Genetic Algorithm in Ab Initio Protein Structure Prediction Using Low Resolution Model: A Review

Author

Hoque, Md Tamjidul, Chetty, Madhu, Sattar, Abdul

Published

2009

Book Title

Biomedical Data and Applications

DOI

10.1007/978-3-642-02193-0_14

Genetic Algorithm in *Ab Initio* Protein Structure Prediction using Low Resolution Model: A Review

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

[1]IIIS, Griffith University, Nathan, QLD-4111, Australia, ~~2~~GSIT, Monash University, Churchill, VIC 3842, Australia.

Tamjidul.Hoque@gmail.com, Madhu.Chetty@ infotech.monash.edu.au, a.sattar@griffith.edu.au

Abstract Proteins are sequences of amino acids bound into a linear chain that adopt a specific folded three-dimensional (3D) shape. This specific folded shape enables proteins to perform specific tasks. The protein structure prediction (PSP) by *ab initio* or *de novo* approach is promising amongst various available computational methods and can help to unravel the important relationship between sequence and its corresponding structure. This article presents the *ab initio* protein structure prediction as a conformational search problem in low resolution model using genetic algorithm. As a review, the essence of twin removal, intelligence in coding, the development and application of domain specific heuristics garnered from the properties of the resulting model and the protein core formation concept discussed are all highly relevant in attempting to secure the best solution.

1 Introduction

Ab initio protein structure prediction (PSP) is an important and very challenging interdisciplinary problem encompassing *biochemistry*, *biophysics*, *structural biology*, *molecular biology* and *computational biology* to give just a couple of examples. Structure prediction, especially in revealing the relationship between sequences and protein folding is the key to combating many diseases and the development of several crucial biotechnological applications and the *ab initio* approach in this regard offers great hope for improving the human condition. More than half of the dry weight of a cell is made up of proteins of various shapes and sizes and protein's specific folded three-dimensional (3D) shape (Fig. 1) enables it to perform specific tasks. From the computing point of view, the exciting investigations concerning proteins is not necessarily about these molecules carrying out

vital tasks but mainly about the process of its acquiring various shapes, i.e. protein folding problem, which enable it to perform the specific tasks. To solve the PSP problem, among other approaches nondeterministic searching approach Genetic Algorithms are found promising [1, 2, 3]. On the other hand, to model and to handle the complexity of the protein folding the low resolution model found [4, 5, 6] to be effective exploring the vast and convoluted search space in a reasonable time scale. The low resolution model aids in providing a valuable theoretical insight which is otherwise often very hard to extract in the high resolution model.

(a) (b) (c)

Fig. 1. Nature's 3D folded protein process (a) primary sequence of amino acid (b) complex folding process (c) folded protein [7].

In this article, we prefer to provide a review to show how novel techniques can improve GA to handle the low resolution based PSP problem, which is yet too complex to be solved. Thus in Section 2, the conformational complexity of protein structure prediction has been discussed. Section 3 describes the modelling issue of the computational protein structure prediction. Section 4 discusses novel computational techniques to cope the low resolution model. The preference of the face-centred-cube (FCC) lattice configuration for the PSP problem has been advocated in Section 5 and in Section 6 a novel model, named hHPNX model, has been presented which can remove some limitations of the existing HP and HPNX model and thus provides better predictions. Finally, Section 7 draws the conclusions.

2 Conformational Complexity

Amongst the 20 different amino acids, any two can join themselves by forming peptide bond thus resulting in an amide plane (Fig. 2). Formation of peptide bonds and properties of amide plane are very important in providing specific shape to a specific polypeptide chain formed from the amino acid concatenation. The amide plane is rigid and dihedral angles, φ and ψ provide flexibility in mobility about 2π, around the N-C_α and C_α-C connecting axis. Each of the amino acids can have large number of torsion angles χ (see Fig. 2) depending on the length of the side chain, however here we assume two per amino acid. To estimate the complexity and to test the feasibility of an exhaustive search algorithm can be considered by all possible combinations of the shape parameters (e.g., dihedral and torsion dis-

crete angles); if there are n numbers of residues in a particular sequence, the total number of conformations (C_{Tot}) can be expressed as:

$$C_{Tot} \approx \left(\varphi_1 \times \varphi_2 \times ... \times \varphi_{(n-1)}\right)\left(\psi_1 \times \psi_2 \times ... \times \psi_{(n-1)}\right)\left(\chi_1 \times \chi_2 \times ... \times \chi_{2n}\right) \qquad (1)$$

However, in practice for sterical disallowance, due to the shape and size of the atoms and their positioning, some reduction in the degree of freedom is possible, which is commonly depicted by the *Ramachandran plot* [8]. Even though, the search space remains astronomically large. For example, with tremendous simplification, assume each amino having only three discrete angles with three degrees of freedom, a 50 residue-long protein sequence will have $\approx 3^{(3\times 50)}$ possible conformations. Now, typically a computer capable of searching ≈ 200 conformations per second would require $\approx 5.8661^{61}$ years to confirm the best search result.

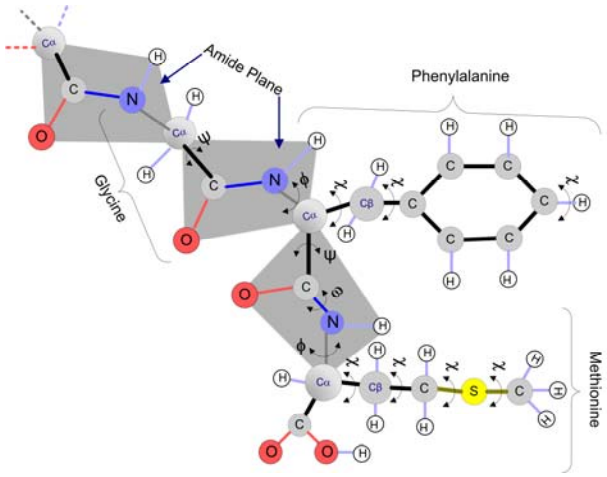


Fig. 2. A schematic of a portion of the [Met]-enkephalin [9] molecule's concatenated amino acid sequence, "…- glycine – phenylalanine - methionine", showing the formation of rigid amide plane (virtually shown using shaded plane) and the side-chains of the corresponding amino-acids. The mobility of the sequence is mostly due to the angles, indicated by φ and ψ over the connection between N-C_α and C_α-C. The side chain torsion angle is shown by χ.

Along with the conformational search complexity, in reality, there are also other forces [10] such as *hydrophobic* interaction, *hydrogen* bonding and *electrostatic* forces together with *Van der Waals* interactions, *disulphate bridge*, so on serve to influence the final 3D conformation. We discuss the existing conformational investigation techniques in two categories next.

2.1 Non-Computational Techniques

Non-computational or experimental techniques such as X-ray crystallography (XC) [11] and nuclear magnetic resonance (NMR) spectroscopy methods are used for PSP. They are very time consuming, expensive and labour intensive [12]. Moreover, the NMR becomes less accurate for longer sequence and the crystallization for XC process may force the protein to have a non-native structure [13].

2.2 Computational Techniques

The computational approaches have the potential to correlate and predict the primary sequence of a protein to its structure thus can overcome the aforementioned difficulties associated with the experimental approaches. So, there has been significant research interest [7] into application of computational approaches for protein structure prediction. Approaches such as homology modelling [14] (which is based on the similarity comparison of the sequence) and threading [15] (which is a process to thread together likely short sub-conformation of the corresponding subsequence) are based on the database of protein sequences and their corresponding structure. However, as these methods depend on the availability of similar sequence samples in the database, their results may become unconvincing for dissimilar sequences [4, 5] and they become less accurate for longer sequences as the formation of the whole conformation derived from its sub-conformations is less likely to match the native conformation because more dissimilarity is added between similar fragments [16, 17].

Consequently, the *ab initio* (meaning 'from the origin') or *de novo* approach predict folded protein's 3D structure from its primary sequence alone [18] based on intrinsic properties (namely, hydrophobic and hydrophilic) of amino acids. The concept of *ab initio* folding is based on the *Anfinsen's thermodynamic hypothesis*, which assumes [19, 20] that the native state of the folded protein is the global free energy minimum. Together with *Levinthal Paradox* which Cyrus Levinthal postulated [21], in what it is popularly known as that, "the proteins fold into their specific 3D conformations in a time-span far shorter than it would be possible for protein molecules to actually search the entire conformational space for the lowest energy state. However, in contrast protein cannot sample all possible conformations while folding, and hence folding cannot be a random process which leads to conclude that folding pathways must exist", which motivates the *ab initio* based computation. However, in practice, as *ab initio* approach is computationally intensive, usually short protein sequences have been simulated at the atomic level, mostly using simplified low-resolution model and simple fitness function. Some methods are hierarchical [9, 19, 22] in that they begin with a simplified lattice representation and end with an atomistic detailed molecular dynamics simulation [23, 24]. With further advancement, the energy functions include atom-based potentials from molecular mechanics packages [25] such as CHARMM, AMBER [26]

and ECEPP [27]. While *ab initio* is the most computationally demanding of the three computational approaches, it conversely also is the most promising in providing reliability, accuracy, usability and flexibility in checking the functional divergence of a protein by modifying its structure and sequence.

