Neurocomputing 73 (2010) 2308–2316

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

DFS-generated pathways in GA crossover for protein structure prediction

Md Tamjidul Hoque [a,b,*], Madhu Chetty [c], Andrew Lewis [b], Abdul Sattar [b], Vicky M. Avery [a]

[a] Discovery Biology, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan, QLD 4111, Australia
[b] Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Nathan, QLD 4111, Australia
[c] Gippsland School of Information Technology (GSIT), Monash University, Churchill, VIC 3842, Australia

ARTICLE INFO

Available online 1 June 2010

Keywords:
Depth-first search
Protein structure prediction
Genetic algorithm
HP lattice
HCP model

ABSTRACT

Genetic algorithms (GAs), as nondeterministic conformational search techniques, are promising for solving protein structure prediction (PSP) problems. The crossover operator of a GA can underpin the formation of potential conformations by exchanging and sharing potential sub-conformations. However, as the optimum PSP conformation is usually compact, the crossover operation may result in many invalid conformations (by having non-self-avoiding walks). Although a crossover-based converging conformation suffers from limited pathways, combining it with depth-first search (DFS) can partially reveal potential pathways and make an invalid crossover valid and successful. Random conformations are frequently applied for maintaining diversity as well as for initialization in many GA applications. The random-move-only-based conformation generator has exponential time complexity in generating random conformations, whereas the DFS-based random conformation generator has linear time complexity and performs relatively faster. We have performed extensive experiments using popular 2D, as well as useful 3D, models to justify our hypothesis empirically.

1. Introduction

Protein structure prediction (PSP) is a problem of determining the natural state of a protein from its primary structure and is of great importance because the three dimensionally folded structures determine the biological function [1] and hence prove very useful in applications such as drug design [2]. Particular folded structures are essential for the functioning of living cells as well as for providing body structure. High-resolution protein modeling is possible, provided a homologue of the target protein exists [3]. The application of the high-resolution model becomes less effective without a homologous template. However, the homologous template is unable to explain how and why a protein adopts a specific structure. Thus, low resolution *ab initio* [4] or *de novo* modeling becomes essential. In an *ab initio* approach, the building of a 3D conformation (structure) is essentially based on the properties of amino acids, since protein is a three dimensionally folded molecule composed of amino acids [5] linked together in a particular order (called the primary structure) specified by the DNA sequence of a gene [6]. In this paper, our efforts are to investigate the *ab initio* protein structure prediction problem.

Lattice protein models introduced by Dill [7] are widely used for investigating the underlying principles of protein folding [8]. Protein conformation as a *self-avoiding walk* in the lattice model has been proven to be *NP-complete* for the 2D square and 3D cube HP models [9,10]. Therefore, a deterministic algorithm for folding prediction is not feasible. Reasonably, a nondeterministic approach with robust strategies that can extract minimal energy conformations efficiently from these models becomes necessary. Nevertheless, this is a very challenging task as there exists an astronomical number of possible conformations even for a very short sequence of amino acids [11,12].

Due to its superior performance the genetic algorithm (GA), with crossover as one of its key operations [13], is often chosen as a vehicle for providing solutions to the PSP problems. Not only within the GA itself but also in many other PSP solving algorithms the core concepts of GAs and their components are often adapted for effectiveness [14–18]. While crossover can be very effective in joining two different potential sub-conformations, it can be repeatedly unsuccessful in converging. This is because, the conformation being compact in nature, it leaves limited pathways available for a valid (i.e., self-avoiding-walk) conformation, thereby causing many potential sub-conformations to be lost. This motivates us to apply partial pathways, based on depth-first search (DFS) [19], to regain potential sub-conformations, leading to effective algorithms and superior conformations, resulting in better PSP solutions.

* Corresponding author at: Discovery Biology, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan, QLD 4111, Australia.
Tel.: +61 7 37356040; fax: +61 7 37356001.
E-mail address: tamjidul.hoque@gmail.com (M.T. Hoque).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.02.021

2. Background and preliminaries

In nature, a protein folds quickly, generally requiring between a tenth of a millisecond and one second, whereas any algorithm on any modern computer is still unable to simulate this task in anything close to real time [13,20]. Current research confronted with the immense complexity of the protein structure prediction problem has led to the manifestation of several important issues and approaches, which are yet to be investigated [13,21,22].

First, the energy function, which is a combination of several factors that determine the free energy of a folded protein, is not fully understood. Therefore, existing formulations for energy functions do not suggest any obvious path to a solution for the PSP problem.

Second, conformational search algorithms are promising approaches for solving this hard optimization problem. However, the PSP problem still needs considerable research to find an effective algorithm. The aim of the search is to identify an optimum conformation within a very vast and convoluted search landscape.

Third, Cyrus Levinthal observed, in what is popularly known as the Levinthal paradox, that proteins fold into their specific 3D conformations in a time span far shorter than it would be possible for the molecule to actually search the entire conformational space for the lowest energy state [23]. As the process of folding constrains proteins to a subset of intermediate conformations (for example, obviously no sequence of intermediate structures can include non-SAW states) the existence of folding pathways, consisting of a set of more probable, possible intermediate conformations, can be postulated.

While focusing on the second issue [24–29], we utilize DFS strategies, developing novel search algorithms in a form to address the pathway hypothesis. The energy landscape of protein folding is very convoluted. Macroscopically it resembles a funnel-shaped energy landscape, where at any point of the surface the energy of conformation drops increasingly with decreasing search space. This is reflected in structure prediction as subsequent crossover failure because the converging, compacting conformations face more collisions between sub-conformations during attempted crossover operations. This can result in the rejection of the potentially useful sub-conformations as being unfit when paired with the available counterpart of the crossover portion (from a dissimilar conformation). Thus, it has been concluded that conformational searching is a major bottleneck in protein folding prediction, with the observed folding rates found to be proportional to the number of microscopic folding routes [30].

2.1. HP lattice model

The simplified HP lattice model [31,32] is based on *hydrophobicity* [33], dividing the amino acids into two different beads—*hydrophobic* (H) and *hydrophilic* (or *polar* (P)). The model allows HP protein sequences to be configured as self-avoiding walks (SAWs) on the lattice path, favoring a low free energy state due to HH interaction. The energy of a given conformation is defined as the number of *topological neighbor* (TN) contacts between those Hs that are not adjacent in the sequence. This contact between two neighboring H residues (or HH contact) is a TN and is assigned a value for the potential, termed *interaction potential,* which is defined as -1 for the HP model. The value is chosen to be negative, since each protein's stable folded state is assumed to correspond to the global minimum free energy [34]. Further, the HP and PP interaction potentials are assigned values of 0, which basically implies that there is no interaction between an H and a P of HP contact or between the Ps of PP contacts.

