Amino Acids Pattern-Biased Spiral Search for Protein Structure Prediction

Mahmood A. Rashid[1,2], Md. Masbaul Alam Polash[1], M.A. Hakim Newton[1], Md Tamjidul Hoque[3], and Abdul Sattar[1,2]

[1] Institute for Integrated & Intelligent Systems, Griffith University
[2] Queensland Research Lab, National ICT Australia
[3] Computer Science, University of New Orleans, USA

Abstract. Proteins are essentially sequences of amino acids. They adopt specific folded 3-dimensional structures to perform specific tasks. The formation of 3-dimensional structures is largely guided by the constituent amino acids. Therefore, the positional presence of amino acids in a sequence might play important roles during the protein folding process. In this paper, we present a new heuristic derived from the positional patterns of amino acids in a sequence. With the help of a biased tabu tenure, we apply this heuristic within a spiral search algorithm. The spiral search is an efficient algorithm to develop hydrophobic core in a protein structure pulling hydrophobic amino acids towards the core centre in a spiral fashion. On a set of standard benchmark proteins, we experimentally show that applying our new heuristic improves the performance of a spiral search algorithm consistently.

Keywords: Protein Structure Prediction, Spiral Search, Local Search, Lattice Models, Amino Acid Patterns.

1 Introduction

Proteins are the sequences of amino acids connected together by peptide bonds. Amino acids which are the constituents of all proteins form the protein chain and interact with each other to fold into a stable three dimensional native structure. However, the stability of the structure might depend on many factors such as the temperature and the solvent.

The protein structure prediction (PSP) problem is to find the 3-dimensional (3D) native structure of a protein from its amino acid sequence. The native structure of a protein has the minimum free energy possible and determines the function of the protein. Protein structure prediction is one of the most important goals pursued in many areas of bioinformatics and theoretical chemistry such as molecular docking, protein-protein docking and combinatorial chemistry. It has important usages in drug design and biotechnology.

There is a large number of existing search algorithms that attempt to solve the PSP problem by exploring feasible structures called *conformations*. However, for the face-centred cubic (FCC) lattice representation, the state-of-the-art

D.-N. Pham and S.-B. Park (Eds.): PRICAI 2014, LNAI 8862, pp. 143–156, 2014.

results have been achieved by local search (LS) methods (Rashid et al., 2013; Shatabda et al., 2013; Dotú et al., 2011). In general, the success of the methods crucially depends on the balance of diversification and intensification of the search. Moreover, these algorithms often get stuck in local minima. As a result, they perform poorly on large proteins. Any further progress to these algorithms requires addressing the above issues appropriately. In this paper, we present a new heuristic derived from the positional patterns of amino acids in a protein. We apply this heuristic within spiral search algorithm using a biased tabu tenure (SS-Tabu). On a set of standard benchmark proteins, we experimentally show that applying our new heuristic improves the performance of a spiral search algorithm consistently.

The rest of the paper is organised as follows: Section 2 presents the preliminaries; Section 3 discusses existing work on protein structure prediction; Section 4 describes our approaches in detail; Section 5 presents the experimental results and analyses; and finally, Section 6 draws the conclusions and outlines the future research.

2 Preliminaries

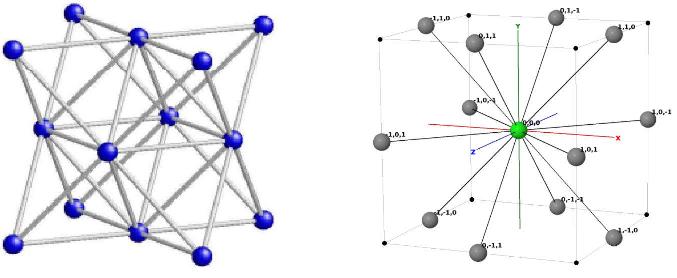
There are three computational approaches for protein structure prediction. These are *Homology modeling* (Zhang and Skolnick, 2005), *protein threading* (Bowie et al., 1991; Torda, 2005) and *ab initio* methods (Simons et al., 1999; Baker and Sali, 2001). However, our work is based on the *ab initio* approach that only depends on the amino acid sequence of the target protein. Levinthal's paradox (Levinthal, 1968) and Anfinsen's hypothesis (Anfinsen, 1973) are the basis of *ab initio* methods for PSP. The idea was originated in 1970 when it was demonstrated that all information needed to fold a protein resides in its amino acid sequence.

2.1 Simplified Model

In this paper, we use 3D FCC lattice points for mapping the conformation and to generate the backbones of the protein structures. We use the HP energy model for conformation evaluation. The 3D FCC lattice, and the HP energy model are briefly described below.

