JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 1, 2009
© Mary Ann Liebert, Inc.
Pp. 85–103
DOI: 10.1089/cmb.2008.0082

Extended HP Model for Protein Structure Prediction

TAMJIDUL HOQUE,[1] MADHU CHETTY,[2] and ABDUL SATTAR[1]

ABSTRACT

This paper describes a detailed investigation of a lattice-based HP (hydrophobic-hydrophilic) model for *ab initio* protein structure prediction (PSP). The outcome of the simplified HP lattice model has high degeneracy, which could mislead the prediction. The HPNX model was proposed to address the degeneracy problem as well as to avoid the conformational deformity with the hydrophilic (P) residues. We have experimentally shown that it is necessary to further improve the existing HPNX model. We have found and solved the critical error of another existing YhHX model. By extracting the significant features from the YhHX for the HPNX model, we have proposed a novel hHPNX model. Hybrid Genetic Algorithm (HGA) has been used to compare the predictability of these models and hHPNX outperformed other models. We preferred 3D face-centered-cube (FCC) lattice configuration to have closest resemblance to the real folded 3D protein.

Key words: protein structure prediction, novel low resolution model, genetic algorithm.

1. INTRODUCTION

FOR AN EFFECTIVE AND FASTER EXPLORATION of the protein structure prediction (PSP) landscape, various types of lattice models are used and are found to be useful for investigations. Usually, a particular lattice model is adopted with the intention of restricting the protein structure space (Wroe et al., 2005) to encodable structures that otherwise would not have been encodable (Alm et al., 2002) in the unrestricted continuous and complex structure space. The usefulness of the low-resolution modeling for solving the *ab initio* PSP problem in practice can be found elsewhere (Baker, 2006; Chivian et al., 2003; Samudrala et al., 1999; Hinds and Levitt, 1994; Koehl and Levitt, 1999; Kolinski et al., 2003; Schueler-Furman et al., 2005; Xia et al., 2000). If high-resolution models are to be used, this can be done for a smaller pool of approximate conformations obtained by selecting the superior solutions of simplified (i.e., low resolution) lattice model from a huge pool of approximate conformations. This two-stage hierarchical paradigm improves the overall computational time required for solving the *ab initio* problem. For instance, in Samudrala et al. (1999), 10,000 fit samples were taken from a pool of a possible 10 million conformations by using the simple tetrahedral lattice model, and then those 10,000 samples were improved for further investigation, which helps scaling down the number of fitter solutions further in the next step.

Among various lattice models based on different numbers of beads, the hydrophobic-hydrophilic (HP) lattice model (being simple) has always played a vital role for research in the PSP problem. The rationale

[1]Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Nathan, QLD, Australia.
[2]Gippsland School of Information Technology (GSIT), Monash University, Churchill, VIC, Australia.

behind the application of lattice-based low-resolution prediction comes from the fact that the direct real structure prediction is too complex to be handled using existing resources. However, inclusion of all possible conformations, even using simplified lattice model for a sequence of small or moderate length, is astronomical (Chen and Lin, 2002; Guttmann, 2005; MacDonald et al., 2000; Schiemann et al., 2005). The prediction based on the simplified lattice has proven to be *NP-complete* (Berger and Leighton, 1998; Crescenzi et al., 1998). Therefore, lattice-based nondeterministic approaches become most feasible in solving PSP problems.

The PSP, using the low-resolution lattice model, has been attracting researchers from several perspectives, for example, development of strategies for nondeterministic approaches (Hoque et al., 2005, 2006a), modeling with varying number of beads or alphabets (Bornberg-Bauer, 1997) and various possible set of interaction values forming different fitness function of the respective models, structure, and the resolution of regular structure such as two-dimensional (2D) square or three-dimensional (3D) cube, tetrahedral (Hinds and Levitt, 1994), triangular (Agarwala et al., 1997), or face-centered-cube (FCC) (Backofen et al., 2000), and so on. Though being popular, the simple but crucial two-bead HP lattice model (Dill, 1985) needed to be extended and modified for two reasons. Firstly, the simplified HP model, having two beads, produces relatively large degeneracy (i.e., different possible conformations with the same energy) (Backofen et al., 1999). Consequently, these redundant lattice conformations are processed by the Genetic Algorithm (GA), making the search very time intensive, and this can also result in misleading the search due to loss of significant conformations in the multitude. Secondly, since the locations of polar segments (i.e., P) are not directly optimized (Guo et al., 2006) when searching for optimal structures, this can result in distorted structures while predicting, especially if these segments are too long or are located at the ends of the sequences.

The possible enhancement of the HP model to avoid unwanted structural deformity involved in the mapping led us to consider the developed HPNX model (Bornberg-Bauer, 1997; Backofen et al., 1999) as a logical extension for reducing the degeneracy problem. Based on our experiments using the HPNX model and reported later in this paper, we observe that it is beneficial to improve the HPNX model further. This paper is devoted to the development of a novel lattice model, referred to as the hHPNX model, in which the H bead of the HPNX model is split into two parts (h and H), emphasizing the properties of two amino acids of the H group of the earlier models (Crippen, 1991). The distinctly different interactions of the two split groups of H, referred as h and H in this paper, are found to be highly consistent for the examined protein data sets we investigated. For PSP, being a computationally intensive as well as NP-complete problem, a nondeterministic search approach is considered to be an appropriate option. As a search technique, we have opted for the heuristic-based Hybrid Genetic Algorithm (HGA) for investigating the predictability of the models. The HGA was earlier presented (Hoque et al., 2005, 2006a, 2006b, 2007b) for the HP models. The HGA has been extended and generalized in this paper for the following models: HP, HPNX, and hHPNX, in 3D FCC lattice configuration.

The remainder of the paper is organized as follows. Section 2 defines the alpha-carbon (C_α) root-mean-square-deviation (cRMSD), which is used for evaluating the model performance by comparing the outcome of the model with the real folded protein. Section 3 provides the background of the simplified lattice model and the interaction potentials, and proposes the novel hHPNX model. In Section 4, heuristics based on domain knowledge have been developed and extended for the hHPNX model and the HPNX model. The search strategy, based on the heuristics, is developed subsequently in Section 5. Simulation results are presented in Section 6. Finally, Section 7 provides the conclusion.

2. MEASURING CONFORMATION RESEMBLANCE

To measure the performance of prediction of the different models, the output of a model can be compared with the output of the corresponding real folded protein. The closeness of the output is measured by using root-mean-square-deviation (RMSD) (Backofen et al., 2000). The RMSD is defined as follows:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(r_{ij}^{\mathrm{model}} - r_{ij}^{\mathrm{real}})^2}{\frac{n(n-1)}{2}}} \tag{1}$$

In the above Equation (1), the term *model* in the equation refers to the conformation generated from the model and $r_{ij}^{\text{model}} = |r_i^{\text{model}} - r_j^{\text{model}}|$, where r_i and r_j indicate the ith and jth residues respectively in a sequence of length n. Similarly, the term *real* refers to the real folded protein obtained from the Protein Data Bank (PDB, 2007) and $r_{ij}^{\text{real}} = |r_i^{\text{real}} - r_j^{\text{real}}|$. n indicates the number of residues in the sequence.

Since the C_α coordinate is considered as the center of an amino acid residue, the measure of resemblance (Equation (1)) can be termed cRMSD. Before applying cRMSD, we must ensure that the average distance of any two consecutive residues in the model, compared to the real protein, be in the same scale and therefore the real folded protein data is normalized before applying cRMSD.

3. LATTICE MODELS

Low resolution models are helpful in effective and fast (Backofen et al., 2000; Dill et al., 1995) exploration of the vast convoluted landscape of the PSP problem. Only these simple models are described here, instead of the complex models. This is because the complex models, which consider the individual interaction value (Miyazawa and Jernigan, 1996; Buchler and Goldstein, 1999) for each pair of amino acids, could be training set dependent and also could cause an extremely slow exploration of the search landscape. In this section, after reasoning the application of the FCC model configuration, the details of the widely used simplified lattice models and their variations, based on different interaction potentials, are presented.

3.1. Face-centered-cube configuration

Before conducting a comparative analysis on the prediction performance of the two simplified models, that is the HP and the HPNX with the proposed hHPNX model, we present the reasons for choosing the 3D FCC orientation for regular lattice models:

(i) With the availability of the full proof of *Kepler Conjecture* (Hales, 2005), it has been proved that the 3D FCC is the densest sphere packing configuration. Therefore, it can provide the densest protein core (Yue and Dill, 1993; Backofen et al., 2000) during the process of protein structure prediction. However, the protein core needs not necessarily be the most compact one (Backofen et al., 1999) for all the cases of the folded 3D or tertiary protein.

