# Protein Folding Prediction in 3D FCC HP Lattice Model Using Genetic Algorithm

Md Tamjidul Hoque[1, 2], Madhu Chetty[1] and Abdul Sattar[2]

*Abstract*— **In most of the successful real protein structure prediction (PSP) problem, lattice models have been essentially utilized to have the folding backbone sampling at the top of the hierarchical approach. A three dimensional face-centred-cube (FCC), with the provision for providing the most compact core, can map closest to the folded protein in reality. Hence, our successful hybrid Genetic Algorithms (HGA) proposed earlier for a square and cube lattice model is being extended in this paper for a 3D FCC model. Furthermore, twins (conformations having similarity with each other), in GA population have also been considered for removal from the search space for improving the effectiveness of GA. The HGA combined with the twin removal (TR) strategy showed best performance when compared with the Simple GA (SGA), SGA with TR, and HGA only versions. Experiments were carried out on the publicly available benchmark HP sequences and results are expressed based on the fitness of the corresponding applied lattice model, which will help any future novel approach to be compared.**

## I. Introduction

REAL protein folding simulation preformed directly from its primary protein sequence on a computer with the aid of current technology is not viable [1]-[5], since it involves immense computational complexity. However, the lattice model servers as a feasible alternative [1]-[5] and has been of significant use in modeling protein folding. This lattice model reduces the computational complexity by:

*i)* presenting each amino acid in a protein sequence mostly as a single residue or bead in the lattice chain.

*ii)* discretizing and restricting the continuum space into a 2D or 3D regular structure, such as square, cubic, triangular, face-centred-cube (FCC)[5][6], or any of the *Bravais Lattices.*

HP lattice [1][2] model in both square (2D) and cube (3D) form, having two beads – *hydrophobic* (H) and *hydrophilic* (or *polar* (P)) and configured as a self-avoiding walk (SAW) on the lattice path favoring an energy free state due to HH interaction, is the most simplified model and hence very popular with the research community [1]-[7], [17]-[19], [21]-[30]. It allows development, validation and comparison of new techniques [17]-[19], [22]-[30] for *protein structure prediction* (PSP) in the first place.

[1] Gippsland School of Information Technology (GSIT), Monash University, Australia

[2] Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Australia.

Corresponding Author: Md Tamjidul Hoque.

E-mails: M. T. Hoque: (Tamjidul.Hoque@infotech.monash.edu.au), M. Chetty (madhu.chetty@infotech.monash.edu.au), A. Sattar (a.sattar@griffith.edu.au ).

However, even if we use this simplified model and that too for modeling short sequence, we have an inordinate number [8]-[11] of valid (i.e., SAW) conformations. For example, for a sequence of $n$ amino acids, the number of valid conformations is proportional [11] to $\mu^n$, where the connective constant or the effective coordinate number $\mu$, is lattice dependent [9]. The prediction of the optimal conformation using lattice model is also an *NP-complete* problem [12]-[13]. However, to predict the backbone conformation of the folded protein from its amino acid sequence based on global interactions such as *hydrophobicity*, the lattice model has been commonly [1]-[5] used for approximation. Most of the successful approaches for *ab initio* prediction presented in recent *Critical Assessment of Structure Prediction* (CASP) [3]-[5], have followed the hierarchical paradigm where the lattice based backbone conformational sampling works at the top of the hierarchy. Therefore, a lattice based backbone modeling is very important as it is effectively creating a blueprint or skeleton on which the final outcome of the real protein structure prediction will depend. With further steps in the hierarchy towards full modeling from lattice, the expanded energy functions include atom-based potentials from molecular mechanics packages [14]-[16] such as CHARMM, AMBER, ECEPP and so on, which is extremely time consuming for developing and checking achieved each lattice based sample.

Compared with the other two PSP approaches i.e., *comparative* or *homology* modeling and *threading* or *fold recognition* approaches, the *ab initio* is the most computationally demanding of the three. However, it is the most promising with regard to providing reliability, accuracy, usability and flexibility in checking the functional divergence of a protein or drug. Our early works resulted in building up of competitive and effective strategies [17]-[19] using HP lattice model towards *ab initio* approach and its success motivates us to expand the strategy towards a suitable opening approach for the real PSP problem, which we prefer to have 3D face-cantered-cube (FCC) lattice orientation for the following reasons:

*i)* Based on the full proof of *Kepler Conjecture* [20], a 3D FCC is proven to be the densest sphere packing orientation. It can provide densest protein core [6]-[7] while predicting a protein structure (though the protein core may not necessarily be the most compact one [22] in all cases).

*ii)* In 3D FCC orientation, a residue can have 12 neighbors in a 3D space (Fig. 1) and 6 neighbors in a hexagonal



1-4244-1340-0/07/$25.00 ©2007 IEEE

pattern in 2D. Such orientation provided by FCC allows maximum excluded volume due to offering densest compactness [6][20], therefore logically inferring, for a region with fixed volume, a FCC model has more option for placing a residue in suitable neighboring position with respect to another residue than any other lattice models. As a rudimentary example, the FCC model is parity [7] problem (odd indexed residue can only be the topological neighbor of the even indexed residue and vise versa) free, whereas the square or the cube lattice is not.

*iii)* Therefore, within the lattice constraints, the FCC lattice can provide maximum degree of freedom and FCC can provide closest resemblance to the real or high resolution folding. With respect to modeling a real protein, FCC orientation can therefore provide the closest conformational alignment [23] amongst the lattice disciplines.