3D FCC Lattice. The FCC lattice has the highest packing density compared to the other existing lattices (Hales, 2005). The hexagonal close packed (HCP) lattice, also has 12 neighbours that correspond to 12 basis vertices with real-numbered coordinates; which causes the loss of structural precision for PSP. In FCC, each lattice point has 12 neighbours as shown in Fig. 1b. The 12 *basis vectors* w.r.t. the origin are presented below denoting as $\boldsymbol{A} \ldots \boldsymbol{L}$.

$$\begin{array}{llll}
\boldsymbol{A} = (1,1,0) & \boldsymbol{B} = (0,1,1) & \boldsymbol{C} = (1,0,1) & \boldsymbol{D} = (-1,1,0) \\
\boldsymbol{E} = (0,-1,1) & \boldsymbol{F} = (-1,0,1) & \boldsymbol{G} = (1,-1,0) & \boldsymbol{H} = (0,1,-1) \\
\boldsymbol{I} = (1,0,-1) & \boldsymbol{J} = (-1,-1,0) & \boldsymbol{K} = (0,-1,-1) & \boldsymbol{L} = (-1,0,-1)
\end{array}$$



(a) (b)

Fig. 1. (a) A unit 3D FCC lattice. (b) A unit 3D FCC lattice with 12 basis vectors on the Cartesian coordinates.

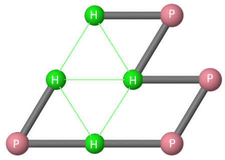
In simplified PSP, conformations are mapped on the lattice by a sequence of basis vectors, or by the *relative vectors* that are relative to the previous basis vectors in the sequence.

HP Energy Model. The 20 amino acid monomers are the building block of protein polymers. These amino acids are broadly divided into two categories based on their hydrophobicity: (a) hydrophobic amino acids (*Gly*, *Ala*, *Pro*, *Val*, *Leu*, *Ile*, *Met*, *Phe*, *Tyr*, *Trp*) denoted by H; and (b) hydrophilic or polar amino acids (*Ser*, *Thr*, *Cys*, *Asn*, *Gln*, *Lys*, *His*, *Arg*, *Asp*, *Glu*) denoted by P.

$$E = \sum_{i<j-1} c_{ij}.e_{ij} \tag{1}$$

Here, c_{ij}= 1 if amino acids i and j are non-consecutive neighbours, otherwise 0; and $e_{ij} = -1$ if ith and jth amino acids are hydrophobic, otherwise 0.

	H	P
H	-1	0
P	0	0



(a) (b)

Fig. 2. (a) HP energy model (Lau and Dill, 1989) and (b) H-H contacts of a random sequence *HPHPPHPH* to calculate energy on the lattice model

In the HP model (Lau and Dill, 1989), when two non-consecutive hydrophobic amino acids become topologically neighbours, they contribute a certain amount of negative energy, which for simplicity is shown as -1 in Fig. 2a. The total energy (E) of a conformation based on the HP model becomes the sum of the contributions of all pairs of non-consecutive hydrophobic amino acids as shown in Equation 1, which is eventually the optimisation function (minimisation for PSP) in HP energy model. In Fig. 2b, the number of such H-H contacts is 5. Therefore, the fitness of the structure of a random sequence *HPHPPHPH* is -5.

3 Related Work

Different types of metaheuristic have been used in solving the simplified PSP problem. These include Monte Carlo Simulation (Thachuk et al., 2007), Simulated Annealing (Tantar et al., 2008), Genetic Algorithms (GA) (Unger and Moult, 1993), Tabu Search with GA (Böckenhauer et al., 2008), Tabu Search with Hill Climbing (Klau et al., 2002), Ant Colony Optimisation (Blum, 2005), Immune Algorithms (Cutello et al., 2007), Tabu-based Stochastic Local Search (Cebrián et al., 2008; Shatabda et al., 2012), and Constraint Programming (Dotú et al., 2011). The Bioinformatics Group, headed by Rolf Backofen, applied Constraint Programming (Backofen and Will, 2006; Mann et al., 2008, 2009) using exact and complete algorithms. Their exact and complete algorithms work efficiently if similar hydrophobic-core exists in the repository. Cebrian *et al.* (Cebrián et al., 2008) used tabu-based local search, and Shatabda *et al.* (Shatabda et al., 2012) used memory-based local search with tabu heuristic and achieved the state-of-the-art results. However, Dotu *et al.* (Dotú et al., 2011) used constraint programming and found promising results but only for small sized ($length < 100$ amino acids) proteins. Besides local search, Unger and Moult (Unger and Moult, 1993) applied population based genetic algorithms to PSP and found their method to be more promising than the Monte Carlo based methods (Thachuk et al., 2007). They used absolute encodings on the square and cubic lattices for HP energy model. Later, Patton (Patton et al., 1995) used relative encodings to represent conformations and a penalty method to enforce the self-avoiding walk constraint. GA has been used by Hoque *et al.* (Hoque et al., 2007) for cubic, and 3D HCP lattices. They used DFS-generated pathways (Hoque et al., 2010) in GA crossover for protein structure prediction. They also introduced a twin-removal operator (Hoque et al., 2011) to remove duplicates from the population to prevent the search from stalling. Ullah *et al.* in (Ullah et al., 2009) and (Ullah and Steinhöfel, 2010) combined local search with constraint programming. They used a 20×20 energy model (Berrera et al., 2003) on FCC lattice and found promising results. In another hybrid approach (Jiang et al., 2003), tabu meta-heuristic was combined with genetic algorithms in two-dimensional HP model to observe crossover and mutation rates over time.

