defined (Fig. 2) and later applied. Two broad categories of sub-sequences are defined; gS_H and gS_P, where $g \in \mathrm{N}$ (N is natural number). These two categories completely cover the *HP mixed layer* including outer kernel. Let S_H and S_P represent segments of H and P respectively. A segment refers to a contiguous string of length g, so $3S_H$ for example means *-PHHHP-*, i.e. $g = 3$ with the two boundary residues being of the opposite type. g is divided into even

g_e and odd g_o numbers. For $g_o > 1$, the category g_oS_P is split into $g_oS_{P_\phi}$ and $g_oS_{P_x}$, where $x \in \{1, 2, 3\}$ which implies the run of P is bounded by an additional H at left ($x = 1$), right ($x = 2$) or both ($x = 3$) sides, while the former category, by ϕ indicates no additional H. For example, $3S_{P_3}$ means a sub-sequence *-HHPPPHH-*. Collectively, they will be called as H-Core Boundary Builder Segments (HBBS) and they are mapped to potential sub-conformations which are named as 'H-Core Boundary Builder sub-Conformation' (HBBC) in this paper. Conformations that are highly likely are chosen where either 'H' is put towards HCC and 'P' is away, or the two Hs contributing TN are encouraged with position HCC as well. According to the similarity of importance, the sub-conformations are grouped as HBBC$_1$, HBBC$_2$ and HBBC$_3$ as indicated in Figure 2. As in Figure 2, the expansion of $2S_H$ is stopped, otherwise it would involve the 'H' of the inner core or the H-core. No particular sub-conformation is defined for $g_0S_{P_\phi}$ since it is taken care of by the sub-sequence $1S_H$.

TABLE II
FORMATION OF PCF

HBBC	Value calculation of PCF	
	Reward	Penalty
HBBC$_1$	-1	1
HBBC$_2$	$\left(-\dfrac{2}{g_e}\right)$	$\left(\dfrac{2}{g_e}\right)$
HBBC$_3$	$\left(-\dfrac{2}{(g_o-1)}\right)$	$\left(\dfrac{2}{(g_o-1)}\right)$

A. *Formation of Probabilistic Constrained Fitness*

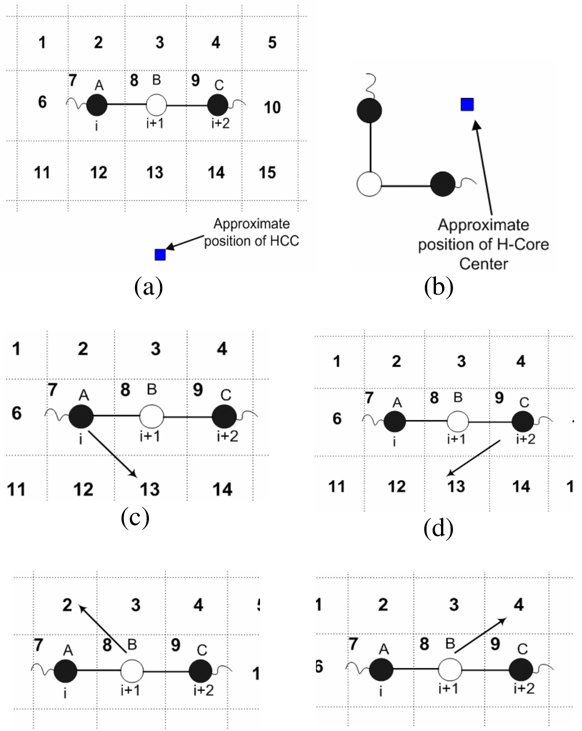
During searching for an optimum conformation if a sub-conformation corresponding to a particular sub-sequence exists for HP mixed layer exists in a developing conformation, it is rewarded, and otherwise it is penalized. This measure of this fitness is named *Probabilistic Constrained Fitness* (PCF) in this paper. If any member of HBBC$_1$ correspond to the related sub-sequence that PCF will be added a '-1' as reward, otherwise a '1'. Table II shows the details.