3 Models for Structure Prediction

The most appropriate approach for protein modeling would be to simulate the actual folding process which occurs in nature [28], such as *molecular dynamics* (MD) (which is based on collaborative motion and energy of the molecules in a protein sequence) [29, 30, 31]. However, this is infeasible for two reasons:

i) The computation time even for a moderately-sized folding transition exceeds the feasible range even using the current best capable machines applying molecular dynamics principles.

ii) The forces involved in the stability of the protein conformation are currently not modeled with sufficient accuracy.

Thus, to handle the complexities for PSP, models of different resolutions are applied, which help transform the continuous large conformational landscape into a reduced and discrete search landscape, reducing the timescale of protein motion and makes the sampling of the landscape more feasible. Also, the modeling chronology from low to high considers the backbone modeling first and then subsequently the side-chain packing and extended modeling. In low resolution models, more atoms are grouped together, especially from the same amino acid and then treated as a single entity. The most simplified paradigm is the lattice model which focuses only upon *hydrophobicity* by dividing the amino acids into two parts: *hydrophobic* (H) and *hydrophilic* or *polar* (P) thereby leads to its popular appellation of the HP model [32, 33]. The lattice can have several regular shapes with varying numbers of neighboring residues either in 2D or 3D, such as square, cubic, triangular, face-centered-cube (FCC) [22, 34], or any of the *Bravais Lattices*. Conversely, the off-lattice model [35, 36] relaxes the regular lattice structure and both lattice and off-lattice normally start with backbone modeling and then increase the resolution, breaking the residues into further smaller constituents or considering the inclusion of side-chains. In the *side-chain-only* [37] (SICHO) approach, the side chains are initially constructed ahead of the main chain, with the argument being that interactions within proteins are due to different characteristics of the side chain, while the interactions of the main chain are rather more generic. CABS (an acronym for $C_{\alpha\text{-}\beta}$ and Side group) [38] is a relatively high resolution lattice model which assumes a lattice confined C_α representation of the main chain backbone, with 800 possible orientations of the C_α–C_α vectors. The lattice spacing of the underlying simple cubic lattice is assumed to be 0.61Å. The model assumes four united atoms (interaction centres) per residue: α-carbon, centre of the virtual C_α–C_α bond (serving as a centre of interactions for the peptide bonds),

C_β (see Fig. 2) and where applicable, the centre of mass of the side-group. While the coordinates of the α-carbons are restricted to the underlying lattice, the coordinates of the remaining united atoms are off-lattice and defined by the C_α-coordinates and the amino acid identities. The force-field of this model consists of several potentials that mimic averaged interactions in globular proteins. Finally, the direct *all-atom* [12, 39] model considers all the atoms including the forces. The finest possible model applies the theories of *Quantum Mechanics* (QM) with the principal simulation paradigm, especially for the all-atom model, being based upon the *thermodynamics hypothesis*, namely that the stable structure corresponds to the global free energy minimum. The computation to find the most stable energy-free state is based on MD [12, 30] using the collaborative motion and energy of the molecules involved from the protein and solvent. In MD simulation [40], the system is given an initial thermal energy and the molecules are allowed to move in accordance with MD principles. After a short time delay, typically 10^{-15} to 10^{-4} seconds, forces are used to calculate the new position of the atoms, which produces the atomic coordinates as a function of time. IBM's *Blue Gene* [40, 41] project involved such an effort with *peta*-FLOP capability (10^{15} floating point operation per seconds). This is still however, many orders of magnitude lower than the requirement for a realistic solution.

With the objective of successfully building an effective computational strategy to unravel the complexities of the sequence-to-folding relationship, even using the well-established HP model, an efficient and robust solution has still to be developed. In highlighting the various computational intelligence approaches for *ab initio* PSP, the next section focuses mainly upon the low resolution HP model.

The HP Model

The HP model introduced by Dill [32, 33] is based on the fact that the *hydrophobic* interactions dominate protein folding. The Hs form the protein core freeing up energy, while the Ps, have an affinity with the solvent and so tend to remain in the outer surface. For PSP, protein conformations of the sequence are placed as a *self-avoiding walk* (SAW) on a 2D or 3D lattice. The energy of a given conformation is defined as a number of *topological neighbouring* (TN) contacts between those Hs, which are not sequential with respect to the sequence.

PSP is formally defined as: for an amino-acid sequence $s = s_1, s_2, s_3, \cdots, s_n$, a conformation c needs to be formed whereby $c^* \in C(s)$, energy $E^* = E(C) = \min\{E(c) \mid c \in C\}$ [42], where n is 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 q then the value of $E(c)$ is defined as $E(c) = -q$ and the *fitness function* is $\mathrm{F} = -q$. The optimum conformation will have maximum possible value of |F|. In a 2D HP square lattice model (Fig. 3. (a)), a non-terminal and a terminal residue, both having 4 neighbours can have a maximum of

2 TNs and 3 TNs respectively. In a 2D FCC HP model (Fig. 3. (b)), a non-terminal and a terminal residue both having 6 neighbours can have a maximum of 4 TNs and 5 TNs respectively.

Many of the successful PSP software such as ROSETTA [4, 43], PROTINFO [44, 45], TASSER [46] use various resolution of models embedded into the hierarchical paradigm [6, 46−49] to cope with the high computational complexity. The low resolution model can be used to determine the backbone of the 3D conformation and can passes it to the next step for further expansion.

(a) (b)

Fig. 3. Conformations in the 2D HP model shown by a solid line. (a) 2D square lattice having fitness = - (TN Count) = -9. (b) 2D FCC lattice having fitness = -15. '●' indicates a hydrophobic and '●' a hydrophilic residue. The dotted line indicates a TN. Starting residue is indicated by '1'.

4 Search Algorithms

The PSP in HP lattice model has been proven to be NP-complete [50, 51], which implies that neither a polynomial time nor an exhaustive search [52−55] methodology is feasible. Thus the non-deterministic search techniques have dominated attempts, of which there are ample approaches such as, *Monte Carlo* (MC) simulation, *Evolutionary* MC (EMC) [56, 57], *Simulated Annealing* (SA), *Tabu search* with *Genetic Algorithm* (GTB) [58] and *Ant Colony Optimization* [42], though because of their simplicity and search effectiveness, *Genetic Algorithm* (GA) [1− 3, 7, 9, 59, 60] is one of the most attractive [2, 59]. Therefore, we focus on GA and we starts with preliminaries on GA associated with PSP problem in low resolution.

4.1 Underlying Principle of Nondeterministic Search and GA Preliminaries

The algorithm shown in Fig. 4. provides a generic framework for the nondeterministic search approaches.

1. Initial random solution generated randomly or, using domain knowledge.
2. Obtain new solution (S_{new}) from the current single solution (S_{curr}) or pool of solutions using special operator/operation defined by individual approaches.
3. Assess the quality or the fitness F of S_{new} .
4. IF F indicates improved solution accept S_{new} , ELSE accept/reject based on special criteria.
5. IF END-OF-SOLUTION is not reached THEN go back to Step 2.

Fig. 4. Template for a nondeterministic search approach.

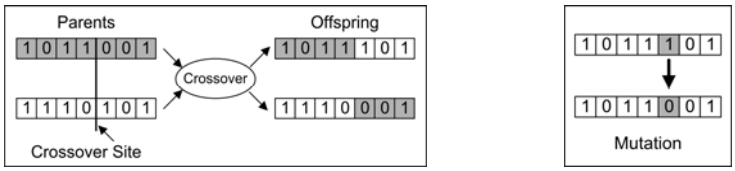
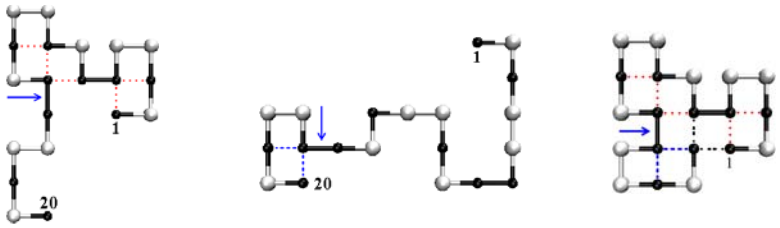


Fig. 5. An example showing (a) 1-point crossover, (b) mutation by 1 bit flipping.