However, for the simplified model (HP energy model and 3D FCC lattice) that is used in this paper, a tabu-based local search algorithm known as spiral search (Rashid et al., 2013) currently produces the state-of-the-art results.

4 Our Approach

The work in this paper is powered by a hydrophobic-core directed spiral search algorithm proposed by Rashid *et al.* in (Rashid et al., 2013). Spiral search belongs to the family of stochastic local search, guided by tabu meta-heuristic and used here to minimise the intermolecular energy of a protein structure. To make this paper self sufficient, we present a brief overview of the spiral search algorithm in the following section.

4.1 Spiral Search Algorithm

Protein structures have hydrophobic cores (H-core) that hide hydrophobic amino acids from water and expose the polar amino acids to the surface to be in contact with the surrounding water molecules (Yue and Dill, 1993). H-core formation is the main objective of HP based PSP.

The spiral search is guided by a tabu meta-heuristic where an unbiased tabu-tenure is used to keep track of the changes of all hydrophobic amino acid positions of a given protein sequence. Therefore, in existing spiral search, every hydrophobic amino acid has the same priority level to participate forming the H-cores. The *pseudocode* of spiral search is shown in Algorithm 1.

Algorithm 1: The Spiral Search Algorithm with Unbiased-Tabu

```
Generate initial solution
Initialise tabu with unbiased tabu-tenures
while not exit condition do
    Select and apply move to generate a solution
    Update unbiased tabu for this move
    Evaluate the generated solution
    Considering the search progress apply random-walk
    Considering the search progress apply relay-restart
```

In spiral search, only the diagonal move operator is used repeatedly (as shown in Fig. 3) in building H-cores. A diagonal move displaces ith amino acid from its position to another position on the lattice without changing the position of its succeeding $(i+1)$th and preceding $(i-1)$th amino acids in the sequence. The move is just a corner-flip to an unoccupied lattice point. The functionalities of the spiral search algorithm are enlisted below:

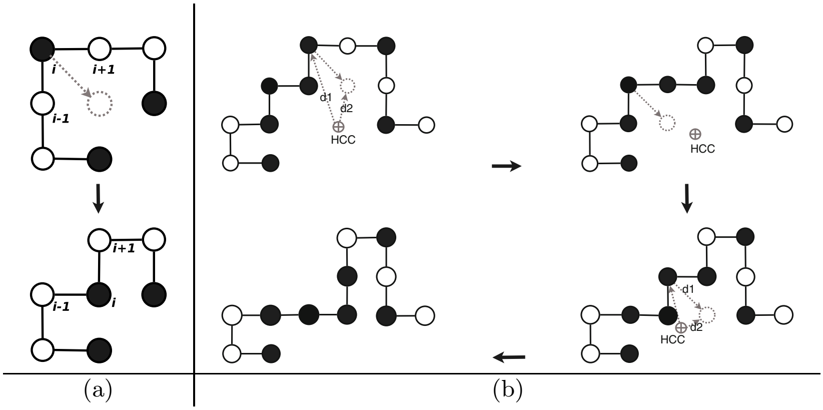


Fig. 3. (a) Diagonal move operator. (b) Spiral search algorithm comprising a series of diagonal moves. For simplification and easy understanding, the figures are presented in 2-dimensional space.

1. **Generating initial solution:** Spiral search algorithm starts with a feasible conformation. An initial feasible conformation is generated following a self-avoiding walk (SAW) on FCC lattice points. During initial solution generation, it always places the first amino acid at $(0, 0, 0)$ and then randomly selects a vector from the 12 basis vectors of FCC lattice space (See Fig. 1) to place the successive amino acid at a neighbouring free lattice point. The mapping proceeds in a backtracking fashion until a self-avoiding walk is found for the whole protein sequence.