(ii) In FCC, a residue can have 12 neighbors in a 3D space (Fig. 1a and 1b) and six neighbors in a hexagonal pattern in 2D space. However, as mentioned earlier, FCC offers the densest compaction amongst all possible lattice configurations. Therefore, for a region of fixed volume of space, logically inferring, FCC can offer highest degree of freedom for placing a residue in suitable neighboring position compared to any other form of lattice configuration. For instance, FCC is parity (Backofen et al., 2000) problem free, whereas a square or a cube lattice is not.

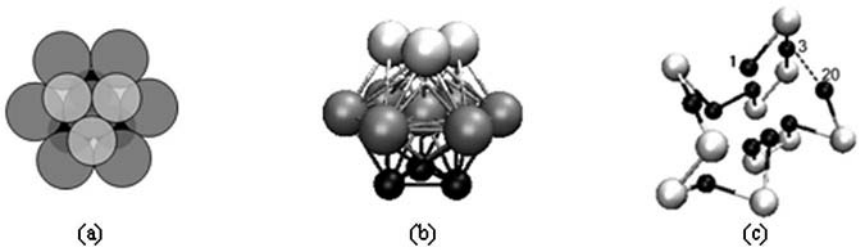

FIG. 1. FCC packing. **(a)** Top view. **(b)** Front view. Layers are separated using colors and size. Dedicated connections provide assistance in conceptualizing. The centered sphere at the second layer has a total of 12 neighbors in the 3D space. It has 3 neighbors at its top layer, 6 more at the middle layer, and 3 more at the bottom layer. **(c)** A conformation using HP model in 3D FCC lattice, where the spheres 'black ball' and 'white ball' indicate hydrophobic and hydrophilic residues, respectively. There are a total of 29 topological neighbors (TNs). Therefore, $F = -(\text{TN Count}) = \text{A}$ dotted line between residue 3 and 20 indicates one TN, for instance. Drawing tool VMD (VMD 2007) has been used with lattice supported file (Potzsch et al., 2006).

(iii) Therefore, within the lattice constraints, FCC lattice can provide maximal degree of freedom and FCC can provide closest resemblance to the real or high-resolution folds.

With respect to modeling the real protein, the FCC model can therefore provide closest conformational alignment (Raghunathan and Jernigan, 1997) among the lattice disciplines with regular configuration or regular lattice orientation.

3.2. Alphabets and potentials

The least complex and yet highly effective (Backofen et al., 2000) representation of lattice models for protein folding investigation is the HP model (Dill et al., 1995), which uses two letter alphabets, namely H and P, where H indicates a *hydrophobic* amino acid and P represents a polar, or *hydrophilic*, amino acid. This simple model has been developed based on the dominating hydrophobic force. The energy function for this HP model is calculated in the following manner. Two residues are *topological neighbors* (TN), by being at the adjacent or immediate neighboring lattice positions and not sequential in the original connectivity, as indicated by the dotted lines in Figure 1c. If both of them involving TN are Hs, then an energy contribution (say ε) termed *interaction potential* is made, where ε in general, is considered to have a value of -1. The sum of ε for HH interactions in a conformation using HP model is referred to as the fitness F of that particular conformation (Fig. 1c). Any valid conformation on a lattice must have a *self-avoiding-walk* (SAW). The conformational search is expected to predict the desired conformation that has a minimal F. The three important variations of the HP model, based on the values of the interaction potential, are shown in Figure 2a–2c.

Figure 2d shows the four-bead HPNX model, the logical extension of the HP model, which was introduced (Bornberg-Bauer, 1997; Backofen et al., 1999) to limit the degeneracy problem. In the HPNX model, the P (polar) monomer of the HP model is split and the splitting is based on the category of the electric charge, namely positive (P), negative (N), and neutral (X). The member amino acids belonging to the P, N, X groups of the HPNX model are given in Table 1.

Backofen et al. (1999) showed that the HPNX model can reduce the degeneracy over the HP model and lead to a more effective protein folding simulation. Based on the structural observation of a protein data set of 57 protein sequences, Crippen (1991) proposed a potential interaction matrix (Fig. 3a), where the amino acids were divided into four different groups. These classifications of amino acids into four groups led to a new four-bead model, which is different from the HPNX model. The four beads or groups were represented using a single-letter amino acid code as follows: 1 = {GYHSRNE}, 2 = {AV}, 3 = {LICMF}, and 4 = {PWTKDQ}. Thus, we will refer to Crippen's matrix in this paper as the *1234* model.

Crippen (1991) emphasized the importance of treating the two amino acids—alanine and valine (shown as group 2 = {AV} in Fig. 3a)—separately, because these two amino acids had interactions in the observed data set which are consistently different when compared with the interactions of other hydrophobic residues. The conclusion on the difference in interactions is based on their geometrical positions in the folded protein. Further, it is to be noted that frequency of occurrence of these two amino acids, on an average, is observed to be the highest amongst all the amino acids in proteins (Jordan et al., 2005).

The YhHX matrix is shown in Figure 3b, which has been derived (Bornberg-Bauer, 1997) from Crippen's 1234 matrix by converting the values of the matrix elements from real numbers into integers but maintained

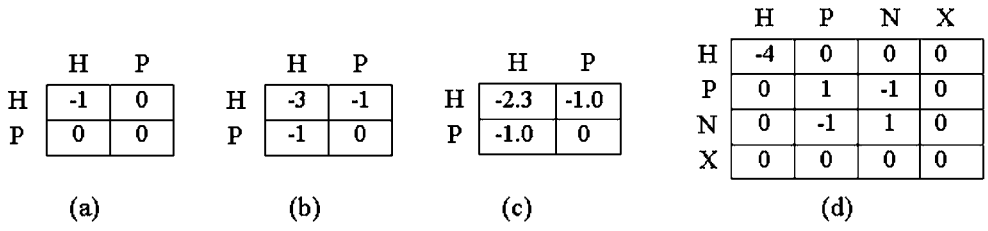

FIG. 2. Various interaction-potential matrices. Negative value indicates reward for being a topological neighbor (TN) in the lattic model, where interaction with positive value represents a penalty. A '0' indicates neutral (or, no) interaction. **(a)** Original HP model (Dill, 1985) is shown. Other variations of the HP model are shown from Bornberg-Bauer (1997) **(b)** and from Li et al. (1996) **(c)**. HPNX (Backofen et al., 1999) matrix is shown in **(d)**.

TABLE 1. MEMBER AMINO ACIDS OF THE GROUPS DEFINED IN THE HPNX MATRIX

Group	*Amino acid*		
	Name	*3-letter code*	*1-letter code*
Positive (P)	Arginine	Arg	R
	Histidine	His	H
	Lysine	Lys	K
Negative (N)	Aspartic acid	Asp	D
	Glutamic acid	Glu	E
Neutral (X)	Asparagine	Asn	N
	Cysteine	Cys	C
	Glutamine	Gln	Q
	Serine	Ser	S
	Threonine	Thr	T
	Tyrosine	Tyr	Y
Hydrophobic (H)	Alanine	Ala	A
	Glycine	Gly	G
	Isoleucine	Ile	I
	Leucine	Leu	L
	Methionine	Met	M
	Phenylalanine	Phe	F
	Proline	Pro	P
	Tryptophan	Trp	W
	Valine	Val	V

approximately the same ratio for the entries. This YhHX model has continued to be followed (Buchler and Goldstein, 1999, 2000; Li et al., 2005) later on, perhaps because of the simplification of having integer quantities.

3.3. Corrections to the YhHX matrix

While investigating the YhHX model, we detected that the conversion of the 1234 matrix to YhHX has been incorrectly done in three different places. The first error is in element (2, 2), highlighted in black in Figure 3a, with its corresponding correct value highlighted in Figure 3b. The second and third errors in the YhHX matrix relate to the occurrence frequencies of the amino acids. These two errors are highlighted in gray in Figure 3b. In the rest of this section, we will elaborate on this erroneous conversion

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

(a)

	Y	h	H	X
Y	0	-1	-1	2
h	-1	-2	-4	2
H	-1	-4	-3	0
X	2	2	0	0
f_q	10	16	36	28

(b)

	Y	h	H	X
Y	0	-1	-1	2
h	-1	2	-4	2
H	-1	-4	-3	0
X	2	2	0	0
f_q	36	16	20	28

(c)

FIG. 3. **(a)** Crippen's matrix, showing classification of amino acid and their interaction-potentials. The indices represent the group of amino acids (presented using single-letter amino acid code): 1 = {GYHSRNE}, 2 = {AV}, 3 = {LICMF}, and 4 = {PWTKDQ}. **(b)** YhHX matrix: as same order as the Crippen's matrix. Here, f_q implies the percentage of occurrence frequencies of amino acid for each of the four groups. **(c)** Corrected YhHX as it should have been considered in Bornberg-Bauer (1997). Blacked and shaded entries are values that are under consideration.

and will provide correction subsequently. In the next two subsections, we will discuss the steps involved in reinstating the correct values. We then produce the correct YhHX matrix in Figure 3c.