B. *Implementation of the Sub-Conformation*

Before going into the details of how the HBBCs are implemented, a number of terms need to be defined. (1) H-Core center (HCC): It is calculated as the arithmetic mean of the coordinates of all H. That is,

$$x_{HCC} = \frac{1}{n_H}\sum_{i=1}^{n_H} x_i \,,\, y_{HCC} = \frac{1}{n_H}\sum_{i=1}^{n_H} y_i \,,\, z_{HCC} = \frac{1}{n_H}\sum_{i=1}^{n_H} z_i \quad (9)$$

Before enforcing any sub-conformation, the HCC (x_{HCC}, y_{HCC}, z_{HCC}) is updated to place 'H' near toward HCC and 'P' as far from HCC as possible. (2) Diagonal Move: If three consequative residues are $\overline{ABC}$ and $AB \perp BC$, then diagonal move implies moving B to $\left(\overline{A} + \overline{C} - \overline{B}\right)$. (3) Pull Move: In this paper we used short Pull Moves as defined in [16]. If a sequence has total m residues and $\overline{ABC}$ are three consecutive points indexed $(i+1)$, i, $(i-1)$ respectively, then B is pulled to diaginally to a free point and then C need to be placed diagonal to A and make $AB \perp BC$. The pull propagates towards the first residue until a valid conformtion is reached. That is, $(i-2)^{th}$ residue will occupy the previous position C or $(i-1)^{th}$ residue, $(i-3)^{th}$ will occupy the previous position of $(i-2)^{th}$ residue, and so on towards the first residue. Short Pull Move imply that it pull stops as soon as a valid conformation is reached. The pull can be in either ways towards the first or the last residues. (4) Tilt Move: Two or more consecutive residues with precondition that all of them on a same line move straight together to adjacent


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
A
B
C
i
i+1
i+2
Approximate position of HCC
Approximate position of H-Core Center

(a) (b)

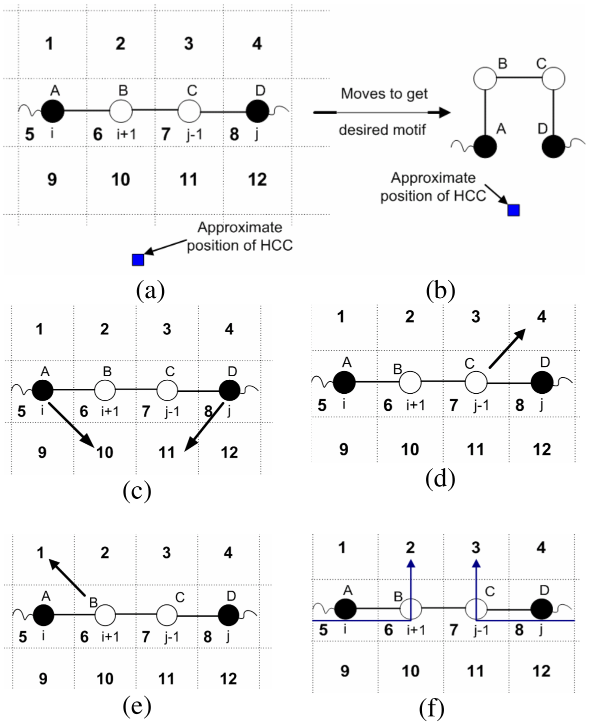
(c) (d)

(e) (f)

Fig. 3. (a) pre-condition and (b) desired post-condition. At (c), if location 13 is free and 12 is either free or already occupied by residue (i-1), then pull moving of A to 13 will get the desired result. At (d) C can be pull moved to 13 if 13 is free and 14 is either free or already occupied by residue (i+3). If (c) and (d) fails, at (e) B can be pull moved to 2 if 2 and 3 are free. In (f), if 3 and 4 are free then placing B to 4 by pull move is possible, and desired conformation is achieved.

next parallel line one lattice distance at a time if free points are available, and the pull propagates towards both ends upto the terminal residues.