2. **Initialise unbiased-tabu:** In spiral search, for hydrophobic amino acids, a tabu-list is used to mark the position of each amino acid whenever a change of position occurrs. A recently changed amino acid is then not moved for a certain number of iterations (called tabu-tenure). In spiral search (Rashid et al., 2013), all hydrophobic amino acids are considered equal regardless of their positions in the protein sequence, and the same weight is put for all H amino acids while defining the tabu-tenure and the value is calculated using Equation 2. In this work, we denote this tabu-tenure of (Rashid et al., 2013) as *unbiased tabu-tenure.*

$$\mathsf{tenure} = \left(10 + \frac{\mathsf{hCount}}{10}\right) \tag{2}$$

   where hCount is the number of H amino acids in the protein sequence.

3. **Move selection:** In move selection, the hydrophobic amino acids get priority in comparison to hydrophilic amino acids. Therefore, during move selection, two different strategies are maintained depending on the hydrophobic property of the amino acids. In spiral search, a virtual hydrophobic core centre (HCC) is calculated by finding arithmetic means of x, y, and z coordinates of all hydrophobic amino acids using Equation 3.

$$x_{\mathrm{hcc}} = \frac{1}{n_{\mathrm{h}}}\sum_{i=1}^{n_{\mathrm{h}}} x_i,\ y_{\mathrm{hcc}} = \frac{1}{n_{\mathrm{h}}}\sum_{i=1}^{n_{\mathrm{h}}} y_i,\ z_{\mathrm{hcc}} = \frac{1}{n_{\mathrm{h}}}\sum_{i=1}^{n_{\mathrm{h}}} z_i \tag{3}$$

   In move selection process, this HCC is used to calculate and compare the distances of amino acids from the H-core.

   (a) *H-move selection:* The H-move selection is guided by the Euclidean distance d_i between HCC and the hydrophobic amino acids in the sequence. d_i is calculated using Equation 4. For the ith hydrophobic amino acid, the common topological neighbours of the $(i-1)$th and $(i+1)$th amino acids are computed. The topological neighbours (TN) of a lattice point are the points at unit lattice-distance apart from it. For 3D FCC lattice space, the maximum number of common TN for the $(i-1)$th and $(i+1)$th amino acids is four. From the common neighbours, the unoccupied points are identified. The Euclidean distance of all unoccupied common neighbours are calculated from the HCC using Equation 4.

$$d_i = \sqrt{(x_i - x_{\mathrm{hcc}})^2 + (y_i - y_{\mathrm{hcc}})^2 + (z_i - z_{\mathrm{hcc}})^2} \tag{4}$$

Then the point with the shortest distance is picked. This point is listed in the possible H-move list for ith hydrophobic amino acid if its current distance from HCC is greater than that of the selected point. When all hydrophobic amino acids are traversed and the feasible shortest distances are listed in H-move list, the amino acid having the shortest distance in H-move list is chosen to apply diagonal move (Fig. 3) on it. The process stops when no improving H-move is found. In this situation, the control is transferred to select and apply P-moves.

(b) *P-Move selection:* For polar amino acids, the same kind of diagonal moves are applied as H-move. For each ith polar amino acid, all free lattice points that are common neighbours of lattice points occupied by $(i-1)$th and $(i+1)$th amino acids are listed. From the list, a point is selected randomly to complete a diagonal move for the respective polar amino acid. No HCC is calculated and no Euclidean distance is measured for P-move. After one try for each polar amino acid the control is returned to select and apply H-moves.

4. **Update unbiased-tabu:** An unbiased tabu list is maintained for each hydrophobic amino acid to control the selection priority amongst them. For each successful move, the tabu list is updated (See Line 5 in Algorithm 1) for the respective amino acid. No tabu list is maintained for P-moves.

5. **Evaluate solution:** After each iteration, the conformation is evaluated by counting the H-H contacts (topological neighbours as shown in Fig. 2) where the two amino acids are non-consecutive. The two amino acids are in contact if the Euclidean distance among them is $\sqrt{2}$ in 3D FCC lattice space.

6. **Handle stagnation:** For hard optimisation problems such as protein structure prediction, local search algorithms often face stagnation. In spiral search algorithm, a random-walk (Rashid et al., 2012) and a relay-restart (Line 7 and Line 8 in Algorithm 1) techniques are applied on an on-demand basis to deal with stagnation.