Among existing approaches, there are some statistical approaches such as *Contact Interaction* (CI) [24] and *Chain Growth* (CG) [25] for the PSP problem are characterized by exhibiting lower accuracy as the sequence length increases and also by being non-reversible in their search. Non-deterministic search techniques have dominated attempts to solve the PSP problem, of which there are ample approaches such as, *Monte Carlo* (MC) simulation, *Evolutionary* MC (EMC) [26], [27], *Simulated Annealing* (SA), *Tabu Search* with *Genetic Algorithm* (GTB) [28] and *Ant Colony Optimization* [29]. Because of simplicity and search effectiveness, *Genetic Algorithms* (GA) [17]-[19], [30]-[33] are attractive and they consistently provided superior performance over MC as well as detailed in [30]-[31]. Since, even with the simple HP model, the discrete search landscape of the protein folding or structure prediction problem is inordinately large [2] [3] [30] and complex having many peaks and troughs, it makes sense to note that the neighboring points of any current solution will not necessarily guide the algorithm towards a better solution when techniques like SA (or MC) and *Hill Climbing* (HC) are employed. In such cases, SA needs more operations than an exhaustive search to guarantee that the best solution can be found by random [30] techniques, although an exhaustive search has been proved to be infeasible [8]-[11]. Similarly for HC, the neighboring information does not provide clues for an enhanced solution. Compared to MC, genetic algorithms by their very nature reduce the necessity for highly accurate and locally insensitive energy functions [30] when applied to the PSP problem, with the crossover operation aiding construction of global conformations from the cooperative combination of many local substructure features. Furthermore, there is a possibility that a particular substructure which is irrelevant for one solution may be useful for another. GAs optimize the effort of testing and generating new individuals if their representation permits development of building blocks (*schemata*), a concept formalized in the *Schemata Theorem* [30],[34]-[37].

While GA performance is often very effective [17]-[19], [30]-[34], it can often stall [37] in a hard optimization problem like PSP and does not ensure the final generation contains an optimal solution, a phenomenon that becomes more prevalent as the sequence length increases [17]-[19], [30]. Proceeding further with the GA search, the similarity is found to be increasing among the population and diversity reduces which leads to stalled condition. Further, as the search progresses, the resulting obtained solution becomes phenotypically (physical or conformational interpretation of the GA chromosome) compact. This phenomenon poses considerable challenges to the existing theory, strategy and modeling to carry on the search with a view to obtain an improved conformation out of the surviving sample conformations. To overcome these problems, we provided remedy to those problems separately [17]-[19], [34], [38] and in this paper we present a combination of these efforts and provide a comparison to identify a superior strategy in 3D FCC HP model.

## II. Definitions

In this section, we define and describe the frequently referred fitness function of the lattice model, the metaphorical view of the protein core concept [21],[17]-[19] and the *H-Core Centre* (HCC).

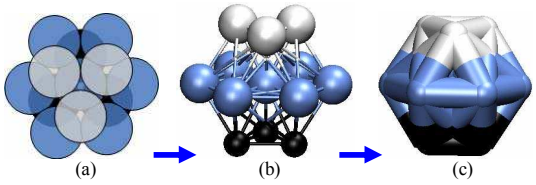


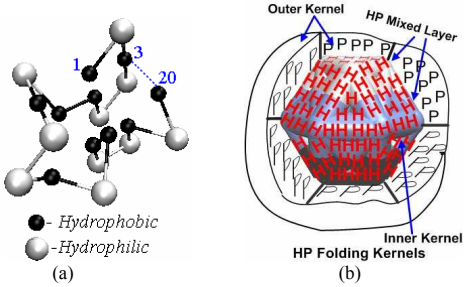

Fig. 1. (a) Top view of the 3D FCC sphere packing (b) Front view of the 3D FCC sphere packing. Layers are separated by both colors and size, and dedicated connections are used to assist the visualization of the concept. The center sphere at layer 2, has 12 neighbors in 3D. It has 3 neighbors from top layer, secondly 6 from middle layer and 3 more from the bottom layer. (c) By compaction the overall shape resembles a cuboctahedron. For drawing (b) and (c), VMD [41] software was used, with lattice supported file format [42].

(1) *The three Dimensional FCC HP Lattice Model*: *Hydrophobicity* is regarded as the major activity [1]-[5] that provides the protein its 3D global or overall shapes from it primary sequence. Based on this observation, the HP model was introduced by Dill [1] with amino acids being represented as a reduced set of *H* (*Hydrophobic* or *Non-Polar*) and *P* (*Hydrophilic* or *Polar*) only. With this, the protein conformations of the sequence can be placed as a self-avoiding walk (SAW) on a 3D FCC model. Then, the energy of a given conformation is defined as a number of topological neighboring (TN) (Fig. 2(a)) contacts between those *H*s, which are not sequential with respect to the sequence. It can be formally defined for structure prediction using lattice as:

Assuming amino-acid sequence is given as $s = s_1, s_2, s_3, \cdots, s_m$, a conformation $c$ needs to be formed

where, $c^* \in C(s)$, energy $E^* = E(C) = \min\{E(c) \mid c \in C\}$ [29].

Here, $m$ represents the total amino acids in the sequence and $C(s)$ is the set of all valid (i.e., SAW) conformations of $s$. If the number of TNs in a conformation $c$ is $k$ then the value of $E(c)$ is defined as, $E(c) = -k$ which is regarded as fitness function and expressed as $F = -k$. In a 3D FCC HP model (Fig. 2, (a)), a non-terminal and a terminal residue both having 12 neighbours can have a maximum of 10 TNs and 11 TNs, respectively.




(a) (b)

Fig. 2. (a) Conformation in 3D FCC HP model. Dotted line indicates one TN for H-H interaction in between residue 3 and 20. Such TNs are total 29, here. Therefore, Fitness $F$= -(TN Count) = -29. (b) 3D metaphoric HP folding kernels for the FCC model.