These moves are used to implement the sub-

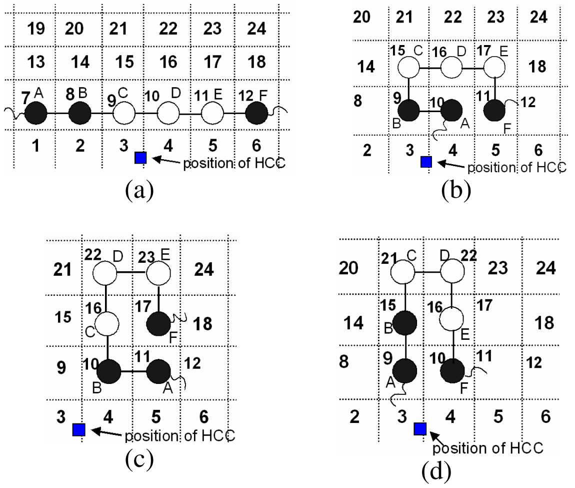


(a) (b) (c) (d) (e) (f)

Fig. 4. (a) precondition and (b) desired post-condition. At (c), it is possible to get the desired motif by pull moving A to 10 and D to 11 if possible. Otherwise at, (d) or (e) either B can be pull moved to 1 or C can be to 4. If 1 and 4 are not free and 2 and 3 are only free then as shown in (f), placing B and C by tilt move to 2 and 3 will bring desired motif.

conformations defined in the early section. During implementation less destruction to the other part of the conformation is desirable. Diagonal move is the least destructive. If the mapping of the sub-conformation has failed using diagonal move then pull move is tried otherwise the tilt move is applied. Tilt move destructs the stability of the other parts more with respect to diagonal or pull move, on the contrary, tilt move needs less number of move and it is easily implementable for complex sub-conformation mapping. Among the HBBCs, sub-conformations are specified only for HBBC$_1$ and HBBC$_3$. Mapping of some of the sub-conformation has been demonstrated from their precondition from Figure 3 to 5.

Note that, HBBSs that correspond to HBBC$_1$ are one to one mapping. But HBBSs that correspond to HBBC$_3$ are 1:3 or 1:6 mapping. In case of HBBC$_3$, probability measure is used to select them for implementation. HBBC$_3$ can be part of line (that is a part of plane or edge) and corners of a core, therefore the probabilities (using (7) and (8)) for constructing *Line Hook*, *Corner Hook* and *Clip Hook* (providing none of them already exist) are assigned as, $\Pr_{LH}=\Pr_{non-corner}$, $\Pr_{CH}=\dfrac{\Pr_{corner}}{2}$ and $\Pr_{IH}=\dfrac{\Pr_{corner}}{2}$ respectively when $x=1$ or $x=2$ in $g_o S_{P_x}$. For $x=3$, there will be a total of six variations, and the probabilities will be $\Pr_{LH_1}=\Pr_{LH_2}=\dfrac{\Pr_{LH}}{2}$, $\Pr_{CH_1}=\Pr_{CH_2}=\dfrac{\Pr_{CH}}{2}$ and $\Pr_{IH_1}=\Pr_{IH_2}=\dfrac{\Pr_{IH}}{2}$. If $g_o S_{P_x}$ corresponds to a particular member of HBBC$_3$ in a sequence and it is adopted by any developing conformation during a particular search, then it may be allowed to be replaced as there are more than one sub-conformations possible. The replacement probability in the subsequent search iterations with any of the remaining candidates is chosen by maximum probability proportion of the remaining probabilities set. Then, the candidate is selected for the replacement based on their probabilities proportion calculated early for the occurrence probabilities.



(a) (b) (c) (d)

Fig. 5. (a) pre-condition and (b), (c) and (d) desired motifs. The desired motifs can be achieved in a number of ways. Also assume a situation that 1 to 6 points are not free or congested. Now, as an example tilt move in (a) can be applied to place C, D, E to 15, 16 and 17. Then A needs to move to the free location at 10. To achieve (c) using tilt move, DE can be shifted twice, first (16, 17) then to (22, 23). A will need to move to 11, pull move will help. Similarly (f) can be achieved by tilt moving C, D twice and finally placed to (21, 22). No additional move required further.