To define PSP formally, assume for an amino-acid sequence $s=s_1,s_2,s_3,\ldots,s_n$, a conformation c needs to be formed where $c^*\in C(s)$, $C(s)$ is the set of all valid (i.e., SAW) conformations of s, n is the total number of amino acids in the sequence and the energy of the conformation is $E^*=(C)=\min\{E(c)|c\in C\}$ [17]. If the number of TNs (for HH contact) in a conformation c is q then the value of $E(c)$ is defined as $E(c)=-1\times q=-q$ and the *fitness function* is $F=-q$. The optimum conformation will have a maximum possible value of $|F|$. With respect to the configurations, the lattice model can be of many types; however 2 different useful forms of configuration will be discussed in Sections 2.1.1 and 2.1.2.

2.1.1. 2D square configuration

A 2D square lattice model (Fig. 1) is widely used within the research community [13,25,31,32,35–41]. Easy to implement, many search algorithms are developed using the 2D square HP lattice model. Any algorithm development in the same configuration allows validation and comparison of new techniques for protein structure prediction [24–26,28,29,42].

In a 2D HP square lattice model, a non-terminal and a terminal residue, each with 4 neighbours, can have a maximum of 2 TNs and 3 TNs, respectively, and the model will have a parity problem. However, results for a 2D square HP lattice model are provided to allow easy comparison and then the algorithm developed has been extended for the 3D hexagonal close packed (HCP) HP configuration as described in Section 2.1.2.

2.1.2. 3D hexagonal close packed (HCP) configuration

In a 3D HCP lattice configuration, a residue can have 12 nearest neighbours (see Fig. 2). As can be seen from Fig. 2, the distributions of the neighbours follow the following arrangement: 6 are in the same plane, 3 are above in an upper plane and the remaining 3 are below in the lower plane. Any two adjacent residues are always a unit distance apart [38,47].

By placing the centre residue at the origin (0, 0, 0), the 6 neighbours form a uniform hexagon with their x, y and z coordinates given by $(1,0,0)$, $(1/2,\sqrt{3},2,0)$, $(-1/2,\sqrt{3}/2,0)$, $(-1,0,0)$, $(-1/2,-\sqrt{3}/2,0)$ and $(1/2,-\sqrt{3}/2,0)$. The coordinates of the upper layer

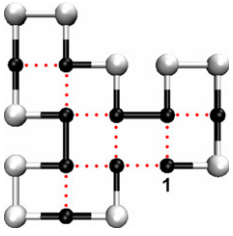

Fig. 1. HP conformation in the 2D HP square model shown by a solid line. 2D square lattice having fitness $=-$(TN Count)$=-9$. ● indicates a hydrophobic and ● indicates a hydrophilic residue. The dotted line indicates a TN. Starting residue is indicated by a '1' in the figure.

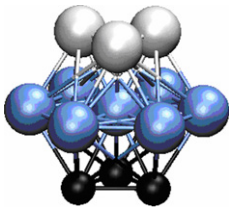

Fig. 2. Hexagonal close packed arrangement. To distinguish the layers are separated by both colours and size, and dedicated connections are used to assist the visualisation of the concept.

of 3 residues are $(0, 1/2, \sqrt{3}/2)$, $(-\sqrt{3}/4, -1/4, \sqrt{3}/2)$ and $(\sqrt{3}/4, \sqrt{3}, -1/4, \sqrt{3}/2)$. For the lower 3 residues the coordinates will be $(0, -1/2, -\sqrt{3}/2)$, $(\sqrt{3}/4, 1/4, \sqrt{3}/2)$ and $(-\sqrt{3}/4, 1/4, -\sqrt{3}/2)$.

The algorithms developed for a 2D HP model are extended to the 3D HCP model in this article. This is done to generate a realistic sample, which can be fed into a high-resolution model in a hierarchical manner [43]. The choice of an HCP model for the 3D lattice is based on the following reasons [44]:

(i) 3D HCP is the densest sphere packing lattice configuration for spheres of equal size [45]. Thus, it *can* provide the most compact protein core [38,39]. However, in the physical structure the protein adopts in nature, the core may not necessarily be the most compact one. The 3D HCP configuration is also parity problem free [46].
(ii) Following (i), for a region of fixed volume of space HCP can offer the highest degree of freedom for placing a residue in a suitable neighboring position.
(iii) Following (ii), HCP can provide the most realistic discrete mapping of the folded protein.

We will continue using the 2D square HP model to explain relevant concepts in subsequent sections. Discussions related to 3D HCP will re-appear with the simulation results in Section 3 on experiments.

2.2. Complexity of the lattice model

Even if we use a simplified lattice model and the sequence length is short, we have an inordinate number of valid (i.e., SAW) conformations [11,12,48]. For instance, in a 2D HP model, for a sequence of n amino acids the number of valid conformations is proportional to μ^n, where the connective constant or the effective coordinate number μ is lattice dependent [12]. Prediction of the optimal conformation using the lattice model is also an NP-complete problem [9,10]. To predict the backbone conformation of the folded protein from its amino-acid sequence based on global interactions such as hydrophobicity, lattice models are used for approximation [31,32,35–37]. For *ab initio* prediction in *critical assessment of structure prediction* (CASP) [35–37], most successful approaches followed a hierarchical paradigm where the lattice-based, backbone conformational sampling works very effectively at the top of the hierarchy. Advancing toward all-atom or full modeling from the lattice, the energy functions include atom-based potentials from molecular mechanics packages such as CHARMM, AMBER and ECEPP [49,50]. Conformational search algorithms built on lattice models, which play a key role in solving PSP, are discussed next.

2.3. Rationale of low resolution model

At the present time, due to the immense computational complexity, a high-resolution model can be applied only when a homologue of the target protein exists. Even when used it lacks the ability to answer how and why a protein adopts its specific structure [3,4]. Thus, *ab initio* modeling is essential for a complete solution to the PSP problem, including the investigation of the physicochemical principles of protein folding, as recent data indicate that the fundamental physics underlying the folding process may be simpler than was previously thought [21]. Many changes in amino-acid sequence often do not vary the overall topology of a protein, which suggests that folding mechanisms depend more on the low resolution geometrical properties of the native state [21] and therefore a simplified model can be applied to understand the physical principles governing the folding processes [51].

Low resolution models can potentially be used to produce an initial approximation of the protein structure, a folding rate that depends on the height of the free-energy barrier, effects of mutations on the folding rate that depend on the region(s) of the protein ordered near the top of the barrier and so on, which in turn allows insight into how adaptation and selection operate among large collections of sequence to structure mappings. Also, an initial coarse sampling of the energy landscape makes the conformational search feasible as well as faster [4]. Thus low resolution models have been found to be promising [21,52–54].

