Imperial College Press
www.icpress.co.uk

FOLD HELICAL PROTEINS BY ENERGY MINIMIZATION IN DIHEDRAL SPACE AND A DFIRE-BASED STATISTICAL ENERGY FUNCTION

HONGZHI LI[*,‡] and YAOQI ZHOU[*,†,§]

Howard Hughes Medical Institute Center for Single Molecule Biophysics
Department of Physiology & Biophysics
State University of New York at Buffalo
124 Sherman Hall, Buffalo, New York 14214, USA

†*Department of Macromolecular Science*
The Key Laboratory of Molecular Engineering of Polymers
Fudan University, Shanghai, China
‡*hli9@buffalo.edu*
§*yqzhou@buffalo.edu*

Statistical energy functions are discrete (or stepwise) energy functions that lack van der Waals repulsion. As a result, they are often applied directly to a given structure (native or decoy) without further energy minimization being performed to the structure. However, the full benefit (or hidden defect) of an energy function cannot be revealed without energy minimization. This paper tests a recently developed, all-atom statistical energy function by energy minimization with a fixed secondary helical structure in dihedral space. This is accomplished by combining the statistical energy function based on a distance-scaled finite ideal-gas reference (DFIRE) state with a simple repulsive interaction and an improper torsion energy function. The energy function was used to minimize 2000 random initial structures of 41 small and medium-sized helical proteins in a dihedral space with a fixed helical region. Results indicate that near-native structures for most studied proteins can be obtained by minimization alone. The average minimum root-mean-squared distance (rmsd) from the native structure for all 41 proteins is 4.1 Å. The energy function (together with a simple clustering of similar structures) also makes a reasonable selection of near-native structures from minimized structures. The average rmsd value and the average rank for the best structure in the top five is 6.8 Å and 2.4, respectively. The accuracy of the structures sampled and the structure selections can be improved significantly with the removal of flexible terminal regions in rmsd calculations and in minimization and with the increase in the number of minimizations. The minimized structures form an excellent decoy set for testing other energy functions because most structures are well-packed with minimum hard-core overlaps with correct hydrophobic/hydrophilic partitioning. They are available online at http://theory.med.buffalo.edu.

Keywords: Energy minimization; knowledge-based potential; helical proteins.

1. Introduction

How to make an accurate prediction of the structures of proteins from their amino-acid sequences is one of the most challenging problems in computational biology. It is challenging because of the existence of astronomically large number of possible structures available to the linear polypeptide chain of a protein and the lack of an accurate free energy function that guides the conformational search and distinguishes the native structure from decoys.

Significant progresses have been made in the past decades in structure prediction with approaches ranging from homology modeling, fold recognition to *ab initio* prediction.[1] Among them, the secondary structure of a protein can now be predicted with a reasonable accuracy (for a recent review, see e.g. Ref. 2). This raises the interest to predict structure by packing secondary structures through energy and/or geometric optimization.[3−9] For example, Zhang *et al.*[6] folded helical proteins by using torsion angle dynamics and predicted restraints. Nanias *et al.*[7] used a global optimization method of a residue-based energy function to pack the helices of helical proteins. Although fixing the secondary structure of a protein dramatically reduces its conformational space, it is still a challenging exercise to predict the tertiary structure because of the lack of an accurate energy function and a large number of possible loop conformations and possible packing arrangement of secondary structure elements.

In this paper, we test a newly-developed all-atom knowledge-based potential by predicting the tertiary structure of a helical protein whose helical region is known. This statistical energy function is based on a physical reference state of the distance-scaled, finite, ideal-gas reference (DFIRE) state. The DFIRE-based energy function has been successfully applied to structure[10] and docking structure selections,[11] loop prediction,[12] prediction of mutation-induced change in stability[10] and binding affinity of protein-protein (peptide),[11] protein-ligand,[13] and protein-DNA complexes.[13] More importantly, the physical reference state of ideal gases appears to make the DFIRE physically more accurate because its performance is mostly independent of the systems with different compositions of amino-acid residues at surface, core,[14] and protein-protein interface[11] and independent of the structural database (α or β proteins) used for energy extraction.[15]

The DFIRE-energy function, like many other statistical energy functions,[16−18] have not yet been tested for structural refinement via energy minimization. This is because a statistical energy function lacks a hard-repulsive core and it is a discrete function. Here, we combine the DFIRE energy function (interpolated by cubic spline) with a simple repulsive interaction and an improper torsion potential. The energy function was implemented in the program TINKER[19] and used to minimize the random initial structures of 41 helical proteins with fixed helical regions in dihedral space. We find that near-native conformations (< 6.5 Å rmsd from the native) can be reached for almost all tested proteins by minimization alone. The successes

and failures of the DFIRE-energy function in selecting near-native structures from minimized structures are discussed.

2. Methods

2.1. *Proteins and their initial structures*

We picked 41 α-helical proteins for this study. Among them, 23 of the targets are those from Ref. 6. Other proteins are chosen based on their sizes and number of residues. The names (PDB IDs) and the sizes of proteins are listed in Table 1. The number of residues of the proteins ranges from 40 to 124; the number of non-helical residues from 7 to 52; and the number of helices from 2 to 6. We also tested the effect of flexible, solvent-exposed terminal regions on energy minimization for eight proteins (PDB IDs ended with a lower case c).

The initial structures of the proteins are generated with random dihedral angles for the residues in the nonhelical regions and native dihedral angles for the residues in the helical regions. We also tested proteins whose initial structures of the helical regions are generated by a Gaussian distribution within a standard deviation of 10 degrees around the ideal right-handed helix parameters for dihedral angles.[20]

For most of the proteins, we generated 2000 initial structures each. We also minimized 45 000 structures for 1GABc and 9000 structures for proteins 1LP1, $1LBU_{1-83}c$, $1A04A_{150-216}c$, and 1B0XAc.

