A New Guided Genetic Algorithm for 2D Hydrophobic-Hydrophilic Model to Predict Protein Folding

Md. Tamjidul Hoque, Madhu Chetty and Laurence S Dooley
Gippsland School of Computing and Information Technology
Monash University, Churchill VIC 3842, Australia
{Tamjidul.Hoque, Madhu.Chetty, Laurence.Dooley }@infotech.monash.edu.au

Abstract- This paper presents a novel Guided Genetic Algorithm (GGA) for protein folding prediction (PFP) in 2D Hydrophobic-Hydrophilic (HP) by exploring the protein core formation concept. A proof of the shape for an optimal core is provided and a set of highly probable sub-conformations are defined which help to establish the guidelines to form the core boundary. A series of new operators including Diagonal Move and Tilt Move are defined to assist in implementing the guidelines. The underlying reasons for the failure in the folding prediction of relatively long sequences using Unger's Genetic Algorithm (GA) in 2D HP model are analysed and the new GGA is shown to overcome these limitations. 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. While the guidelines do not force particular conformations, the result is a number of conformations for particular putative ground energy and superior prediction accuracy, endorsing the improved performance compared with other well established nondeterministic search approaches.

1 Introduction

Protein Folding Prediction (PFP) involves the folding of a linear chain of amino acids into a three dimensional (3D) structure. To achieve this, the 2D Hydrophobic-Hydrophilic (HP) model (Dill 1985) can be applied. In this model, copolymer chains of H (hydrophobic) and P (polar or hydrophilic) monomers are configured as self-avoiding walks favouring HH interaction in two dimensions. HH interactions can be divided into two types:- 1) connected H or covalent bonded, which is represented by a solid line in Figure 1(a); 2) non-connected H but separated by a unit lattice distance from each other, namely a *Topological Neighbour* (TN) shown by a dotted line in Figure 1(a). As the conformation with dominant HH interaction is desirable (Yue *et al*, 1993) and given the covalent bond interaction is always fixed, only the TNs need to be computed in order to measure the fitness function F. Each TN count is an indication of *free energy* and is notionally assigned the value -1, so the best conformation therefore has the highest number of TNs and F has the maximum negative value.

It is well known (Dill 1985) 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 visualises the folded protein as a two-layered kernel shown in Figure 1(b). The inner kernel, called the H-Core (Yue *et al* 1993, Flebig *et al* 1993), 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 2). 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*.

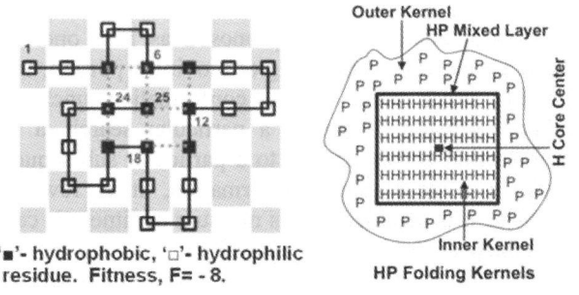


Figure 1: (a) A typical 2D HP model, showing a sequence of amino acids connected by solid line. (b) Metaphoric HP Folding Kernels.

In the recent past, for PFP, various non-deterministic search procedures incorporating several outstanding concepts have been applied. Unger's GA set the frame work for PFP (Unger 1993a and 1993b) and the sub-sequent variation comes with the other non-deterministic search approaches, such as a number of versions of Monte Carlo (MC), Evolutionary MC (EMC) (Bastolla *et al* 1998; Liang 2001 *et al*), Simulated Annealing (SA), and Tabu Search with GA (GTB) (Jiang 2003). PFP based on HP model proved to be an NP-Complete problem (Berger 1998), that makes the prediction more challenging. Statistical approaches such as Contact Interaction (CI) (Toma 1996) and Chain Growth (CG) (Bornberg 1997) have also been applied to PFP, however these techniques are all characterised by the fact that as the sequence length increases, the accuracy reduces (see Table 2), except for exhaustive manoeuvres such as in (Lesh 2003).

This paper, investigates why the Unger's GA becomes trapped in local minima for longer sequences, with one

0-7803-9363-5/05/$20.00 ©2005 IEEE.

reasons being that the *HP mixed layer* moves inside the H-Core or sometimes the part of the boundary needs to be shifted in order to improve the fitness function F. As the search converges, the crossover operation becomes ineffective (Liang 2001), leading to the generation of invalid conformation, that is a *non self-avoiding walk* or *collision*. One of the solutions would be to design a new and effective operator. The mutation operation also, at the same time, becomes ineffectual with collisions. To solve this, one way is to predict a complex set of sequences of mutations that would be required to reach the optimum conformation from a sub-optimal conformation, which is difficult to generalize and hard to predict for an unknown goal.