Recalling the discussions on Crippen's matrix from the previous section and also the matrices in Figure 3a–3c, we note that the corresponding elements of the matrices represent the same interactions. Hence, element (2, 2) of Figure 3a and 3b represent the interactions within group $2 \equiv \{\text{AV}\}$ in Figure 3a or the hh interaction in Figure 3b. We identified that the hh interaction as the highlighted element in Figure 3b is shown incorrectly as "−2," whereas it should have been correctly recorded as "2" (details in next section). Because of the significance of this interaction in protein folding, its incorrect conversion as element (2, 2) of matrix in Figure 3b cannot be ignored.

The second and third conversions are related to the frequencies indicated by f_q in the matrix of Figure 3b. The frequency of occurrence (f_q), which should add up to a total of 100%, is adding up to only 90%. Further, the two matrix elements highlighted in gray in Figure 3b indicate the frequency of occurrence of groups Y and H. As shown in this figure, the two values are erroneously recorded, respectively, as 10% and 36%. Instead, as we will show in the following section, these values should be 36% and 20%, respectively.

The underlying analysis to obtain corrections for the identified problems is presented in Sections 3.3.1 and 3.3.2, reinstating the corrected matrix shown in Figure 3c.

3.3.1. Interaction value of hh in YhHX matrix. To justify our claims about the incorrect recording of the elements in the YhHX matrix in the previous section, we begin by verifying Crippen's particular statement about the different interaction behavior of the two amino acids (alanine and valine) of the h group. For verification, we repeated Crippen's experiment (Hoque et al., 2007a) by choosing the same dataset of 57 protein sequences that were selected in Crippen (1991), but by taking these sequences from the most recent version (in 2007) of the publicly available source, Protein Data Bank (PDB, 2007). However, we had to reduce the dataset to 56 sequences since the sequence *2mt2* had been omitted from the recent PDB. To ensure that Crippen's selection of sequences were without any bias, we selected 10 protein sequences randomly from the PDB to form a second data set which is used for cross-validation. For both the datasets, only the first polypeptide chain was taken into account: that is, the chain is read up to the first break point, as followed in Crippen (1991). For example, in case of the protein sequence having PDB ID *2pka*, the first 77 residues were taken into account. The first group of our experimental protein data set is presented in the format *current PDB ID* (previous ID).

The 56 protein sequences mentioned above are as follows: *2mlt* (1mlt), *1ppt*, *1crn*, *7rxn* (3rxn), *1dur* (1fdx), *2ovo*, *4pti*, *3ebx* (2ebx), *2sn3* (1sn3), *2abx* (1abx), *3icb* (2icb), *2pka*, *351c*, *1cc5*, *1hip*, *1cyo* (2b5c), *2gn5*, *4fxc* (3xfc), *2pcy*, *5cyt* (4cyt), *5fd1* (2fd1), *2cdv*, *1rei*, *5cpv* (3cpv), *1ccr*, *3c2c*, *2hmq* (1hmq), *2rhe*, *2cy3* (1cy3), *155c*, *1pp2*, *1bp2*, *3rn3*, *2ccy*, *2aza* (1aza), *1lzl*, *1ecd*, *2fox* (4fxn), *2mhb* (1mhb), *2hhb*, *2sns*, *1fxl*, *2lhb*, *2sod*, *1lh1*, *5mbn* (3mbn), *4dfr*, *1zlm*, *9wga* (3wga), *4gcr* (1gcr), *2buk* (2stv), *7fab* (3fab), *1ppd*, *2act*, *2cna*, and *1tim*.

The 10 additional protein sequences, *1pjf*, *2crt*, *8ame*, *1tuk, 2ptl*, *2ffw*, *1ae3, 1gh1*, *1a7d*, and *1eca*, given in order of their increasing length, form the second group.

To facilitate the presentation in a tabular form and present it in two tables (Tables 2 and 3); the first group of 56 sequences is broken into two sets, keeping the same sequential order given above. These tables, including Table 5, show the minimum average Euclidian distance (see Equation (2)) among several groups from their geometrical positions in a folded protein, which can be considered as the measure of the potential interactions amongst different groups.

To define the minimum average Euclidian distance, assume the interactions between any two groups of amino acids, say group X and group Y, is represented as XY. Further, assume the group members to be denoted by x and y where $x \in \text{X}$ and $y \in \text{Y}$. The cardinality of X and Y is represented respectively as |X| and |Y|.

Now, the minimum average Euclidian distance, $\overline{d}^{E}_{\min}$, between X and Y can be expressed as:

$$\overline{d}^{E}_{\min} = \frac{1}{|\text{X}|} \sum_{i=1}^{|\text{X}|} \min\{(x_i, y_1), (x_i, y_2), (x_i, y_3), \ldots, (x_i, y_{|\text{Y}|})\} \tag{2}$$

TABLE 2. DATA SET CONSISTS OF THE FIRST 28 SEQUENCES FROM THE FIRST DATA SET OF 56 SEQUENCES

PDB ID	*hh*	*hH*	*HH*	*hY*	*HY*	*XY*	*hX*
2mlt (1mlt)	**6.367593**	4.177397	4.985412	4.84788	4.80347	4.77909	4.65215
1ppt	**11.62076**	5.533692	5.332277	4.627	4.17016	4.36981	4.20055
1crn	**9.408218**	4.780548	6.239725	4.22786	5.04039	4.57178	5.3477
7rxn (3rxn)	**8.552146**	5.624427	7.60697	5.25907	4.8047	5.41156	5.93622
1dur (1fdx)	**8.262073**	4.513595	6.219922	4.12937	4.68521	4.5928	5.33631
2ovo	6.527874	5.079051	7.937867	5.52703	4.99916	4.96608	**6.77048**
4pti	**12.01579**	5.779846	7.472582	3.79219	4.3062	4.30998	5.25532
3ebx (2ebx)	**11.31001**	3.101884	8.184653	5.18959	3.42753	4.40254	2.6345
2sn3 (1sn3)	**10.69952**	4.290174	8.28926	4.65826	4.38104	4.195	5.72338
2abx (1abx)	**8.65277**	5.283757	7.715188	5.33383	4.56445	4.83966	4.53347
3icb (2icb)	**8.023476**	4.990781	5.404907	5.00787	4.45831	4.18621	4.25396
2pka	7.614279	4.997368	**9.454919**	4.92551	4.54538	4.99525	4.32554
351c	6.417271	5.376686	**7.840636**	4.60088	4.76186	4.73341	5.54269
1cc5	**8.526686**	5.86846	6.567001	4.68708	4.21577	4.31785	4.59607
1hip	6.772224	**7.104508**	6.1309	4.6833	5.53784	4.77482	5.11517
1cyo (2b5c)	**9.40202**	6.405175	7.05977	4.59436	4.46357	4.37149	4.19428
2gn5	**9.815458**	5.642486	8.088104	4.42251	5.0498	4.72413	4.10944
4fxc (3xfc)	**8.504008**	5.097056	7.969018	4.3765	4.97746	4.74396	5.8899
2pcy	**8.543231**	4.892625	7.495968	4.50292	4.42545	4.23634	4.81599
5cyt (4cyt)	**9.709897**	5.293058	6.853845	4.64932	4.75179	4.53952	4.24634
5fd1 (2fd1)	**9.0402**	4.740451	6.545186	4.8958	4.85518	4.94919	5.31106
2cdv	**8.642787**	7.034808	7.434779	5.49078	4.46056	5.60071	4.2835
1rei	**11.001**	5.736049	8.454562	4.41909	4.62238	4.66647	4.42297
5cpv (3cpv)	**6.955895**	4.938366	6.023143	5.50954	5.06136	5.0797	5.48504
1ccr	**10.66491**	5.397467	7.004012	4.72032	4.3823	4.69493	4.76048
3c2c	**8.186079**	6.240735	6.475626	4.44438	4.78102	4.66128	5.28316
2hmq (1hmq)	**9.360979**	6.117882	6.962379	4.24041	4.53634	4.51052	3.87187
2rhe	**10.20137**	5.667569	8.269964	4.61681	4.25666	4.56032	4.81806

Several interaction pairs have been checked, such as hh, hH, HH, and so on. For these interaction measures, the average of the minimum Euclidian distance is taken between the members of any interaction set, based on their geometrical positions in the folded protein. The distance of two residues ith and jth are taken *iff* $|i - j| > 4$ according to Brocchieri and Karlin (1995). The chain length ranges from 27 to 114.

Here, x and y represent, respectively, the ith and jth residues in the group X and Y. The term (x_i, y_j) indicates the Euclidian distance between x_i and y_j, and the 'min' indicates the minimum of the Euclidian distances.