2.4. Nondeterministic conformational search algorithms

For solving *ab initio* PSP using the lattice model, numerous nondeterministic approaches have been investigated: Monte Carlo (MC) simulation, evolutionary MC (EMC) [14,15], simulated annealing (SA), tabu search with genetic algorithm (GTB) [16], ant colony optimization [17] and immune algorithm (IA) based on an artificial immune system (AIS) [55]. Due to their simplicity and search effectiveness, genetic algorithms (GAs) [13,28,34,56–59] are the most attractive. They also provided superior performance over MC [57,58]. The concepts of GAs are also widely adapted within these algorithms. For instance, a new MC algorithm [14] adopted the population-based cut-and-paste (i.e., crossover) operation to achieve higher fitness. The evolutionary Monte Carlo (EMC) [15] algorithm incorporated the evolutionary features of genetic algorithms, such as a population that is updated by crossover and mutation operations. Jiang et al. [16] applied the GA with a tabu search (GTB) to solve PSP using lattice models. Also, the conformational space annealing (CSA) algorithm [18,60] is based on a concept similar to that of GA, where a "bank" in CSA is equivalent to the "population" in GA.

2.5. Focus of the paper

Given the widespread adoption of GAs for PSP, improving the essence of GA, i.e., its crossover operation, by combining it with DFS, may have a positive impact on solving the PSP problem. In solving PSP using a conventional GA, where the optimum conformation is mostly physically compact (see Fig. 3), a crossover-based converging conformation suffers from limited pathways and the algorithm thus increasingly generates invalid conformations. Our hypothesis is that the combination of DFS with crossover can instead reveal potential pathways in solving PSP. Thus, by using DFS a repeatedly failing crossover having congested but potentially useful sub-conformation can be allowed for a limited number of pathways as a possible candidate for crossover counterparts, if there exists at least one path.

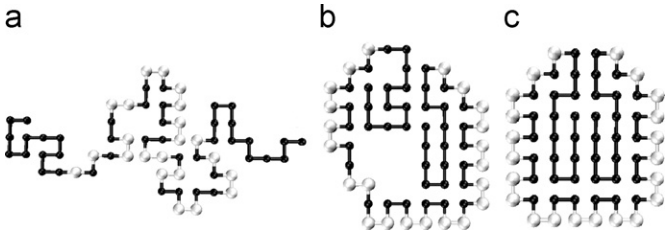

Fig. 3. As the search proceeds the conformation gets more compact. For a typical run, conformations at generation 1, 1434 and 5646 have been shown in (a)–(c) respectively, showing the fitter conformation is relatively more compact: (a) fitness = −15, (b) fitness = −33 and (c) fitness = −42.

2.6. Defining the GA operators for PSP problem

Here, we define the GA operators for the PSP problem based on the HP lattice model:

Crossover operation: for PSP, this aids the construction of global solutions by the cooperative combination of many local substructures [13]. We particularly follow the commonly used crossover operation pioneered by Unger and Moult [57], as illustrated in Fig. 4, a single-point crossover. We follow this single-point crossover since otherwise the converging conformation, being compact in nature, would generate more collisions or invalid conformations [15]. In addition, the ability to rotate before joining within the crossover provides a mutation-equivalent operation especially when *relative encoding* is followed [42]. For example, if we emulate the crossover in Fig. 4 without the rotation, we can write using relative encoding that:

Crossover (a:'LFLLRRLRLLFLRFRLFL', b:'RFFFRFRFLFLRFRLLFL') →would output, c': 'LFLLRRLRLLFL*RLLFL' without the rotation before joining. (Here, '*' indicates an undefined move in relative encoding and here it indicates a non-SAW move.) But, with rotation, the conformation can have SAW, i.e., c: 'LFLLRRLRLLFLRRLLFL'.

Comparing c':'LFLLRRLRLLFL*RLLFL' and c:'LFLLRRLRLLFLRRLLFL', it becomes clear that the '*' is replaced by an 'R' after the rotation, which is genotypically a single-point mutation.

Crossover failure: this implies that before joining two parts, all possible rotated positions at the joining point have been tried but failed to produce at least one valid conformation (i.e., a SAW).

Combination of crossover and DFS: for generating a conformation this implies that a DFS-generated random and partial path has been joined with the first half of the sub-conformation.

DFS after crossover failed: this implies that 'combination of crossover and DFS' has been performed after an occurrence of 'crossover failure'.

Mutation operation: this involves pivot rotation (Fig. 5) as basically pioneered by Unger and Moult [57]. We employ single-point mutation to avoid more collisions.

Ordinary random conformation generation: this implies the generation of a SAW conformation based on random-move-only (RMO). In a 2D square lattice model *left*, *right* and *forward* moves are permissible but *backward* moves are prohibited. For a conformation, once a path search has failed after looking in the three possible degrees of freedom the whole process restarts.

Random conformation generation by DFS: this implies that we apply DFS to generate a SAW conformation. As the DFS proceeds, it stores the possible pathways using a stack memory [19] and, on total failure after trying all possible degrees of freedom on a particular location (i.e. lattice point), it can backtrack to restart from the stored options instead of restarting the creation of the whole conformation.

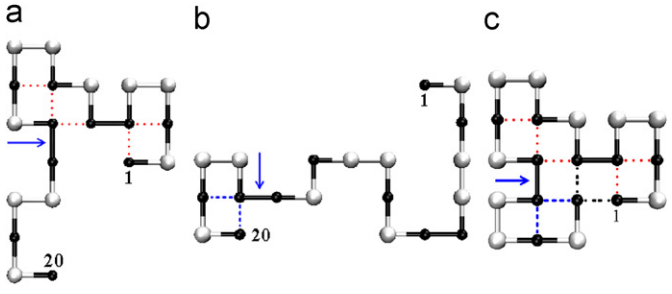

Fig. 4. An example of the crossover operation [57]. Conformations are randomly cut and pasted with the cut point chosen randomly between residues 14 and 15. The first 14 residues of (a) are rotated first as needed (as allowed by the degree of freedom of the model configuration) and then joined with the last 6 residues of (b) to form (c), where fitness, $F=-9$; '→' indicates the crossover positions.

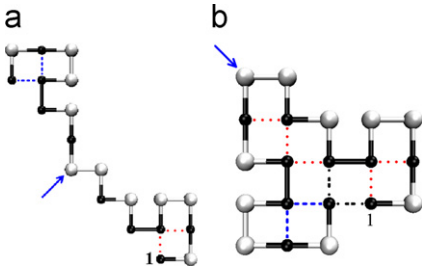

Fig. 5. An example of the mutation operation [57]. Dotted lines indicate TN. Residue number 11 is chosen randomly as the pivot. For the move to apply, a 180° rotation (among a number of possible degree of freedom defined by the model configuration) alters (a) with $F=-4$ to (b) $F=-9$; '→' indicates the mutation residue.

1. Initialize fixed size current population (Pop_z) of randomly generated conformations.
2. Obtain new solution (S_{new}) from the current population by using ***Crossover*** and ***Mutation*** operations at the pre-specified rates (p_c and p_m respectively).
3. Assess quality or fitness, F, of S_{new} .
4. Promote the obtained S_{new} , and elite and untouched chromosomes, to the next generation and assign the new generation as the current population.
5. IF END-OF-SOLUTION is not reached THEN repeat from Step 2.

Fig. 6. Genetic algorithm for solving PSP problem. Terms in *italic* is explained in Section 2.6.