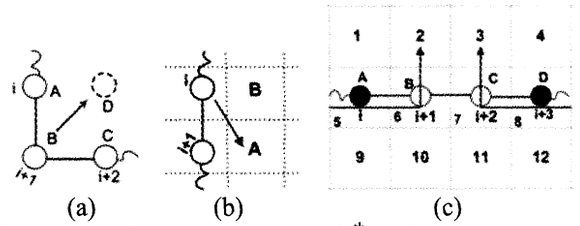
This paper proposes core formation concept as a guideline for GA. Short sub-sequence is mapped to very likely sub-conformation so building up the *HP mixed layer* using GA. The defined mapping sets are finite and in order to implement the guidelines, GA is preferred, because it is better able to handle imprecise specifications compared with other non-deterministic approaches. The fundamental basis of a sub-conformation is that the H needs to be placed as near as possible to the HCC, while P needs be placed far from HCC. If the *HP mixed layer* forms a proper cavity, it can guide the Hs to form the inner core, hence maximising the fitness. To map sub-sequences to sub-conformations, a number of *move sets* (defined in Section 2) are proposed as new operators, which are capable of mapping or moving a portion, while retaining other parts of the conformation unchanged without collision. If during a particular search, a sub-conformation corresponding to a particular sub-sequence exists in a developing conformation, it is rewarded, otherwise it is penalized. This measure of fitness is called *Probabilistic Constrained Fitness* (PCF) and the total fitness TF is thus defined as a weighted combination of PCF and F. A novel *swing* function is then used, which applies alternative positive and negative phases for varying the weights of both PCF and F, thereby guiding the GA to explore potential conformations by oscillating the search for the best F to avoid becoming trapped in local minima.

Now, the remainder of the paper is organized as follows. In Section 2, the move sets and the HCC are defined, while Section 3 provides a proof of the optimal shape of the H-Core. Section 4 describes the fine set of possible sub-conformations for the *HP mixed layer* and the construction of the PCF. Section 5 discusses the theoretical aspects of using a GGA for PFP, while Section 6 shows the GGA with a detailed implementation and results. Finally, Section 7 draws some conclusions.

2 Definitions

This sections defines both the move sets and the HCC. In particular new operators are defined.

(1) *Diagonal Move*: For any residue, if the covalent bonded two neighbours are diagonally positioned with respect to each other, then by diagonal move the said residue can be moved to its diagonal position provided the position is not occupied. (2) *Pull Move*: Pull moves (Lesh 2003) are diagonal moves, where at least two residues are moved. In Figure 2(b), prior to pull move, if $(i-1)^{th}$ residue is already at B, then the pull move would be a diagonal move. Pulling can occur in either direction towards the first or last residue. (3) *Tilt Move*: Two consecutive residues move together to two free locations, unit lattice distance apart, with the connecting line of those residues being parallel to previous positions. The pull for this move progresses to both ends by positioning $(i-1)^{th}$ residue to occupy the previous position of i^{th} location until $(i=1)$ and place $(i+2)$ to $(i+1)$ until $\{i=(last\#-1)\}$, thus shifting all the residues. As the example in Figure 2(c) shows, *Tilt Move* places $(i+1)^{th}$ residue from location 6 to 2 and from i^{th} to connected all the residues upto 1 are shifted to its previous residue's position. Similarly, the $(i+2)^{th}$ residue moves from location 7 to 3 and pulls the rest of the chain until last residue is reached. *Diagonal Move* is less destructive in the sense that it only moves one residue. *Pull Move* operates on at least two or more residues and stops as soon as a valid conformation is achieved. Although *Tilt Move* moves all the residues, it is very effective in a congested situation where *Pull Move* or *Diagonal Move* does not fulfil the pre-conditions to move (Hoque 2005). *Tilt Move* is used as a last preference to apply while *Diagonal Move* is the first preference.



(a) (b) (c)

Figure 2: (a) By diagonal move $(i+1)^{th}$ residue can be moved from B to D. (b) Before the Pull Move A and B need to be freed. (c) Effects of Tilt Move are indicated by arrows.

(4) *HCC Coordinate*: These are calculated as the arithmetic mean of the coordinates of all H. That is,

$$x_{HCC} = \frac{1}{n_H}\sum_{i=1}^{n_H} x_i \text{ and } y_{HCC} = \frac{1}{n_H}\sum_{i=1}^{n_H} y_i \quad (1)$$

Before enforcing any sub-conformation using the aforementioned moves, the HCC (x_{HCC} , y_{HCC}) is updated, the detailed procedures for which, are explained in (Hoque 2005).

3 Shape of 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 is a variation of the proof presented in (Yue *et al* 1993). In 2D HP Model, every H has four sides and therefore has a maximum of four

possible neighbours. Even if an H is positioned inside the core, all its neighbours (sides) may or may not be within the core depending on its position.