In the experiment, using Equation (2), we compared the average of the minimum Euclidian distances of each member of the Y, h, H, and X sets of the YhHX matrix (Bornberg-Bauer, 1997) in order to estimate the interaction potentials of the sets hh, hH, HH, hY, HY, XY, and hX, based on their geometrical positions in a 3D folded protein. Coordinates of C_α of a residue has been used to measure the distances. The measuring distance for any two residues, say ith and jth, is taken iff $|i - j| > 4$, as it was in (Brocchieri and Karlin, 1995). Table 4 summaries the findings of this experiment for the first set of 56 sequences, showing the average and the standard deviation (row wise) of the normalized value. In Table 5, the result obtained for the second dataset also provides a summary of the experimental results at the last two rows of the same table. Comparing these two summaries, we can cross validate that the hh interaction is consistently high compared to other interactions. In other words, the members of the hh group (i.e., Ala and Val) consistently keep themselves away from each other, exhibiting repulsion between them. Therefore, the matrix entry for hh interaction should be positive (i.e., $+2$) instead of negative (i.e., -2) because the positive value indicates repulsion and the negative value indicates affinity.

3.3.2. Occurrence frequency of the groups in YhHX matrix. The occurrence frequencies of the amino acids distributed for four different groups of the YhHX matrix are shown in the last row of the matrix

TABLE 3. DATA SET CONSISTS OF THE REMAINING 28 DATA FROM THE FIRST GROUP OF 56 SEQUENCES

PDB ID	*hh*	*hH*	*HH*	*hY*	*HY*	*XY*	*hX*
2cy3 (1cy3)	**9.281676**	5.969045	7.627191	4.65379	4.56772	5.28935	5.68037
155c	**0.165587**	0.122048	0.170465	0.09695	0.10256	0.10763	0.11873
1pp2	**9.555322**	5.062259	7.814553	4.89304	4.45955	4.87718	4.69848
1bp2	**9.680847**	4.566841	7.706119	5.06508	4.50433	4.69837	5.42968
3rn3	**7.430872**	5.10297	**7.43667**	4.73607	4.12831	4.67952	4.69784
2ccy	6.23967	5.304916	**7.688208**	5.59853	4.80733	5.2063	6.55376
2aza (1aza)	**7.972992**	5.768998	6.874242	4.36565	4.72901	4.59575	4.40732
1lzl	**6.409071**	5.128551	6.863231	4.6679	4.6469	5.06841	5.16111
1ecd	**7.724681**	5.257093	6.256951	4.18826	4.48101	4.88548	4.748
2fox (4fxn)	**8.356373**	4.773615	6.449546	4.64961	4.48901	4.26363	5.03024
2mhb (1mhb)	**7.298212**	6.133656	6.258017	4.34404	4.23733	4.83695	4.71941
2hhb	**6.746687**	6.055006	6.252882	4.39142	4.28591	4.90959	4.79548
2sns	**7.41029**	5.215863	6.865676	4.68796	4.92437	4.74831	4.87645
1fxl	**8.336546**	5.491308	7.889971	4.09093	4.20343	4.71917	5.32687
2lhb	**7.13039**	5.906971	6.748795	5.24624	4.73701	4.91157	5.18915
2sod	7.339638	5.295761	**8.200556**	5.15008	4.38795	4.40993	4.87777
1lh1	**6.803112**	5.401621	6.770869	5.06381	4.6661	4.98484	4.8559
5mbn (3mbn)	**7.286236**	5.3647	6.2805	4.48687	4.36297	4.48058	4.74125
4dfr	**0.119374**	0.066601	0.09756	0.06099	0.05581	0.05697	0.06011
1zlm	5.904595	4.775562	**7.04619**	4.46834	4.07612	4.28061	5.42924
9wga (3wga)	**12.0906**	4.82203	5.864122	3.86875	4.15949	4.3295	6.06623
4gcr (1gcr)	**11.20687**	4.886947	8.660066	4.07504	3.54564	4.14054	5.15644
2buk (2stv)	**8.916063**	5.693849	7.986415	4.50824	4.40876	4.78408	4.92073
7fab (3fab)	**8.506969**	6.024638	8.28134	4.65237	4.66051	4.58765	4.59252
1ppd	7.341831	5.584384	**7.433267**	4.53251	4.37745	4.33457	4.93837
2act	**0.138244**	0.094384	0.128596	0.08	0.07849	0.07592	0.07431
2cna	**8.291953**	6.052063	7.040712	4.38955	4.37844	4.37124	4.81956
1tim	**8.65277**	5.283757	7.715188	5.33383	4.56445	4.83966	4.53347

The same measures as in Table 2 are continued here. In this table, the chain length ranges from 118 to 247.

of Figure 3b, and for verifying the occurrence frequencies these have been compared with other sources mentioned in Table 6.

As noted earlier, these occurrence frequencies should add up to 100% but instead, add up to only 90%. Further, the frequency of Y and H are given in the matrix as 10% and 36%, respectively. We confirm for the correctness of these two values by comparing them with the values given by Bornberg presented in Table 6, along with the values for Y, h, H, and X from other sources, included for comparison and referred to as *group-wise frequencies*. The table also includes *group-wise average frequencies* that indicate the average frequency of a member of a group.

By comparing the three sets of values (Table 6) for frequency, it is clear that the three different values for frequencies of h (that is, 16%, 14.2%, and 15.03%) are reasonably close. Also, the frequency values for X from the three different sources (28%, 29.3%, and 25.45%) can be considered to be reasonably close. Due to this closeness, the values for h and X (16% and 28%) proposed by Bornberg can be treated as correct.

TABLE 4. DATASET OF THE FIRST GROUP OF 56 SEQUENCES

	hh	*hH*	*HH*	*hY*	*HY*	*XY*	*hX*
Avg.	0.990372	0.673772	0.800059	0.654984	0.638136	0.685766	0.679997
SD	0.030638	0.097631	0.124991	0.10993	0.1164	0.112036	0.100794

Data shown as the average (avg.) and standard deviation (SD) of the normalized values, measured after implementing the row-wise normalization from Tables 2 and 3.

TABLE 5. DATASET OF ALL 10 SEQUENCES FROM THE SECOND GROUP

PDB ID	*hh*	*hH*	*HH*	*hY*	*HY*	*XY*	*hX*
1pjf	**11.33825**	7.155584	6.584539	7.21642	4.80852	4.30745	11.0103
2crt	**8.993926**	4.409069	7.837169	7.74627	5.16126	6.55286	5.10119
8ame	6.832036	**6.987313**	6.309292	4.61328	5.0806	5.24707	5.00406
1tuk	**7.511582**	4.785005	6.538986	5.27015	4.53898	4.46744	5.29134
2ptl	**7.669825**	5.713559	6.33671	4.24463	4.73833	4.3857	4.34729
2ffw	12.04664	**12.33088**	7.57963	5.11552	5.36316	6.25133	4.75762
1ae3	**9.891208**	5.778435	7.595174	4.45339	5.11109	4.63095	3.8606
1gh1	**8.987582**	4.924321	7.215585	3.81885	4.25337	4.10654	4.7508
1a7d	**8.431158**	5.530274	7.330131	4.82805	4.46102	4.34255	4.22724
1eca	**7.719121**	5.264312	6.25885	4.17917	4.48065	4.88557	4.75049
Avg.	**0.994314**	**0.693585**	**0.769422**	**0.711646**	**0.665243**	**0.736740**	**0.703401**
SD	**0.01798**	**0.132698**	**0.176015**	**0.155437**	**0.156858**	**0.176874**	**0.125742**

The same process has been applied here as in Table 2. In this table, the chain length ranges from 47 to 136. The last two rows show the average (avg.) and standard deviation (SD) of the normalized values, after taking row-wise normalization of the rest of the rows.

However, the interaction of 10% for Y (see also the YhHX matrix in Fig. 3b) is given in Table 6 for the other two sources as 36.1% and 34.47%. Since these two values from two other sources are consistent with each other, this value (i.e., approximately 35% or so) should be treated as the correct value. This is the value which Bornberg's YhHX model should have considered. The closest value available from Bornberg is 36% corresponding to the H interaction. Hence, we can infer that the Y interaction value for YhHX model from Bornberg should have been 36%, a value that can be achieved if this value of 36% (H interaction) is swapped with the value of 10% (Y interaction). Again, if the swapping of the values of Y and H is implemented, we are left with the values of 10% for H in Bornberg's YhHX model. Recalling that the occurrence frequencies in this model add up to only 90% (instead of 100%; Table 6) leads us to conclude that the H interaction value should be corrected as 20%. It is to be noted that with this correction, the corrected H interaction value (i.e., 20%) is close to its corresponding values (20.5% and 25.05%) from the other resources given in Table 6. Hence, we conclude that the frequency of H in Figure 3b must be 20%, which is finally reinstated in Figure 3c.