   (a) *Random-walk:* Premature H-cores are observed at local minima. To escape local minima, a random-walk (Rashid et al., 2012) algorithm (Rashid et al., 2012) is applied. This algorithm uses pull moves (Lesh et al., 2003) to break the premature H-core. During pulling, energy level and structural diversification are observed to maintain a balance among these two. The algorithm allows energy level to change within 5% to 10% with changes in the structure from 10% to 75% of the original. It accepts the conformation that is close to the current conformation in terms of energy level but is diverse in terms of structure.

   (b) *Relay-restart:* Instead of using a fresh restart or restarting from the current best solution (Cebrián et al., 2008; Shatabda et al., 2012), spiral search uses a relay-restart technique when the search stagnation situation arises. It uses relay-restart when random-walk fails to escape from local minima. In spiral search, an improving solution list is maintained that contains all the improving solutions after initialisation. When a solution with energy level better than the current global best is found, the solution is added to the top of the list pushing existing solutions back. For relay-restart, a random conformation from the top 10% solutions of the list is selected to restart with. The selected solution is then sent back to the bottom of the list to keep it away from the scope of reselection in very near future.

4.2 Spiral Search with Biased-Tabu

The more the value of biasing factor, the more the value of tabu-tenure and the less the chance to get involved in the move of position. Formation of hydrophobic-core is one of the prime objectives for the HP based Protein Structure Prediction. In this work, we tried to give more chances to the H amino acids having adjacent H amino acids to participate in forming hydrophobic cores. We implemented the spiral search algorithm as shown in Algorithm 2. The difference between Algorithm 2 and Algorithm 1 is that in Algorithm 1, an unbiased tabu-tenure is used (as shown in Line 2 Algorithm 1) whereas in Algorithm 2, a biased tabu-tenure is used (as shown in Line 2 Algorithm 2).

Algorithm 2: The Spiral Search Algorithm with Biased-Tabu

```
1 Generate initial solution
2 Initialise tabu with biased tabu-tenures
3 while not exit condition do
4     Select and apply move
5     Update biased tabu
6     Evaluate solution
7     Considering the search progress apply random-walk
8     Considering the search progress apply relay-restart
```

In the unbiased case, intuitively, tabu-tenure is calculated based on the number of hydrophobic amino acids (hCount) in the sequence using the formula in Equation 2. Therefore, the size of tabu-tenure is the same for all H amino acids in a particular sequence.

Biased Tabu-Tenure. In this paper, based on the orientation of H amino acids within a protein sequence, we define five bias factors $(1 - 5)$ for eight different patterns as presented in Table 1. The patterns are described below:

1. **-H-H-H-:** In this pattern, the H amino acid under consideration is connected with two other H amino acids on both the sides. The biasing factor of this pattern is taken as the basis which is equal to the standard biasing factor (i.e., $\mathsf{bf} = 1$).
2. **-H-H-P- or -P-H-H-:** In this pattern, the H amino acid under consideration is connected with H amino acid in one side and a P amino acid in the

Table 1. Bias-factors for tabu tenure(TT) based on the orientation of hydrophobic and polar amino acids in the sequence. The factors in the table correspond to the Hs underlined.

HP-pattern	*H-position*	**Biasing factors (bf)**		**Tabu-tenure (tt)**	
		Standard	*Biased*	*Standard*	*Biased*
-H-$\underline{\text{H}}$-H-	Middle	1	1	1×tenure	1×tenure
-H-$\underline{\text{H}}$-P- or -P-$\underline{\text{H}}$-H-			2		2×tenure
-P-$\underline{\text{H}}$-P-			3		3×tenure
$\underline{\text{H}}$-H-*- or -*-H-$\underline{\text{H}}$	Terminal	1	4	1×tenure	4×tenure
$\underline{\text{H}}$-P-*- or -*-P-$\underline{\text{H}}$			5		5×tenure

other side. The biasing factor of this pattern is two times of the pattern in 1 (i.e., bf = 2).

3. **-P-$\underline{\text{H}}$-P-:** In this pattern, the H amino acid under consideration is connected with two other P amino acids on both the sides. The biasing factor of this pattern is three times of the pattern in 1 (i.e., bf = 3).
4. **$\underline{\text{H}}$-H-*- or -*-H-$\underline{\text{H}}$:** In this pattern, the H amino acid under consideration is on the terminal and connected with another H amino acid with the sequence. The biasing factor of this pattern is four times of the pattern in 1 (i.e., bf = 4).
5. **$\underline{\text{H}}$-P-*- or -*-P-$\underline{\text{H}}$:** In this pattern, the H amino acid under consideration is on the terminal and connected with a P amino acid with the sequence. The biasing factor of this pattern is five times of the pattern in 1 (i.e., bf = 5).