The positioning of H inside the core is categorized based upon the number of neighbours covered within the core as shown in Figure 3. The positioning of H at corner, edge and interior will respectively have 2, 3 and 4 H neighbours (sides) inside core. The objective is now to determine the shape of the H-Core that will maximize the total number of sides within the core. Only circular, triangular and rectangular shapes are considered, with all other possibilities rejected because they increase the number of corners. A shape with fewer corners is preferable otherwise a number of H sides may be outside the core. Creating a shape in a 2D HP model can be compared with raster scanning (see Figure 4) using a digital differential analyser or Bresenham's Algorithm (Xiang 2001) to verify these facts.

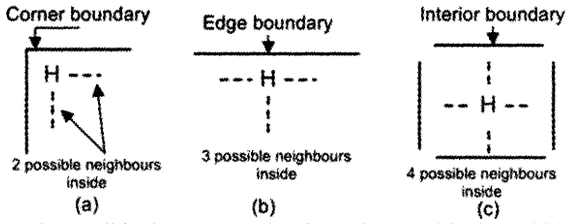


Figure 3: Possible three categories depend on positioning within the H-Core.

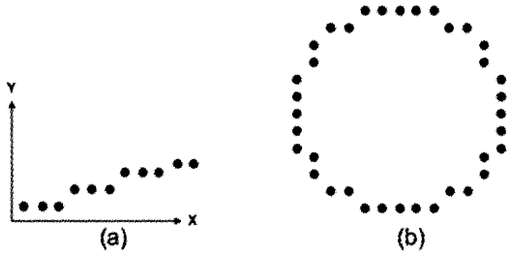


Figure 4: Scan-converting a (a) line (b) circle.

For a circle, the number of corners increases with radius, while even for a right-angled triangle with one arm horizontal and other vertical, the third arm does not, so beyond three corners; the third arm has an increasing number of corners with increasing length. However, a rectangle always has four corners, so both the circle and triangle cannot be chosen, leaving the rectangle as the best choice. The properties of the rectangle-shaped core (Table 1) are now investigated. For length l and width w, the total number of Hs inside the core equals the Core Area ($A_{\text{rectangle}}$), that is,

$$A_{\text{rectangle}} = lw \tag{2}$$

Table 1: Total H sides inside the core

Position	Count	H Sides	Total H Sides
Corner	*4*	*2*	*8*
Edge (1): Length	*2 (l-2)*	*3*	*6l-12*
Edge (2): Width	*2(w-2)*	*3*	*6w-12*
Interior	*(l-2)(w-2)*	*4*	*4lw-8(l+w)+16*
Total inside core sides or bonding, $B=$			*4lw-2(l+w)*

$$B = 4lw - 2(l + w) \tag{3}$$

Substituting (2) in (3), gives,

$$B = 4A_{\text{rectangle}} - 2\left(l + \frac{A_{\text{rectangle}}}{l}\right) \tag{4}$$

To maximize B, (4) is differentiated with respect to l and with $A_{\text{rectangle}}$ = constant. Using $\frac{dB}{dl} = 0$, it is obtained,

$$l = \sqrt{A_{\text{rectangle}}} \tag{5}$$

Equation (5) will maximize B since, $\frac{d^2B}{dl^2} \le 0$.

Substituting (2) in (5), $w = \sqrt{A_{\text{rectangle}}}$ (6)

Thus, a square area is the best core cavity. Since, the square root as a real number is not applicable in HP model, hence it is referred to as a *maximal rectangle*. In other words, the core will tend to be a maximal rectangular in shape. Again, let, n_H be the number of total H in a sequence. The probability of those H immediately covalent bonded with P and being on edge can be expressed as, $\Pr_{\text{edge}} = 1 - \frac{1}{\lfloor\sqrt{n_H}\rfloor - 1}$ (7)

The probability of H being in the corner is,

$$\Pr_{\text{corner}} = \frac{1}{\lfloor\sqrt{n_H}\rfloor - 1} \tag{8}$$

These measures given above are referred in the following sections.

4 Possible Sub-Conformation for HP Mixed Layer

It is straightforward to form the cavity by placing the P of a *-HP-* segment in Figure 5(a) on the opposite side of H with respect to the developing HCC, while searching for the desired conformation.

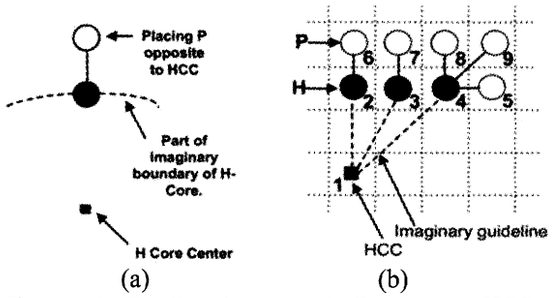


Figure 5: (a) Possible placement of a P opposite to HCC with respect to the covalent bonded H. (b) If HCC is at location 1, placing P at 6 will be exactly opposite to H, provided H is at 2. Similarly, location 7 is approximately opposite with respect to H having position at 3. If H is at 4, placing P at 9 is invalid. And placing P at 5 or 8 is equally likely.