Fig. 6. An example of mutation operation [2]. Dotted lines indicate TN. Residue number 11 is chosen randomly as the pivot. For the move to apply, a 180° rotation alters (a) with F = -4 to (b) F = -9. '→' indicates mutation residue.

Nondeterministic approaches can vary base on the steps shown in Fig. 4. For instance, Hill Climbing approach [61] starts (step 1) with a random bit string and then obtains (in step 2) a set of neighboring solutions by single bit flipping of the current solution. Then, the best is keep as the new current solution and the process is repeated until the stop criterion is met. SA uses the same framework, but differs in its acceptance criteria (step 4): When the new solution is not better than the current, the algorithm can still accept it based upon some randomly defined criteria. GA uses a pool of solution (step 2), named population and obtains new solution by crossover (see Fig. 5 (a)) and mutation (see Fig. 5 (b)) operators. In the PSP context, mutation is a pivot rotation (see Fig. 6) which is also followed in crossover

operation (see Fig. 7).



(a) (b) (c)

Fig. 7. An example of crossover operation [2]. Conformations are randomly cut and pasted with the cut point chosen randomly between residues 14 and 15. The first 14 residues of (a) are rotated first and then joined with the last 6 residues of (b) to form (c), where fitness, F = -9. '→' is indicating crossover positions.

GAs optimize the effort of testing and generating new individuals if their representation permits development of building blocks (*schemata*), a concept formalized in the *Schema Theorem* [1, 61−70]. In each generation of a GA, the fitness of the entire population is evaluated by selecting multiple individuals from the current population based on their fitness before crossover is performed to form a new population. The i^{th} chromosome C_i is selected based on the fitness f_i with the proportionate selection $\left(f_i / \bar{f}\right)$, where $\bar{f}$ is the average fitness of the population. Parents then produce off-spring by crossover at a rate p_c for the population of size Pop_z, thus forming the next generation. Mutation is applied on the population of generated off-spring at a rate p_m and the selection probability of any off-spring or chromosome is again $\left(f_i / \bar{f}\right)$. A small percentage, typically between 5% and 10% of elite chromosomes (those having higher fitness), are copied to the next generation to retain potential solutions. The remaining chromosomes (if they exist), which are unaffected by crossover, mutation or elitism operations are then moved to the next generation.

Throughout this article, a short sequence will imply a sequence with $n < 25$ (n indicates the number of residues in a sequence or the protein length), a moderate length will imply $25 \le n < 50$ and long sequences will imply $n \ge 50$.

4.2 Incorporating Intelligence into the GA

The fundamental basis of the GA, the schema theorem, supports that schema fitness with above average values in the population will more likely be sustained as generations proceed and as a consequence the similarity [61, 64, 71−73] of chromosomes grows within the population, thus grows *twins* (same or similar chromo-

somes) leading lower variations within the population. The existence of twins 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 [72]. 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 decline relatively quickly for long converging problems. Also, [61] 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.

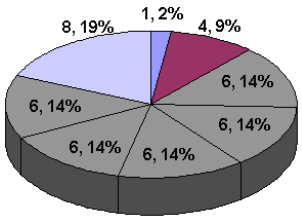


Fig. 8. The probability of a chromosome C_k (with fitness f_k) being selected by roulette wheel selection, is $p_k = f_k / \sum_{i=1}^{n} f_i$. So, for a population of eight chromosomes having fitnesses 8, 6, 6, 6, 6, 6, 4 and 1 for example, the proportionate selection probability of the first chromosome will be $p_1 = (8/43)$, and similarly $p_2 = (6/43)$, …, $p_8 = (1/43)$. The fallacy is, from the pie-chart, we see the fitness 6 occupies 68% in total (assume chromosomes having the same fitness are identical), so the effective selection probability is, $p_{effective_2} = \sum_{i=2}^{6} p_i = 30/43$ or, 70% instead of 14%.

To solve, the need for twin removal was originally highlighted in [73]. The study however, was confined to the detection and removal of identical chromosomes only. Recently, in [71], the notion of *twins* was broadened by introducing *chromosome correlation factor* (CCF) [71] which defines the degree of similarity existing between chromosomes, and it was shown that by removing chromosomes having a similarity value greater than or equal to specific value of CCF during the search process enables the GA to continue seeking potential PSP solutions to provide superior results and helps overcome fallacious effect (see Fig. 8) of the selection procedure.

Table 1. Benchmark protein sequences for 2D HP model

Length	Sequences	Ref.
50	H2(PH)3PH4PH(P3H)2P4H(P3H)2PH4P(HP)3H2	[74]
60	P2H3PH8P3H10PHP3H12P4H6PH2PHP	[74]
64	H12(PH)2(P2H2)2P2HP2H2PPHP2H2P2(H2P2)2(HP)2H12	[74]
85	4H4P12H6P12H3P12H3P12H3P1H2P2H2P2H2P1H1P1H	[75]
100a	6P1H1P2H5P3H1P5H1P2H 4P2H2P2H1P5H1P10H1 P2H1P7H11P7H2P1H1P3H6P1H1P2H	[75]
100b	3P2H2P4H2P3H1P2H1P2H1P4H8P6H2P6H9P1H1P2H1P11H2P3H1P2H1P1 H2P1H1P3H6P3H	[75]

'H' and 'P' in the sequence indicate hydrophobic and hydrophilic amino acid, respectively.

Outcome of the Twin Removal

Simulations were undertaken both with (CCF ≤ 1) and without the twin (WT) removal strategy implemented in the population, with in the former case, the twin removal being performed after the crossover and mutation operations. In every generation, twins were removed in all runs for a range of CCF settings from $r = 1.0$ (identical chromosomes) down to $r = 0.5$ (the least chromosome similarity, i.e., $0.5 \leq$ CCF≤ 1.0) in steps of 0.1. Following twin removal from a population, the gap was filled by randomly generated chromosomes. The default GA parameters [71] for all experiments were set for population size, crossover, mutation and elitism rates as *200, 0.8, 0.05* and *0.05,* respectively, and the 2D square HP lattice model was applied to the various benchmark sequences (Table 1). The corresponding results are displayed in Table 2, indicate that twin removal with $r = 0.8$, i.e., having *80%* and above similarity being removed, has obtained the best performance. Introduction of the twin removal helps improved generically.

Table 2. Run results for 5 iterations of PSP for various sequences using GA. Each iteration has maximum generation = 6000, the average fitness of the runs is shown below.

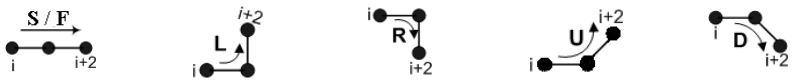
Length	WT	r =1.0	r = 0.9	r = 0.8	r = 0.7	r = 0.6	r = 0.5
60	-29.4	-32.6	-33.4	-33.8	-32.2	-32.4	-32.6
64	-29.4	-34.2	-35	-37	-35.4	-34	-32.2
85	-42.2	-45	-47	-46.8	-46.2	-45	-44.4
100a	-38.6	-39.4	-42.4	-43	-42.4	-42.4	-40.8
100b	-37.4	-40.4	-42.6	-44.8	-42.8	-42.2	-42

4.3 Intelligence in Chromosomal Encoding

The encoding used in the HP lattice models was mainly isomorphic, which add unwanted variations for the same solution (conformation). Non-isomorphic encoding scheme [76] further constrains the search space, aids convergence and similarity comparisons are made easier while applying a twin removal and removes implicit controlling of the crossover and mutation rates (see Fig. 12), thus provides superior results.

Fig. 9. Absolute moves (a) 2D square lattice based representation and (c) 3D cube lattice based representation. (b) Coordinate frame used for encoding.

In the literature, four different encoding strategies have been reported [76]: *i)* Direct coordinate presentation, *ii)* Absolute encoding, *iii)* Relative encoding and *iv)* Non-isomorphic encoding. Rather than using a binary string, preference to use conformations themselves is known as *direct coordinate presentation*.



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

Fig. 10. The relative moves in 3D, namely (a) Straight / Forward (S or F) (b) Left (L) (c) Right (R) (d) Up (U) and (e) Down (D). However, Backward (B) move does not need a self avoiding walk.