2.2. *Energy function and minimization*

The DFIRE energy function is a knowledge-based statistical potential.[10] The DFIRE-based atom-atom potential of the mean force $u^{\mathrm{DFIRE}}(i, j, r)$ between atom types i and j that are distance r apart is given by[10]

$$u^{\mathrm{DFIRE}}(i,j,r) = \begin{cases} -\eta RT \ \ln \dfrac{N_{\mathrm{obs}}(i,j,r)}{\left(\frac{r}{r_{\mathrm{cut}}}\right)^{\alpha} \left(\frac{\Delta r}{\Delta r_{\mathrm{cut}}}\right) N_{\mathrm{obs}}(i,j,r_{\mathrm{cut}})}, & r < r_{\mathrm{cut}}, \\ 0, & r \geq r_{\mathrm{cut}}, \end{cases} \tag{1}$$

where $\eta = 0.0157$ for the energy unit of kcal/mole, R is the gas constant, $T = 300\,\mathrm{K}$, $\alpha = 1.61$, $N_{\mathrm{obs}}(i,j,r)$ is the number of (i,j) pairs within the distance shell r observed in a given structure database, $r_{\mathrm{cut}} = 14.5\,\text{Å}$, and $\Delta r(\Delta r_{\mathrm{cut}})$ is the bin width at $r(r_{\mathrm{cut}})$. ($\Delta r = 2\,\text{Å}$, for $r < 2\,\text{Å}$; $\Delta r = 0.5\,\text{Å}$ for $2\,\text{Å} < r < 8\,\text{Å}$; $\Delta r = 1\,\text{Å}$ for $8\,\text{Å} < r < 15\,\text{Å}$.) The exponent α for the distance dependence was obtained from the distance dependence of the number of pairs of ideal gas points in finite spheres (finite ideal-gas reference state). The value of the prefactor η was obtained so that the correlation slope is one between experimental measured and theoretical predicted changes in stability due to mutation. Residue specific atomic types were used (167 atomic

types).[17,18] The number of observed atomic (i, j) pairs with the distance shell r $[N_{\text{obs}}(i, j, r)]$ was obtained from a structural database of 1011 non-homologous (less than 30% homology) proteins with resolution $< 2\,\text{Å}$, which was collected by Hobohm *et al.*[21] (http://chaos.fccc.edu/research/labs/dunbrack/culledpdb.html). This database provides sufficient statistics for most distance bins (except near the hard repulsive van der Waals regions between atoms). The average number of observed atomic pairs per bin is 655. The sufficiency of statistics is also reflected from the fact that the results for structural discrimination are insensitive to the size[10] and the type[15] of structural databases used to generate the potential.

All the statistical potentials do not have an appropriate hard-repulsive core due to the lack of statistics in this region. The DFIRE energy function is supplemented with an arbitrary repulsive potential. The modified DFIRE energy function is as follows:

$$u^{\text{MDFIRE}}(i, j, r) = u^{\text{DFIRE}}(i, j, r) + \begin{cases} f_V \eta \left[\left(\dfrac{\sigma_{ij}}{r} \right)^6 - \left(\dfrac{\sigma_{ij}}{\sigma_{ij}^c} \right)^6 \right], & r < \sigma_{ij}^c \\ 0, & r \geq \sigma_{ij}^c \end{cases} \tag{2}$$

where σ_{ij} is van der Waals diameter[22] and $f_V = 1$ and $\sigma_{ij}^c = \sigma_{ij}$ for two atoms that are not in the neighboring residues. A weaker repulsive term is used for two atoms in the neighboring residues (i.e., residue I and residue $I \pm 1$) with $f_V = 0.1$ and $\sigma_{ij}^c = 2.5\,\text{Å}$. Note that we have used the same prefactor $f_V \eta$ for all the atomic pairs for simplicity. This repulsive term is weaker than the repulsive portion of the Lennard-Jones potential which decays in r.[12] We did not study the effect of using different forms of repulsive hard cores in this work. The effect is likely small because hard repulsive regions contribute little to the energy of a well minimized structure.

The DFIRE energy function is further supplemented with the improper torsion energy that maintains the chirality and planar shape of some atoms. The parameters for improper dihedral angles (ω_0) and force constants (k_ω) are from the CHARMM19 parameter set.[23] The final equation for the energy function is

$$E = 3 \sum_{\text{improper}} k_\omega (\omega - \omega_0)^2 + \sum_{i<j} u^{\text{MDFIRE}}(i, j, r_{ij}) \tag{3}$$

where a factor of 3 is used to strengthen the ability of the energy function to keep improper torsion angles in their original values. We find that the improper torsion energy does not change much during most minimizations.

TINKER (http://dasher.wustl.edu/tinker/), a software tool for protein simulations, is modified to integrate the DFIRE-based potential into its default energy function. In order to calculate the energy derivation, the discrete DFIRE potential is fitted to a continuous function by cubic spline interpolation. The MINIROT program in the TINKER package, which performs a limited memory BFGS quasi-Newton nonlinear minimization[24] in dihedral space, has been used to minimize the initial structure. The minimization is stopped when the change in rms gradient

during unidimensional line search is less than 0.001. The MINIROT program also has a selection tool that allows users to fix specified regions (helical regions in this work).

2.3. *Clustering*

The minimized structures are clusterized based on pairwise backbone rmsd values. One common method to cluster structures is to use a pre-determined cutoff rmsd value. We find that it is difficult to set one cutoff value for all 41 proteins because this will lead to too few clusters for some proteins and too many clusters for the others. Thus, we use a protein-dependent cutoff value at which the top three largest clusters contain 5% of all structures to be clustered (i.e. 100 structures; out of a total of 2000 structures). This method is used so that 1) the top three largest clusters have a statistically meaningful number of structures; and 2) the number of clusters is large enough to reflect the diversity of the structures. The rmsd cutoff values range from $0.5\,\text{Å}$–$8\,\text{Å}$, depending on the sizes of the nonhelical portions of the proteins.

We use two methods to cluster structures. One is based on energy. The first cluster contains the structures that are within the cutoff rmsd value from the lowest energy structure. The next cluster is based on clustering around the lowest energy structure from the remaining structures. The procedure is repeated until the top five clusters are obtained. In the second method, the structures are clusterized around the structure with most structures within the same cluster, rather than the structure with the lowest energy. For convenience, we called the second method the size-based method to distinguish it from the energy-based method. We find that the two clustering methods give essentially the same result for average rmsd values ranked by the cluster sizes. Thus, in this paper, we only report the result based on the first method because the the size-based method searches the center of the cluster on the pairwise level and is, thus, more time-consuming than the energy-based method.

3. Results

The results of energy minimization in dihedral space are summarized in Table 1. The results are reported in terms of the best structure (the structure with the smallest rmsd from the native structure among 2000 decoys); the best structure in the top five structures selected based on energy (after clustering); and the best structure in the top five structures selected based on the size of clusters. The DFIRE energy provides a reasonably efficient sampling. With only 2000 decoys, the average of the best decoy structures for 41 proteins is $4.1\,\text{Å}$. However, the DFIRE energy function often does not select the best near-native structure as the lowest energy structure. The average rmsd value for the best structure within the top five structures are $6.8\,\text{Å}$ for the energy-based selections and $6.4\,\text{Å}$ for the cluster-size-based selections. For some proteins, the top structures are more than $10\,\text{Å}$ away from the native.

Table 1. Conformational sampling and structure selections by energy minimization of 41 helical proteins with fixed helical regions and random initial loop regions.