The importance of the h group, even if it consists of only two amino acids (i.e., alanine and valine), is obvious due to its average high occurrence frequency (Table 6). Therefore, the incorrect incorporation of the hh interaction cannot be overlooked. The significance of the h $\equiv$ {AV} group also emphasizes the need of a new model that incorporates the correct interactions to model the protein sequence in a simple but realistic manner. For this, we can reconsider the HPNX model, which is the logical extension of the HP model, but must hybridize it by incorporating separately the strong and consistent properties of the h group, as done in the YhHX matrix.

TABLE 6. COMPARISON OF THE FREQUENCY DISTRIBUTIONS OF THE AMINO ACIDS GROUPS: Y, h, H, AND X OF THE YhHX MATRIX FOUND IN DIFFERENT SOURCES

	Frequency distribution of amino acids								
	Group-wise total (in %)					*Group-wise average (in %)*			
Source	*Y*	*h*	*H*	*X*	*Total*	*Y*	*h*	*H*	*X*
Bornberg-Bauer, 1997	10.0	16.0	36.0	28.0	**90.0**	1.43	**8.0**	7.2	4.66
Beals et al., 1999	36.1	14.2	20.5	29.3	100.1	5.16	**7.1**	4.1	4.88
Jordan et al., 2005	34.47	15.03	25.05	25.45	100.0	4.92	**7.51**	5.01	4.24

The reference Beals et al. (1999) is a web reference based on two articles in well-known journals, one by Dyer (*Journal of Biological Education* 5, 15–24, 1971) and the second by King et al. (*Science* 164, 788–798, 1969). Reference Jordan et al. (2005) is a more recent publication from *Nature*.

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

FIG. 4. Interaction-potentials for the proposed hHPNX matrix.

This hybridization by amalgamation of two important model characteristics results in a novel hHPNX model. This model is formed by considering the logical extension of the P bead of the HP model to be represented by three groups (i.e., P, N, X) and the H bead to be represented by splitting the H group (i.e., h and H). The interaction potentials of the hHPNX matrix are shown in Figure 4, for which we use the interaction potential of the HPNX (Fig. 2c) model and the corrected value of h group from the YhHX model.

4. COMPUTATIONAL SEARCH USING HEURISTICS

As shown in Section 1, the complex landscape can be searched by using a suitably designed genetic algorithm. The HGA developed elsewhere (Hoque et al., 2005, 2006a, 2006b, 2007b) was found to be very effective. Hence, we have carried on the investigations using HGA for the HPNX and the hHPNX models. In a manner similar to our earlier work (Hoque et al., 2005, 2006a, 2006b, 2007b), for the application of the sub-conformation-based heuristics, we have formulated the *Probabilistic Constrained Fitness* (PCF) function based on the core formation concept (Yue and Dill, 1993; Hoque et al., 2005), presented next.

4.1. Building likely sub-conformations

Our optimal core formation concept (Yue and Dill, 1993), as explained elsewhere (Hoque et al., 2005, 2006a, 2006b, 2007b), conceptualizes the protein folding concept by visualizing the tertiary protein as 3-layered kernels. Briefly, in this conceptualization, the inner kernel, called H-Core, is formed mostly by the Hs that form the protein's core. The Ps, due to their affinity with the solvent, tend to remain on the outer surface and form the outermost kernel. In between the inner and outer kernels, a thin composite kernel or layer formed by the covalent bonded H and P can exist: this is referred to as the HP-mixed-layer. In the case of the HPNX model, the H forming the core remains the same but instead of P (polar) we have to consider P (positive), N (negative), and X (neutral) because the P (polar) residue has been split. For the proposed hHPNX model, the inner core consists of h and H type residues, with P (polar) replaced by P (positive), N (negative), and X (neutral), as has been done for the HPNX model. The thin layer between the outer and inner layer basically comprises the consecutive pairs of "$\grave{\lambda}\wp$," where $\grave{\lambda} \in \{H, h\}$ and $\wp \in \{P, X, N\}$. This thin layer is important because its proper shape helps to accommodate the maximal bonding among the $\grave{\lambda}$s inside the inner core, which will ultimately result in the conformation having global minima. Found in the general H-Core Center (HCC) (which is the mean of the coordinates of $\grave{\lambda}$s), a center of these $\grave{\lambda}$s guides the core formation around it, for the HPNX model as well as the hHPNX model. As part of our strategy, local interactions for this thin layer are applied carefully, such as applying highly likely sub-conformations and replacing unlikely sub-conformations of the thin layer, to help reform the cavity towards its optimal capacity.

In order to implement the strategy, we extend the likely sub-conformations (which are either three or four residues long) for the HPNX and hHPNX model and also include the HP model (for the sake of generalization) and for the 3D FCC configuration. The likely sub-conformations are formed based on the following axioms:

Axiom 1: TN is formed if a TN is feasible within a short sub-conformation having a fitness value which reflects favorable local interaction.

Axiom 2: As the Hs or hydrophobic residues stay away from solvent and form the core, they need to be placed towards HCC (i.e., towards the inner side of a conformation), while the polar (i.e., hydrophilic) residue should be placed away from HCC (i.e., towards the outer side of a conformation).

Axiom 3: Polar residue can be either neutral, positively charged, or negatively charged. Therefore, a short sub-conformation will accordingly reflect either a repulsive or an attractive force (i.e., local interaction, see Axiom (1) above) for the polar residue.

Using these axioms, we next build the likely sub-conformations for the intermediate thin layer for the HP, HPNX, and hHPNX models. Further, those likely sub-conformations cover the outer layer of the HPNX and hHPNX models.

The sub-conformations for the three models are defined next:

(i) *Sub-conformations for HP model*. Figure 5 shows the likely sub-conformation for the HP model in the 3D FCC configuration of the lattice model. For "–PHP–," Figure 5a is formed based on Axiom (2) above. Similarly, Axiom (2) is true for Figures 5d and 5e. Axioms (1) and (2) correspond, respectively, to Figures 5b and 5c.

(ii) *Sub-conformations for HPNX model*. The concept for -PHP- sub-conformation given above is similarly extended for the HPNX model. The "–$\wp$H$\wp$–" sub-conformation, where $\wp \in$ {P, X, N}, corresponds to Figure 5a according to Axiom (2). The "–H$\wp$H–" sub-conformation corresponds to Figure 5b by applying Axioms (1) and (2). The sub-conformation for "–H$\wp\wp$H–" corresponds to Figure 5c according to Axioms (1) and (2).

Finally, sub-conformation for "–$\wp$HH$\wp$–" can be further divided into the following three sub-categories:

- "–PHHP–" and "–NHHN–" correspond to Figure 5e, due to Axioms (2) and (3) for repulsive force.
- "–XHHX–," "–XHHP–," and "–XHHN–" correspond to any of the alternatives of Figures 5d or 5e by Axiom (2) only (as having no local interaction within the sub-conformation due to the presence of the neutral group X).
- "–NHHP–" corresponds to Figure 5d by Axioms (2) and (3) for attractive force.

(iii) *Sub-conformations for hHPNX model*. Sub-conformation for the sub-sequence "–$\wp\grave{\wp}\wp$–," where $\grave{\wp} \in$ {H, h} and $\wp \in$ {P, X, N}, corresponds to Figure 5a by Axiom (2). Sub-conformation corresponding to "–$\grave{\wp}\wp\grave{\wp}$–" corresponds to Figure 5b by Axioms (1) and (2). However, the "–h$\wp$h–" of "–$\grave{\wp}\wp\grave{\wp}$–" is the same as in Figure 6a, due to Axioms (2) and (3) for the observed "hh" interaction. For sub-sequence "–$\grave{\wp}\wp\wp\grave{\wp}$–," the sub-conformation corresponds to Figure 5c by Axioms (1) and (2), but again "–h$\wp\wp$h–" corresponds to Figure 6b by Axioms (2) and (3) due to the repulsive tendency (from Tables 4 and 5 of hh interaction). Finally, sub-conformations for "–$\wp\grave{\wp}\grave{\wp}\wp$–" correspond exactly to "–$\wp$HH$\wp$–" in the HPNX model in all cases, including the three sub-categories mentioned earlier. This can easily be visualized by considering H or h of the hHPNX matrix as equivalent to the H of the HPNX model.