Absolute encoding [34, 42, 77−79] replaces the direct coordinate presentation with letters representing directions with respect to the lattice structure. The permitted moves for absolute encoding are: *f* (forward), *l* (left), *r* (right), *b* (back), *u* (up) and *d* (down) (see Fig. 9), while *u* and *d* indicate +*z* and –*z* direction respectively. A conformation *c* in 2D with *n* residues could be $c \in \{f, l, r, b\}^{n-1}$ while in 3D it would be $c \in \{f, l, r, b, u, d\}^{n-1}$. Alternatively, in relative encoding the move direction is defined relative to the direction of the previous move as shown in Fig. 10, rather than relative to the axis defined by the lattice. These moves are lattice automorphic [34], with the initial move always expressed by *F* (forward). A conformation *c* of *n* residues in 2D and 3D could then be $c \in \{F, L, R\}^{n-2}$ and

$c \in \{F, L, R, U, D\}^{n-2}$, respectively. Relative encoding (Fig. 10) was developed with a view to improving presentation over absolute encoding with pivot mutation being represented as the single locus or character alteration of a chromosome as shown in Fig. 11.

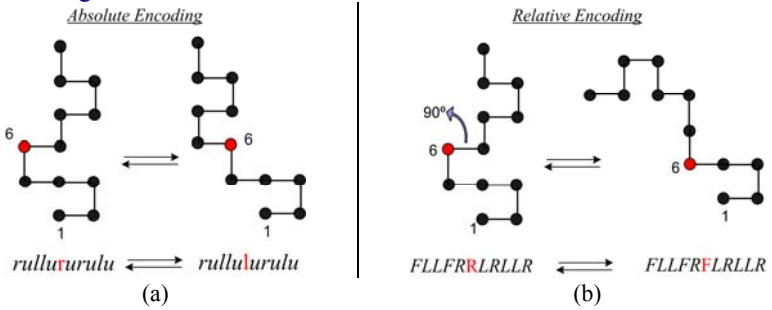


Fig. 11. (a) Single mutation at residue number 6 (red colour) using absolute encoding using changes in genotype and in the corresponding phenotype is not a pivot rotation (b) Single mutation at residue 6 using relative coding results in true pivot rotation.

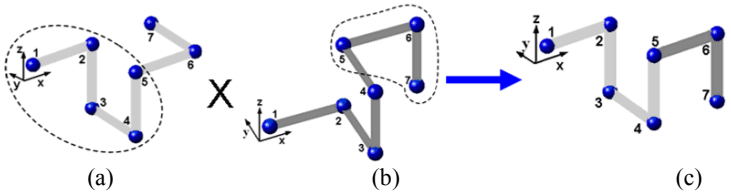


Fig. 12. The cross-exchange indicated by the dotted contour in the identical conformations (a) and (b) result conformation in (c), which can also be constructed from (a) or (b) by applying mutation (i.e. pivot rotation) at residue number 5. Hence, the crossover can result equivalent to the mutation operation for identical parents.

It is clear that the coordinates of a rotating object change so direct coordinate presentation is inherently isomorphic. Moreover as shown in Fig. 13 and Fig. 14, absolute and relative encodings are also isomorphic. Thus, a non-isomorphic encoding algorithm is essentially proposed in [76] by assigning a fixed directions for a growing chain based upon the first occurrences of the move in a particular dimension. The direction from first residue towards the second is marked '1' and the reverse is marked '2', which defines the complete move in 1-dimension. The first occurrence of a direction perpendicular to the 1-dimension is marked as '3' and the reverse as '4', which completes the moves in 2-dimensions. The first occurrence of the move perpendicular to the plane formed by '1' and '3' moves is marked as '5' and the reverse as '6', which finally defines the moves in 3-dimensions.

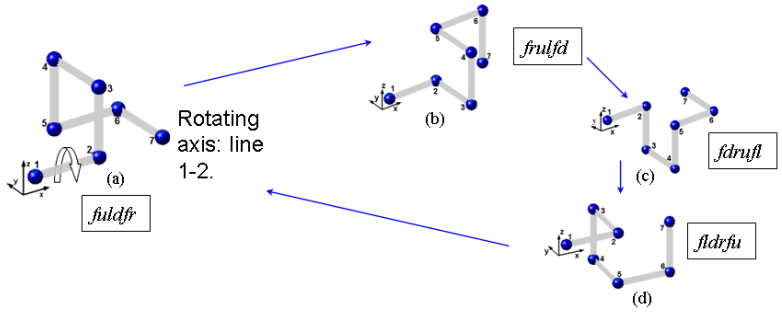


Fig. 13. Absolute encoding is isomorphic. For six directions, namely $+x$, $-x$, $+y$, $-y$, $+z$ and $-z$, $24\,(=6\times4)$ different genotypes are possible for a given 3D conformation.

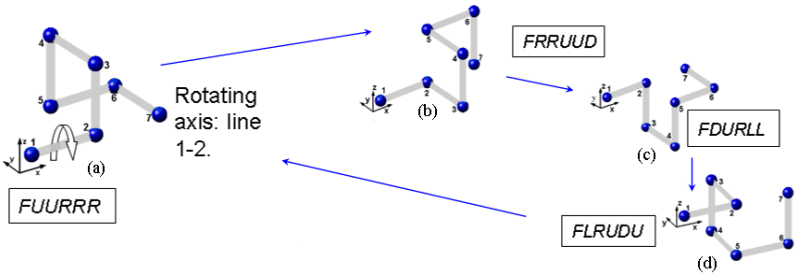
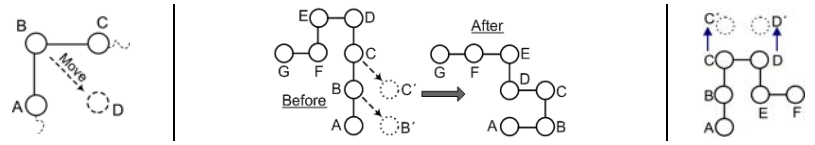


Fig. 14. Relative encoding is isomorphic. Four variations shown in 3D by rotating around axis formed by 1-2 connecting line, but no variation achieved by the change in direction (x or y or z).

4.4 Domain Knowledge Based Heuristic

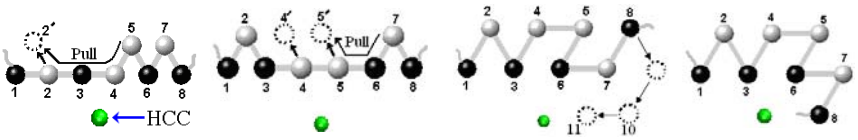
Despite the aforementioned improvements, PSP remains an intractable problem because during the search, the solution becomes phenotypically more compact, thereby increasing the number of collisions [3, 57]. To solve, alternate operators and move sets have also been applied [75]. An operator that is able to move the intended portion of the converging conformation with a predefined target, while concomitantly having minimal impact on the stable portion, exhibits considerable promise. One such operator, short pull move, or pull move was proposed by Lesh in the square 2D lattice model [75], which subsequently extended by Hoque *et al.* [3], with the introduction of the *tilt move,* which is applicable when other moves fail due to congestion. The tilt move however can disturb the stability more than

the pull move. A selective implementation of the move sets based on current scenario could represent a powerful combination such as for instance, firstly attempting a *diagonal move* [3] and if this cannot be performed to reach a predefined goal then next applying a pull move and then a tilt move if the pull move perchance fails. Fig. 15 describes these moves in further detail.



(a) (b) (c)

Fig. 15. Various move operators (a) if 'D' is free, then 'B' can be move to 'D' via a diagonal move. (b) Before and after applying pull move is displayed. In first case 'B' can be pulled to 'B′' if 'C′' is free or 'C' is already at the position of 'C′' and the rest of the chain upto one end can be pulled until a valid conformation is reached. (c) Tilt move, 'C' and 'D' can be moved to 'C′' and 'D′' respectively and pull will propagate towards both ends.