PDB[a]	N_h[b]	N_{res}[c]	N_{res}^{Free}[d]	$T35$[e]	$rmsd_{min}$[f]	Top 5 E[g]	Top 5 C[h]
1G6U	2	48	7	0.5	0.4	1.2(E2)	1.2(C2)
2ERL	3	40	15	2.8	2.5	3.6(E1)	3.6(C1)
1LP1	3	55	15	3.0	2.5	3.0(E2)	3.8(C5)
1EZ3A	3	124	18	5.1	3.2	8.2(E1)	4.7(C1)
1LVFA	3	106	13	3.4	1.7	2.3(E2)	2.3(C1)
1BW6	3	56	21	4.0	1.8	3.5(E4)	4.3(C1)
1DV0	3	45	22	3.7	2.5	3.4(E1)	3.4(C4)
1EDK	3	56	16	3.1	1.8	3.0(E1)	3.0(C2)
1EF4	3	55	31	5.0	4.6	5.2(E1)	5.2(C1)
1IDY	3	54	25	4.6	4.6	5.7(E5)	5.7(C1)
1BDD	3	60	25	4.2	4.0	5.5(E4)	5.5(C4)
1MBE	3	53	23	4.4	3.3	4.8(E3)	4.8(C5)
1PRB	3	53	23	4.2	3.9	7.6(E1)	7.6(C3)
1PRU	3	56	28	4.9	5.1	7.6(E5)	6.3(C5)
2SPZ	3	58	15	2.9	1.6	3.1(E1)	3.1(C1)
2HOA	3	68	31	7.0	6.9	9.0(E1)	9.0(C4)
1CKTA	3	71	26	5.5	4.2	10.7(E2)	10.7(C5)
1DV5	3	80	42	6.8	3.8	3.9(E1)	8.9(C3)
1GAB	3	53	23	3.9	4.3	7.2(E1)	6.5(C3)
1LBU$_{1-83}$	3	83	52	7.6	6.1	9.1(E4)	9.1(C2)
1LEA	3	72	38	6.3	5.4	8.6(E4)	7.9(C2)
1LRE	3	81	28	5.2	4.1	5.3(E1)	5.3(C1)
2OCCH	3	79	39	6.2	5.9	9.9(E4)	7.4(C3)
4-helix[25]	4	106	32	7.5	4.7	9.5(E3)	6.8(C5)
1A04$A_{150-216}$	4	67	27	5.5	4.8	6.8(E2)	8.0(C2)
1A6S	4	87	33	7.4	4.6	9.3(E5)	7.5(C4)
1C5A	4	65	20	4.8	2.9	4.8(E3)	4.8(C1)
1FFH$_{2-88}$	4	87	26	5.2	2.7	8.4(E1)	3.6(C1)
1NKL	4	78	24	5.4	2.7	4.8(E1)	4.8(C1)
2ABD	4	86	31	6.9	5.2	9.5(E5)	8.1(C2)
1AIS$B_{1108-1205}$	5	98	34	7.9	5.4	8.9(E2)	10.0(C3)
1B0NA_{1-68}	5	68	29	6.0	4.2	7.7(E3)	7.8(C5)
1B0XA	5	72	27	6.2	3.9	4.0(E2)	4.0(C1)
1UNKA	5	87	42	7.0	5.2	10.1(E1)	6.4(C5)
1CTJ	5	89	42	7.8	5.4	8.5(E3)	8.5(C3)
1KDXA	5	81	26	6.1	4.5	7.4(E2)	7.7(C2)
1BMT$A_{651-740}$	5	90	21	6.0	3.2	6.8(E1)	6.8(C3)
1QC7A	6	101	31	8.2	7.7	11.9(E3)	11.0(C4)
1BXM	6	98	44	8.2	6.3	10.5(E4)	10.5(C1)
1NGR	6	85	34	7.5	5.7	8.7(E4)	7.4(C1)
1RZL	6	91	30	7.1	5.6	9.6(E1)	9.6(C2)
Ave.	3.8	74.2	27.5	5.5	4.1	6.8(E2.4)	6.4(C2.6)

[a]PDB code. [b]The number of helices. [c]The number of residues. [d]The number of residues that are not fixed. [e]The cutoff value of rmsd for cluster (in Å). See context for the definition. [f]The minimum rmsd value from the native structure in 2000 minimized structures (in Å). [g]The minimum rmsd value (and the rank) from the top five structures ranked by energy after clustering. [h]The minimum rmsd value (and the rank) from the top five structures ranked by cluster size after clustering.
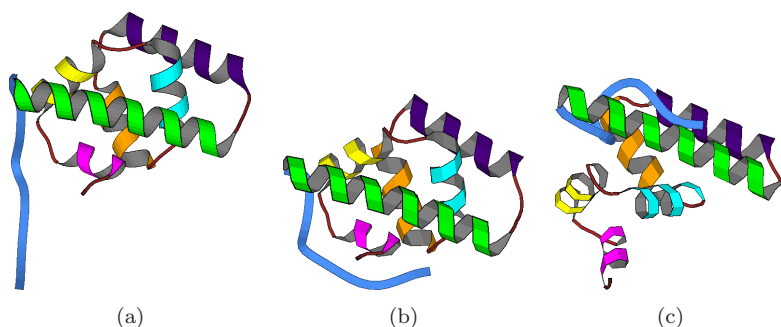
Fig. 1. Protein 1QC7A: (a) The native structure, (b) minimized native structure, and (c) the best structure in top five ranked by cluster size.

What is the source for the failure of the DFIRE energy function for detecting the best near-native structures? We find that one reason is the existence of a long flexible terminal region for many proteins. Figure 1 shows the native structure, the minimized native structure, and the best structure in the top five ranked by cluster sizes for protein 1QC7A. The minimized native structure is 5.0 Å rmsd from the native but only 1.5 Å if the flexible terminal region is excluded in calculating the rmsd values. The rmsd value for the best size-ranked structure is also reduced significantly from 11.0 Å to 6.6 Å. Thus, it appears that a more accurate way to analyze structures sampled by the DFIRE-based energy function is to exclude the flexible region in evaluating rmsd values because the actual position of the region is likely very dynamic.

Table 2 shows the results after the rmsd values are recalculated for 31 proteins with significant flexible regions (longer than three residues based on visual inspection of the native structures). The rmsd values for the best decoy structure, the best structure in the top five structures ranked by energy or cluster sizes all decrease significantly by one to two Å on average. The final average of the best decoy structures for all 41 proteins is 2.9 Å. The average rmsd value for the best structure within the top five structures is 5.3 Å for the energy-based selections and 5.0 Å for the cluster-size-based selections.