3. Experiments and results

We carried out experiments to empirically verify our hypothesis that combining DFS with crossover will be advantageous. The simple GA (SGA) applied for PSP is outlined in Fig. 6 and the crossover variations with the possible implementation have been shown in Fig. 7.

As shown in Fig. 7, we have experimented with four variations of the crossover operation. *Crossover* (a) (see Fig. 7(a)) represents a conventional crossover operation for PSP without DFS. *Crossover* (b) (see Fig. 7(b)) applies DFS-based partial path generation with the sub-conformation, immediately the sub-conformation fails to join with its counterpart sub-conformation after trying all possible degrees of freedom. *Crossover* (d) (see Fig. 7(d)) is similar to *Crossover* (b) in operation but allows more time for a failed crossover to search for a suitable counterpart sub-conformation to match. *Crossover* (c) is a variation of *Crossover*(d) where, instead of a sub-conformation looking for its counterpart sub-conformation in the population, *Crossover*(c) directly uses DFS to generate the rest of the path to complete the conformation. This alternative was investigated to determine an effective rate of DFS.

The default GA parameters for all experiments were set as a population size (Pop_z) of 200, crossover rate (p_c) of 0.85 or 85%, mutation rate (p_m) of 5% and preservation of an elite of 5% [61,62].

The folding processes for longer PSP sequences generally have complex energy landscapes [32,63–68], and hence these sequences would normally require longer times to converge. Accordingly, we chose these longer benchmark sequences (see Table 1) to highlight the true impact of this approach.

3.1. Experiments using 2D HP square model

A maximum of 2000 generations was allocated for each of the 10 iterations carried out per sequence and in each category of experiments. The benchmark PSP sequences used are shown in Table 1 for the 2D square HP lattice model [7], with the length ranging from 50 to 100 [69,70]. The results are shown in Table 2.

In Table 2, we include two other algorithms in their generic form, namely Unger's GA (UGA [57]) and conformational space annealing (CSA) algorithm [18,60] along with our proposed algorithm. It may be noted that UGA has been reported to have outperformed many MC variations [13,57]. We emulated UGA in our experiment without changing the original parameter for cooling. The initial cooling temperature was set to 2 and was decreased by 0.99 in every 200,000 steps until the temperature reached 0.15.

The format of column entries is 'average/minimum' and indicates fitness function values. The *X* implies *Crossover* operation. Thus, *X*(a) indicates *Crossover* (a) as described above and so on. CSA and UGA indicate conformational space annealing algorithm [18] and Unger's GA [57], respectively. *Bold* entries indicate the row-wise best values obtained.

We abstracted the general form of the CSA algorithm by removing the heuristic-based special moves, keeping the generic form intact, to provide a fair comparison in our experiment. Comparison with the CSA algorithm is particularly important for our work, since the CSA approach has recently been applied in the PSP software ROSETTA [35,71–74]. Both UGA and CSA ran 2000 GA generation equivalent runs per iteration.

a

1. DO single-point ***Crossover.***
2. IF '***Crossover failure***' = TRUE then
3. REPLACE one of the parents.
4. DO single-point ***Crossover***.
5. END IF

b

1. DO single-point ***Crossover.***
2. IF '***Crossover failure***' = TRUE then
3. DO '***DFS after crossover failed***'.
4. END IF

c

1. DO single-point '***Combination of crossover and DFS***'.

d

1. DO apply option: (a).
2. IF no improvement for 5 consecutive generations,
3. DO apply option: (b).
4. END IF

Fig. 7. Crossover operation and variation details. Terms in *italic* is explained in Section 2.6.

3.2. Experiments using 3D HP HCP model

As mentioned earlier in Section 3, we kept the same value for GA parameters, i.e., population size 200, crossover rate 85%, mutation rate 5% and elitism of 5%. We have also used the same set of benchmark HP sequence shown in Table 1 as used for the 2D square HP lattice model. For each sequences, the simulation ran for 10 iterations; however each run has been executed a maximum of 1500 GA generations. The results obtained are shown in Table 3.

The format of column entries is 'average/minimum'. The *X* implies *Crossover* operation. Thus, *X*(*i*) indicates *Crossover* (*i*) where *i*='*a*'–'*d*'. CSA and UGA indicate conformational space

Table 1
Benchmark protein sequences for the 2D HP model.

Len.	Sequences	Ref.
50	H2(PH)3PH4PH(P3H)2P4H(P3H)2PH4P(HP)3H2	[70]
60	P2H3PH8P3H10PHP3H12P4H6PH2PHP	[70]
64	H12(PH)2(P2H2)2P2HP2H2PPHP2H2P2(H2P2)2(HP)2H12	[70]
85	4H4P12H6P12H3P12H3P12H3P1H2P2H2P2H2P1H1P1H	[69]
100	3P2H2P4H2P3H1P2H1P2H1P4H8P6H2P6H9P1H1P2H1P11H2P3H1P2H1P1H2P1H1P3H6P3H	[69]

'*H*' and '*P*' in the sequence indicate hydrophobic and hydrophilic amino acids, respectively. Len. indicates length.

Table 2
Run results of 10 iterations on each PSP sequence (see Table 1 for the sequences). GA runs with four different crossover options (shown in Fig. 7), have been compared.

Len	*X*(a)	*X*(b)	*X*(c)	*X*(d)	CSA	UGA
50	−17.3/−**20**	−17.6/−**20**	−14.5/−17	−**18**/−**20**	−17/−19	−16.6/−18
60	−29.2/−**32**	−29.8/−**32**	−27.8/−31	−**30.5**/−**32**	−30.4/-**32**	−29/−31
64	−29.1/−31	−29.3/−31	−25.2/−29	−**32**/−**35**	−29/−30	−27.8/−31
85	−39.4/−44	−39.6/−45	−34.5/−38	−**43.4**/−**46**	−43.2/−**46**	−41.4/−**46**
100	−37.1/−39	−37.6/−41	−30.2/−37	−**38.5**/−**42**	−37.2/−38	−37.4/−40

Table 3
Average and minimum-fitness-value from run-results of 10 iterations on each PSP sequence using 3D HCP model.