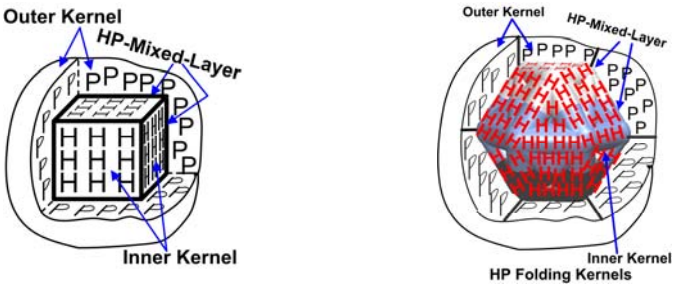


(a) (b) (c) (d)

Fig. 16. The subsequence -123- in (a) need to remap to sub-conformation corresponds to –HPPH-. If the position $2'$ is free then 2 can be placed at $2'$ and a pull (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 leaving (b). The |fitness| in (b) is increased by 1. In (b) assume, $4'$ and $5'$ are free positions and the segment 3 to 6 can be recognized as –PHHP-. To enforce a mapping to highly probable sub-conformation, 4 and 5 can be shifted to $4'$ and $5'$ respectively applying a pull move which results (c). In (c), 8 can pass through position 9, 10, 11 and results (d) and increases |fitness| by 1 further. The position of H-Core centre (HCC) ('●') is the arithmetic mean of the coordinates of all Hs.

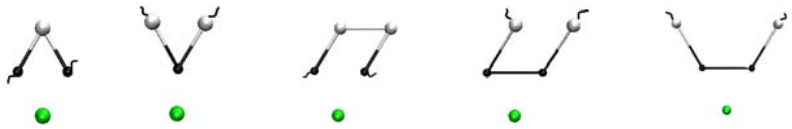
Lesh's experiment demonstrates the superior performance in achieving the minimum energy conformation for longer sequences using the pull move in moving phonotypically compact conformation, but it also provides lessons that random application of the move can consume significant computational resources. Hoque et al, has subsequently proven that incorporating domain specific knowledge [3, 80−82] with the move and their combinations afford considerable promise. As illustrated in Fig. 16, the pull move in both 2D and 3D FCC model helps to improve the fitness. Furthermore, as the parity problem is absent in the FCC model, the pull move does not need to be moved diagonally [81, 82] to start as in an ordinary

pull because with more neighbours, the model is likely to get a valid conformation without the need to propagate the pull often upon the terminal residue.



(a) (b)

Fig. 17. Metaphoric HP folding kernels for (a) Cube 3D Lattice (b) 3D FCC lattice.



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

Fig. 18. Potential sub-conformation in 3D space for the subsequence. Sub-conformation in (a) relates to –HPH- ($1S_p$), (b) relates to –PHP- ($1S_H$) and (c) related to -HPPH- ($2S_P$). Further, both (d) and (e) relate to -PHHP- ($2S_H$). Symbol ●, ○ and ●, respectively indicate an H, a P and the approximate position of HCC.

Further, Hoque *et al.* [3, 80−82] conceptualised the folded protein as a three-layered kernel (Fig. 17). The inner kernel, called the H-Core, is assumed compact and mainly formed of Hs while the outer kernel consists mostly of Ps. The H-Core [83] Centre is named HCC. The composite thin layer between the two kernels consists of those Hs that are covalent-bonded with *P*s and is referred to as the HP-mixed-layer. To integrate domain knowledge, Hoque et al, showed that the optimal core for a square 2D [3], cube 3D [80], 2D FCC [81] and 3D FCC [82] lattice are square, cube and regular hexagon respectively, which concludes the optimal core that maximizes the |fitness| can be predicted based upon the properties and dimension of the model. To form the cavity of H-Core Hoque *et al.* further introduced, motifs or sub-conformations based approach which are highly probable to a sub-sequence (defined in Fig. 13 for 2D FCC) are forced to re-map. The rationale is to form immediate TN and place P as far away as possible from HCC while concomitantly placing H as near as possible to the HCC. For the mapping, two broad categories of sub-sequences are defined; gS_H and gS_P, where $g \in \mathrm{N}$, where N is a 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, e.g. $2S_H$ means -PHHP-, so $g = 2$ with the two boundary residues being of the opposite type. g is divided into even g_e and odd g_o numbers. For $1S_p$, $1S_H$, $2S_P$ and $2S_H$, there are only a few possible sub-conformations, so only highly potential sub-conformations (Fig. 18) are chosen, based on embedded TN and core formation [83, 84] concepts. Collectively they are referred to as *H-Core Boundary Builder Segments* (HBBS) [3] and are mapped to potential sub-conformations which are known as *H-Core Boundary Builder sub-Conformation* (HBBC). HBBC forms part of a corner (especially when $g = 1$ and through the composition with other group having $g = 2$) and an edge (especially when $g = 2$ and with the composition of the former group) of the H-Core boundary. The selection for mapping HBBC into HBBS is probabilistically applied while searching.

Formulation of Multi-Objectivity

Combining the moves with domain knowledge, Hoque *et al*., formulated the prediction into multi-objective optimization [3, 80−82] by combining an additional *probabilistic constrained fitness* (PCF) measure along with the original fitness. When searching for an optimum conformation, if any member of a HBBC corresponds to the related sub-sequence exists PCF rewards otherwise penalizes the search.

Implementation of the Heuristics in a way to Enable Backtracking Capacity

Here aforementioned heuristics are combine strategically as: The conformational search process is divided into two alternative phases namely, *Phase 1* (see (4)) in which F dominates PCF and starts building the core. In the alternate *Phase 2* (see (4)), PCF dominates which covers the formation of an HP-mixed-layer, i.e. the Core boundary. The enforcement of HBBC is also performed in *Phase 2*, since PCF helps to sustain and stabilize any applied change. The HBBC mapping is performed only if they are not found according to the likely sub-conformations for the corresponding sub-sequences. This may reduce the achieved fitness F, but it is expected that it will help reformulate a proper cavity that will maximize the *H* bonding inside the core, while shifting to the favorable *Phase 1* will maximize $|\mathrm{F}|$.

As the phases alternate during the search process (using (3)), the impact becomes such that F and PCF come up with common goal that is more likely to be optimal. The total or combined fitness is defined as:

$$Total\ Fitness = \alpha(t) \times F + \beta(t) \times PCF \quad (2)$$

where t is t^{th} generation while search is carried out by the GA. To adjust the weights α and β to dominate F and PCF over each other, the oscillatory function $\delta(t)$ shown in Fig. 17, is introduced. The setup maintains a variation in the amplitude (A).

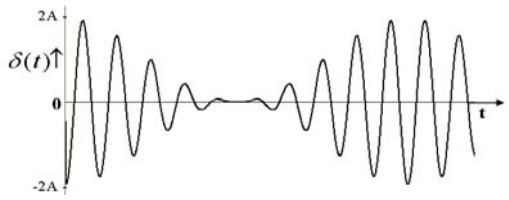


Fig. 17. Plot of $\delta(t)$ function.

$$\delta(t) = A(1 + \cos \omega_m t) \cos \omega_0 t \quad (3)$$

where $\omega_m << \omega_0$ and t = number of generations. The assignment of α and β are as:

$$Phase\ 1\text{: } \alpha(t) = \delta(t), \beta(t) = 1 \text{, when } \delta(t) > 0 \quad (4)$$

$$Phase\ 2\text{: } \alpha(t) = 1, \beta(t) = -\delta(t) \text{, when } \delta(t) < 0 \quad (5)$$

$$Transient\ Phase\text{: } \alpha(t) := 1, \beta(t) := 1 \text{, when } \delta(t) = 0 \quad (6)$$

Typical parameter values for the $\delta(t)$ function (see plot in Fig. 17) were set as follows: A = 30, ω_m = 0.004 and ω_0 = 0.05. The choice of A came from $2A \geq \max\left(|F_l|, |PCF_l|\right)$ where F_l and PCF_l respectively imply the upper bounds of F and PCF, which is predictable from the chosen model. The lower bound of F can be defined by (7) for 2D square and 3D cube HP lattice model and (8) for 2D FCC and 3D FCC model.

$$F_l = -2 * \dim * (\min\{E[Seq], O[Seq]\}) + n_{T_H} \quad (7)$$

$$F_l = -\left(\dim \times n_H + n_{T_H}\right) \quad \textbf{(8)}$$

where in (7), $E[Seq]$ and $O[Seq]$ indicate the number of even and odd indexed H residues in the sequence and n_{T_H} indicates number of terminal H residue, where $0 \le n_{T_H} \le 2$. The value of *dim* in (7) for 2D square and 3D cube HP lattice model is 1 and 2 respectively (and in (8) for 2D FCC and 3D FCC the values are 2 and 5 respectively). The 'min' implies 'minimum of'. The n_H in (8) indicates the total number of hydrophobic residues. Note, the minimum value of both $|\alpha(t)|$ and $|\beta(t)|$ is 1 and so never becomes zero in (4), (5) and (6), thereby preserving the sub-conformation or schema possessing good features, that may have been created in the alternate phase. The search uses a simple GA (SGA) paradigm which is hybridized (see Algorithm I) with the aforementioned move sets, PCF etc with a population size of 200 for all sequences. The elite rate = 0.1, crossover rate = 0.85, mutation rate = 0.5 and single point mutation by pivot rotation was applied. The implementation of both crossover and mutation operations were as in [2], but without any special treatment such as cooling. The *roulette wheel* selection procedure was used.