To further determine the effect of the flexible terminal regions, we performed minimizations for several proteins with flexible terminals that are removed. The results are shown in Table 3. Again, the rmsd values for the best decoy structure, the best top five structures ranked by energy or cluster sizes decrease further from the rmsd values calculated without the flexible regions by about another 1 Å in average. For example, the removal of the flexible region in 1QC7A leads to a reduction in number of unfixed residues from 31 to 16 and in the best rmsd value from 7.7 Å (5.1 Å if flexible region is not used in calculating the rmsd value) to 2.9 Å.

Despite the improvement in accuracy with the removal of flexible regions, the rmsd value of the best near-native structure within the top five predicted by energy or clustering continues to be about 2 Å greater than that of the best near-native

Table 2. Re-analyzing conformational sampling and structure selections of 31 helical proteins by excluding their flexible terminal regions in the rmsd calculation.

PDB[a]	rmsd$_{min}$ (Å)[b]	Top 5 E (Å)[c]	Top 5 C (Å)[d]	$N_{res}^{Exclude}$[e]
2ERL	1.5	2.5(E1)	2.5(C1)	5
1LP1	1.2	1.9(E2)	1.9(C3)	5
1BW6	1.3	2.4(E4)	3.3(C1)	9
1DV0	1.4	3.0(E1)	2.5(C4)	9
1EDK	0.7	1.1(E2)	1.1(C2)	8
1EF4	1.8	2.0(E1)	2.0(C1)	21
1IDY	2.3	3.9(E4)	3.9(C1)	9
1BDD	1.5	2.2(E4)	2.2(C2)	13
1MBE	1.1	2.8(E1)	2.8(C2)	13
1PRB	2.0	3.7(E5)	4.7(C5)	9
1PRU	1.6	3.5(E1)	3.5(C1)	16
2SPZ	0.8	2.3(E5)	2.7(C1)	6
2HOA	3.1	3.8(E1)	3.8(C1)	18
1CKTA	3.6	9.0(E2)	9.0(C3)	7
1GAB	2.5	4.0(E4)	4.0(C5)	9
1LBU$_{1-83}$	4.6	6.5(E5)	7.9(C4)	23
1LEA	2.4	3.3(E4)	4.3(C5)	25
1LRE	1.9	2.6(E1)	2.6(C1)	16
2OCCH	2.3	6.0(E1)	3.3(C4)	27
1A04A$_{150-216}$	2.8	6.7(E2)	6.0(C3)	8
1NKL	2.0	3.6(E5)	3.7(C1)	6
2ABD	5.0	9.3(E5)	7.8(C2)	4
1AISB$_{1108-1205}$	4.6	7.5(E2)	10.3(C2)	9
1B0NA$_{1-68}$	3.3	6.7(E3)	6.7(C5)	8
1B0XA	3.1	3.2(E2)	3.2(C2)	10
1UNKA	3.8	8.3(E3)	4.8(C4)	15
1CTJ	5.0	8.1(E3)	8.1(C5)	8
1KDXA	2.9	4.7(E2)	6.6(C1)	15
1QC7A	5.1	9.4(E1)	6.6(C4)	15
1BXM	6.3	10.1(E4)	10.1(C3)	4
1NGR	4.4	8.6(E4)	6.0(C4)	9
Ave.	2.8	4.9(E2.7)	4.8(C2.7)	11.6
(Original)[f]	4.4	6.9(E2.5)	6.7(C2.7)	

[a]PDB code. [b]The minimum rmsd value from the native structure in 2000 minimized structures (in Å). [c]The minimum rmsd value (and the rank) from the top five structures ranked by energy after clustering. [d]The minimum rmsd value (and the rank) from the top five structures ranked by cluster size after clustering. [e]The number of flexible residues excluded in rmsd calculations. [f]The average values for the 31 proteins in which all residues are used in rmsd calculations.

structure sampled (Table 3). There is, however, a significant linear correlation between the lowest rmsd value in 2000 decoys and the best rmsd value within the top five structures ranked by energy (see Fig. 2). The correlation coefficients for 41 proteins are 0.89 (0.84 if the flexible terminal regions are included in the rmsd calculations). One moderate outlier is caused by 1FFH. In eight proteins for which the flexible terminal regions are removed during minimizations (Table 3), the rmsd value of the best structure in the top five ranked by energy for either protein 1AIS

Table 3. Comparison of sampling for several proteins by different methods.

PDB	$rmsd_{min}^a$			Top 5 E[b]			Top 5 C[c]		
	Original[d]	Re-analyzed[e]	Cut[f]	Original[d]	Re-analyzed[e]	Cut[f]	Original[d]	Re-analyzed[e]	Cut[f]
1PRB	3.9	2.0	1.8	7.6(E1)	3.7(E5)	1.8(E2)	7.6(C3)	4.7(C5)	2.7(C2)
1GAB	4.3	2.5	1.9	7.2(E1)	4.0(E4)	2.7(E1)	6.5(C3)	4.0(C5)	2.7(C2)
1LBU	6.1	4.6	4.4	9.1(E4)	6.5(E5)	5.9(E4)	9.1(C2)	7.9(C4)	7.5(C4)
1LEA	5.4	2.4	2.3	8.6(E4)	3.3(E4)	4.5(E2)	7.9(C2)	4.3(C5)	4.5(C1)
1A04	4.8	2.8	2.3	6.8(E2)	6.7(E2)	6.7(E2)	8.0(C2)	6.0(C3)	4.6(C5)
1AIS	5.4	4.6	2.7	8.9(E2)	7.5(E2)	7.4(E4)	10.0(C3)	10.3(C2)	7.4(C5)
1B0X	3.9	3.1	2.3	4.0(E2)	3.2(E2)	2.3(E1)	4.0(C1)	3.2(C2)	2.3(C1)
1QC7	7.7	5.1	2.9	11.9(E3)	9.4(E1)	2.9(E4)	11.0(C4)	6.6(C4)	2.9(C4)
Ave	5.2	3.4	2.6	8.0(E2.4)	5.5(E3.1)	4.3(E2.5)	8.0(C2.5)	5.9(C3.8)	4.3(C3.0)