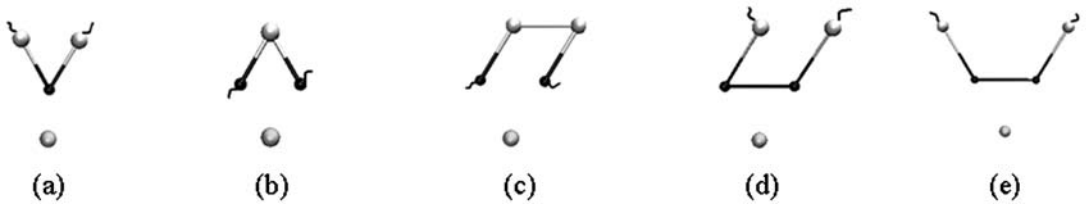

FIG. 5. In the HP model, potential sub-conformation in 3D space for the sub-sequences –PHP–, –HPH–, and –HPPH– are shown in **(a)**, **(b)**, and **(c)**, respectively. For –PHHP– there are two alternatives, as shown in **(d)** and **(e)**. Spheres 'black ball,' 'white ball' and 'gray ball,' respectively, indicate an H (and h for the hHPNX model), a P (and P/N/X for both the HPNX and hHPNX models), and the approxiate position of HCC.

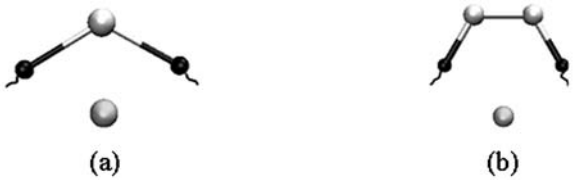

FIG. 6. In the hHPNX model, two potential subconformations in 3D space for the subsequences, in (a) for –h℘h– and in (b) for –h℘℘h–, where ℘ ∈ {P, X, N}.

5. SEARCH STRATEGY

In our previous work (Hoque et al., 2005), it was shown that in the search for optimal conformation, the phenotypical compactness of the conformation deteriorates the prediction performance of the search algorithm and makes the crossover and mutation (pivot rotation) operators in GA ineffective. To overcome this problem, the two major strategies followed are as follows:

(i) Applying effective move operators (Hoque et al., 2005) such as diagonal move, pull move and tilt move
(ii) Formulating the novel objective function PCF and appending it to the existing fitness function F.

In the search for an optimum conformation, if a sub-conformation corresponding to a particular subsequence exists in the thin or intermediate layer for a developing conformation, it is rewarded; otherwise, it is penalized. If any member of a defined sub-conformations corresponds to the related sub-sequence and the Hs are nearer to HCC than the Ps, then PCF is decreased by 2 (double the HH interaction value) as a reward; the same case is rewarded by decreasing PCF by 8 in both the HPNX and the hHPNX models (with is also double the magnitude of the respective model's hydrophobic interaction). On the other hand, the sub-sequence corresponding to a non-desired sub-conformation is penalized by an increase of PCF by 1 for the HP model, and the same case is rewarded by assigning PCF with a value of 4 for both HPNX and hHPNX models. Similarly, there will be an increase of 2 (8 for both HPNX and hHPNX) for a proper shape which is opposite to the desired direction (i.e., violating Axiom (2)).

The assigned values are proportional to the magnitude of the maximum interaction value of the hydrophobic residues. For example, for the HP, HPNX, hHPNX matrices these interaction values are -1 (Fig. 2a, HH interaction), -4 (Fig. 2d, HH interaction), and -4 (Fig. 4, hH interaction), respectively. The values are proportionately weighted by multiples of integers to express the importance of one sub-conformation over other.

While searching and applying these likely sub-conformations, their occurrence is not guaranteed, and therefore these likely conformations basically provide possible guidelines to lead the algorithm towards an optimal conformation.

To achieve this, a *Total Fitness* (TF) function is formed as:

$$\mathrm{TF} = \alpha(t) * F + \beta(t) * \mathrm{PCF} \tag{3}$$

where α and β are positive weightings and t is the number of generation. The values α and β are chosen by considering two alternative phases, namely *phase 1* and *phase 2*, as the generations of GA progress with t. In phase 1, α is made to vary and $\alpha > \beta$ while in phase 2, β is made to vary and $\alpha < \beta$. A likely sub-conformation is enforced to replace a less likely sub-conformation of the converging conformation when $\alpha < \beta$ if it is not already chosen, and PCF dominates over F to adopt the change. To alter the weight of α and β of Equation (3) in order to enable the dominance of F and PCF over each other alternatively, a swing function (see Equation (4) and Fig. 7) is used.

$$\delta(t) = A(1 + \cos \omega_m t) \cos \omega_0 t \tag{4}$$

where $\omega_m \ll \omega_0$. The term t denotes the generator number and has the same meaning as in Equation (3). The assignments of the values for α and β are as shown in the three equations, Equation (5)

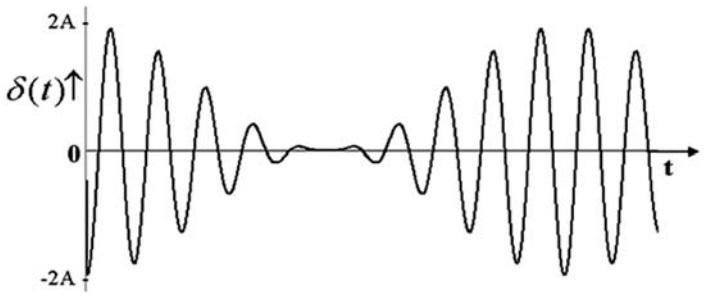

FIG. 7. Plot of $\delta(t)$ function.

to Equation (7), below:

$$\text{Phase 1:} \qquad \alpha(t) = \delta(t), \quad \beta(t) = 1, \quad \text{if } \delta(t) > 0 \tag{5}$$

otherwise,

$$\text{Phase 2:} \qquad \alpha(t) = 1, \quad \beta(t) = -\delta(t), \quad \text{if } \delta(t) < 0 \tag{6}$$

otherwise,

$$\text{Transient Phase:} \qquad \alpha(t) := 1, \quad \beta(t) := 1 \tag{7}$$

Note that the minimum values of $|\alpha(t)| = 1$ and $|\beta(t)| = 1$ are always maintained, and these are never set to zero. This is to preserve the sub-conformation or schema with good features developed in each of the two alternating phases. The oscillatory nature of the function controls the dominance of F and PCF over each other in a non-monotonous manner and provides a range of combinations to cope with the irregular search space of the PSP problem. Further, the oscillatory nature of the function allows any forced sub-conformations applied in the phase 2, to be altered if contradicting with the effective fitness during the occurrence of phase 1. This means that we are influencing or guiding the search, but not forcing it towards a particular conformation.

For the simulation, the values of $\delta(t)$ parameters (see Equation (4)) are set as follows: $\omega_m = 0.004$ and $\omega_0 = 0.05$. The value of A is selected as:

$$A > \frac{\max\{|F_l|, |\text{PCF}_l|\}}{2} \tag{8}$$

This ensures dominance of either F or PCF over the corresponding alternative part in the maximum possible length. In Equation (8) above, PCF_l indicates the lower limit of PCF, which can be calculated by summing the assigned rewards (see discussion earlier in this section) for intermediate layer building sub-sequences that exist for the particular sequence. F_l indicates the lower limit of F.

For the 3D FCC HP model, F_l can be generalized and can be written as shown below in Equation (9).

$$F_{l,\text{3D FCC HP}} = -(5n_H + n_{T_H}) \tag{9}$$

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

In the 3D FCC model, any non-terminal H can have a maximum of 10 TNs out of 12 neighbors. Also, as a TN involves two H residues, the maximum contributing TN count per H will be 5 multiplied by the interaction value, which is "−1" for the HP model (Figure 2b). The terminal H can have one more TN than a non-terminal H.

Similarly for HPNX, as the maximum hydrophobic interaction value is "−4," therefore the maximum contributing TN count per non-terminal H will be 5 multiplied by this interaction value of "−4." Hence, the can be written as given by the following equation.

$$F_{l,\text{3D FCC HPNX}} == \{20n_H + 5\min(n_P, n_N) + 4n_{T_H} + n_{T_P} + n_{T_N}\} \tag{10}$$

where n_H, n_P, and n_N indicate the number of H, P, and N residues respectively in the sequence for the HPNX model. Further, n_{T_H}, n_{T_P}, and n_{T_N} indicate the number of terminal residues of type H, P, and N where $0 \leq n_T \leq 2$ is true for all the types.

Similarly, for the hHPNX model, we can write,

$$F_{l,\text{3D FCC hHPNX}} = -\{20n_H + 5\min(n_P, n_N) + 4n_{T_H} + n_{T_P} + n_{T_N}\} \tag{11}$$

Here, n_H, n_P, and n_N indicate the number of H, P, and N residues.

It is to be noted here that the n_H considered for HPNX in Equation (10) is different from the n_H used for hHPNX in Equation (11), as the group H of HPNX includes member amino acids of both the groups h and H of the hHPNX model.

An idea of the implementation of the likely sub-conformation, based on example precondition and use of move operators to achieve the desired conformation, has been demonstrated in Figures 8–11. For any move, the least destructive move out of the three move operators is given a priority in application unless it fails due to an infeasible precondition of the conformation. In such a case, a less destructive move operator is then selected.