With this placement, the cavity will tend to form a circular shape, which is undesirable. Furthermore, the

placement is not easy because the locations in lattice model are discrete (Figure 5(b)). To address these problems, 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-*. To summarise, the total categories are g_eS_H, g_oS_H, g_eS_P, g_oS_P ($g_o = 1$ only), $g_oS_{P_\phi}$ ($g_o > 1$) and $g_oS_{P_x}$. Collectively, they will be called as H-Core Boundary Builder Segments (HBBS) and they are mapped to potential sub-conformations. These potential sub-conformations are named as 'H-Core Boundary Builder sub-Conformation' (HBBC).

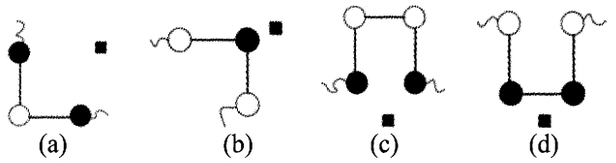


Figure 6: Potential sub-conformation for (a) $1S_p$ (b) $1S_H$ (c) $2S_p$ (d) $2S_H$. ●, ○ and ■ indicate an H, a P and approximate position of HCC respectively.

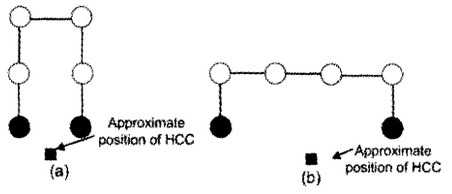


Figure 7: Among others, 2 variations of $4S_P$ are shown in (a) and (b).

For $1S_p$, $1S_H$, $2S_P$ and $2S_H$, there are few possible sub-conformations, so only highly potential sub-conformations (shown in Figure 6) are chosen, based on embedded TN and core formation concepts are chosen. These four HBBCs are grouped as HBBC_1.

HBBC_1 forms part of corner and line of the H-Core boundary. Each member of HBBC_1 will contribute '-1' to PCF. For g_oS_H and g_eS_H, only g_o=1 and g_e= 2 are considered because others variations are mostly the inner part of the H-Core. For $g_oS_{P_\phi}$ ($g_o > 1$), adjacent residue of the boundaries are either null (terminal) or P. In case of P, the boundary will be taken care of by g_oS_P ($g_o = 1$ only), which will give equivalent desired result.

For g_eS_P, where g_e>2, more than one sub-conformations possible with Figure 7 showing two examples. Therefore, it is not possible to conclude about any particular sub-conformation. However, for this group, the Hs at the boundary have a chance to form TN. The chance decreases with increasing values of g_e. Let this group be referred as HBBC_2. Further, to encourage a TN for a member of this group, a step value $\left(-\frac{2}{g_e}\right)$ is assigned to PCF.

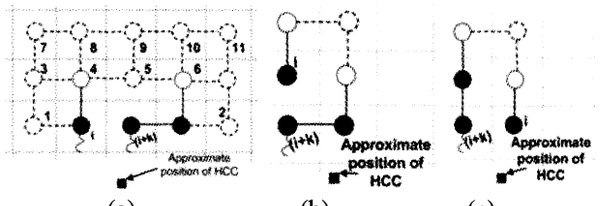


Figure 8: (a) Formation of line hook having the base (Hs) on a line with g_o Ps in between Hs. Fixing P's at '4' and '6' for k=5 (i.e. number of P =3), the other P will be positioned at '5'. For k=7 (i.e. number of P=5), the remaining 3 P's can be positioned at any valid manner from 1 to 11, similarly k=9, 11, etc can be considered. Possible corner with 3 Ps, (b) Corner hook and (b) Clip hook. g = 5, 7, ... can be extended in any manner maintaining the relative position of Hs.

$g_oS_{P_x}$ now corresponds to a number of important sub-conformations. With an additional H, the boundary Hs form three different important sub-conformations with $x = 1$ or 2. With $x = 3$, there will be total six sub-conformations, which will be referred to as *Hook Patterns*. The three different specific sub-conformations of *Hook Pattern* will be referred as *Line Hook*, *Corner Hook* and *Clip Hook*. In *Line Hook* (Figure 8(a)) the relative position of the 3 Hs form part a of line while both the *Corner Hook* (Figure 8(b)) and *Clip Hook* (Figure 8(c)) can form a corner of the H-Core boundary. These three sub-conformations have an embedded TN each. The chance of having TN decreases with increasing values of g_o. So, to encourage a TN the step value $\left(-\frac{2}{(g_o - 1)}\right)$ is assigned to PCF and this group is referred as HBBC_3.