Len.	*X*(a)	*X*(b)	*X*(c)	*X*(d)	CSA	UGA
50	−70.3/−77	−69.8/−76	−59.6/−69	−78.4/−**82**	−72.4/−77	−70.2/−74
60	−145.8/−157	−149.8/−**161**	−115.1/−121	−152.7/−**161**	−140.8/−149	−148.1/−158
64	−138.2/−**147**	−139.6/−145	−119.2/−130	−140.8/−**147**	−129/−135	−133.3/−141
85	−191.6/−204	−193.1/−203	−169.8/−187	−197.9/−**212**	−188.5/−200	−189.6/−202
100	−180.3/−190	−185.6/−193	−142.7/−154	−191.1/−**204**	−182.8/−193	−171.2/−183

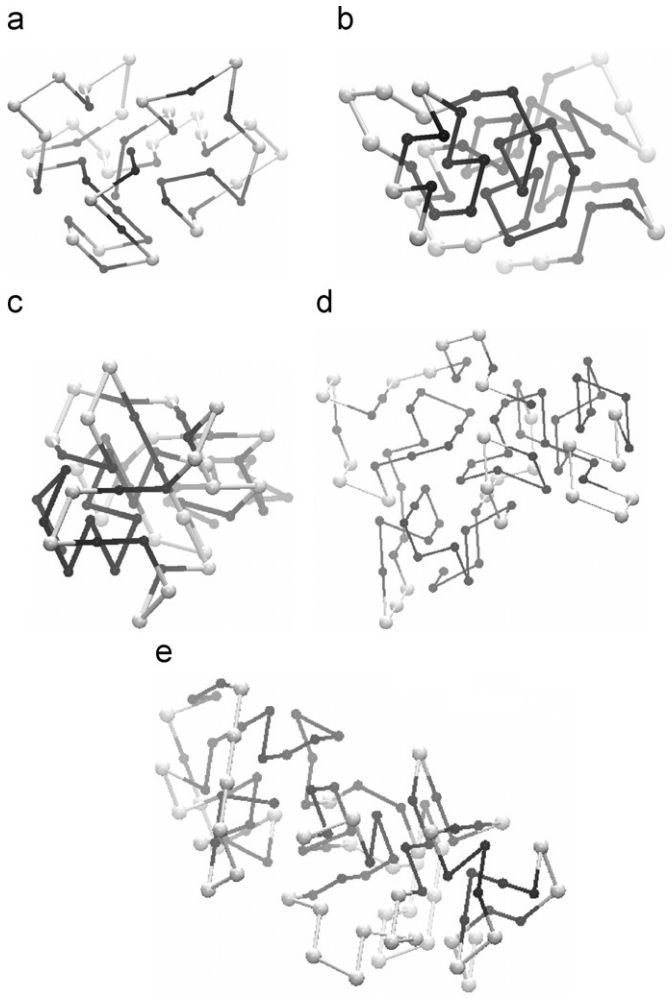

Fig. 8. Minimum conformations found in the X(d) simulation using the 3D HCP configuration. The figures from (a) to (e) correspond to the sequence length 50–100 (see Table 1) in the same order. The figures are drawn using VMD tool with 'orthographic' and 'depth cueing' options chosen.

annealing algorithm and Unger's GA, respectively. Bold entries indicate the row-wise best values obtained.

The performance using the 3D HCP model remains consistent when compared with the previous experiments performed using the 2D HP model in Section 3.1. The X(d) algorithm consistently performed the best and X(c) performed the worst and slowest amongst all other search algorithms. The minimum conformations achieved in this experiment using 3D HCP model have been shown in Fig. 8.

4. Discussion of the experimental results

We have introduced the concept of finding potential partial pathways using a depth-first search (DFS) strategy with a converging potential sub-conformation if a crossover failed to find a matching counterpart to produce a valid (i.e., a self-avoiding-walk) conformation. Crossover variation X(c) gave the worst results (see Table 2 and Table 3). X(c) involves applying DFS constantly at the same rate as the crossover operation to generate the other half of the crossover portion, which is misguiding the optimum results more than guiding them compared to the other strategies applied. X(a) represents the crossover-only approach, that is, crossover with DFS, and X(b) is the variant where DFS is applied whenever a crossover fails. X(b) demonstrates a slight

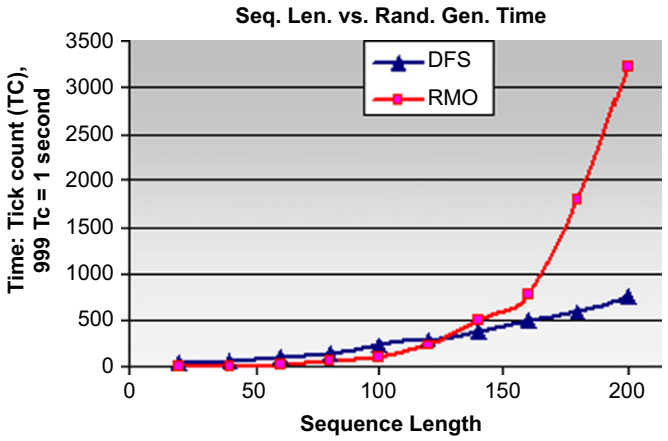

Fig. 9. Random conformation generation. DFS approach versus random-move-only (RMO) approach. An average of 100 iterations is taken for a particular length of a single random conformation generation (not from benchmark sequences).

improvement over X(a). X(d) performed the best, with results comparable to that of UGA and CSA algorithms. This is most likely due to the fact that in X(d), crossover was applied exhaustively by allowing a failed crossover to search for more counterparts to match and, when there was no improvement at all in the whole population for a few consecutive generations, the failed crossover is combined with DFS to generate possible potential pathways. It is interesting to note that in our experiment we find DFS has *zero* failure in finding pathways. Thus, a constantly failing sub-conformation in a crossover operation, which is likely to have few possible pathways, can be salvaged using DFS to unravel the hidden paths effectively. As an alternative to DFS, breadth-first search (BFS) [19] could have been used; however, BFS is both memory and time intensive.

The idea of the conformational space annealing (CSA) algorithm seemed appealing as it maintains the diversity based on maintaining distinguished conformations in a separate bank of population, implying the division of the search space into a manageable finite number of banks (say 50), which can represent possible distinguished (non-overlapped) regions of the fitness landscape and for each region the best representative is maintained. However, with the immense complexity and extent of the fitness landscape associated with the PSP problems, dividing the landscape into a finite number of regions (say 50) does not make any difference—considering a very large number divided by 50, then the numbers of possible conformations within each division remain unmanageably large. Therefore, the real effort would be to find the best representative for each region with the large and convoluted search landscape, which is again the main goal of the entire search. Thus, for the convoluted nature of the PSP problem, it is not convincing that CSA can reasonably maintain the region representative sufficiently well. This may be the reason as to why it has not performed well.

5. Supplementary applications of DFS in PSP

It is important to remember that *ordinary random conformation generation*[1] takes exponential time, fitted to a curve given by the following Equation:

$$y = 2.8723\,e^{0.0326x} \quad (1)$$

[1] Terms in *italic* is explained in Section 2.6.

Table 4
For various PSP sequences, entries indicate percentage of average chromosome removal per generation while percentage in column header (except #1) indicates equal or greater percentage of similarity between two chromosomes for which one of them is removed from the population to maintain diversity.

Len.(%)	100	90	80	70	60	50
50	17.2	33.6	39.5	45.4	53.2	76.9
60	11.8	25.7	31.1	38.3	45.0	67.9
64	13.4	23.0	28.5	36.7	43.7	64.5
85	6.2	16.7	24.5	32.0	40.4	50.9
100	6.2	19.7	25.4	33.6	41.2	49.8

Len. indicates length.