(2) *Metaphorical protein core concept*: According to the core formation concept [21], the *H*s form the protein core, thereby freeing up the energy while the *P*s, have an affinity with the solvent and so tend to remain in the outer surface. Based on this concept, we visualize the folded protein through the 3D FCC HP model as a three-layered kernel (Fig. 2(b)). A centre of the Hs, called H-Core Centre (HCC) (defined next), guides the core formation around it. The inner most layer of the defined kernel, called the *H-Core* [21], [17]-[19], is assumed to be compact and having high concentration of *H* residues, while the outer kernel forms mostly of *P*s. In between the inner and outer layers, a thin composite layer assumed to exist which is formed by the covalent bonded *H* and *P*, that referred as the HP-Mixed-Layer [17]-[19]. This intermediate layer helps outline the core. Therefore, this thin layer is very important in the sense that its proper shape would accommodate the maximal bonding among the *H*s inside the core, which ultimately helps result into maximally free energy conformation. As part of our effective strategy, local interactions for this thin layer are applied carefully, such as applying highly likely sub-conformations replacing unlikely sub-conformation of the HP-Mixed-Layer to help reform the cavity towards its optimal capacity.

(3) *H-Core Centre (HCC)*: It is calculated as the arithmetic mean of the coordinates of all *H*s. That is,

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

In this paper, a specific set of moves (Sec. IV.B) is applied to a predefined highly likely sub-sequence correspond to a sub-sequence, before enforcing any sub-conformation, the HCC ( $x_{HCC}$, $y_{HCC}$, $z_{HCC}$ ) is updated which works as a *dynamic nuclei* that helps placing *H* towards HCC and *P* as far away from HCC as possible and implement the sub-conformation.

## III. Three-Dimensional H-Core

The proof of *Kepler Conjecture* [20] emphasizes that a FCC orientation is the most dense packing and in this orientation a residue has 12 neighbors. If the orientation of a residue with its 12 neighbors covered with a thin outer layer, the overall structure resembles a cuboctahedron [39] (transition of Fig. 1(b) to 1(c)). A cuboctahedron has 14 faces, 6 of them are square and 8 of them are equilateral triangle and it has 12 corners of vertices.

Therefore, if a typical sequence is considered composed of *H*s only, it is likely to form a cuboctahedron shape to have maximal fitness. The positioning of the *H*s inside the core can be categorized as *H* at the corner, *H* on the edge, *H* on the face and *H* inside the interior. As in [17]-[19], we are interested to work on the *HP mixed layer*, which concern the *H*s in *HP mixed layer* which is basically the outer most layer of the core. Our concern is to compute the probability of an *H* to be appearing at a corner and on an edge or a face of the cuboctahedron to determine the placement of a highly likely sub-conformation described in the Section IV. It is to be noted that for a packing shape cuboctahedron of any size, the number of corner residues will remain 12, but the number of residues on edge increase with the increasing size of the cuboctahedron, i.e. with the increasing number of *H*s. Let, the length of an edge of a cuboctahedron be $x$, therefore the volume can be written as [39]:

$$V_1 = \frac{5\sqrt{2}}{3}x^3 \qquad (2)$$

Further let us assume that excluding the one residue equivalent thickness volume from the outer surface of the cuboctahedron, the remaining volume of the inner cuboctahedron is $V_2$, which can be written as:

$$V_2 = \frac{5\sqrt{2}}{3}(x-a)^3 \qquad (3)$$

Here, $a$ is the average reduction of the edge of the outer cuboctahedron with respect to the inner cuboctahedron. Assume, $n_H$ is the number of *H*s, where $n_H \geq 13$. Minimal cuboctahedron can be formed when, $n_H = 13$ with 12 of the residues will be on the surface of the outer cuboctahedron and 1 will be inside. Based on this fact and applying it in equation (1), we get, $x \approx 1.7668236$ and can write (4):

$$V_1 - V_2 \Rightarrow \frac{5\sqrt{2}}{3}\left\{x^3 - (x-a)^3\right\} = 12 \qquad (4)$$

From (4) considering $a$ as a variable we can write,

$$f(a) = a^3 - 3a^2x + 3ax^2 - \frac{36}{5\sqrt{2}} \qquad (5)$$

and,

$$f'(a) = 3a^2 + 3x^2 - 6ax \qquad (6)$$

Using this value of $x$ in (5) and (6), considering $a$ as variable and then using *Newton-Raphson* method for

approximation, we get $a \approx 1.0154105$. Putting the value of $a$ in (4), the following equations (7) and (8) are achieved, with which we compute the probabilities of an $H$ being at the corner or non-corner position to remap likely sub-conformation (Fig 3). Now, for a sequence where the number of $H$s is $n_H$, the probability of an $H$ residue being at a corner is,

$$\Pr_{corner} = \frac{12}{\frac{5\sqrt{2}}{3}\left\{\frac{3n_H}{5\sqrt{2}} - \left(\sqrt[3]{\frac{3n_H}{5\sqrt{2}}} - 1.0154105\right)^3\right\}} \quad (7)$$

And, the probability of an H on surface but at non-corner is

$$\Pr_{non-corner} = 1 - (1/\Pr_{corner}) \quad (8)$$

provided that $n_H \geq 13$.

## IV. Sub-conformations for Building *HP Mixed Layer*

To form the optimal cavity to accommodate the optimal *H-Core*, intuitively the *P* of a -HP- segment needs to be on the opposite side of H while the other side of the *H* is facing the current HCC. To handle the placement, motif or sub-conformation that is highly probable to a sub-sequence (sited in Fig. 3) is forced to remap. It is also preferred in choosing the motif with readily available TN of H-H interactions and other preferences of placing *P* away from HCC while placing *H* as near as possible to HCC. Small sub-conformations of sizes 3 to 4 have been chosen. Within this size, highly likely conformations are easy to identify.