Algorithm-I: HGA for PSP

Input : Sequence S.
Output: Conformation with best fitness, F.
COMPUTE: PCF, *A*. $t = 0$, $F = 0$ /* *Gen*. count and fitness initialization */
Populate with random (valid) conformations based on *S*.
WHILE NOT Terminate Condition
{ $t = t + 1$, COMPUTE $\delta(t)$, $\alpha(t)$, $\beta(t)$, *TF*
CROSSOVER and then MUTATION
IF $\delta(t) < 0$ THEN
{ FOR i =1 to *population_size* DO
Check $chromosome_i$ for any miss mapping of HBBC based on Model.
IF miss-mapping = TRUE THEN
{Re-map the sub-sequence to corresponding HBBC using move-sets.}}
COMPUTE: *TF*, SORT, KEEP *Elite*
F ← Best fitness found from the population. }
END.

The experiment results were very impressive (see Table 3) and outperformed in all available low resolution models including square 2D [3], cube 3D [80], 2D FCC lattice [81] and 3D FCC lattice model [82]. This general concept is referred to as guided GA [3, 80] or hybrid GA [12, 81], and it importantly provides a intelligent backtracking capability if any local minimum is assumed. Combining HGA with twin removal (as mentioned in Section 4.1) having $r = 0.8$, it was shown in

[82] to obtain best performance over the form of *i*) SGA, *ii*) SGA + $r = 0.8$ and *iii*) HGA - for the 3D FCC lattice model**.**

Table 3. Performance comparison of nondeterministic search approaches [12] using 2D HP square lattice model.

Length / Sequence	GGA [3]	GTB [58]	EMC [57]	GA [2]	MC [2]	CI [85]
20 / (HP)2PH(HP)2(PH)2HP(PH)2	-9	-9	-9	-9	-9	-9
24 / H2P2HP2HP2(HPP)4H2	-9	-9	-9	-9	-9	-9
25 / P2HP2H2P4H2P4H2P4H2	-8	-8	-8	-8	-8	-8
36 / P3(H2P2)2P2H7P2H2P4H2P2HP2	-14	-14	-14	-12	-13	-14
48 / (P2H)2(HP2)2P4H10P6(H2P2)2HP2H5	-23	-23	-23	-22	-20	-23
50 / H2(PH)3PH4PH(P3H)2P4H(P3H)2PH4P(HP)3H2	-21	-21	-21	-21	-21	-21
60 / P2H3PH8P3H10PHP3H12P4H6PH2PHP	-36	-35	-35	-34	-33	-35
64 / 12(PH)2(P2H2)2P2HP2H2PPHP2H2P2(H2P2)2 (HP) 2H12	-42	-39	-39	-37	-35	-40

As there could be a number of possible lattice structure or orientations [34], we next justify the preferred on for PSP problem (in Section 5) and modify the two bead HP model further to improve the prediction in Section 6.

5 Preferred Lattice Structure for PSP

A number of lattice models are used for studying the PSP problem. However, towards preferring a lattice structure or orientation in 3D for effectively mapping the real folded protein, we advocate the preference of the 3D face-centred-cube (FCC) orientation for the following reasons:

i) Based on the full proof of Kepler Conjecture [86], a 3D FCC is proven to be the densest sphere packing orientation. It can provide densest protein core [87] while predicting a protein structure (though the protein core may not necessarily need to be in the most compact form [88]).

ii) In 3D FCC orientation, a residue can have 12 neighbours in a 3D space [82]. Such orientation allows maximum excluded volume for offering densest compactness [3, 80, 81]. Therefore logically inferring, for a region with fixed volume, an FCC model has more option for placing a residue in suitable neighbouring position with respect to another residue than any other lattice models. A rudimentary example is, the FCC model is parity [88] problem 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 [3, 75, 80, 81].

In the FCC orientation, if its 12 neighbours are assumed to be covered with a thin outer layer, the overall structure resembles to a cuboctahedron [3, 80, 82] (see the shape of the inner kernel in Fig. 17 (b)), where a cuboctahedron has 14 faces, 6 of them are square and 8 of them are equilateral triangle and it has 12 corners of vertices.

6 hHPNX – an Extension of the HP Model

For an effective and faster exploration of the PSP landscape, the lattice models are indispensable. However, this crucial HP model (i.e. for interaction potential, see Fig. 18 (a)) having two beads, produces relatively large number of degeneracy [89] (i.e., the chance of different possible conformations but having same ground state energy), consequently which can result in useful conformations being lost in the multitude. Second, the positions of polar segments (i.e. P) are not optimized [90], can result in deformed structures, especially if the segment is too long or located at the end of the sequences. Thus necessarily a modification and an extension to the HP model, keeping simplicity as much as possible, lead to proposing the HPNX model (for interaction potential, see Fig. 18 (b)), where a logical extension of the HP model being proposed [79, 89]. In the HPNX model, the splitting of P (polar) monomer of HP model is actually based on the variations of electric charge, namely positive (P), negative (N) and neutral (X) among amino acids.

	H	*P*
H	-1	0
P	0	0

(a)

	H	*P*	*N*	*X*
H	-4	0	0	0
P	0	1	-1	0
N	0	-1	1	0
X	0	0	0	0

(b)

	h	*H*	*P*	*N*	*X*
h	2	-4	0	0	0
H	-4	-3	0	0	0
P	0	0	1	-1	0
N	0	0	-1	1	0
X	0	0	0	0	0

(c)

Fig. 18. Interaction potential matrixes of (a) HP (b) HPNX [89] and (c) hHPNX model. Negative entry indicates reward for being topological neighbors (TN) in the lattice model, whereas interaction for TN with positive value represents a penalty, '0' indicates neutral (i.e., no) interaction.

However, based on many structural observation of a protein data sets Crippen proposed [91] a new potential interaction matrix as shown in Fig. 19 (a), where the

amino acids where divided into four different groups. Crippen emphasized the small set of particular group for the certainty of their distinguishing properties, namely Alanine (Ala or A) and Valine (Val or V). It has been detected [92] that this particular element of the matrix highlighted in Fig. 19 (a) was converted with few wrong entries by Bornberg [79] as shown in the matrix of Fig. 19 (b). and named YhHX matrix.

	1	*2*	*3*	*4*
1	-0.012	-0.074	-0.054	0.123
2	-0.074	**0.123**	-0.317	0.156
3	-0.054	-0.317	-0.263	-0.010
4	0.123	0.156	-0.010	-0.004

(a)

	Y	*h*	*H*	*X*
Y	0	-1	-1	2
h	-1	**-2**	-4	2
H	-1	-4	-3	0
X	2	2	0	0
fq.	10	16	36	28

(b)

	Y	*h*	*H*	*X*
Y	0	-1	-1	2
h	-1	**2**	-4	2
H	-1	-4	-3	0
X	2	2	0	0
fq.	36	16	20	28

(c)

Fig. 19. (a) Crippen's matrix [91]; classifies amino acid contacts, presented using single letter: 1 = {GYHSRNE}, 2 = {AV}, 3 = {LICMF} and 4 ={PWTKDQ}. (b) YhHX matrix as converted by Bornberg in [79] from Crippen's matrix. Here, *fq.* implies the percentage of occurrence frequencies of amino acid for each of the four groups. (c) Corrected YhHX as it should have been considered in [79]. Blacked and shared entries in (a), (b) and (c) are the problem area.

The emphasised [91] small group {A, V} has highest frequency among proteins on an average compared to the occurrence frequencies of all the amino acids [93], and hence it is important to amend the incorrect conversion of the element (2, 2) of matrix in Fig. 19 (a) to element (2, 2) of matrix in Fig. 19 (b). The element depicts the 'hh' interaction of the YhHX matrix of Fig. 19 (b). Note that h ≡ {A, V}, should have been recorded as '2' instead of this highlighted element being incorrectly shown as '-2' in Fig. 19 (b) which can be easily observed comparing rest of the entries of the original matrix in Fig. 19 (a) with entries of the matrix in Fig. 19 (b). Further, the frequencies, indicated by '*fq.*' and the shaded elements shown in Fig. 19 (b), also need to be swapped. Moreover, the "10%" mentioned in the YhHX matrix needs to be corrected as 20%. The corrected matrix, incorporating all necessary changes, is shown in Fig. 19 (c). To incorporated further the essence of the HPNX model with the aforementioned correction, an hHPNX model has been proposed (see interaction potential, Fig. 18 (c)) [92]. In this hHPNX model basically the H of HP or HPNX model has been split into two by indicating h ≡ {A, V}, leaving the rest of the members of the H group as it is.