The square of the coefficient of determination of Eq. (1), $R^2=0.9832$, with increasing sequence length uses the random-move-only (RMO) approach. In contrast, the run-time for *random conformation generation by DFS* remains a quadratic fitted curve (see Fig. 9), as shown in Eq. (2):

$$y = 0.02x^2 - 0.5717x + 54.789 \quad (2)$$

with $R^2=0.9996$ (for Eq. (2)). Although Eq. (2) is mathematically quadratic, the coefficient of x^2 in Eq. (2) is close to zero, and the actual relationship can be considered to be almost linear.

The application of *random conformation generation by DFS* may have a generally lower impact because totally random conformations are generated only for initialization of the population. However, to maintain diversity many GA approaches replenish a considerable portion of the population at frequent intervals [75,76]. For example, Hoque et al. [75,77] have shown that removal of chromosomes having 80–90% or greater similarity from a GA population helps it to perform better. After removal it is necessary to replenish the population by random conformations of 16–40% in each generation as indicated in Table 4. Thus, in such a case, for longer sequences, *random conformation generation by DFS* would make the GA search far more efficient.

6. Conclusions

A depth-first search (DFS) strategy at a low rate has been applied in combination with a powerful crossover operation. Together they revealed convoluted and microscopic pathways for solving the protein structure prediction problem. Experiments using a variety of longer, standard benchmark sequences from literature have demonstrated the efficacy and improved performance characteristics of this approach empirically using two different HP model configurations. The search strategy developed was inspired by the pathway hypothesis. Further work will be directed at exploring the biological significance and relevance of this novel approach.

Acknowledgment

Support from Australian Research Council (Grant no. DP0557303) is gratefully acknowledged.

References

[1] J. Pietzsch, The importance of protein folding, Nature (2009) 〈http://www.nature.com/horizon/proteinfolding/background/importance.html〉 (accessed: Jan (2009)).
[2] S.Petit-Zeman, Treating protein folding diseases, Nature (2009) 〈http://www.nature.com/horizon/proteinfolding/background/treating.html〉 (accessed: Jan (2009)).
[3] J. Lee, S. Wu, Y. Zhang, *Ab initio* protein structure prediction, in: D.J. Rigden (Ed.), From Protein Structure to Function with Bioinformatics., Springer, Netherlands, 2009, pp. 3–25.
[4] D. Chivian, T. Robertson, R. Bonneau, D. Baker, Ab initio methods, in: P.E. Bourne, H. Weissig (Eds.), Structural Bioinformatics, Wiley-Liss, Inc., 2003, pp. 547–557.
[5] F. Allen, et al., Blue gene: a vision for protein science using a petaflop supercomputer, IBM System Journal 40 (2001) 310–327.
[6] M.T. Hoque, M. Chetty, A. Lewis, A. Sattar, D.F.S. Based, in: Partial Pathways in GA for Protein Structure Prediction Pattern Recognition in Bioinformatics (PRIB). LNCS, Springer, Melbourne, Australia, 2008 41–53.
[7] K.A. Dill, Theory for the folding and stability of globular proteins, Biochemistry 24 (1985) 1501–1509.
[8] R. Backofen, S. Will, A constraint-based approach to fast and exact structure prediction in three-dimensional protein models, Constraints Journal 11 (2006) 5–30.
[9] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, M. Yannakakis, On the complexity of protein folding, in: Proceedings of the Second Annual International Conference on Computational Molecular Biology. ACM, 1998, pp. 597–603 (extended abstract).
[10] B. Berger, T. Leighton, Protein folding in the hydrophobic–hydrophilic (HP) model is NP-complete, Journal of Computational Biology 5 (1998) 27–40.
[11] R. Schiemann, M. Bachmann, W. Janke, Exact enumeration of three-dimensional lattice proteins, Computer Physics Communications, 166 (2005) 8–16.
[12] A.J. Guttmann, Self-avoiding walks in constrained and random geometries, in: B.K. Chakrabarti (Ed.), Statistics of Linear Polymers in Disordered Media, Elsevier, 2005, pp. 59–101.
[13] R. Unger, J. Moult, on the applicability of genetic algorithms to protein folding. in: Proceedings of the 26th Hawaii International Conference on System Sciences, vol. 1, 1993, pp. 715–725.
[14] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, W. Nadler, Testing a new Monte Carlo algorithm for protein folding, PROTEINS: Structure, Function, and Genetics 32 (1998) 52–66.
[15] F. Liang, W.H. Wong, Evolutionary Monte Carlo for protein folding simulations, Journal of Chemical Physics 115 (2001) 3374–3380.
[16] T. Jiang, Q. Cui, G. Shi, S. Ma, Protein folding simulations of the hydrophobic–hydrophilic model by computing tabu search with genetic algorithms, Journal of Chemical Physics 119 (2003) 4592–4596.
[17] A. Shmygelska, H.H. Hoos, An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem, BMC Bioinformatics 6 (30) (2005) 1–22.
[18] J. Lee, Conformational space annealing and a lattice model protein, Journal of the Korean Physical Society 45 (2004) 1450–1454.
[19] T.H. Cormen, C.E. Leiserson, R.L. Rivest, in: Introduction to Algorithms, MIT Press, 1998.
[20] L. Toma, S. Toma, Folding simulation of protein models on the structure based cubo-octahedral lattice with the Contact interactions algorithm, Protein Science 8 (1999) 196–202.
[21] D. Baker, A surprising simplicity to protein folding, Nature 405 (2000) 39–42.
[22] V.S. Pande, D. Rokhsar, Folding pathway of a lattice model for proteins, Proceedings of the National Acadamy. of Sciences of the United States of America, Biochemistry 96 (1999) 273–1278.
[23] C. Levinthal, Are there pathways for protein folding? Journal of Chemical Physics 64 (1968) 44–45.
[24] M.T. Hoque, M. Chetty, L.S. Dooley, A. Guided, genetic algorithm for protein folding prediction using 3D hydrophobic–hydrophilic model. Special Session in WCCI/IEEE Congress on Evolutionary Computation (CEC) (2006), pp. 2339–2346.
[25] M.T. Hoque, M. Chetty, L.S. Dooley, Significance of hybrid evolutionary computation for *ab initio* protein folding prediction, in: C. Grosan, A. Abraham, H. Ishibuchi (Eds.), Hybrid Evolutionary Algorithms, Vol. 75, Springer-Verlag, Berlin, 2007, pp. 241–268.
[26] M.T. Hoque, M. Chetty, L.S. Dooley, A hybrid genetic algorithm for 2D FCC hydrophobic–hydrophilic lattice model to predict protein folding. in: Proceedings of the 19th ACS Australian Joint Conference on Artificial Intelligence, LNAI, Springer, 2006, pp. 867–876.
[27] M.T. Hoque, M. Chetty, L.S. Dooley, Fast computation of the fitness function for protein folding prediction in a 2D hydrophilic-hydrophobic model, International Journal of Simulation Systems, Science and Technology(special issue) 6 (2005) 27–37.
[28] M.T. Hoque, M. Chetty, L.S. Dooley, A. new guidedgenetic algorithm for 2D hydrophobic–hydrophilic model to predict protein folding. IEEE Congress on Evolutionary Computation, CEC, Edinburgh, UK, 2005, pp. 259–266.
[29] M.T. Hoque, M. Chetty, A. Sattar, Protein folding prediction in 3D FCC HP lattice model using genetic algorithm, in: Proceedings of theBioinformatics Special Session, IEEE Congress on Evolutionary Computation, CEC, Singapore, 2007, pp. 4138–4145.
[30] K. Ghosh, S.B. Ozkan, K.A. Dill, The ultimate speed limit to protein folding is conformational searching, Journal of American Chemical Society 129 (2007) 11920–11927.
[31] K.F. Lau, K.A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins Macromolecules 22 (1989) 3986–3997.