To implement the same strategies in FCC HP model, various moves [40], [17]-[19] are further simplified and merged which is the benefit of the FCC model over the HP-square or cube model into the 3D space of the FCC model.

For the sub-sequences, -HPH-, -PHP-, -HPPH- and –PHHP-, there are few possible sub-conformations, so only highly likely sub-conformations (Fig. 3) are chosen, based on embedded TN and core formation concepts. The selections for mapping these sub-conformations are based on construction of building corners and non-corner (such as face or edge of the assumed cuboctahedron shaped *H-Core*) based on the predicted probabilities as shown in equation (7) and (8). Another advantage using FCC model is that it is parity problem free, which helps to reduce the number of likely sub-conformations and reduce their numbers compared to [17]-[18].

### A. Move Sets and Implementation

As crossover and mutation become ineffectual [17]-[19], [27], [34] because of the phenotypically congested conformation, strategically designed move sets specially when used in conjunction with domain knowledge helps keep the search effective [17]-[19]. We apply three different move sets and they are chosen to apply first, based on the preference that makes less destruction to the already achieved highly fitted subpart while moving the other subpart with a view to increase the fitness further. If the preferred move set fails in a critically congested region, a move with more chance of destruction is then preferred and the resultant conformation provides more chance to reconstruct the conformation and lead it closer to a globally optimal conformation. This helps explore further potential region of the landscape.

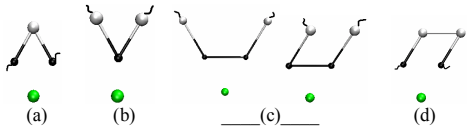



Fig. 3. Potential sub-conformation in 3D space for the subsequence (a) –HPH- (b) –PHP- (c) -PHHP- (d) -HPPH-. ●, ○ and ●, respectively indicate an *H*, a *P* and the approximate position of HCC. For (c) two alternates have been chosen.

The complete details of the manner in which we have applied the moves for remapping the likely sub-conformation (Fig. 3), from all possible preconditions is beyond the scope of this paper. However, few demonstrative simple examples are given in Fig 6 and Fig.7. The *move sets* are described next.

(1) *Diagonal Move*: Assume $\overline{A}$, $\overline{B}$, $\overline{C}$ are three consecutive vector points in the 3D space. By diagonal move, $B$ is moved to $(\overline{A} + \overline{C} - \overline{B})$.

(2) *Pull Move:* Pull move has been shown to be very effective in [19] especially for enforcing any sub-conformation [17]-[19] to apply the domain knowledge. This move is used to implement the sub-conformations or motifs with less distortion of other parts due to the required pulling, which may be in an optimal position as demonstrated in Fig. 4. For the FCC model we used the redefined pull move for the same purpose. Since, the parity problem is absent in FCC model, the pull move does not need to be moved diagonally [40] to start. Therefore an ordinary pull to the next neighboring position performs the same.

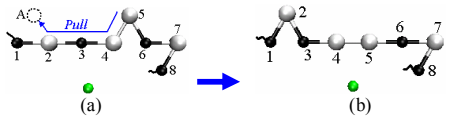



Fig. 4. The subsequence -123- in (a) need to remap to likely sub-conformation for –HPH-. If the position A is free then 2 can be placed at A using a pull (as indicated in (a)) applied towards the higher indexed end. The pull moves 3 to 2, 4 to 3 and 5 to 4 and then finds a valid conformation without pulling further which saves the TN in between residue 6 and 8. The pull move results (b). The |fitness| of (b) is increased by 1. Due to the newly formed TN in between residues at 1 and 2.

This is mainly because FCC orientation relatively has more neighbors and it is the parity problem which helps it to get a valid conformation after a pull without the requirement to propagate the pull often up to the terminal residue, so destruction is further less compared to the 2D square or 3D cube HP lattice model.

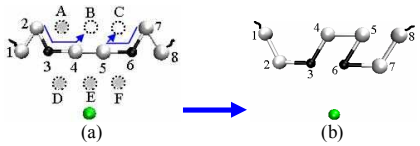



Fig. 5. The subsequence -3456- in (a) need to remap to likely sub-conformation for –PHHP-. And assume position A, B, C, D, E, F are available position near by, but only B and C are not occupied by any other residue. By tilt move residues at 4 and 5 can be positioned to their next parallel positions B and C. The pulls applied in tilt move are indicated by arrows which are pulling from residue 4, 3, 2, 1 sequentially towards B and another pulling 5, 6, 7, 8 sequentially towards C in the 3D space. Finally by this move it result (b), which gains a TN in between residue 3 and 6.

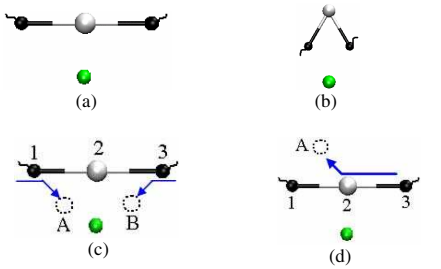



Fig. 6. (a) Pre-condition and (b) desired post-condition. At (c), if location A and B are free then two pull moves (indicated by arrow) can place residue 1 and 2 at A and B respectively to have desired sub-conformation. At (d), if position A is free a single pull can place residue 2 to A and can have the desired sub-conformation in (b)

(3) *Tilt Move*: Two or more consecutive residues which can be connected by a straight line are moved to immediate lattice positions which can also be connected by straight together with the condition that it is parallel to the previous line and unit lattice distance away. It is implemented by pulling towards both ends as demonstrated in Fig. 5.