6. SIMULATION AND RESULTS

The search procedure based on 3D FCC-based HGA for PSP is given in Figure 12. The algorithm in Figure 12 is based on the simple GA but incorporates the heuristics. As the GA parameters are not under investigation, we followed the parameter settings as given in Digalakis and Margaritis (2002). Thus, the population size is set to 200 for all the cases, with elite rate = 0.10, crossover rate $p_c = 0.85$, and mutation rate $p_m = 0.5$. To avoid generating more collision (i.e., non-self-avoiding-walk), single-point mutation was applied which is a pivot rotation (Unger and Moult, 1993; Hoque et al., 2005, 2006a, 2006b, 2007b), and the implementation of crossover was also single-point. The roulette wheel was applied as the selection procedure. The simulation was carried out for 10 additional protein sequences (see Section 3.3.1) for the 3D FCC-based models, namely HP, HPNX, and hHPNX. The runs for all these models were simultaneously implemented on a number of machines of the same capacity.

The simulation was stopped when any one of these runs converged or became too slow to improve. This is because, from the model design, analysis, and Tables 2–5, it is obvious that the hHPNX model is superior, since the hHPNX model considers the missing fact over the other models. Therefore, as we found in the experiment HP stops improving early (as expected) and then allows HPNX and hHPNX to run until they slow down in convergence, showing superiority of hHPNX ultimately for the aforementioned obvious reasons. However, here through the simulation, we are able to compare the developed prediction strategies (covered in Sections 4 and 5) for each model to run for the same amount of time to have a fair go. Thus, all the simulations were stopped when any one of these runs converged for the same sequence.

To implement the stop criterion, along with the simulation program, we also kept a record of the clock-time, which was updated as soon as any of the runs converged. In that way, we could compare the predictions from other models which ran for an equal length of time. Since each sequence was run for 30 iterations, the total runs were for 900 different instances, since ($models \times sequences \times iterations = 3 \times 10 \times 30 = 900$).

To compare the predicted results from the lattice models for verifying their closeness with real folded proteins from PDB, the PDB proteins are first normalized with the average C_α Euclidian distances and then cRMSD is taken to measure the closeness as discussed in Section 2. The final result is given in Table 7: in terms of closeness (i.e., the lowest value in a row), the hHPNX clearly outperformed the other two models HP and HPNX. The superiority in the performance is because the different interaction of the hydrophobic group h has not been considered by either HPNX or by HP model but is accounted for by the hHPNX model. From Table 7, it is also clear that, in general, that HPNX has outperformed the HP model. This is because the polar or hydrophilic residue in a HP predicted folding remains without interaction values to take care of, whereas in both the HPNX model and the hHPNX model the polar residue has been further categorized as either positively charged or negatively charged residue, or as neutral, presented through the interaction values.

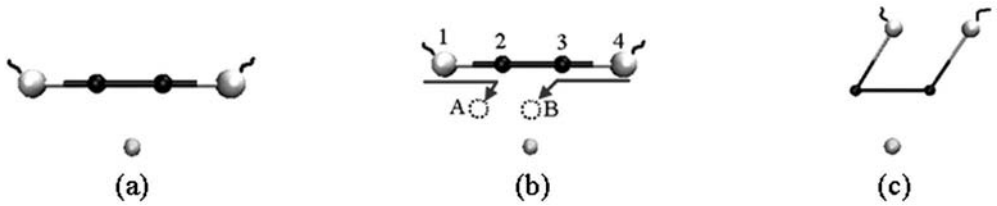

FIG. 8. **(a)** A pre-condition. **(b)** If location 'A' and 'B' are free, placing residue '2' and '3' to 'A' and 'B,' respectively, by a tilt move will have the desired sub-conformation, as in **(c)**.

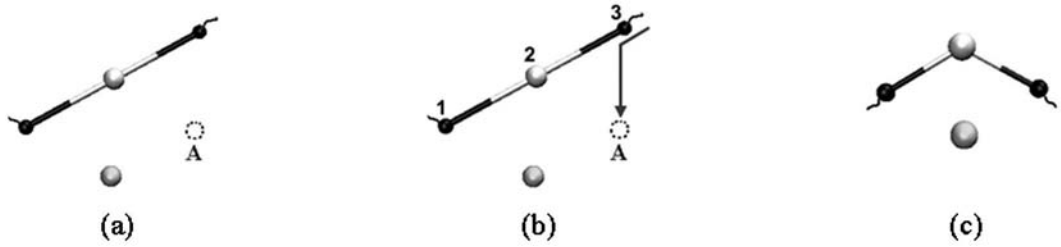

FIG. 9. **(a)** A pre-condition. **(b)** If location 'A' is free, then placing residue '3' to 'A' by a pull move (indicated by arrow from residue '3' to 'A') will have the desired sub-conformation, as in **(c)**.

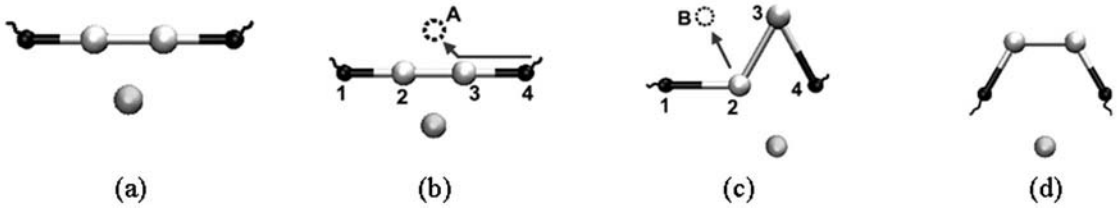

FIG. 10. **(a)** A pre-condition. **(b)** If location 'A' is free, then placing residue '3' to 'A' by a pull move (indicated by arrow from residue '3' to 'A') will have the sub-conformation, as in **(c)**. If position 'B' is found available from (c), then residue '2' can be moved there by a diagonal move to have the desired sub-conformation, as in **(d)**.

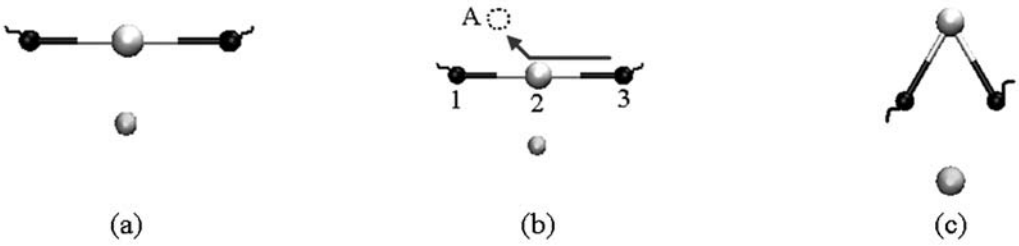

FIG. 11. **(a)** A pre-condition. **(b)** If location 'A' is free, placing residue '2' to 'A' by a pull move (indicated by arrow) will have the desired sub-conformation, as in **(c)**.

Fig. 12. 3D FCC based Hybrid Genetic Algorithm (HGA).

```
INPUT:  Sequence S,
OUTPUT: Conformation with best fitness, F.
BEGIN
1. COMPUTE: PCF; COMPUTE: A
2. t = 0, F = 0                          /* Gen. count and fitness initialization */
3. Populate with random (valid) conformations for S.
4. While <Terminate> NOT ON THEN
5.  { t = t + 1,
6.  COMPUTE δ(t), α(t), β(t), TF
7.  CROSSOVER and then MUTATION
8.  IF δ(t) < 0 THEN
9.    { FOR i = 1 to population_size DO
10.  Check chromosome_i for any miss mapping of highly likely sub-conformations
       based on Model.
11.     IF miss-mapping = TRUE THEN
12.          {Re-map the sub-sequence to corresponding likely sub-conformations
           based on model using move-sets. }}
13.    COMPUTE: TF
14.    SORT, KEEP Elite
15.    F ← Best fitness found from the population. }
END.
```

FIG. 12. Three-dimensional face-centered-cube (FCC)-based Hybrid Genetic Algorithm (HGA).

TABLE 7. AVERAGE OF MINIMUM VALUES OF cRDMS, COLLECTED FROM 30 ITERATIONS FOR EACH PDB SEQUENCE PER MODEL

PDB ID	*HP*	*HPNX*	*hHPNX*
1pjf	3.516033082	*3.604867003*	**3.228616595**
2crt	*2.005523982*	1.974233255	**1.953833002**
8ame	*1.803999617*	1.61129892	**1.56987175**
1tuk	*1.391392007*	1.321471091	**1.324636426**
2ptl	*2.618440351*	2.391839518	**2.3255697**
2ffw	*3.323760067*	3.040873033	**2.42754827**
1ae3	*2.51664787*	2.490903541	**2.448412833**
1gh1	1.78315438	*1.788712216*	**1.777429276**
1a7d	2.521329903	2.45409665	**2.435961871**
1eca	3.055854953	2.424334282	**2.054491181**

The distances between any two consecutive alpha-carbon normalized; thus, the outcome is unit-less. The bold entry indicates best value, whereas the italic entry indicates worst value for the simulation run for same time span.