5 Experiments and Analyses

We implemented the spiral search algorithm in Java (J2EE). We ran our experiments on a multi-computer cluster. The cluster consists of a number of identical Dell PowerEdge R415 computers, each equipped with 2×AMD 6-Core Opteron 4184 processors, 2.8GHz clock speed, 3M L2/6M L3 Cache, 64 GB memory and running Rocks OS (a Linux variant for cluster). For each protein, we ran each algorithm 50 times.

5.1 Benchmark Sequences

The experimental results on HP benchmarks are presented in Table 2. Amongst the sequences, F180 and R instances are taken from Peter Clote laboratory website[1]. These instances have been used in (Dotú et al., 2011; Shatabda et al., 2012; Rashid et al., 2013) for evaluating different algorithms. Moreover, we use

[1] Peter Clote Lab: `http://bioinformatics.bc.edu/clotelab/FCCproteinStructure/`

other six larger sequences that are taken from the CASP[2] competition. The corresponding CASP target IDs for proteins *3mse*, *3mr7*, *3mqz*, *3no6*, *3no3*, and *3on7* are *T0521*, *T0520*, *T0525*, *T0516*, *T0570*, and *T0563*. These CASP targets are also used in (Shatabda et al., 2012). To fit in the HP model, the CASP targets are converted to HP sequences based on the hydrophobic properties of the constituent amino acids. The lower bounds of the free energy values (in Column *LBFE* of Table 2) are obtained from (Shatabda et al., 2012); however, there are some unknown values (presented as n/a) of lower bounds of free energy for large sequences.

5.2 Experimental Results

The experimental results of our new approach are presented in Table 2. There are two different sets of results obtained using spiral search algorithm. One set of results (Columns under *Reference(r)*) are obtained from (Rashid et al., 2013) where an unbiased tabu is used as a meta-heuristic, however, the other set of results (Column under *Target(t)*) are obtained from the same algorithm where a biased tabu is applied. The Column *RI* represents the relative improvements calculated using Equation 5. The results are calculated over 50 different runs with identical settings for each approach. The bold-faced values in Column Avg are the winners for the respective proteins, and that of Column RI are the minimum and maximum values of relative improvements. Column *LBFE* presents the lower bounds of free energy. The running time is mentioned in the last Column of Table 2 corresponds to the protein sequences.

5.3 Relative Improvement

The difficulty of improving energy level is increased as the improved energy level approaches to the lower bound of free energy. For example, if the lower bound of free energy of a protein is −100, the efforts to improve energy level from −80 to −85 are much less than that to improve energy level from −95 to −100 though the change in energy is the same (−5). Relative Improvement (RI) explains how close our predicted results to the lower bound of free energy with respect to the energy obtained from the state-of-the-art approaches.

$$\mathsf{RI} = \frac{E_t - E_r}{E_l - E_r} * 100\% \tag{5}$$

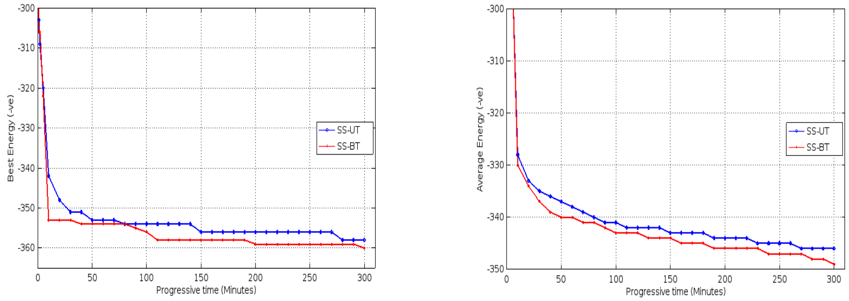
In Table 2, we also present a comparison of improvements (%) on average conformation quality (in terms of free energy levels). We compare the results obtained by using biased tabu (target) with the results obtained by using unbiased tabu (references). For each protein, the *RI* of the target (t) w.r.t. the reference (r) is calculated using the formula in Equation 5, where E_t and E_r denote the average energy values achieved by the target and the reference respectively, and E_l is the lower bound of free energy for the protein in the HP

[2] CASP website: `predictioncenter.org/casp9/targetlist.cgi`

Table 2. Comparison of performances of spiral search with and without biased tabu-tenure. The best and average energy values are presented in the Columns Best and Avg for the respective approaches. The values are calculated over 50 different runs of identical settings. The Column Run time shows the running time of both approaches.