### *B. Formulation of Multi-objective Optimization Criterion*

While predicting conformation, it was experienced [17]-[19], phenotypical compactness of the conformation resists the predictability of the search algorithm. We formulate and add constraint fitness function named *Probabilistic Constrained Fitness* (PCF) with existing fitness function F. While searching for an optimum conformation, if a sub-conformation corresponding to a particular sub-sequence exists in the HP-Mixed-Layer for a developing conformation, it is rewarded, otherwise penalized. This measure of fitness is referred to as the *Probabilistic Constrained Fitness* (PCF), so if any member of a defined sub-conformation corresponds to the related sub-sequence and the *H*s are nearer to HCC than the *P*s, then PCF will be decreased by 2 as a reward, otherwise it will be penalized by an increase of 1 for a non-desired sub-conformation and 2 for a proper shape but having opposite of the desired directions. The composition of *F* with PCF is such that, the impact of one of them will have higher weight and impact over another for a while and then vice versa and this keep on oscillating in non-monotonous manner, which has been described in Sec. IV.

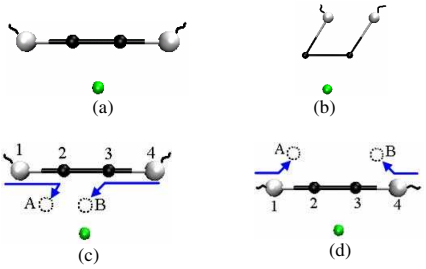



Fig. 7. (a) Pre-condition and (b) desired post-condition. At (c), if location A and B are free, placing residue 2 and 3 to A and B respectively by a tilt move will have the desired sub-conformation. For pre-condition (d), if position A and B are free, two pull moves as indicated by arrows can have desired effect as Fig. 3 (c), first instance.

## V. TWIN REMOVAL IN GA

We are shown in [34] that removal of twin chromosomes, i.e. chromosomes' similarity based on 80% and above (*TR 0.8*, in short) correlation, from the population keeps GA in the optimal search condition, without which (e.g. in case of Simple GA or SGA) it often stalls due to reduction of diversity.

The existence of *twins* (same or similar chromosomes) and the requirement for their removal in a GA is not new, as their growth was considered in evaluating the cost of duplicate or identical chromosomes in [44]. It suggested starting each chromosome with different patterns to avoid twins, but if twin growth is inherent in a GA search, then the effect of initialization using different patterns will relatively quickly decline for long converging problems like PSP. Also, [35] advocated that if a population comprised all unique members, tests need to be continually applied to ensure identical chromosomes did not breed. If chromosome similarities within population do not grow, then the GA may not converge as the search process effectively remains random rather than stochastic, while if similarities grow, then finding a non-similar chromosome to mate with clearly becomes more scarce because of the inevitable occurrence of twins, and the increasingly high cost of finding dissimilar chromosomes in a lengthy convergence process. A direct consequence of this is the creation of two distinct chromosome groupings; a large collection of highly correlated chromosomes and a much smaller set of dissimilar ones, with mating restricted exclusively to members of the respective groups so producing different offspring. These will however, become more similar as they inherit communal features from each group, i.e. equivalent to two dissimilar parents breeding an entire next generation that inherits much commonality.

The need for twin removal was originally highlighted in [45] which emphasized that duplicate chromosomes (*twins*) reduce diversity and ultimately lead to poorer performance. The study however, was solely confined to the detection and removal of identical chromosomes that were unique to each other, with no consideration being given to the removal and impact of similar or highly correlated chromosomes. The situation of having very similar (say, 95% to 99%)

chromosomes within a population rather than restricting it to exact matches (100% similar) has almost the same effect. Therefore in [34], the notion of *twins* was initially reviewed before being broadened to not only include identical, but also similar (highly correlated) chromosomes in the population. A *chromosome correlation factor* (CCF) [34] defined the degree of similarity existing between chromosomes, and it was shown that by removing chromosomes having a similarity value greater than or equal to CCF =0.8 or 80% during the search process enables the GA to continue seeking potential PSP solutions and ultimately provide superior results.

TABLE I
PERFORMANCE COMPARISONS

| Length / Sequence [17] | SGA | SGA+TR 0.8 | HGA | HGA+TR 0.8 | Conformation for best *F*, Fig. |
|---|---|---|---|---|---|
| 20/ HPHPPHHPHPPHPHHPPHPH | -26 | -27 | -29 | -29 | 7a |
| 24/ HHPPHPPHPPHPPHPPHPPHPPH H | -26 | -26 | -28 | -28 | 7b |
| 25/ PPHPPHHPPPPHHPPPPHHPPPP HH | -24 | -24 | -25 | -25 | 7c |
| 36/ P3(H2P2)2P3H7P2H2P4H2P2HP2 | -41 | -47 | -50 | -51 | 7d |
| 48/ P2(HP2H)2HP510HP5(H2P2)2HP2 H5 | -59 | -64 | -65 | -69 | 7e |
| 50/H(HP)4H4PHP3HP3HP3HP4H P3HP3HPH4(PH)4H | -55 | -56 | -59 | -59 | 7f |
| 60/P2H3PH8P3H10PHP3H12P4H6 PH(HP)2 | -97 | -112 | -114 | -117 | 7g |
| 64/H12(PH)2((P2H2)2P2H)3PHPH12 | -81 | -90 | -98 | -103 | 7h |
| 20/HHHPPHPHPHPPHPHPHPPH | -23 | -27 | -29 | -29 | 7i |

Predictability of i) SGA ii) SGA with TR0.8 iii) HGA and iv) HGA with TR0.8. Achieved maximum |fitness| from 15 runs is shown. Total independent runs are $(4\times15\times9)$ or, 540, using 30 different machines of same capacity.