[a]The minimum rmsd value from the native structure in 2000 minimized structures (in Å). [b]The minimum rmsd value (and the rank) from the top five structures ranked by energy after clustering (in Å). [c]The minimum rmsd value (and the rank) from the top five structures ranked by cluster size after clustering (in Å). [d]All the residues in the native structures are used in minimization and calculation of rmsd values. [e]Protein flexible terminal regions are included in minimization, but excluded in rmsd calculation. [f]Protein flexible terminal regions are removed in minimization.
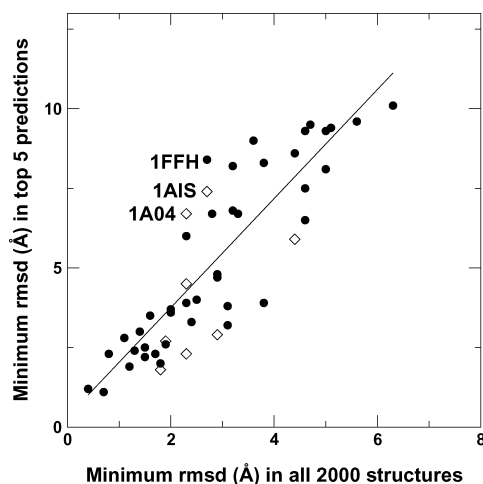
Fig. 2. The rmsd values of the best structure within the top five selected based on the DFIRE energy as a function of the rmsd value of the best structure in 2000 decoys. Solid line is from linear regression of 41 proteins (filled circles) with a correlation coefficient of 0.89. Open diamonds denote proteins whose flexible terminal regions are removed during minimization.

or protein 1A04 does not change much, although there is a significant reduction of the minimum rmsd values in 2000 minimized structures. Proteins 1FFH, 1A04A and 1AIS studied here are all a fragment of much larger proteins. The missing portions of the structures (or, interactions) in these three proteins may have led to the DFIRE energy function selecting the structures further away from the near-native ones. The same stronger correlation (with a correlation coefficient of 0.89) is also found between the lowest rmsd value in the 2000 decoys and the best rmsd value within the top five structures ranked by cluster size. The correlation between the lowest rmsd values in the 2000 decoys and the best predicted values suggests that structure selections may be improved by more comprehensive conformational sampling.

To verify the importance of sampling in structures selections, we increase the number of minimizations to 9000 for five proteins. These five proteins are chosen based on their sizes and the computational time requirement for completing 9000 minimizations. Table 4 compares the results obtained with 9000 minimizations with those with 2000 minimizations for five proteins. Increasing the number of minimizations by a factor of 4.5 (from 2000 to 9000 minimizations) leads to a reduction of the average of the minimum rmsd values for five proteins from 2.7 Å to 2.0 Å. More importantly, a better sampling improves the discrimination power of the energy or cluster-based methods for three out of five proteins. (The accuracy of the predicted structures is the same for 1B0XA and is worse for 1LBU.) The rmsd value for the best structure within the top five averaged over all five proteins decreases from 4.3 Å to 3.2 Å for energy-based selections. We further performed 45 000 minimizations for 1GAB. The rmsd value of the best near-native structure remains the same

Table 4. Conformational sampling and structure selections by energy minimization for five proteins at 2000 and 9000 decoys each (and 45000 decoys for 1GABc).

PDB	rmsd$_{min}$[a]			Top 5 E[b]			Top 5 C[c]		
	2000[d]	9000[d]	45000[d]	2000[d]	9000[d]	45000[d]	2000[d]	9000[d]	45000[d]
1LP1	2.5	1.1		3.8(E5)	1.8(E1)		3.0(C2)	1.8(C1)	
1LP1[e]	1.2	0.9		1.9(E2)	1.8(E1)		1.9(C3)	1.8(C1)	
1LBU$_{1-83}$c	4.4	3.2		5.9(E4)	7.9(E3)		7.5(C4)	7.9(C3)	
1A04$A_{150-216}$c	2.3	1.8		6.7(E2)	2.3(E5)		7.5(C4)	2.3(C1)	
1B0X*Ac*	2.3	2.1		2.3(E1)	2.3(E1)		2.3(C1)	2.3(C1)	
1GAB*c*	1.9	1.8	1.8	2.7(E1)	1.8(E5)	1.8(E4)	2.7(C2)	2.7(C1)	1.8(C4)
Ave	2.7	2.0		4.3(E2.6)	3.2(E3)		4.6(C2.6)	3.4(C1.4)	

[a]The minimum rmsd value from the native structure (in Å). [b]The minimum rmsd value (and the rank) from the top five structures ranked by energy after clustering (in Å). [c]The minimum rmsd value (and the rank) from the top five structures ranked by cluster size after clustering (in Å). [d]The number of initial structures. [e]The flexible terminal region was not included in rmsd calculations. This result is not used in average.

(1.8 Å), while the rank of the best near-native structure continues to improve from rank 5 to 4 by energy and rank 10 to 4 by cluster size. The difficulty of reducing the minimum rmsd further for 1GAB may signal the existence of an energetic barrier that makes the native structure unreachable by minimization alone.

The ability to sample conformational space for a given protein is limited by the size of the conformational space of the protein. The latter is determined by the degree of freedom for that protein. Indeed, the best rmsd values in 2000 decoys have a significant correlation with the number of residues that are not fixed (In Fig. 3 the residues not used in rmsd calculations are treated as fixed residues in this figure.). The correlation coefficient is 0.84 (0.82 if all the residues present in the minimization are used in the rmsd calculations). (In comparison, there is no significant correlation between the minimum rmsd values and the total number of residues.) There is one significant outlier (1QC7A) which is no longer an outlier if the flexible terminal region of 1QC7A is removed in minimization.

One interesting question is: What are those structures whose energies are close and even lower than the native or near-native structures? Figure 4 plots the energies of the minimized structures for 1GAB with the flexible terminal region removed, as a function of their rmsd values (1GABc). There are two dominant structures (A and B) in the 9000 minimized structures that represent the two largest structural clusters with 145 and 127 structures each. These two structures are shown in Fig. 5. The main difference between these two structures is the arrangement of three helices in a clockwise manner (Structure A) or a counter-clockwise manner (structure B). Both structures have an energy value that is lower than the energy of the native structure (but higher than the energy of the minimized native structure which is 1.0 Å rmsd away from the un-minimized native structure). Structure A is a near-native structure with a rmsd value of 2.7 Å, while structure B with a rmsd value of 8.2 Å is far from the native structure.
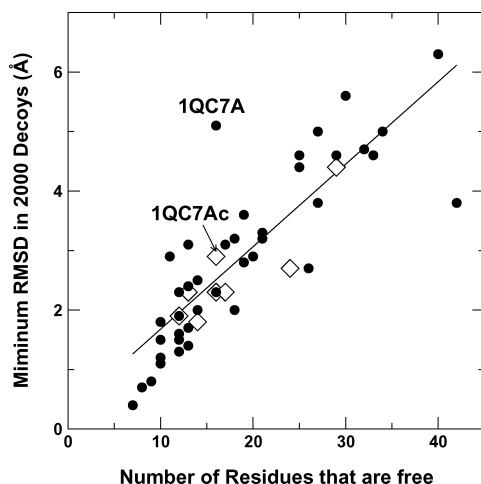
Fig. 3. The lowest rmsd value in 2000 decoys as a function of the number of unfixed residues. Open diamonds are those proteins with flexible regions removed in the minimization. Solid line is from linear regression.
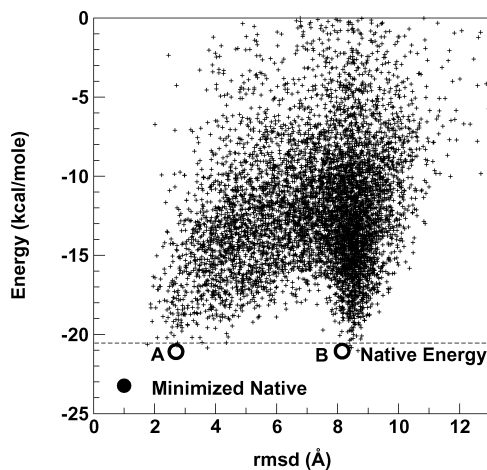