As these sub-conformations are less likely to be created automatically, the corresponding sub-sequences (if they exist) are forced to map to HBBC_3. HBBC_3 can be part of line 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, $\Pr_{LH} = 1 - \frac{1}{\lfloor \sqrt{n_H} \rfloor - 1}$, $\Pr_{CH} = \frac{1}{2(\lfloor \sqrt{n_H} \rfloor - 1)}$ and $\Pr_{IH} = \frac{1}{2(\lfloor \sqrt{n_H} \rfloor - 1)}$ 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 $\Pr_{LH_1} = \Pr_{LH_2} = \frac{\Pr_{LH}}{2}$, $\Pr_{CH_1} = \Pr_{CH_2} = \frac{\Pr_{CH}}{2}$ and $\Pr_{IH_1} = \Pr_{IH_2} = \frac{\Pr_{IH}}{2}$. 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 of the remaining probabilities set. Then, the candidate is selected for the replacement based on their probabilities proportion calculated early for the occurrence probabilities among the *Hook Patterns*.

5 Search Using Genetic Algorithm (GA)

GA computation is based on schema theorem. Schema theorem for PFP states (Unger 1993b) 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. Let, G_t be the population with various conformations at generation number t, having fixed length (say, length = l) chromosome. In the context of fitness function, a schema S fits if each character in S fit. The order of schema $o(S)$ indicates the number of fixed position which is without the positions of wild card '*' and $\delta(S)$ indicates the length of the schema, i.e. the difference between the first and last fixed position. $m(S,t)$ is the number of conformation in G_t that fit the schema S, f_i is the fitness of the i[th] conformation and $\bar{f}$ is the average fitness of the whole population. $f(S)$ is the averaged fitness of the conformations that match the schema S. p_m indicates the individual independent mutation probability of each position of the conformation, p_c is the crossover probability, according to schema theorem it can be written as,

$$m(S,t+1) = m(S,t)\left(\frac{f(S)}{\bar{f}} \times (1 - p_c \frac{\delta(S)}{l-1}) \times (1-p_m)^{o(S)}\right) \quad (9)$$

Now, as the search converges, the conformation becomes compact, a corollary of which is that the crossover and mutation success rates for higher fitness factors tend to zero. Hence, the equation (9) can be re-written as;

$$m(S,t+1) = m(S,t)\left(\frac{f(S)}{\bar{f}}\right) \quad (10)$$

Equation (10) signifies the stuck condition. The intention is therefore to ensure that before being trapped in local minima, the convergence is guided using the H-Core formation concepts. The protein conformation search can be viewed as a concatenation of favourable schemata or sub-structures or schema. A schema in this case can be presented as a string of {0, 1, 2,*}, where 0, 1, 2 may indicate the direction *Left*, *Right* and *Forward* positioning of the current H with respect to the previous two residues, and the wild card (*) signifying no particular goal may be assigned to P as the parsing of the schema through the fitness function does not reward the P bonding directly. So, fitness F does not care where P is positioned and it is assumed to be automatically taken care of (Unger 1993b). However, as the generation converges, the effectiveness of crossover and mutation (Pivot Move in Unger 1993a) is weakened in the case of PFP, as by increasing the compact folded structure, the failure of crossover operation also increases due to collision (Liang 2001). Further, without complex sequences of mutation, there will be often invalid conformations due to collisions within compact conformation, so that, during the search as the conformation convergences, there are fewer options and less potential in the population to replace the near optimal solution with optimal one.

To clarify this finding further, let it be assumed that for the sub-conformation *-HHHHP-*, a U-shape is favoured for the 4H string, but then what happens to the P? The portion *–HHP-* may be on the same line or even form a right angle. If a hypothetical H-Core edge crosses the *-HP-* covalent bond line as shown in Figure 9, then it is highly probable that position of P will not occupy a potential position inside the H-Core and thereby reduce the value of total |F|. Otherwise there is a chance that |F| will be reduced, leading to either local minima or near optimal solution.

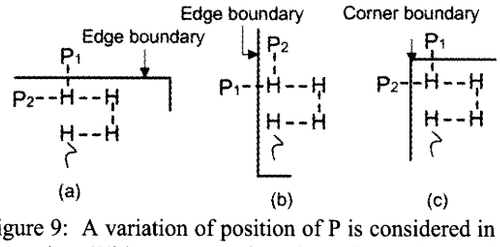


Figure 9: A variation of position of P is considered in a sub-conformation. With respect to the sub-conformation, position P_1 is highly probable instead of P_2 in instances (a) and (b), in (c) both are equally likely.