## VI. EXPERIMENTAL SEARCH IN 3D FCC HP MODEL AND RESULTS

To the best of our knowledge, using 3D FCC HP lattice model on benchmark or publicly available HP sequences prediction results are not available *in terms of fitness* to compare the various prediction strategies for accuracy purpose. Rolf Backofen et al [6][7], have reported on 3D FCC HP lattice model. Though, in [6] a combined approach using GA and MC execution in 3D FCC HP lattice has been reported and the search prediction was based on fitness function, but the results were not based on fitness as a refinement measure was applied and finally a *root mean square deviation* (RMSD) based comparisons were carried out on the dataset which, to the best of our knowledge, is not publicly available. Further, in [7] an HP sequence was reported, but the results were based on timing for fast computation. But, our target is to verify the accuracy of our approach which is challenging, since the problem has been proven *NP-complete* [12][13] therefore a faster algorithm without a deterministic approach is not likely to be effective. So, to verify the predictability of our approach we ran four different GA based setup using publicly available HP sequences [17]. Variations are: *i*) Simple GA (SGA), then *ii*) SGA with optimal (having similarity ≥ 80% within population is removed [34]) *twin removal* (TR with 0.8), the combination is expressed as "SGA+TR0.8". Third, the main topic elaborated in this paper, i.e., the heuristic based *iii*) hybrid GA (HGA) ran as well as *iv*) HGA with TR0.8 (HGA+TR0.8 in short).

In Sec. VII, methods are discussed along with the achieved results. The additional constraint (PCF) formulates the multi-objectivities criterion (Sec. IV.B) and the implementation is such that it ultimately maximizes the goal of original fitness. The two objectives, namely PCF and F get alternately emphasized during the execution of the algorithm. For eventually achieving a global optimal solution, we therefore ensure that the best solution obtained during the phase when F is emphasized is not lost. The total or combined fitness (*TF*) is defined as,

$$TF = \alpha(t) * F + \beta(t) * PCF \tag{9}$$

where $t$ is $t^{th}$ generation of GA. For alternating the weight of $F$ and $PCF$, through α and β, a non-monotonous oscillating function has been used as given by equation (10):

$$\delta(t) = A(1+\cos\omega_m t)\cos\omega_0 t \tag{10}$$

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

$$\alpha(t) = \delta(t),\ \beta(t) = 1,\ when\ \delta(t) > 0 \tag{11}$$

$$\alpha(t) = 1,\ \beta(t) = -\delta(t),\ when\ \delta(t) < 0 \tag{12}$$

$$\alpha(t) = 1,\ \beta(t) = 1,\ when\ \delta(t) = 0 \tag{13}$$

For the typical value of $\delta(t)$ parameters are set as follows: amplitude $A$=30, $\omega_m$ = 0.004 and $\omega_0$ =0.05. The value of $A$ is selected as, $2A \geq \max\left(|F|, |PCF|\right)$ where the upper limit of $|F|$ is set using (14), which has been extended from [19].

$$F = -(5n_H + n_T) \tag{14}$$

Here, $n_H$ is the total number of hydrophobic residues in a sequence and $n_T$ is the number of hydrophobic residues at the terminal positions and $0 \leq n_T \leq 2$.

The search procedure is given in Algorithm-I. A simple GA which is hybridized with population size of 200 is 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 [30] is applied. The implementation of crossover and mutation is the same as in [17]-[19],[30] but without any special treatment (e.g. cooling). Roulette wheel is used for selection procedure.

The Algorithm-I described for *HGA+TR 0.8* is a superset for the other approaches compared here. The steps 2-5, 8, 15 and 16 in Algorithm-I constitute SGA run. (*SGA+TR 0.8*) additionally includes the step # 6 for TR 0.8. If the run is only HGA, it will include all the steps except #6. For the similarity comparison used in *TR*, non-isomorphic encoding

approach was developed for 3D FCC in [42] following the similar approach in [38]. The measure could have been approached by RMSD, but, the complexity of RMSD is high and as a frequent step in search, the progress of the GA gets extremely slow. The time complexity of non-isomorphic coding is $O(n)$, where as for RMSD it is $O(n^2)$. Moreover, RMSD does not provide any scale to measure the percentage of similarity, only 100% similarity can be concluded from it, when RMSD comparison results zero.

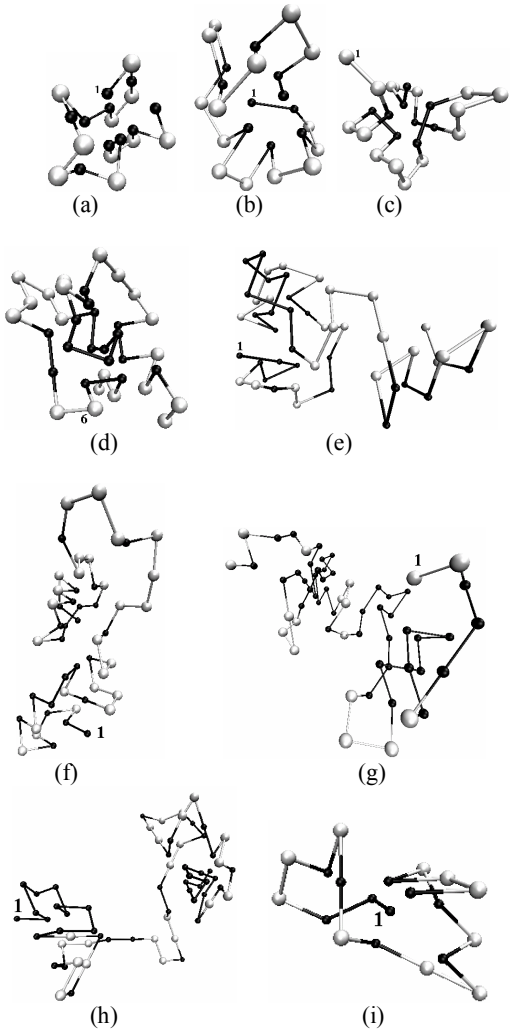



Fig. 7. (a) to (i) correspond to the conformation with minimal *fitness* achieved as indicated in Table-1 (column 5 in this case).