V. Implementation of the Prediction

Though the added constraint (PCF) has been defined for a conformation, the ultimate goal is to maximize the fitness $|F|$. The search process is divided into two alternate phases. In phase one, *F* dominates over *PCF* and the core starts building. In the alternate phase (say phase two), *PCF* dominates over *F*. This phase take care the proper formation of the HP mixed layer. Further, the HBBCs implantation is done here since *PCF* would favor the change. The HBBCs implantation is done only if they are not found for the corresponding sub-sequences. This action may reduce the already achieved fitness *F*. But hopefully it would help reform proper cavity that would maximize the H bonding inside core. As the phases altered through search, the impact is such that F and PCF come up with common goal – which is highly likely the optimum one. The total or combined fitness is defined as

$$TF = \alpha(t) * F + \beta(t) * PCF \qquad (10)$$

where t is tth generation while search is carried out by GA. To alter the weight of α and β to dominate *F* and *PCF* over each other, a swing function (equation (12) is used.

$$\delta(t) = A(1 + \cos\omega_m t)\cos\omega_0 t \qquad (11)$$

where $\omega_m << \omega_0$, t = number of generation. The assignment of α and β is as,

$$\begin{array}{ll} \textit{if } \delta(t) > 0 \textit{ then} & \alpha(t) = \delta(t), \beta(t) = 1 \\ \textit{elseif } \delta(t) < 0 & \alpha(t) = 1, \beta(t) = -\delta(t) \\ \textit{else} & \alpha(t) = 1, \beta(t) = 1 \\ \textit{endif} & \end{array}$$

Algorithm-I: GGA for PFP using 3D HP

```
Input:   Sequence S, Target Fitness of the
         Sequence (Target_F)
Output:  3D Folding of the given sequence.
  COMPUTE PCF; COMPUTE A
   t=0        /* Generation count */
   F=0        /* Best fitness found from the search */
   Fillup the population with random (valid)
              conformation possible for S.
While F < > Target_F THEN
{  t = t + 1
   COMPUTE δ(t), α(t), β(t), TF
   Crossover
   Mutation
   IF δ(t) < 0 THEN
     { FOR i =1 to population_size DO
        Check chromosome_i for any miss-
          mapping of HBBC_1 or HBBC_3
         IF miss-mapping true then
   { Re-map the sub-sequence to corresp-
              onding HBBC using move-sets. }}
   COMPUTE TF
   IF no improvement of the best solution for long
      {Remove twins from population}
   Sort and Keep Elite.
   F ← Best fitness found from the population.
}
END.
```

A typical value set for $\delta(t)$ is, A=30, ω_m = 0.004 and ω_0 =0.05. The value of A (amplitude) is selected as, $2A \geq \max(|F|, |PCF|)$, where the upper limit *F* can be predicted by $F = -2 * \min\{E[Seq], O[Seq]\}$ [22]. $E[Seq]$ and $O[Seq]$ indicate the number of even and odd indexed H residues in the sequence. Note that, minimum value of $|\alpha(t)| = 1$ and $|\beta(t)| = 1$ are maintained, and never set to zero. This is to preserve the sub-conformation or schema developed in the alternate phase with good features.

The detail search procedure is given in Algorithm-I. A simple GA is used with population size of 200 was chosen for all sequences, the elite rate = 0.10, p_c = 0.85, p_m = 0.5 and a single point mutation by pivot rotation. The implementation of crossover and mutation is same as in [18] [19] but without any special treatment such as cooling. Selection procedure was based on roulette wheel.

TABLE III
PERFORMANCE COMPARISONS

SL	*E*	GGA	X	CI	CG	HZ
1	-32	-32 (3286)	-32	-32	-32	-31
2	-34	-34 (14288)	-34	-33	-34	-34
3	-34	-34 (4655)	-34	-32	-34	-31
4	-33	-33 (15898)	-33	-32	-33	-30
5	-32	-32 (7934)	-32	-32	-32	-30
6	-32	-32 (19208)	-32	-30	-32	-29
7	-32	-32 (21084)	-32	-30	-32	-29
8	-32	-32 (5053)	-32	-30	-32	-29
9	-34	-34 (9872)	-34	-32	-33	-31
10	-33	-33 (7246)	-33	-32	-33	-33

Performance comparison for 3D PFP [12]. X implies PERM, CHCC and ACO algorithm, E indicates the putative ground energy, and the format for GGA is: "maximum fitness achieved (minimum generation)". GGA results are from two iterations only. Results of other algorithms such as CI, CG and HZ are given.