To compare, HP, HPNX and hHPNX model, developed HGA (reported in Section 4.3) was applied on sequences taken arbitrarily from Protein Databank (PDB) [94], measuring the models' output using 'alpha-carbon (C_α) root-mean-square-deviation' (cRMSD) [34]. As expected, hHPNX performed the best [92].

6 Conclusions

The *ab inito* protein structure prediction (PSP) is an important yet extremely challenging problem. It urges to involve a considerable amount of computational intelligence. Low resolution or simplified lattice models are very helpful in this regard to explore the search landscape of astronomical size in a feasible time scale. Due to the nature of the complex PSP problem, nondeterministic approaches such as genetic algorithm (GA), especially for its potential operators found to be relatively promising for conformational search. However, even GA often fails to provide reasonable outcome especially for longer sequences and also without the effectiveness in the conformational search in low resolution, the full-fledged prediction, which encompasses low to high resolution modelling in a hierarchal system, would suffer later on. Therefore, a way to improve the nondeterministic search (such as GA) for PSP, has been reviewed in the context of a twin removal within population, intelligent encoding for problem presentation, so on, which become indispensable for providing further effectiveness. Domain knowledge based heuristics are shown very useful. Moreover, in the modelling point of view, simplified model can be made further effective by preferring a lattice orientation, beads and contact potential that can map real folded protein closely possible.

Acknowledgments Support from Australian Research Council (grant no DP0557303) is thankfully acknowledged.

References

1. Unger R, Moult J (1993) On the Applicability of Genetic Algorithms to Protein Folding, The Twenty-Sixth Hawaii International Conference on System Sciences, pp. 715-725.
2. Unger R, Moult J (1993) Genetic Algorithms for Protein Folding Simulations. Journal of Molecular Biology 231 75-81.
3. Hoque M T, Chetty M, Dooley L S (2005) A New Guided Genetic Algorithm for 2D Hydrophobic-Hydrophilic Model to Predict Protein Folding., IEEE Congress on Evolutionary Computation (CEC), Edinburgh, UK.
4. Bonneau R, Baker D (2001) AB INITIO PROTEIN STRUCTURE PREDICTION: Progress and Prospects. Annu. Rev. Biophys. Biomol. Struct. 30 173-189.
5. Chivian D, Robertson T, Bonneau R, Baker D (2003) Ab INITIO METHODS. in: Bourne, P.E., H. Weissig, (Eds.), Structural Bioinformatics, Wiley-Liss, Inc.
6. Samudrala R, Xia Y, Levitt M (1999) A Combined Approach for ab initio Construction of Low Resolution Protein Tertiary Structures from Sequence Pacific Symposium on Biocomputing (PSB) 4 505-516.
7. Corne D W, Fogel G B (2004) An Introduction to Bioinformatics for Computer Scientists. in: Fogel, G.B., D.W. Corne, (Eds.), Evolutionary Computation in Bioinformatics pp. 3-18.
8. Berg J M, Tymoczko J L, Stryer L, Clarke N D (Eds.) (2002) Biochemistry, W. H.

Freeman and Company.
9. Takahashi O, Kita H, Kobayashi S (1999) Protein Folding by A Hierarchical Genetic Algorithm, 4th Int. Symp. AROB.
10. Kuwajima K, Arai M (Eds.) (1999) Old and New Views of Protein Folding, ELESEVIER.
11. Pietzsch J (2007) Protein folding technology, http://www.nature.com/horizon/proteinfolding/background/technology.html, July
12. Hoque M T, Chetty M, Dooley L S (2006) Significance of Hybrid Evolutionary Computation for Ab Inito Protein Folding Prediction in: Grosan, C., A. Abraham,H. Ishibuchi, (Eds.), Hybrid Evolutionary Algorithms, Springer-Verlag, Berlin.
13. Lamont G B, Merkie L D (2004) Toward effective polypeptide chain prediction with parallel fast messy genetic algorithms. in: Fogel, G., D. Corne, (Eds.), Evolutionary Computation in Bioinformatics pp. 137- 161.
14. Guex N, Peitsch M C (2007) Principles of Protein Structure: Comparative Protein Modelling and Visualisation. http://swissmodel.expasy.org/course/course-index.htm,April.
15. Jones D T, Miller R T, Thornton J M (1995) Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins: Structure, Function, and Genetics 23 387-397.
16. Sánchez R, Šali A (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. PNAS 95 13597-13602.
17. Jones D T (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology 287 797-815.
18. Wikipedia (2007) De novo protein structure prediction, http://en.wikipedia.org/wiki/De_novo_protein_structure_prediction, July.
19. Xia Y, Huang E S, Levitt M, Samudrala R (2000) Ab Initio Construction of Protein Tertiary Structures using a Hierarchical Approach. J. Mol. Biol. 300 171-185.
20. Anfinsen C B (1972) Studies on the Principles that Govern the Folding of Protein Chains, http://nobelprize.org/nobel_prizes/chemistry/laureates/ ... .
21. Levinthal C (1968) Are there pathways for protein folding? Journal of Chemical Physics 64 44-45.
22. Backofen R, Will S (2006) A Constraint-Based Approach to Fast and Exact Structure Prediction in Three-Dimensional Protein Models. Constraints Journal 11.
23. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D (2005) Progress in Modeling of Protein Structures and Interactions. Science 310 638-642.
24. Hirst J D, Vieth M, Skolnick J, Brook C L (1996) Predicting leucine zipper structures from sequence. Protein Engineering 9 657-662.
25. Roterman I K, Lambert M H, Gibson K D, Scheraga H (1989) 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 7 421-453.
26. Cornell W D, Cieplak P, Bayly C I, Gould I R, Jr K M M, Ferguson D M, Spellmeyer D C, Fox T, Caldwell J W, Kollman P A (1995) A second generation force field for the simulation of proteins and nucleic acids. J. Am. Chem. Soc. 117 5179-5197.
27. Nemethy G, Gibson K D, Palmer K A, Yoon C N, Paterlini G, Zagari A, Rumsey S, Scheraga H A (1992) Improved geometrical parameters and non-bonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. Journal of physical chemistry 96 6472-6484.
28. Heureux P L et al., (2002) Knowledge-Based Prediction of Protein Tertiary Structure. Computational Methods for Protein Folding: Advances in Chemical Physics 120.
29. Ercolessi F (1997) A molecular dynamics primer, ICTP, Spring College in Computational Physics.