Fig. 4. The energies of minimized structures (in kcal/mole) as a function of their rmsd values for 9,000 decoys of 1GABc. The energy of the native structure is shown by the dashed horizontal line. The points A and B represent the top two largest clusters with 145 and 127 decoys respectively.

Can a non-native structure frequently sampled by DFIRE be the structures for other proteins? To address this question, we compare the misfolded structure B described above with the 8049 representative structures in the protein databank by using the structural alignment program CE.[26] The representative structures are obtained from the template library used in the fold recognition methods SPARKS[27] and SP.[3,28] The library was built by using the 40% representative domains of SCOP

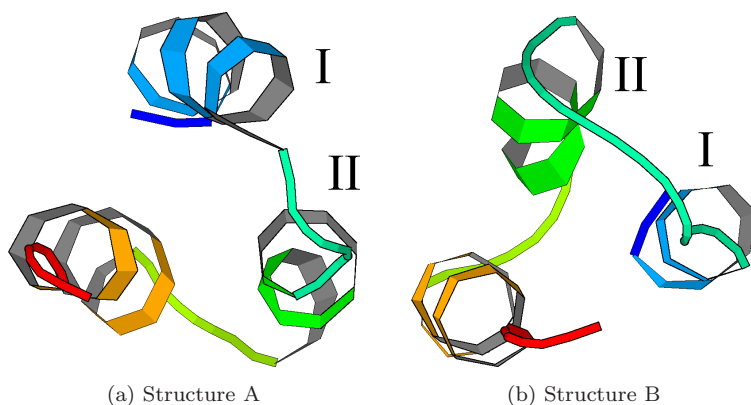(a) Structure A   (b) Structure B

Fig. 5. Structures A and B of the two dominant clusters for 1GABc. The three helices (I and II are labeled) in structures A and B are arranged clockwised and counterclockwised, respectively.

1.61[29] and then updated with new proteins released after SCOP 1.61 if they have less than 40% sequence identity with the sequences already in the library. (This was done by protein sequence culling server PISCES.[30]) We find that the decoy structure B has a close match to the PDB structure 1AKHa with a rmsd value of 2.7 Å. (All the residues in the decoy structure matched to the residues in 1AKHa and only three gaps are inserted in the decoy sequence for the match.)

Protein 1GABc, however, is a simple three-helix bundle. To further confirm the above finding, we examine the best predicted model of a six-helix protein 1BXM. The structure has a rmsd value of 10.5 Å from the native structure of 1BXM (see Table 1), but only a 4.4 Å rmsd from the PDB structure 1GTEa (90 out 98 residues of the decoy matched to the sequence of 1GTEa). Thus, a non-native structure frequently sampled by DFIRE may well be a structural fold for another protein.

Most results reported above based on the dihedral angles at the helical regions are fixed around the native values. This allows us to concentrate on the effect caused by the non-helical regions. The effect of using idealized helical structures are examined in three proteins (1PRB, 1GAB, and 1KDXA). Here an idealized helical structure refers to the helix built with its dihedral angle value that fluctuated around its ideal value. Results are shown in Table 5. In general, the use of idealized helical structures decreases the accuracy of the sampled structures (based on the best near-native structures) and the accuracy of the predicted structures (based on the best structures in the top five structures selected by energy or cluster size). This is somewhat expected. The difference in rmsd values for the best near-native structures is only 0.3 Å for 1PRB, 0.3 Å for 1GAB, and 1.0 Å for 1KDXA. The difference in rmsd values of the best structure within the top five ranked by energy is 0.9 Å for 1PRB, −0.7 Å for 1GAB and 0.3 Å for 1KDXA. Thus, at least for these three proteins, it is possible to provide a reasonable structure prediction even with the idealized helical structures.

Table 5. The results of conformational sampling via minimization for three proteins with native dihedral angles and idealized dihedral angles for the helical regions.

PDB	rmsd$_{min}$[a]		Top 5 E[b]		Top 5 C[c]	
	Native[d]	Ideal[e]	Native[d]	Ideal[e]	Native[d]	Ideal[e]
1PRBc	1.8	2.1	1.8(E2)	2.7(E5)	2.7(C2)	3.2(C1)
1GABc	4.3	4.6	7.2(E1)	6.5(E4)	6.5(C3)	6.5(C3)
1KDXA	4.5	5.5	7.4(E2)	7.7(E1)	7.7(C2)	7.7(C1)

[a]The minimum rmsd value from the native structure in 2000 minimized structures (in Å). [b]The minimum rmsd value (and the rank) from the top five structures ranked by energy after clustering (in Å). [c]The minimum rmsd value (and the rank) from the top five structures ranked by cluster size after clustering (in Å). [d]Helical regions with native helical dihedral angles. [e]elical regions with idealized helical dihedral angles.

4. Discussion

In this paper, we performed energy minimizations for 41 helical proteins in a dihedral space based on a modified all-atom DFIRE-based statistical energy function. The best near-native structures of all 41 proteins except 2HOA and 1QC7A are less than 6.5 Å from their corresponding native structures in 2000 minimizations. The minimum rmsd values for 2HOA and 1QC7A are reduced dramatically from 6.9 Å to 3.1 Å and from 7.7 Å to 5.1 Å, respectively, if their flexible terminal regions are not used in the rmsd calculations. We further demonstrated that the removal of the flexible terminal regions in minimization can further improve the quality of the sampled conformations (see Table 3).