7. CONCLUSION

To handle involved immense complexity of the protein structure prediction, a hierarchical is used, where a blueprint or the outline of the conformation is generated, initially using simplified model to ensure exploration of the convoluted search space sufficiently. A simplified model such as the HP model can be easily implemented, but it can result in high degeneracy by generating a large number of conformations with same putative ground energy. The HPNX model, an extension of the HP model, also shows limitations and it can also cause high degeneracy, resulting in extensive computation time while searching. Therefore, with the need to have an error-free model but also to keep it simple, we proposed the hHPNX model. To do so, we have established that the YhHX matrix proposed by Bronberg was obtained by an erroneous conversion of Crippen's 1234 matrix. Corrections to the YhHX matrix for implementing appropriate interaction value have been provided, along with the correction of the stated frequency distribution of the amino acids. We have also highlighted that the hh interaction is significantly important, as confirmed by the aid of an additional set of sequences. Although the group size of h is small, it is observed that the occurrence frequency of the members of this group is the highest on average, making it necessary that the h group should be separated from the H of HP or HPNX.

We have evaluated the prediction performances of the three models (HP, HPNX, and hHPNX) in 3D FCC configuration using the heuristic-based HGA extended for the HPNX and hHPNX models and using cRMDS measurement for establishing the closeness of the model to the real protein data. The simulation results show a superior performance of the proposed hHPNX over the two other existing models.

ACKNOWLEDGMENTS

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

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Agarwala, R., Batzoglou, S., Dancik, V., et al. 1997. Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model. *J. Comput. Biol.* 4, 275–296.

Alm, E., Morozov, A.V., Kortemme, T., et al. 2002. Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* 322, 463–476.

Backofen, R., Will, S., and Bornberg-Bauer, E. 1999. Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *Bioinformatics* 15, 234–242.

Backofen, R., Will, S., and Clote, P. 2000. Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. *Pac. Symp. Biocomput.* 5, 92–103.

Baker, D. 2006. Prediction and design of macromolecular structures and interactions. *Phil. Trans. R. Soc. B* 361, 459–463.

Beals, M., Gross, L., and Harrell, S. 2007. *Amino acid frequency 1999*. Available at: *www.tiem.utk.edu/∼gross/bioed/webmodules/aminoacid.htm*. Accessed October 15, 2008.

Berger, B., and Leighton, T. 1998. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. Comput. Biol.* 5, 27–40.

Bornberg-Bauer, E. 1997. Chain growth algorithms for HP-type lattice proteins. *Proc. RECOMB 1997*.

Brocchieri, L., and Karlin, S. 1995. How are close residues of protein structures distributed in primary sequences? *Proc. Natl. Acad. Sci. USA* 92, 12136–12140.

Buchler, N.E.G., and Goldstein, R.A. 1999. Effect of alphabet size and foldability requirements on protein structure designability. *Proteins* 34, 113–124.

Buchler, N.E.G., and Goldstein, R.A. 2000. Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: a consensus. *J. Chem. Phys.* 112, 2533–2547.

Chen, M., and Lin, K.Y. 2002. Universal amplitude ratios for three-dimensional self-avoiding walks. *J. Phys. A Math. Gen.* 35, 1501–1508.

Chivian, D., Kim, D.E., Malmström, L., et al. 2003. Automated prediction of CASP-5 structures using the Robetta server. *Proteins* 53, 525–533.

Crescenzi, P., Goldman, D., Papadimitriou, C., et al. 1998. On the complexity of protein folding [extended abstract]. Presented at Second Annual International Conference on Computational Molecular Biology.

Crippen, G.M. 1991. Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry* 30, 4232–4237.

Digalakis, J.G., and Margaritis, K.G. 2002. An experimental study of benchmarking functions for genetic algorithms. *Int. J. Comput. Math.* 79, 403–416.

Dill, K.A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24, 1501–1509.

Dill, K.A., Bromberg, S., Yue, K., et al. 1995. Principles of protein folding—a perspective from simple exact models. *Protein Sci.* 4, 561–602.

Guo, Y.-Z., Feng, E.-M., and Wang, Y. 2006. Exploration of two-dimensional hydrophobic-polar lattice model by combining local search with elastic net algorithm. *J. Chem. Phys.* 125, 154102.

Guttmann, A.J. 2005. *Self-Avoiding Walks in Constrained and Random Geometries*. Elsevier, New York.

Hales, T.C. 2005. A proof of the Kepler conjecture. *Ann. Math.* 162, 1065–1185.

Hinds, D.A., and Levitt, M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* 243, 668–682.

Hoque, M.T., Chetty, M., and Dooley, L. 2006a. A guided genetic algorithm for protein folding prediction using 3D hydrophobic-hydrophilic model. Presented at WCCI/IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, Canada.

Hoque, M.T., Chetty, M., and Dooley, L.S. 2005. A new guided genetic algorithm for 2D hydrophobic-hydrophilic model to predict protein folding. *Proc. CEC*.

Hoque, M.T., Chetty, M., and Dooley, L.S. 2006b. A hybrid genetic algorithm for 2D FCC hydrophobic-hydrophilic lattice model to predict protein folding. *Lect. Notes Artif. Intellig.*

Hoque, M.T., Chetty, M., and Sattar, A. 2007a. Correction of the YhHX model: GSIT. Technical report, TR-2007/2. Monash University.

Hoque, M.T., Chetty, M., and Sattar, A. 2007b. Protein folding prediction in 3D FCC HP lattice model using genetic algorithm. Presented at Bioinformatics Special Session, IEEE Congress on Evolutionary Computation (CEC), Singapore.

Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., et al. 2005. A universal trend of amino acid gain and loss in protein evolution. *Lett. Nat.* 433, 633–638.

Koehl, P., and Levitt, M. 1999. A brighter future for the protein structure prediction. *Nat. Struct. Biol.* 6, 2.

Kolinski, A., Gront, D., Pokarowski, P., et al. 2003. A simple lattice model that exhibits a protein-like cooperative all-or-none folding transition. *Biopolymers* 69, 399–405.

Li, H., Helling, R., Tang, C., et al. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* 273, 666–669.

Li, Z.R., Liu, G.R., and Mi, D. 2005. Quantifying the parameters of Prusiner's heterodimer model for prion replication. *Physica A* 346, 459–474.

MacDonald, D., Joseph, S., Hunter, D.L., et al. 2000. Self-avoiding walks on the simple cubic lattice. *J. Phys. A Math. Gen.* 33, 5973–5983.

Miyazawa, S., and Jernigan, R.L. 1996. Residue-residue potential with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256, 623–644.

PDB (Protein Data Bank). 2007. *Protein Data Bank 2007*. Available at: *www.rcsb.org/pdb/*. Accessed October 15, 2008.

Potzsch, S., Scheuermann, G., Stadler, P.F., et al. 2006. Visualization of lattice-based protein folding simulations. Presented at Information Visualization (IV'06).

Raghunathan, G., and Jernigan, R.L. 1997. Ideal architecture of residue packing and its observation in protein structures. *Protein Sci.* 10, 2072–2083.

Samudrala, R., Xia, Y., and Levitt, M. 1999. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac. Symp. Biocomput.* 4, 505–516.

Schiemann, R., Bachmann, M., and Janke, W. 2005. Exact enumeration of three-dimensional lattice proteins. *Comput. Phys. Commun.* 166, 8–16.

Schueler-Furman, O., Wang, C., Bradley, P., et al. 2005. Progress in modeling of protein structures and interactions. *Science* 310, 638–642.

Unger, R., and Moult, J. 1993. Genetic algorithms for protein folding simulations. *J. Mol. Biol.* 231, 75–81.

VMD (Visual Molecular Dynamics). 2007. *Visual molecular dynamics 2007*. Available at *www.ks.uiuc.edu/Research/vmd/*. Accessed October 15, 2008.

Wroe, R., Bornberg-Bauer, E., and Chan, H.S. 2005. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophys. J.* 88, 118–131.

Xia, Y., Huang, E.S., Levitt, M., et al. 2000. Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* 300, 171–185.

Yue, K., and Dill, K.A. 1993. Sequence-structure relationships in proteins and copolymers. *Phys. Rev. E* 48, 2267–2278.

Address reprint requests to:
Dr. Tamjidul Hoque
IIIS, Griffith University
170 Kessels Road
Nathan, QLD-4111, Australia

E-mail: Tamjidul.Hoque@gmail.com