Protein sequence information			**Spiral search algorithm**				Relative improvement on avg energy (%)	Run time (mins)
			Reference(r)		**Target(t)**			
			Unbiased Tabu (Rashid et al., 2013)		Biased Tabu (New)			
ID	*Size*	*LBFE*	*Best*	*Avg*	*Best*	*Avg*	*RI*	
F180_1	180	-378	-357	-340	-354	**-341**	2.63%	300
F180_2		-381	-359	-345	-359	**-347**	5.56%	
F180_3		-378	-362	-353	-364	-353	**0.00%**	
R1	200	-384	-359	-345	-358	**-347**	5.13%	300
R2		-383	-358	-346	-360	**-349**	**8.11%**	
R3		-385	-365	-345	-364	**-346**	2.50%	
3mse	179	-323	-289	-280	-292	-280	**0.00%**	300
3mr7	189	-355	-328	-313	-328	**-315**	4.76%	
3mqz	215	-474	-420	-403	-426	**-406**	5.56%	
3no6	229	-455	-411	-391	-411	**-396**	7.81%	
3no3	258	-494	-412	-393	-417	**-397**	3.96%	
3on7	279	*n/a*	-512	-485	-510	**-489**	*n/a*	

Note: *n/a* denotes unknown.



(a)

(b)

Fig. 4. Search progress for protein *R2* with time for two approaches on (a) best energy values and (b) on average energy values. The results are calculated over 50 different runs with identical settings for each approach.

model. We present the relative improvements only for the proteins having known lower bounds of free energy values. We test our new approach on 12 different proteins of various length and our biased tabu based approach wins on 9 out of 11 proteins.

5.4 Search Progress

We compare the search progresses of two different variants of spiral search algorithm over progressive time. These are unbiased tabu based spiral search (SS-UT) and biased tabu based spiral search (SS-BT). Fig. 4a shows progresses on best energy values whereas Fig. 4b shows the average energy values obtained with times by the two approaches for protein R2. We observe that both approaches achieve very good progress initially, but with increasing time, our biased tabu spiral search makes more progress than unbiased one.

6 Conclusion

The formation of 3D structures of a protein largely depends on the constituents amino acids. Therefore, the positional presence of amino acids in a sequence plays important roles during the protein folding process. In this paper, we present a new heuristic derived from the positional patterns of amino acids in a sequence. We apply this heuristic with the help of a biased tabu tenure in spiral search algorithm. On a set of standard benchmark proteins, we experimentally show that applying our new heuristic improves the performance of spiral search algorithm consistently.

Acknowledgments. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Anfinsen, C.B.: The principles that govern the folding of protein chains. Science 181(4096), 223–230 (1973)

Backofen, R., Will, S.: A Constraint-based approach to fast and exact structure prediction in three-dimensional protein models. Constraint 11(1), 5–30 (2006)

Baker, D., Sali, A.: Protein structure prediction and Structural Genomics. Science 294(5540), 93–96 (2001)

Berrera, M., Molinari, H., Fogolari, F.: Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics 4(1), 8 (2003)

Blum, C.: Ant colony optimization: Introduction and recent trends. Physics of Life Reviews 2(4), 353–373 (2005)

Böckenhauer, H.-J., Dayem Ullah, A.Z.M., Kapsokalivas, L., Steinhöfel, K.: A local move set for protein folding in triangular lattice models. In: Crandall, K.A., Lagergren, J. (eds.) WABI 2008. LNCS (LNBI), vol. 5251, pp. 369–381. Springer, Heidelberg (2008)

Bowie, J.U., Luthy, R., Eisenberg, D.: A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016), 164 (1991)

Cebrián, M., Dotú, I., Van Hentenryck, P., Clote, P.: Protein structure prediction on the face centered cubic lattice by local search. In: The 23rd National Conference on Artificial Intelligence, vol. 1, pp. 241–246. AAAI Press (2008)

Cutello, V., Nicosia, G., Pavone, M., Timmis, J.: An immune algorithm for protein structure prediction on lattice models. IEEE Transactions on Evolutionary Computation 11(1), 101–117 (2007)

Dotú, I., Cebrián, M., Van Hentenryck, P., Clote, P.: On lattice protein structure prediction revisited. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2011)

Hales, T.C.: A proof of the Kepler conjecture. The Annals of Mathematics 162(3), 1065–1185 (2005)

Hoque, M.T., Chetty, M., Lewis, A., Sattar, A.: Twin removal in genetic algorithms for protein structure prediction using low-resolution model. Transactions on Computational Biology and Bioinformatics 8(1), 234–245 (2011)