The DFIRE energy function also provides a reasonable ranking (although it is far from perfect) for detecting near-native structures in the minimized structures. The average rmsd value for the best structure in the top five clusters, ranked by energy is 6.8 Å. The ranking by the size of the clusters can further improve the average rmsd value to 6.4 Å. These two values become significantly smaller (5.3 and 5.0 Å, respectively) if the flexible terminal regions are not used in the rmsd calculations. The improvement due to the size-based ranking highlights the importance of entropy in the structure selections (see e.g. Ref. 31). What is encouraging is that there is a significant correlation between the minimum rmsd value of the sampled conformations and the best rmsd value of the top five ranked structures. This suggested that improving the sampling can further improve the accuracy of the predicted structures. The improvement is subsequently verified by performing 9000 minimizations for five proteins.

The reasonable but limited performance in ranking by the DFIRE energy function is due to several factors. First, the DFIRE energy function, like many other statistical energy functions, is a pairwise energy function that only depends on distance. As a result, minimization tends to make structure even more compact than native ones. This in turn leads to the energies of some minimized structures that are comparable to near-native structures. For example, the top ranked structure for 1QC7A (Fig. 1) involves significant non-native interaction of the flexible

terminal region with other parts of the protein. Second, the solvent contribution to protein stability is only implicit in most statistical energy functions including the DFIRE energy function. Consequently, the energy function cannot handle the flexible region for which the direct interaction with solvent is essential. Work is in progress to search the solution for these problems associated with statistical potentials. Possible solutions include the use of multibody interactions[32–36] and orientation-dependent energy functions.[37,38]

Recently, Nanias *et al.*[7] minimized C_α-based model proteins interacting with simple Lennard-Jones potential with Miyazawa-Jernigan contact energy parameters. A global minimization technique was used. The secondary structure was determined by applying the DSSP algorithm[39] to the native protein. During minimizations, helices are treated as rigid bodies with fixed ideal parameters, terminal regions are removed, and 10 000 minimizations are performed for each protein. As shown in Table 6, our method consistently gives lower minimum rmsd values than the method used by Nanias *et al.*, although we often use longer pieces of proteins and a smaller number of minimizations (2000). Certainly, our method has benefited somewhat from the use of native helical dihedral angles (Table 5). The advantage of their method is the use of a coarse-grained model that allows computationally efficient sampling. Thus, it is of interest to use the residue-level DFIRE energy function[40] for conformational sampling. The residue-level DFIRE energy function[40] was found to be one of the best statistical energy functions for structure selections from decoys.

In another study, Zhang *et al.* sampled conformational space of the helical proteins with torsion angle dynamics and predicted restraints.[6] Their method is based on contact-map prediction and simulated annealing. The secondary structures are from the prediction of PSIPRED.[41] Their method gives an average of 4.6 Å (4.3 Å) over 23 proteins for the minimum rmsd value among 500 structures (after bootstrapping). The corresponding results for the DFIRE-guided minimizations on the same set of proteins is 5.4 Å for the first 500 structures and 4.8 Å for the 2000

Table 6. The minimum rmsd values given by two different methods.

PDB[c]	N_{res}[a]		$rmsd_{min}$(Å)[b]	
	Ref. 7	This work	Ref. 7	This work
1BW6	43	56	2.7	1.8
1FFH	83	87	3.0	2.7
1BMTA	79	90	3.7	3.2
1CTJ	82	89	5.4	5.4
1QC7A	74	86	5.5	2.9[d]
1BXM	92	98	6.4	6.3

[a]The number of residues used in minimization for a given protein used in Ref. 7 and in this work, respectively. [b]The minimum rmsd value in 10 000 structures[7] or in 2000 structures (this work), respectively. [c]Only proteins whose difference in number of residues given by two methods are less than 20. [d]After the flexible terminal is cut (see Table 3).

structures. Thus, their method is more efficient in reaching the high quality of near native structures. One possible reason for their success is the use of simulated annealing rather than simple minimization used in this study. Simulated annealing which searches for a global minimum, allows one to overcome energetic barriers to reach lower energy minima. In this work, we focus on the effect of minimization only. Application of global minimization techniques to the DFIRE energy function will be the subject of a separate study that is in progress.

The 2000 decoys for 41 proteins obtained here are likely a challenging all-atom decoy set for testing the ability for an energy function to predict structure. This is because we have shown that structures with a large rmsd value from the corresponding native structure are possible structural folds for other proteins, rather than artificial structures generated from random sampling. Moreover, the structures generated by the DFIRE energy function satisfy the principle that hydrophobic residues tend to be buried inside, while the hydrophilic residues tend to be on the surface.[14] This can be illustrated by comparing the hydrophobic and hydrophilic distributions in decoys and their corresponding native structure.

As an example, Fig. 6 compares residue depths given by the native structure, structure A (rmsd = 2.7 Å), and structure B (rmsd = 8.2 Å from 1GABc but only 2.7 Å from another protein 1AKHa) generated from 1GABc minimizations. Residue depth,[42] the distance of a residue away from the nearest possible solvent molecules, is one method to characterize the exposure of a residue to solvent. As described above, structures A and B are the representative structures of the top two structural clusters. For all three structures, it is clear that hydrophobic residues are
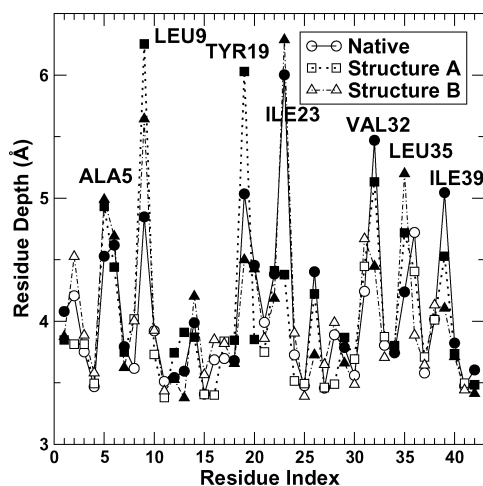

Fig. 6. The residue depths of the native structure, structure A, and structure B for the 1GAB as labeled. The open and close symbols denote hydrophilic and hydrophobic residues, respectively. For all three structures, essentially the same hydrophilic residues (open symbols) are on the surface (with small residue depths).

mostly in the core, while hydrophilic residues are mostly on the surface. There are significant correlations between the residue depths of the native structure and those of structure A (R = 0.75) and between the residue depths of the native structure and those of structure B (R = 0.80). Figure 6 indicates that the surface residues of the three structures are more or less the same, whereas the hydrophobic cores are made of slightly different hydrophobic residues. The core (residues with a residue depth greater than 5 Å) is made of ILE23, VAL32, TYR19 and ILE39 for the native structure, LEU9, TYR19 and VAL32 for structures A, and LEU9, ILE23, and LEU35 for structure B.