Thus, with the H-Core formation concept, the focus is on considering those Ps that are covalent bonded with H. Sub-conformation is enforced temporarily for a while to replace the wild card (*) with any of {0, 1, 2} that is highly likely for positioning P. Those Ps that are covalent bonded with Hs need to be placed in such a way that they either remain (approximately) on the opposite side of the H with respect to the developing HCC or be placed outside the H-Core. By using this approach there will be a greater chance that a part of the proper cavity will survive and eventually form the whole, so the resulting conformation will have a maximal |F|.

For any sub-sequences in the HP mixed layer, a PCF is now formulated. Though corresponding sub-conformations are highly likely, the occurrence can not be guaranteed. What is achievable is to use them as guidance towards an optimal conformation. To this end, a, Total Fitness (TF) function is formed as: TF= $\alpha(t)$ * F + $\beta(t)$ * PCF, where α and β are positive weightings, whose values are chosen by considering by two alternate phases as the generation passes, namely a positive phase and a negative phase, where t is the number of generation. In the former, α varies and $\alpha > \beta$ while in the latter β varies and $\alpha < \beta$. A sub-conformation is enforced when $\alpha < \beta$, if it is not already chosen and PCF dominates over F to adopt the change.

Consider at the positive phase (i.e. with respect to F), a favourable schema had fitness f_t (at time t). With the highly probable sub-conformation enforcement, those TNs that contradict the enforcement are broken (in the worst case) and get fitness f_{t+k} where $|f_{t+k}| < |f_t|$, and k is a positive constant. After a number of generations, $\alpha < \beta$ becomes $\alpha > \beta$ by steps and F predominates PCF and the fitness of the schema becomes f_{t+k+r}, where r is another positive constant. It is highly likely that $|f_{t+k+r}| > |f_{t+k}|$ and if the enforcement is adopted, then it is expected that $|f_{t+k+r}| > |f_t|$, with a success. Otherwise, the schema will be destroyed with exponential decay. In this way, all highly likely sub-conformations are randomly selected and adopted with eventually a proper cavity being formed which has a maximal |F|.

Hence if a sub-conformation is reinforced during the negative phase, it will break the contradictory TNs and help reform the conformation with fewer contradictory TNs. If the sub-conformation is inappropriate (which is not likely) its effect will disappear in the positive phase with the reinforcement of the building TNs. Otherwise, if it is not contradictory, it will exist (which is highly likely) and will help eliminate the possible stuck condition at the local minima. Also, the reinforcement of the sub-conformation at positive phase will help in getting out from the local minima. In practise, even in a positive phase, sub-conformations are enforced if convergence is relatively slow, to get out of any possible local minima.

6 Implementation of GA and Results

With respect to chromosome presentation of the population, crossover and mutation operations, our new GA is similar to Unger's implementation. Although no cooling scheme is employed, but rather the search has been extended by incorporating the proposed guidelines.

In section 4, the members of the three different types of HBBCs were defined and it was further elaborated on how their existence would help in forming the PCF. If sub-sequences corresponding to $HBBC_1$ exist, and are not being adopted, then $HBBC_1$ members are chosen randomly and enforced to map. If sub-sequence corresponding to the members of $HBBC_3$ exists in the sequence, then they are treated as one choice equivalent to single member of $HBBC_1$ for enforcement selection. For example, for the sequence length of 64 (Table 2), $PCF = -25$ and the combined fitness or the Total Fitness is,

$$TF = \alpha(t) * F + \beta(t) * PCF \tag{11}$$

A population size of 200 was chosen for all sequences, the elite rate = 0.10, p_c = 1.0, p_m = 0.9 and a single point mutation was assumed. After every five generations, all residue mutations are checked for higher fitness per chromosome. Selection procedure was based on roulette wheel. For changing weights of α and β alternatively, the following oscillating function was used,

$$\delta(t) = A(1 + \cos\omega_m t)\cos\omega_0 t \tag{12}$$

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

if $\delta(t) > 0$ *then* $\alpha(t) = \delta(t), \beta(t) = 1$
elseif $\delta(t) < 0$ $\alpha(t) = 1, \beta(t) = -\delta(t)$
else $\alpha(t) = 1, \beta(t) = 1$
endif

A typical value set for $\delta(t)$ is, A=30, ω_m = 0.004 and ω_0 =0.05. The plot for $\delta(t)$ is given in Figure 10.

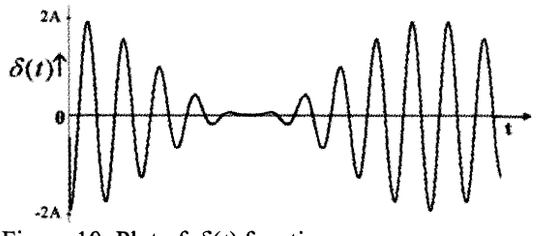


Figure 10: Plot of $\delta(t)$ function.