Hoque, M.T., Chetty, M., Lewis, A., Sattar, A., Avery, V.M.: DFS-generated pathways in GA crossover for protein structure prediction. Neurocomputing 73(13-15), 2308–2316 (2010)

Hoque, M.T., Chetty, M., Sattar, A.: Protein folding prediction in 3D FCC HP lattice model using genetic algorithm. In: IEEE Congress on Evolutionary Computation, vol. 2007, pp. 4138–4145 (2007)

Jiang, T., Cui, Q., Shi, G., Ma, S.: Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. The Journal of Chemical Physics 119, 4592 (2003)

Klau, G.W., Lesh, N., Marks, J., Mitzenmacher, M.: Human-guided tabu search. In: The Eighteenth National Conference on Artificial Intelligence (AAAI 2002), Edmonton, AB, Canada (2002)

Lau, K.F., Dill, K.A.: A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22(10), 3986–3997 (1989)

Lesh, N., Mitzenmacher, M., Whitesides, S.: A complete and effective move set for simplified protein folding. In: Research in computational molecular biology (RECOMB), pp. 188–195. ACM (2003)

Levinthal, C.: Are there pathways for protein folding? Journal of Medical Physics 65(1), 44–45 (1968)

Mann, M., Smith, C., Rabbath, M., Edwards, M., Will, S., Backofen, R.: CPSP-web-tools: a server for 3D lattice protein studies. Bioinformatics 25(5), 676 (2009)

Mann, M., Will, S., Backofen, R.: CPSP-tools – Exact and complete algorithms for high-throughput 3D lattice protein studies. BMC Bioinformatics 9(1), 230 (2008)

Patton, A.L., Punch III, W.F., Goodman, E.D.: A standard GA approach to native protein conformation prediction. In: The 6th International Conference on Genetic Algorithms, CA, USA (1995)

Rashid, M.A., Newton, M.A.H., Hoque, M.T., Shatabda, S., Pham, D., Sattar, A.: Spiral search: a hydrophobic-core directed local search for simplified PSP on 3D FCC lattice. BMC Bioinformatics 14(suppl. 2), S16 (2013)

Rashid, M.A., Shatabda, S., Newton, M.A.H., Hoque, M.T., Pham, D.N., Sattar, A.: Random-walk: a stagnation recovery technique for simplified protein structure prediction. In: BCB, pp. 620–622. ACM (2012)

Shatabda, S., Newton, M.A.H., Pham, D.N., Sattar, A.: Memory-based local search for simplified protein structure prediction. In: The 3rd ACM Conference on Bioinformatics, Computational Biology and Biomedicine, BCB 2012, Orlando, FL, USA. ACM (2012)

Shatabda, S., Newton, M.A.H., Rashid, M.A., Pham, D., Sattar, A.: The road not taken: retreat and diverge in local search for simplified protein structure prediction. BMC Bioinformatics 14(suppl. 2), S19 (2013)

Simons, K.T., Bonneau, R., Ruczinski, I., Baker, D.: *Ab initio* protein structure prediction of CASP III targets using ROSETTA. PROTEINS: Structure, Function, and Bioinformatics (suppl. 3), 171–176 (1999)

Tantar, A.-A., Melab, N., Talbi, E.-G.: A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. Soft Computing-A Fusion of Foundations, Methodologies and Applications 12(12), 1185–1198 (2008)

Thachuk, C., Shmygelska, A., Hoos, H.H.: A replica exchange Monte Carlo algorithm for protein folding in the HP model. BMC Bioinformatics 8(1), 342 (2007)

Torda, A.E.: Protein threading. In: The Proteomics Protocols Handbook, pp. 921–938 (2005)

Ullah, A.D., Kapsokalivas, L., Mann, M., Steinhöfel, K.: Protein folding simulation by two-stage optimization. In: Cai, Z., Li, Z., Kang, Z., Liu, Y. (eds.) ISICA 2009. CCIS, vol. 51, pp. 138–145. Springer, Heidelberg (2009)

Ullah, A.D., Steinhöfel, K.: A hybrid approach to protein folding problem integrating constraint programming with local search. BMC Bioinformatics 11(suppl. 1), S39 (2010)

Unger, R., Moult, J.: A genetic algorithm for 3D protein folding simulations. In: The 5th International Conference on Genetic Algorithms, p. 581. Morgan Kaufmann Publishers (1993)

Yue, K., Dill, K.A.: Sequence-structure relationships in proteins and copolymers. Physical Review E 48(3), 2267 (1993)

Zhang, Y., Skolnick, J.: The protein structure prediction problem could be solved using the current PDB library. The National Academy of Sciences of the United States of America 102(4), 1029 (2005)