Simulations are carried out on benchmark sequences [Table-1, Column-1]. For each of the sequences, aforementioned 4 different experiments were run parallely for an equal amount of time. The runs were terminated if any of the three approaches became non-progressive. The goal has been to compare the predictability of the developed HGA approach in combination with TR 0.8. Results are shown in Table-1 with the conformations corresponding to the minimal *fitness* achieved. *HGA+TR0.8* outperformed other approaches.

Algorithm-I: *"HGA+TR0.8"* for PSP, in 3D FCC Model

**Input**: Sequence *S*,
**Output**: Found maximum |Fitness|, *F* for the run.
1. COMPUTE: *PCF*; COMPUTE: *A* (amplitude)
2. *t*:=0, *F*:=0 /* Gen. count and fitness initialization */
3. Fillup the population with random (valid) conformation possible for *S*.
4. While ($F \neq TargetValue$) AND <NOT TERMINATE> THEN
5. { t: = t + 1
6. Remove Twins from population for 80% and above similarity. Fill the gap with random conformation.
7. COMPUTE $\delta(t)$, $\alpha(t)$, $\beta(t)$, *TF*
8. CROSSOVER and then MUTATION
9. IF $\delta(t) < 0$ THEN
10. { FOR i: =1 to *population_size* DO
11. Check $chromosome_i$ for any miss mapping of highly likely sub-conformations based on probabilities from eqn. (7) and (8).
12. IF miss-mapping = TRUE THEN
13. { Re-map the sub-sequence to corresponding likely sub-conformation using move-sets. }}
14. COMPUTE *TF*
15. Sort, Keep Elite.
16 *F* := Best fitness found from the population. }
END.

## VII. Discussion and Conclusion

There are two major drawbacks with SGA when applied for PSP. The GA computation is based on schema theorem, which states [10] that short, flexible schemata with above average performance will have a higher survival chance in the subsequent generations and schemata with below-average performance will decay very quickly in a nonlinear manner. Hence, an obstacle in using SGA is that the similarity within the population grows very quickly which leads to a stall or stuck condition, since crossover will most likely occur between twins or similar chromosomes with high fitness. Those which are mutated are likely to be heavily dissimilar and would therefore be rejected by the selection process. To address this problem, twin removal was applied and extended [34]. Elitism was also used only to keep those solutions which were found best in each iteration. Although PSP is a convoluted optimization due to its search landscape [2] and conformational physical compactness, the search is time intensive, even applying TR. Further, as the optimum conformation is relatively compact, crossover and mutation confront more increasing collision and produce invalid conformation. Our specific implantation procedure of highly likely sub-conformation with special *move operator*, i.e., HGA approach, moves the compact conformation with less destruction to preserve the already achieved higher fitness. This arrangement creates scope for probable reformation of the *H-Core* cavity to maximize the *H-sides* or bonding inside the *H-Core*. Hence, this approach also aids in removing the stall condition by introducing variation in such a way, that it enhances the chance of retaining and exploring

highly fitted conformations, which can be further strengthened by combining TR of 0.8.

Further, the overall approach has been extended for 3D FCC HP model. However, it is shown that beyond removing the parity problem, FCC further simplifies the pull moves and reduces the need of higher number of sub-conformations for remapping compared to approaches in [17]-[18]. Due to the unavailability of any previous works with direct fitness based results on 3D FCC, we have restricted our comparison to the proposed HGA with the various combinations of standard SGA and TR of 0.8. The experiment results show significant improvement over predictability in favor of the combination of HGA and TR 0.8, especially for the longer sequences.

## ACKNOWLEDGMENT

Support from Australian Research Council (grant no DP05573035) is thankfully acknowledged.