The value of A (amplitude) is selected as,

$$A > \frac{\max\{|F|, |PCF|\}}{2} \tag{13}$$

to ensure dominance of either F or PCF, over other part. The upper bound of F is used in (13) and is predicted from (Newman 2002) as,

$$F = -2 * \min\{E[Seq], O[Seq]\} \tag{14}$$

where, $E[Seq]$ and $O[Seq]$ indicate the number of even and odd indexed H residues in the sequence.

By oscillating, the PCF is prioritized over F and vice versa. As the minimum value of $|\alpha(t)| = 1$ and $|\beta(t)| = 1$ are maintained and they are never set to zero. This is to preserve the sub-conformation or schema developed in the alternate phase with good features. The amplitude, A is varied to cover a large number of schema or sub-structures to be built and to be tested for a potential

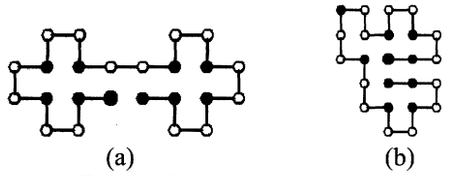
(or, a part of) solution. Throughout the search, as the value of the two weights change alternatively, track is kept of the best solution with respect to F. The putative ground energy is targeted and the simulation is allowed to run upto that target. In the vast majority of cases, the output converged speedily but in few cases (namely sequence 48 and 60 in Table 2) convergence was slower with respect to the sequence length. At least 5 iterations of each of the following sequences are taken and the minimum generations are shown in Table 2. The complete Guided GA (GGA) strategy for PFP is detailed in Algorithm-1.

Using this approach means that any forced sub-conformations built in the negative phase are allowed to be force free during the positive phase. This effectively means influencing or guiding the search, but not forcing it towards a particular conformation, so different configurations having the same putative ground energy (see Figure 11 and 12) are obtained thus confirming that the enforcement was not to reach a particular structure, but rather to explore potential conformations.

Algorithm -1: Guided GA for PFP.

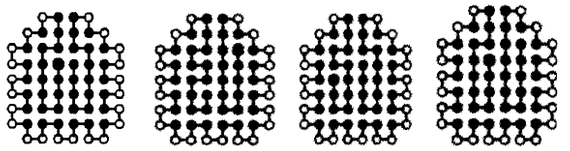
Input: Sequence *S*, Target Fitness of the Sequence (*Target_F*)
Output: 2D Folding of the given sequence.

```
COMPUTE PCF and A for S
t=0          /* Generation count */
F=0          /* Best fitness found from the search */
Fillup the population with random (valid)
        conformation possible from 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 respect to HBBC1 or HBBC3
      IF miss-mapping true then
        { Re-map the sub-sequence to corresp-
          onding HBBC using move-sets.
        }
    }
  COMPUTE TF
  Sort and Keep Elite.
  F ← Best fitness found from the population.
}
END.
```



(a) (b)

Figure 11: Two different conformation instances (a) and (b) for a sequence length of 24, with fitness = -9.



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

Figure 12: For sequence length 64, with 5 separate run, 4 different instances are achieved.

The following snap-shots in Figure 13, show how the new PFP strategy works.

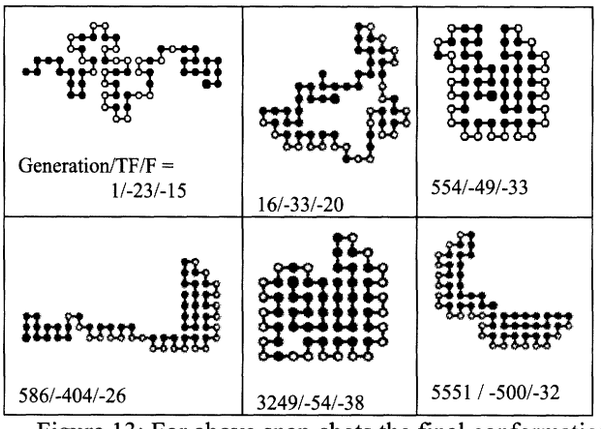


Figure 13: For above snap-shots the final conformation (see Figure 12(a)) is achieved at generation 5646. Used parameters are, A=30, $\omega_m = 0.005$, $\omega_0 = 0.5$, G_{max} =5646.

Table 2: Comparison of proposed Guided GA (GGA).