30. Schlick T (2002) Molecular Modeling and Simulation, Springer.
31. Stote R (2006) Theory of Molecular Dynamics Simulations, http://www.ch.embnet.org/MD tutorial/, March.
32. Dill K A (1985) Theory for the Folding and Stability of Globular Proteins. Biochemistry 24 1501-1509.
33. Dill K A, Bromberg S, Yue K, Fiebig K M, Yee D P, Thomas P D, Chan H S (1995) Principles of protein folding – A perspective from simple exact models. Protein Science 4 561-602.
34. Backofen R, Will S, Clote P (2000) Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. Pacific Symp. On Biocomputing 5 92-103.
35. Schöppe G,Heermann D W (1999) Alternative off-lattice model with continuous backbone mass for polymers. Physical Review E 59 636-641.
36. Chen M, Huang W (2006) Heuristic Algorithm for off-lattice protein folding problem. Journal of Zhejiang Univ Science B 7 7-12.
37. Skolnick J, Kolinski A (2002) A Unified Approach to the prediction of Protein Structure and Function. Computational Methods for Protein Folding: Advances in Chemical Physics 120.
38. Kolinski A, Gront D, Kmiecik S, Kurcinski M, Latek D (2006) Modeling Protein Structure, Dynamics and Thermodynamics with Reduced Representation of Conformational Space. John von Neumann Institute for Computing (NIC) Series 34 21-28.
39. Duan Y, Kollman P A (2001) Computational protein folding: From lattice to all-atom. IBM Systems Journal 40.
40. Allen F et al., (2001) Blue Gene: A vision for protein science using a petaflop supercomputer. IBM System Journal 40.
41. Germain R S, Fitch B, Rayshubskiy A, Eleftheriou M, Pitman M C, Suits F, Giampapa M,Ward T J C (2005) Blue Matter on Blue Gene/L: Massively Parallel Computation for Bio-molecular Simulation, ACM.
42. Shmygelska A, Hoos H H (2005) An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinformatics 6.
43. Chivian D, Kim D E, Malmström L, Schonburn J, Rohl C A,Baker D (2005) Prediction of CASP6 Structures Using Automated Robetta Protocols. PROTEINS: Structure, Function, and Genetics 7 157-166.
44. Hung L-H, Samudrala R (2003) PROTINFO: secondary and tertiary protein structure prediction. Nucleic Acids Research 31 3296-3299.
45. Hung L H, Ngan S C, Liu T, Samudrala R (2005) PROTINFO: new algorithms for enhanced protein structure predictions. Nucleic Acids Research 33 77–80.
46. Zhang Y, Arakaki A K, Skolnick J (2005) TASSER: An Automated Method for the Prediction of Protein Tertiary Structures in CASP6. PROTEINS: Structure, Function, and Bioinformatics 7 91–98.
47. Baker D (2000) A surprising simplicity to protein folding. Nature 405 39-42.
48. Baker D (2006) Prediction and design of macromolecular structures and interactions. Phil. Trans. R. Soc. B 361 459–463.
49. Zhang Y (2006) Protein Structure Prediction by I-TASSER at CASP7.
50. Crescenzi P, Goldman D, Papadimitriou C, Piccolboni A, Yannakakis M (1998) On the complexity of protein folding (extended abstract), the second annual international conference on Computational molecular biology, ACM, pp. 597-603.
51. Berger B, Leighton T (1998) Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-Complete. Journal of Computational Biology 5 27-40.
52. Chen M, Lin K Y (2002) Universal amplitude ratios for three-dimensional self-avoiding walks. Journal of Physics A: Mathematical and General 35 1501-1508.
53. Schiemann R, Bachmann M, Janke W (2005) Exact Enumeration of Three – Dimen-

sional Lattice Proteins. Computer Physics Communications, Elsevier Science. 166.
54. MacDonald D, Joseph S, Hunter D L, Moseley L L, Jan N, Guttmann A J (2000) Self-avoiding walks on the simple cubic lattice. Journal of Physics A: Mathematical and General 33 5973-5983.
55. Guttmann A J (2005) Self-avoiding walks in constrained and random geometries, Elsevier.
56. Bastolla U, Frauenkron H, Gerstner E, Grassberger P, Nadler W (1998) Testing a new Monte Carlo Algorithm for Protein Folding. National Center for Biotechnology Information 32 52-66.
57. Liang F, Wong W H (2001) Evolutionary Monte Carlo for protein folding simulations. J. Chem. Phys 115.
58. Jiang T, Cui Q, Shi G, Ma S (2003) Protein folding simulation of the hydrophobic-hydrophilic model by computing tabu search with genetic algorithms, ISMB, Brisbane Australia.
59. Unger R, Moult J (1993) Genetic Algorithm for 3D Protein Folding Simulations, 5th International Conference on Genetic Algorithms, pp. 581-588.
60. König R, Dandekar T (1999) Refined Genetic Algorithm Simulation to Model Proteins. Journal of Molecular Modeling 5.
61. Michalewicz Z (1992) Genetic Algorithms + Data Structures = Evolution.
62. Holland J H (2001) Adaptation in Natural And Artificial Systems The MIT Press, Cambridge, Massachusetts London, England.
63. Schulze-Kremer S (1996) Genetic Algorithms and Protein Folding
64. Whitley D (2001) An Overview of Evolutionary Algorithms. Journal of Information and Software Technology 43 817-831.
65. Goldberg D E (1989) Genetic Algorithm Search, Optimization, and Machine Learning Addison-Wesley Publishing Company.
66. Vose M D (1999) The Simple Genetic Algorithm, The MIT Press, Cambridge, Massachusetts London, England.
67. Fogel D B (2000) EVOLUTIONARY COMPUTATION Towards a new philosophy of Machine Intelligence, IEEE Press.
68. Davis L (1991) Handbook of Genetic Algorithm, VNR, New York.
69. Yao X (1999) EVOLUTIONARY COMPUTATION Theory and Application, World Scientific.
70. Wikipedia (2007) Genetic Algorithm, http://en.wikipedia.org/wiki/Genetic_algorithm, July.
71. Hoque M T, Chetty M, Dooley L S (2007) Generalized Schemata Theorem Incorporating Twin Removal for Protein Structure Prediction, Pattern Recognition in Bioinformatics, LNBI 4774, Springer, Singapore, 84–97.
72. Haupt R L, Haupt S E (2004) Practical Genetic Algorithms.
73. Ronald S (1998) Duplicate Genotypes in a Genetic algorithm. IEEE World Congress on Computational Intelligence 793-798.
74. Hart W E, Istrail S (2005) HP Benchmarks, http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html, August.
75. Lesh N, Mitzenmacher M, Whitesides S (2003) A Complete and Effective Move Set for Simplified Protein Folding, RECOMB, Berlin, Germany.
76. Hoque M T, Chetty M, Dooley L S (2006) Non-Isomorphic Coding in Lattice Model and its Impact for Protein Folding Prediction Using Genetic Algorithm, IEEE Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, Toronto, Canada.
77. Patton A L, Punch W F, Goodman E D (1995) A Standard GA approach to Native Protein Conformation Prediction, 6th International Conference on Genetic Algorithms,

pp. 574-581.
78. Krasnogor N, Hart W E, Smith J, Pelta D A (1999) Protein Structure Prediction With Evolutionary Algorithms, Genetic and Evolutionary Computation Conference (GECCO-99).
79. Bornberg-Bauer E (1997) Chain Growth Algorithms for HP-Type Lattice Proteins, RECOMB, Santa Fe, NM, USA.
80. Hoque M T, Chetty M, Dooley L (2006) A Guided Genetic Algorithm for Protein Folding Prediction Using 3D Hydrophobic-Hydrophilic Model, Special session in WCCI / IEEE Congress on Evolutionary Computation (CEC).
81. Hoque M T, Chetty M, Dooley L S (2006) A Hybrid Genetic Algorithm for 2D FCC Hydrophobic-Hydrophilic Lattice Model to Predict Protein Folding, 19th ACS Australian Joint Conference on Artificial Intelligence, LNAI, Springer.
82. Hoque M T, Chetty M, Sattar A (2007) Protein Folding Prediction in 3D FCC HP Lattice Model Using Genetic Algorithm Bioinformatics special session, IEEE Congress on Evolutionary Computation (CEC). Singapore.
83. Yue K, Dill K A (1993) Sequence-structure relationships in proteins and copolymers Phys. Rev. E 48 2267 - 2278.
84. Bonneau R, Strauss C, Baker D (2001) Improving the Performance of Rosetta Using Multiple Sequence Alignment Information and Global Measures of Hydrophobic Core. PROTEINS: Structure, Function, and Genetics 43 1-11.
85. Toma L, Toma S (1996) Contact interactions methods: A new Algorithm for Protein Folding Simulations. Protein Science 5 147-153.
86. Backofen R, Will S (2005) A Constraint-Based Approach to Fast and Exact Structure Prediction in Three-Dimensional Protein Models. Kluwer Academic Publishers.
87. Raghunathan G, Jernigan R L (1997) Ideal architecture of residue packing and its observation in protein structures. Protein Sci. 10 2072-2083.
88. Wikipedia (2007) Cuboctahedron http://en.wikipedia.org/wiki/Cuboctahedron, February.
89. Backofen R, Will S, Bornberg-Bauer E (1999) Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets Bioinformatics 15 234-242.
90. Guo Y Z, Feng E M, Wang Y (2006) Exploration of two-dimensional hydrophobic-polar lattice model by combining local search with elastic net algorithm J. Chem. Phys. 125.
91. Crippen G M (1991) Prediction of Protein Folding from Amino Acid Sequence over Discrete Conformation Spaces. Biochemistry 30 4232-4237.
92. Hoque M T, Chetty M, Sattar A (2007) Extended HP model for Protein Structure Prediction. JOURNAL OF COMPUTATIONAL BIOLOGY 16 1-19.
93. Jordan I K, Kondrashov F A, Adzhubei I A, Wolf Y I, Koonin E V, Kondrashov A S, Sunyaev S (2005) A universal trend of amino acid gain and loss in protein evolution. Letter to Nature 433.
94. PDB (2007) Protein Data Base, http://www.rcsb.org/pdb/, April.