TABLE IV
PERFORMANCE COMPARISONS WITH UNGER'S GA [19]

ID	NEW CONFORMATION	E_{UGA}	E_{GGA} (GEN)
643d.2	13232616155354164453251163646422 3131352626633555416224233154453	-29	-30 (14186)
643d.3	13223132623164154615426454136363 3553554641461323161424632546614	-35	-38 (493)
643d.4	13251425263636463154646425535241 4423232611414135413536162324552	-34	-36 (27450)
643d.6	13235264444236353635355246325426 1614252411355146641413354236323	-29	-30 (1899)
643d.9	13252624451364626132536154131462 3135232461641422635235553141525	-32	-34 (14848)
643d.10	13231615535514141623245441632423 3624146363515333622411114452424	-24	-25 (362)

Performance comparison between Unger's GA and our approach (GGA). The sequence IDs are same as in [19]. E_{UGA} is the achieved lowest energy in [19] and E_{GGA} (Gen) implies achievement in GGA and Gen implies generation at which the E_{GGA} is achieved.

A. Results

Simulations were carried out for bench mark 3D problems [23] (see appendix) with the target set equal to putative ground energy. Results are give in Table III. The comparisons given in [12], are time based on program running time. This may not be appropriate comparing non-deterministic search with exhaustive or enumerating or statistical approaches. For GGA, we involved visulation and saved the full trace including several analysis. Non-deterministic approach (such as GGA) employ general technique whereas other special approaches employ special technique based on sequences which may not perform uniformly in every cases and hard to generalised.

Further, to compare the performane between similar Algorithm (Unger's GA[19]), simulations were also carried out with some of the large sequences with one iteration having target set to 2 more of the achieved results in [19]. In all the cases, we found better result given in Table-IV. New conformations are presented using our *Self Direction Dependent* Coding (SDD-Coding) scheme which produces same encoded output for same conformation (i.e. non-isomorphic) whereas the traditional coding scheme [12][14] does not always produce same coding for same conformation (i.e. isomorphic). In this SDD-Coding, the direction of first points towards the second is marked '1' and the reverse is '2'. Direction of first 90 degree (in any co-ordinate based direction) turn is coded '3' and the reverse is '4'. Then, applying right-hand rule from '3' to '1' , the thumb direction is coded as '5' and the reverse is '6'.

VI. DISCUSSION

In this section, we discuss the theoretical aspect of the approaches presented in this paper. We begin with simple GA and analysed the reasons for its possible failure to predict an optimum conformation. Two major drawbacks were observed. GA computation is based on schema theorem. And schema theorem for PFP states [4] that short, flexible schemata with above average performance will receive exponentially increasing survival chance in the subsequent generations while those schemata with below-average performance will decay exponentially. Therefore one obstacle using simple GA's is that similarity within population grows quickly which leads to a suboptimal results and gets stuck. As the similarity grows the crossover will most likely happen between similar chromosome and off-spring will inherit the same properties. Therefore crossover becomes ineffectual and the same is true for mutation. Since, in the midst of similar chromosome even if mutation produces a higher fitted chromosome, it will be lost due to the selection procedure. To address this first problem we apply elitism (Algorithm-I) to ensure that solutions having higher fitness (*F*) are not lost and removed twins from population with similarity from 100 to 80% specially if we observe that for a long time there is no progress from the on going search.

The second obstacle is the operation of crossover and mutation may not be effective unless the output is a SAW conformation and follows the sequential change through the lattice points. Now, as the optimum conformation is relatively compact, crossover and mutation confront more collision or produce invalid conformation increasingly as the search is going on. Our specific implantation procedure of HBBCs moves the compact conformation without collision and the introduced move operators are less destructive to the already gained fitness. The move creates probable reformation of the H-core cavity to maximize the H-sides inside the H-Core. Hence, this approach makes change of the non-progressive situation in such a way that it can enhance the chances of gaining higher fitted conformations.