Length	Putative ground energy	GGA (G_{min}, A, ω_m, ω_0)	Results from other methods				
			GTB (Jiang 2003)	EMC (Liang 2001)	GA (Unger 1993a)	MC (Unger 1993a)	CI (Toma 1996)
20	-9	-9 (2, 30, 0.04, 0.5)	-9	-9	-9	-9	-9
24	-9	-9 (83, 30, 0.04, 0.5)	-9	-9	-9	-9	-9
25	-8	-8 (124, 30, 0.04, 0.5)	-8	-8	-8	-8	-8
36	-14	-14 (814, 30, 0.005, 0.9)	-14	-14	-12	-13	-14
48	-23	-23 (3876, 60, 30, 0.005, 0.9)	-23	-23	-22	-20	-23
50	-21	-21 (720, 120, 0.005, 0.9)	-21	-21	-21	-21	-21
60	-36	-36 (5734, 60, 0.005, 0.9)	-35	-35	-34	-33	-35
64	-42	-42 (5646, 30, 0.005, 0.5)	-39	-39	-37	-35	-40

For the values of (ω_m, ω_0), three sets of values were tested, namely (0.04, 0.5), (0.005, 0.5) and (0.005, 0.9). It is observed that smaller sequences reach target relatively quickly when the swing function changes phase in a faster manner. This is because, shorted sequences being less

compact are less vulnerable to be stuck and with frequent phase shifting it covers possible potential schema quickly. Further, it is noted that sequences having less ratio of P than H, perform better with higher value of amplitude, A. This is because, P being fewer in number, the effect of guideline is less and convergence can be lead by Fitness F mostly. Therefore, amplitude being high helps F lead for a relatively greater amount of time.

7 Conclusions and Future Work

This paper has presented a novel search strategy incorporating guidelines for applying Genetic Algorithms (GA) in Protein Folding Prediction (PFP) applications. Two new move operators, namely *Diagonal Move* and *Tilt Move* have been proposed, which ensure the Guided GA (GGA) outperforms other non-deterministic search approaches, by exploring more potential schema or substructures. The GGA strategy worked better with targeted accuracy that is rarely achieved in other methods, even without incorporating some of the already established features of other methods within GGA. Within the achieved results, for most of the sequences the search converged speedily. It is contemplated that the population being populated with highly likely sub-conformations embedded rather than being randomly orientated, would have quicker convergence. We intend analyse the proposed strategies further by incorporating adaptive parameters. Also, when the search is on, if a probable cavity is kept fixed for a while and Hs are moved separately for computing better fitness – the process would make the search process faster. It is planned to conduct such experiments next and also extend the algorithm to 3D HP model.

8 References

Bastolla, U., Frauenkron, H., Gerstner, E., Grassberger, P. and Nadler, W. (1998) "Testing a new Monte Carlo Algorithm for Protein Folding", National Center for Biotechnology Information, Vol. 32, No. 1, pp.52-66.

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

Bornberg-Bauer, E. (1997) "Chain Growth Algorithms for HP-Type Lattice Proteins", RECOMB, Santa Fe, NM, USA.

Dill, K.A. (1985) "Theory for the Folding and Stability of Globular Proteins", Biochemistry, Vol. 24, No. 6, pp.1501-1509.

Flebig K.M. and Dill, K.A. (1993), "Protein Core Assembly Processes", J. Chem. Phys., Vol. 98, No. 4, pp. 3475-3487.

Hoque, M.T., Chetty, M. and Dooley, L.S. (2005) "Guided Genetic Algorithm for 2D Hydrophobic-Hydrophilic Model for Protein Folding Prediction Using Move Sets", Tech. Report TR-2005/2, GSCIT, Monash University.

Jiang, T., Cui, Q., Shi, G. and Ma, S. (2003) "Protein folding simulation of the hydrophobic-hydrophilic model by computing tabu search with genetic algorithms", ISMB, Brisbane Australia.

Lesh, N., Mitzenmacher, M. and Whitesides, S. (2003) "A Complete and Effective Move Set for Simplified Protein Folding", RECOMB, Berlin.

Liang, F. and Wong, W.H. (2001) "Evolutionary Monte Carlo for protein folding simulations", J. Chem. Phys. Vol. 115, No. 7.

Newman, A. (2002), "A new algorithm for protein folding in the HP model", Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete Algorithms.

Toma, L. and Toma, S. (1996) "Contact interactions methods: A new Algorithm for Protein Folding Simulations", Protein Science, Vol. 5, No. 1, pp.147-153.

Unger, R and Moult, J. (1993a) "Genetic Algorithms for Protein Folding Simulations", Journal of Molecular Biology, Vol 231, pp. 75-81.

Unger, R. and Moult, J. (1993b) "On the Applicability of Genetic Algorithms to Protein Folding", IEEE, pp. 715-725.

Xiang, Z. and Plastock, R. (2001) "Computer Graphics", Schaum's Outline Series, 2nd edition, reprinted, pp. 25-35

Yue, K. and Dill, K.A. (1993) "Sequence-Structure relationships in proteins and copolymers", Physical Review E, Vol 48, No. 3, pp. 2267-2278.

Yue, K. and Dill, K.A. (1995) "Forces of tertiary structure organization in globular proteins", Biophysics, Vol. 92, No. 1, pp. 146-150.