## REFERENCES

[1] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins", Macromolecules, 1989, 22, 3986-3997.
[2] K. A. Dill *et al.*, "Principles of protein folding -- A perspective from simple exact models", Protein Science, 1995, 4(4):561-602.
[3] D. Baker, "Prediction and design of macromolecular structures and interactions", Phil. Trans. R. Soc., 2006, 361: 459–463.
[4] O. Schueler-Furman *et al.*, "Progress in Modeling of Protein Structures and Interactions", 2005 VOL 310 SCIENCE.
[5] Y. Xia *et al.*, "Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach", J. Mol. Biol., 2000, 300: 171-185.
[6] R. Backofen , S. Will, P. Clote, "Algorithmic approach to quantifying the hydrophobic force contribution in protein folding", Pacific Symp. On Biocomputing, 2000, 5:92-103.
[7] R. Backofen, S. Will, "A Constraint-Based Approach to Fast and Exact Structure Prediction in Three-Dimensional Protein Models", Constraints Journal, 2006, 11 (1).
[8] M. Chen and K.Y. Lin, "Universal amplitude ratios for three-dimensional self-avoiding walks", J. of Phys. A, 2002, 35: 1501-1508.
[9] R. Schiemann, M. Bachmann and W. Janke, "Exact Enumeration of Three – Dimensional Lattice Proteins", 2005, Elsevier Science.
[10] D. MacDonald *et al.*, "Self-avoiding walks on the simple cubic lattice", J. Phys. A: Math. Gen., 2000, 33:5973-5983.
[11] A.J. Guttmann, "Self-avoiding walks in constrained and random geometries": Eds. B. Chakrabarti, Elsevier, 2005, 59-101.
[12] P. Crescenzi *et al*.,"On the complexity of protein folding (extended abstract)", ACM, 1998:597-603.
[13] B. Berger and T. Leighton, "Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-Complete", J. of Comp. Bio, 1998, *Spring*, 5(1): .27-40.
[14] I K Roterman *et al.*, "A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. Phi-psi maps for N-acetyl alanine N'-methyl amide: comparisons, contrasts and simple experimental tests". J. Biomol. Struct. Dynamics, 1989, 7(3):421-453.
[15] W D Cornell *et al.*, "A second generation force field for the simulation of proteins and nucleic acids". J. Am. Chem. Soc., 1995, 117:5179-5197.
[16] G Némethy *et al.*, "Energy parameters in peptides. Improved geometrical parameters and non-bonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides". J. Phys. Chem., 1992, 96:6472-6484.
[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 CEC, pp. 259-266.
[18] M. T. Hoque, M. Chetty and L. S. Dooley, "A Guided Genetic Algorithm for Protein Folding Prediction Using 3D Hydrophobic-Hydrophilic Model", 2006 IEEE WCCI, 8103-8110.
[19] M. T. Hoque, M. Chetty, L. S. Dooley, "A Hybrid Genetic Algorithm for 2D FCC Hydrophobic-Hydrophilic Lattice Model to Predict Protein Folding, 2006, LNAI.
[20] T. C. Hales, "A proof of the Kepler conjecture. Annals of Mathematics", 2005, 162(3): 1065–1185.
[21] K. Yue and K. A. Dill, "Sequence-Structure relationships in proteins and copolymers", Physical Review E, 1993, 48(3): 2267-2278.
[22] R. Backofen, S. Will, E. Bornberg-Bauer, "Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets", Bioinformatics, 1999, 15(3):234-242.
[23] G. Raghunathan *et al.*, "Ideal architecture of residue packing and its observation in protein structures", Protein Sci., 1997, 10:2072-83.
[24] L. Toma *et al.*, "Contact interactions methods: A new Algorithm for Protein Folding Simulations", Protein Science, 1996, 5(1):147-153.
[25] E. Bornberg-Bauer, "Chain Growth Algorithms for HP-Type Lattice Proteins", 1997, RECOMB, Santa Fe, NM, USA.
[26] U. Bastolla *et al.*, "Testing a new Monte Carlo Algorithm for Protein Folding", *NCBI*, 1998, 32(1):52-66.
[27] F. Liang, W. H. Wong, "Evolutionary Monte Carlo for protein folding Simulations". J. Chem. Phys., 2001, 115 (7).
[28] T. Jiang *et al.*, "Protein folding simulation of the hydrophobic-hydrophilic model by computing tabu search with genetic algorithms", 2003, ISMB.
[29] A. Shmygelska, H. H. Hoos, "An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem". BMC Bioinformatics, 2005, 6(30).
[30] R. Unger, J. Moult, "Genetic Algorithms for Protein Folding Simulations", J. Mol. Biology, 1993, 231:75-81.
[31] G.B. Fogel, . D.W. Corne (Editors), "Evolutionary Computation in Bioinformatics", 2004, Elsevier Science, USA.
[32] R. König, T. Dandekar, "Refined Genetic Algorithm Simulation to Model Proteins", 1999, Journal of Molecular Modeling.
[33] O. Takahashi, H. Kita, S. Kobayashi, "Protein Folding by a Hierarchical Genetic Algorithm", 4th Int. Symp. 1999, AROB.
[34] M. T. Hoque, M. Chetty and L. S. Dooley, "Critical Analysis of the Schemata Theorem: The Impact of Twins and the Effect in the Prediction of Protein Folding using Lattice Model", 2005, Tech. Report TR-2005/8, GSIT, MONASH University.
[35] Z. Michalewicz, "Genetic Algorithms + Data Structures = Evolution Programs", 1992, New York: Springer-Verlag.
[36] D. Whitley, "An Overview of Evolutionary Algorithms", Journal of Information and Software Technology, 2001, 43: 817-831.
[37] D.B. Fogel, "EVOLUTIONARY COMPUTATION Towards a new philosophy of Machine Intelligence", 2000, IEEE Press.
[38] M. T. Hoque, M. Chetty and L. S. Dooley, "Non-Isomorphic Coding in Lattice Model and its Impact for Protein Folding Prediction Using Genetic Algorithm", 2006 IEEE CIBCB.
[39] Cuboctahedron, [online] http://en.wikipedia.org/wiki/Cuboctahedron, access Feb, 2007
[40] N. Lesh, M. Mitzenmacher, S. Whitesides, "A Complete and Effective Move Set for Simplified Protein Folding", 2003 RECOMB.
[41] VMD, [Online] http://www.ks.uiuc.edu/Research/vmd/
[42] S. Pötzsch *et al.*, "Visualization of Lattice-Based Protein Folding Simulations", IV 2006, 0-7695-2602-0/6, IEEE.
[43] M. T. Hoque, M. Chetty and A. Sattar, "Non-Isomorphic Chromosomal Encoding Algorithm in GA for 3D FCC Lattice Model", Tech. Report TR-2007/4, GSIT, MONASH University.
[44] R. L. Haupt et al., "Practical Genetic Algorithms", 2nd Edition, 2004.
[45] S. Ronald, "Duplicate Genotypes in a Genetic algorithm", IEEE World Congress on Computational Intelligence, 1998, pp793-798.