The energy landscape of the protein folding, even using HP model is very critical [24]-[28]. Due to this, while searching for putative ground fitness, the on going progress in achieving better fitness becomes increasing difficult. This implies that a converging search progress becomes extremely slow and often gets stuck (Table III and Table IV) before reaching to the putative ground energy since the effect of increasingly compact conformation. The lack of effective move operator with associated intelligence to switch from near optimum to optimum, was absent in most of the methods applied early. Therefore, strategy like GGA helps achieve improvement close to putative ground energy even for single fitness, is of great importance.

VII. CONCLUSION

In this paper, the novel strategies using GGA reported earlier for the 2D HP model have been extended for 3D HP model. Using the proposed algorithm, we reached the targeted putative ground benchmark conformation. Also, we achieved optimum conformation compared to the early implementation of GA for 3D HP model by Unger and Moult [19]. Our strategies use new operators associated with domain knowledge, where it maps the sub-sequences of HP-mixed layer into highly probable sub-conformations. The mapping is finite, short and generic and free from the scaling effect of the complexity of the sequences length. Additional constraint associated with the overall implementation through the swing function offers a way of testing more potential conformations while the conformation becomes compact rather than getting stuck. With the implantation of HBBCs, formation of line (as part of an edge or plane) and corner has been considered for the optimum cubic core in 3D HP model. The granularity level of HBBC's forming part of line to compute the probability of it being a part of either an edge or a plane can be extended for future work which would further speedup the search of an optimum conformation. Also, an appropriate parameter prediction of the swing function and implementation would improve it further. However, the overall approaches are robust enough to remove the causes of failures of GA and can be extended for application on a real Protein Folding Prediction. However, designing of an efficient operator would be an important issue for this challenging and hard optimization problem.

APPENDIX

Benchmark [23] 3D sequences

SL.1: $HPH_2P_2H_4PH_3P_2H_2P_2HPH_3PHPH_2P_2H_2P_3HP_8H_2$
SL.2: $H_4PH_2PH_5P_2HP_2H_2P_2HP_6HP_2HP_3HP_2H_2P_2H_3PH$
SL.3: $PHPH_2PH_6P_2HPHP_2HPH_2(PH)_2P_3H(P_2H_2)_2P_2HPHP_2HP$
SL.4: $PHPH_2P_2HPH_3P_2H_2PH_2P_3H_5P_2HPH_2(PH)_2P_4HP_2(HP)_2$
SL.5: $P_2HP_3HPH_4P_2H_4PH_2PH_3P_2(HP)_2HP_2HP_6H_2PH_2PH$
SL.6: $H_3P_3H_2PH(PH_2)_3PHP_7HPHP_2HP_3HP_2H_6PH$
SL.7: $PHP_4HPH_3PHPH_4PH_2PH_2P_3HPHP_3H_3(P_2H_2)_2P_3H$
SL.8: $PH_2PH_3PH_4P_2H_3P_6HPH_2P_2H_2PHP_3H_2(PH)_2PH_2P_3$
SL.9: $(PH)_2P_4(HP)_2HP_2HPH_6P_2H_3PHP_2HPH_2P_2HPH_3P_4H$
SL.10: $PH_2P_6H_2P_3H_3PHP_2HPH_2(P_2H)_2P_2H_2P_2H_7P_2H_2$

REFERENCES

[1] Allen, et al., “Blue Gene: A vision for protein science using a petaflop supercomputer”, IBM System Journal, 2001,Vol-40, No 2.

[2] J. Pietzsch,, “The importance of protein folding”, http://www.nature.com/horizon/proteinfolding/background/importance.html.

[3] S. Petit-Zeman, “Treating protein folding diseases”, http://www.nature.com/horizon/proteinfolding/background/treating.html.

[4] R. Unger, and J. Moult, “On the Applicability of Genetic Algorithms to Protein Folding”, IEEE, 1993, pp. 715-725.

[5] G.B. Fogel and D. W. Corne (Editors), “Evolutionary Computation in Bioinformatics”, Elsevier Science, 2004, USA.