[32] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, H.S. Chan, Principles of protein folding—A perspective from simple exact models, Protein Science 4 (1995) 561–602.
[33] K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, V.A. Voelz, The protein folding problem: when will it be solved? Current Opinion in Structural Biology 17 (2007) 342–346.
[34] D.W. Corne, G.B. Fogel, An introduction to bioinformatics for computer scientists. in: G.B. Fogel, D.W. Corne, (Eds.),: Evolutionary Computation in Bioinformatics, 2004 pp. 3–18.
[35] D. Baker, Prediction and design of the macromolecular structures and interactions, Philosophical Transactions of Royal Society B 361 (2006) 459–463.
[36] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, D. Baker, Progress in modeling of protein structures and interactions, Science 310 (2005) 638–642.
[37] Y. Xia, E.S. Huang, M. Levitt, R. Samudrala, Ab initio construction of protein tertiary structures using a hierarchical approach, Journal of Molecular Biology 300 (2000) 171–185.
[38] R. Backofen, S. Will, P. Clote, Algorithmic approach to quantifying the hydrophobic force contribution in protein folding, in: Pacific Symposium On Biocomputing 5 (2000) 92–103.
[39] K. Yue, K.A. Dill, Sequence–structure relationships in proteins and copolymers, Physical Review E 48 (1993) 2267–2278.
[40] L. Toma, S. Toma, Contact interactions methods: a new algorithm for protein folding simulations, Protein Science 5 (1996) 147–153.
[41] E. Bornberg-Bauer, in: Chain Growth Algorithms for HP-Type Lattice Proteins, Recomb, Santa Fe, NM, USA, 1997 47–55.
[42] M.T. Hoque, M. Chetty, L.S. Dooley, Non-isomorphic coding in lattice model and its impact for protein folding prediction using genetic algorithm, in: Proceedings of the IEEE Computational Intelligence in Bioinformatics and Computational Biology, CIBCB, IEEE, Toronto, Canada 2006, pp. 1–8.
[43] R. Samudrala, Y. Xia, M. Levitt, A combined approach for ab initio construction of low resolution protein tertiary structures from sequence, in: Pacific Symposium on Biocomputing (PSB) 4, (1999), pp. 505–516.
[44] T. Hoque, M. Chetty, A. Sattar, Extended HP model for protein structure prediction, Journal of Computational Biology 16 (2009) 85–103.
[45] T.C. Hales, A proof of the Kepler conjecture, Annals of Mathematics 162 (2005) 1065–1185.
[46] R. Backofen, S. Will, E. Bornberg-Bauer, Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets, Bioinformatics 15 (1999) 234–242.
[47] O.S. University, The structure of crystalline solids online lectures, Materials Science and Engineering, Ohio State University, 〈http://www.matsceng.ohio-state.edu/mse205/lectures/chapter3/index_chap3.htm〉 (accessed: December 2008).
[48] M. Chen, K.Y. Lin, Universal amplitude ratios for three-dimensional self-avoiding walks, Journal of Physics A: Mathematical and General 35 (2002) 1501–1508.
[49] I.K. Roterman, M.H. Lambert, K.D. Gibson, H. Scheraga, 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, Journal of Biomolecular Structure and Dynamics 7 (1989) 421–453.
[50] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould Jr, K.M.M., D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A second generation force field for the simulation of proteins and nucleic acids, Journal of the American Chemical Society 117 (1995) 5179–5197.
[51] Y. Duan, P.A. Kollman, Computational protein folding: from lattice to all-atom, IBM Systems Journal 40 (2001) 297–309.
[52] T. Head-Gordon, S. Brown, Minimalist models for protein folding and design, Current Opinion in Structural Biology 13 (2003) 160–167.
[53] R. Wroe, E. Bornberg-Bauer, H.S. Chan, Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness if the superfunnel paradigm, Biophysical Journal 88 (2005) 118–131.
[54] R. Santana, P. Larrañaga, J.A. Lozano, Protein folding in simplified models with estimation of distribution algorithms, IEEE Transactions on Evolutionary Computation 12 (2008) 418–438.
[55] V. Cutello, G. Nicosia, M. Pavone, J. Timmis, An immune algorithm for protein structure prediction on lattice models, IEEE Transactions on Evolutionary Computation 11 (2007) 101–117.
[56] O. Takahashi, H. Kita, S. Kobayashi, Protein folding by a hierarchical genetic algorithm. in: Proceedings of the Fourth International Symposium, AROB, 1999, pp. 334–339.
[57] R. Unger, J. Moult, Genetic algorithms for protein folding simulations, Journal of Molecular Biology 231 (1993) 75–81.
[58] R. Unger, J. Moult, Genetic algorithm for 3D protein folding simulations, in: Proceedings of the Fifth International Conference on Genetic Algorithms, 1993, pp. 581–588.
[59] R. König, T. Dandekar, Refined genetic algorithm simulation to model proteins, Journal of Molecular Modeling 5 (1999) 317–324.
[60] J. Lee, H.A. Scheraga, S. Rackovsky, New optimization method for conformational energy calculations on polypeptides: conformational space annealing, Journal of Computational Chemistry 18 (1997) 1222–1232.
[61] R.L. Haupt, S.E. Haupt, Practical Genetic Algorithms (2004).
[62] J.G. Digalakis, K.G. Margaritis, An experimental study of benchmarking functions for genetic algorithms, International Journal of Computer Mathematics 79 (2002) 403–416.
[63] S.D. Flores, J. Smith, Study of fitness landscapes for the HP model of protein structure prediction, in: Proceedings of the IEEE, CEC (2003) 2338–2345.
[64] N. Mousseau, G.T. Barkema, Exploring high-dimensional energy landscape, Computing in Science & Engineering, IEEE 1 (1999) 74–80 82.
[65] U.H.E. Hansmann, Protein folding in silico: an overview, IEEE Computing in Science and Enginering 5 (2003) 64–69.
[66] J. Skolnick, A. Kolinski, Computational studies of protein folding, IEEE Computing In Science and Engineering 3 (2001) 40–50.
[67] Y. Cui, W.H. Wong, E. Bornberg-Bauer, H.S. Chan, Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes, Proceedings of the National Acadamy of Sciences of the United States of America 99 (2002) 809–814.
[68] K. Schreiner, Distributed project tackle protein mystery, Computing in Science and Engineering, IEEE 3 (2001) 13–16.
[69] N. Lesh, M. Mitzenmacher, S. Whitesides, in: A complete and effective move set for simplified protein folding, in: Proceedings of RECOMB, Berlin, Germany, 2003 188–195.
[70] W.E. Hart, S. Istrail, HP Benchmarks, 〈http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html〉, (accessed August 2005).
[71] Yiliu, Rosetta 2.1.0., The Rosetta Commons, 2007–2008 〈http://www.rosettacommons.org/tiki/tiki-index.php?page=Change+Log〉, (accessed March 2008).
[72] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C.E.M. Strauss, D. Baker, Rosetta in CASP4: progress in ab initio protein structure prediction, Proteins: Structure, Function, and Genetics 5 (2001) 119–126.
[73] P. Bradley, D. Chivian, J. Meiler, K.M.S. Misura, C.A. Rohl, W.R. Schief, W.J. Wedemeyer, O. Scueler-Furman, P. Murphy, J. Schonbrun, C.E.M. Strauss, D. Baker, D. Rosetta, Predictions in CASP5: success, failure, and prospects for complete automation, Proteins: Structure, Function, and Genetics 53 (2003) 457–468.
[74] K.T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III target using ROSETTA, Proteins: Structure, Function, and Genetics 3 (1999) 171–176.
[75] M.T. Hoque, M. Chetty, L.S. Dooley, in: Generalized schemata theorem incorporating twin removal for protein structure prediction. in: Pattern Recognition in Bioinformatics, Springer, Singapore, 2007 84–97.
[76] V.K. Koumousis, C.P. Katsaras, A. Saw-Tooth, Genetic algorithm combining the effects of variable population size and reinitialization to enhance preformance, IEEE Transaction on Evolutionary Computation 10 (2006) 19–28.
[77] M.T. Hoque, M. Chetty, A. Lewis, A. Sattar, Twin removal in genetic algorithms for protein structure prediction using low resolution model, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2009, 10.1109/TCBB.2009.34.