[6] K. A. Dill, “Theory for the Folding and Stability of Globular Proteins”, Biochemistry, 1985, Vol. 24, No. 6, pp.1501-1509.

[7] P. Crescenzi and et al.,“On the complexity of protein folding (extended abstract)”,ACM, Proceedings of the second annual international conference on Computational molecular biology, 1998, Pp.597-603.

[8] B. Berger, and T. Leighton, “Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-Complete”, Journal of Computational Biology, 1998, Spring; Vol. 5, No. 1, pp.27-40.

[9] U. Bastolla, and et al.,(1998) “Testing a new Monte Carlo Algorithm for Protein Folding”, National Center for Biotechnology Information, 1998, Vol. 32, No. 1, pp.52-66.

[10] F. Liang, and W.H. Wong, “Evolutionary Monte Carlo for protein folding simulations”, J. Chem. Phys., 2001, Vol. 115, No. 7.

[11] T. Jiang, and et al., “Protein folding simulation of the hydrophobic-hydrophilic model by computing tabu search with genetic algorithms”, ISMB, 2003, Brisbane Australia.

[12] A. Shmygelska and H.H. Hoos, “An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem”, BMC Bioinformatics 2005, 6:30.

[13] L. Toma and S. Toma, “Contact interactions methods: A new Algorithm for Protein Folding Simulations”, Protein Science, 1996, Vol. 5, No. 1, pp.147-153.

[14] E. Bornberg-Bauer,“Chain Growth Algorithms for HP-Type Lattice Proteins”, RECOMB, 1997, Santa Fe, NM, USA.

[15] K.Yue and K.A. Dill, "Forces of Tertiary Structural Organization in Globular Proteins", Proc Natl Acad Sci USA 1995, Vol-92, pp.146-150.

[16] N. Lesh, M. Mitzenmacher and S. Whitesides, “A Complete and Effective Move Set for Simplified Protein Folding”, RECOMB, 2003, Berlin.

[17] M. T. Hoque, M. Chetty and L. S. Dooley, “A New Guided Genetic Algorithm for 2D Hydrophobic-Hydrophilic Model to Predict Protein Folding”, 2005 IEEE Congress on Evolutionary Computation (CEC), pp. 259-266, Edinburgh.

[18] R. Unger and J. Moult, “Genetic Algorithms for Protein Folding Simulations”, Journal of Molecular Biology, 1993, Vol-231, pp. 75-81.

[19] R. Unger and J. Moult, “Genetic Algorithm for 3D Protein Folding Simulations”, 5th International Conference on Genetic Algorithms, 1993, pp. 581-588.

[20] K. Yue and K. A. Dill, “Sequence-Structure relationships in proteins and copolymers”, Physical Review E, 1993, Vol-48, No.3, pp. 2267-2278.

[21] K. M. Flebig and K. A. Dill, “Protein Core Assembly Processes”, J. Chem. Phys., 1993,Vol-98, No. 4, pp. 3475-3487.

[22] A. Newman, “A new algorithm for protein folding in the HP model”, Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete Algorithms, 2002.

[23] W. Hart, and S. Istrail, “HP Benchmarks”, http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html

[24] K. A. Dill et al., “Principles of protein folding – A perspective from simple exact models”, Protein Science, 1995, Vol-4, pp: 561-602.

[25] A. Runner and E. Bornberg-Bauer, “EXPLORING THE FITNESS LANDSCAPE OF LATTICE PROTEINS”, [On-line], http://helix-web.stanford.edu/psb97/renner.pdf

[26] S. D. Flores and J. Smith, “Study of Fitness Landscapes for the HP model of Protein Structure Prediction”, Accepted for the Congress of Evolutionary Computation 2003, Australia.

[27] N. Mousseau and G. T. Barkema, “Exploring High-Dimensional Energy Landscape”, Computing in Science & Engineering, IEEE, 1999, Vol. 1, Issue 2, pp. 74-80, 82.

[28] N. E.G. Buchler and R. A. Goldstein, “Effect of Alphabet Size and Foldability Requirements on Protein Structure Designability”, Proteins: Structure, Function, and Genetics, Vol. 34, Issue 1 , pp. 113 - 124, Wiley-Liss, Inc, 1999.