Md Tamjidul Hoque received both his B Sc. Eng. and M.Sc. Eng. degrees in Computer Science and Engineering (CSE) from the Bangladesh University of Engineering and Technology in 1998 and 2002, respectively, and received his Ph.D. degree in IT from Monash University (Australia) in 2008. He was a lecturer at the CSE department, Ahsanullah University of Science and Technology, 1998–1999. He was in the technical management being IT incharge and DGM at Bashundhara Group, Dhaka, Bangladesh from December 1999 to 2004. Currently he is a research fellow at Griffith University (Australia) in Discovery Biology, Eskitis, and a member of IIIS. His research focus is on '*ab initio* protein structure prediction' and 'high content image analysis and algorithm development'. His research interests include evolutionary computation, bioinformatics, networking, communication, database system, complier design, automata theory, distributed systems and parallel computing, computer architecture, petri net theory, and security and operating system.

Dr. Madhu Chetty has been with Monash University, Australia since 1995 and is currently the Deputy Head of Gippsland School of Information Technology. His research interests include bioinformatics, optimization, computational intelligence and modeling complex systems. Dr Chetty has authored over 100 scientific articles, which include book chapters and articles in journals and international conferences. He is a Senior Member of IEEE and Fellow, Institution of Engineers (India). He is currently serving as Chair of Technical Committee (TC-20) of International Association for Pattern Recognition (IAPR) on Bioinformatics and was General Chair of the PRIB'08 (Pattern Recognition in Bioinformatics) conference. He has also served as Vice Chair of the IEEE CIS Technical Committee on Bioinformatics and Bioengineering. He is serving as the Associate Editor of the Elsevier's Neurocomputing journal and is on the editorial board of three other journals in bioinformatics. Prior to his career at Monash, Dr. Chetty worked at VRCE (now VNIT), Nagpur India (1980–1993), and University of Melbourne (1993–1995).

Dr. Andrew Lewis is a Senior Research Specialist in Research Computing Services and an Adjunct Senior Lecturer in ICT at Griffith University. Prior to this appointment he worked in industrial applied research with BHP Billiton. His research interests include parallel optimisation algorithms for large numerical simulations, including gradient descent, direct search methods, evolutionary programming, particle swarm and ant colony systems, multi-objective optimisation techniques for engineering design, and parallel, distributed and grid computing methods. He has numerous publications in the area of optimisation algorithms and applications.

Professor Abdul Sattar is the Founding Director of the Institute for Integrated and Intelligent Systems (IIIS) and a Professor of Computer Science and Artificial Intelligence at Griffith University. He is also a Research Leader at National ICT Australia (NICTA) Queensland Research Lab (QRL), where he has held the positions of QRL Education Director (2006–2008) and Leader of the Smart Applications for Emergencies (SAFE) project (2005–2008), and is currently leading the QRL node of NICTA's largest project, Advanced Technologies for Optimisation and Modelling in Constraints (ATOMIC). He has been an academic staff member at Griffith University since February 1992 as a lecturer (1992–1995), senior lecturer (1996–99), and professor (2000–present) within the School of Information and Communication Technology. Prior to his career at Griffith University, he was a lecturer in Physics in Rajasthan, India (1980–1982), a research scholar at Jawaharlal Nehru University, India (1982–1985), the University of Waterloo, Canada (1985–1987) and the University of Alberta, Canada (1987–1991).

Associate Professor Vicky Avery obtained her Ph.D. in 1994 from the Flinders University of South Australia and was awarded an Australian NHMRC Postdoctoral Fellowship which was undertaken at the University of Adelaide. Between 1998 and 2004, A/Prof Avery was based at Active Biotech AB, Sweden, and she held several positions, including Section Head for Protein Interaction and Drug Discovery; Scientific Project Leader to identify the molecular target of 'Laquinimod', a novel oral treatment for relapsing multiple sclerosis, that has successfully concluded Phase IIb trials and Director of Biochemistry and Molecular Biology and Director, Business Development. Also of significance, she was responsibility for the development of assays for FDA to assess efficacy of a cholera vaccine designed and developed assays to identify immuno-modulatory compounds against CD80, which led to RhuDex®, an oral treatment for RA in clinical trials. A/Prof Avery specializes in high-throughput and high content screening. As the Head of Discovery Biology for the AstraZeneca–Griffith University collaboration, she was responsible for more than 49 HTS campaigns conducted between 2004 and 2007. The Discovery Biology team has also successfully designed and implemented HTS assays for Malaria (MMV) and African Sleeping Sickness (DnDi), being awarded MMV Project of the Year (2007) for innovative use of technology to identify new anti-malarials. As a Programme Leader for the recently established CRC Cancer Therapeutics, she has played an active role in the acquisition of funds and establishment of the Bioactive Discovery (HTS) programme.