Similar partitioning of hydrophobic and hydrophilic residues in the three structures of 1GABc raises an interesting question: What drives nature to favor one structure over the other one? The fact that nearly identical surface with slightly different core arrangements for the three structures points to interaction in the core rather than interaction on the surface (i.e. hydrophilic residues with the water solvent) in this case. While the DFIRE energy function is capable of selecting the near-native structure A, the energy difference between structures A and B (Fig. 4) is too small to be certain that structure A is near-native while B is not. One possible reason for structure A being a near-native is it has a small rmsd value from the native structure. The cluster-size-based rank of the best near-native structure (with a rmsd value of 1.8 Å) increases as the number of minimized structures increases from 2000 to 9000, then to 45 000 for 1GABc (Table 4). Another possible reason is that multi-body cooperative effect neglected in the DFIRE statistical energy function is essential for the formation of hydrophobic core with right residues.[33,34,36]

References

1. Moult J, Fidelis K, Zemla A, Hubbard T, Critical assessment of methods of protein structure prediction (CASP) — Round V, *Proteins* **53**:334–339, 2003.
2. Rost B, Review: protein secondary structure prediction continues to rise, *J Struc Biol* **134**:204–218, 2001.
3. Cohen FE, Richmond TJ, Richards FM, Protein folding — evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example, *J Mol Biol* **132**:275–288, 1979.
4. Mumenthaler C, Braun W, Predicting the helix packing of globular proteins by self-correcting distance geometry, *Protein Sci* **4**:863–871, 1995.

5. Huang ES, Samudrala R, Ponder JW, Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based score functions, *J Mol Biol* **290**: 267–281, 1999.
6. Zhang C, Hou J, Kim S, Fold prediction of helical proteins using torsion angle dynamics and predicted restraints, *Proc Natl Acad Sci USA* **99**:3581–3585, 2002.
7. Nanias M, Chinchio M, Pillardy J, Packing helices in proteins by global optimization of a potential energy function, *Proc Natl Acad Sci USA* **100**:1706–1710, 2003.
8. Fain B, Levitt M, Funnel sculpting for in silico assembly of secondary structure elements of proteins, *Proc Natl Acad Sci USA* **100**:10700–10705, 2003.
9. Li X, Jacobson MP, Friesner RA, High-resolution prediction of protein helix positions and orientations, *Proteins* **55**:368–382, 2004.
10. Zhou H, Zhou Y, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Sci* **11**:2714–2726, 2002.
11. Liu S, Zhang C, Zhou H, Zhou Y, A physical reference state unifies the structure-derived potential of mean force for protein folding and binding, *Proteins* **56**: 93–101, 2004.
12. Zhang C, Liu S, Zhou Y, Accurate and efficient loop selections using DFIRE-based all-atom statistical potential, *Protein Sci* **13**:391–399, 2004.
13. Zhang C, Liu S, Zhu Q, Zhou Y, A unified knowledge-based energy function for protein-ligand, protein-protein and protein-DNA complexes, *J Med Chem* **48**: 2325–2335, 2005.
14. Zhou H, Zhou Y, Quantifying the effect of burial of amino acid residues on protein stability, *Proteins* **54**:315–322, 2004.
15. Zhang C, Liu S, Zhou H, Zhou Y, The dependence of all-atom statistical potentials on training structural database, *Biophys J* **86**:3349–3358, 2004.
16. Skolnick J, Kolinski A, Ortiz A, Derivation of protein-specific pair potentials based on weak sequence fragment similarity, *Proteins* **38**:3–16, 2000.
17. Samudrala R, Moult J, An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, *J Mol Biol* **275**: 895–916, 1998.
18. Lu H, Skolnick J, A distance-dependent atomic knowledge-based potential for improved protein structure selection, *Proteins* **44**:223–232, 2001.
19. Ponder JW, Richards FM, An efficient Newton-like method for molecular mechanics energy minimization of large molecules, *J Comput Chem* **8**:1016–1026, 1987.
20. Devlin T, *Textbook of Biochemistry with Clinical Correlations*, Wiley-Liss, Inc, New York, 4th edition, 1997.
21. Hobohm U, Scharf M, Schneider R, Sander C, Selection of representative protein data sets, *Protein Sci* **1**:409–417, 1992.
22. Ramachandran GN, Sasisekharan V, Conformation of polypeptides and proteins, *Adv Protein Chem* **23**:283–438, 1968.
23. Brooks B, Bruccoleri R, Olafson B *et al.*, A program for macromolecular energy, minimization, and dynamics calculation, *J Comput Chem* **4**(2):187–217, 1983.
24. Lui D, Nocedal J, On the limited memory BFGS method for large scale optimization, *Math Prog* **45**:503–528, 1989.
25. Chu R, Takei J, Knowlton R *et al.*, Redesign of a four-helix bundle protein by phage display coupled with proteolysis and its structural characterization by multidimensional NMR and x-ray crystallography, *J Mol Biol* **323**:253–262, 2002.
26. Shindyalov IN, Bourne P, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng* **11**:739–747, 1998.

27. Zhou H, Zhou Y, Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, *Proteins* **55**:1005–1013, 2004.

28. Zhou H, Zhou Y, Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments, *Proteins* **56**:accepted, 2004.

29. Murzin AG, Brenner SE, Hubbard T, Chothia C, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol* **247**:536–540, 1995.

30. Wang G, Dunbrack Jr. RL, PISCES: a protein sequence culling server, *Bioinformatics* **19**:1589–1591, 2003.

31. Zhang Y, Kolinski A, Skolnick J, TOUCHSTONE II: A new approach to *ab initio* structure prediction, *Biophys J* **85**:1145–1164, 2003.

32. Kocher J-PA, Rooman MJ, Wodak SJ, Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches, *J Mol Biol* **235**: 1598–1613, 1994.

33. Kolinski A, Galazka W, Skolnick J, On the origin of the cooperativity of protein folding: Implication from model simulations, *Proteins* **26**:271–287, 1996.

34. Liwo A, Oldziej S, Pincus MR *et al.*, A united-residue force field for off-lattice protein-structure simulations, I. Functional forms and parameters of long-range sidechain interaction potentials from protein crystal data, *J Comput Chem* **18**:849–872, 1997.

35. Munson PJ, Singh RJ, Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment, *Protein Sci* **6**:1467–1481, 1997.

36. Gan HH, Tropsha A, Schlick T, Lattice protein folding with two and four-body statistical potentials, *Proteins* **43**:161–174, 2001.

37. Kortemme T, Morozov A, Baker D, An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes, *J Mol Biol* **326**:1239–1259, 2003.

38. N-V Buchete JS, Thirumalai D, Development of novel statistical potentials for protein fold recognition, *Curr Opin Struct Biol* **14**:225–232, 2004.

39. Kabsch W, Sander C, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**:2577–2637, 1983.

40. Zhang C, Liu S, Zhou H, Zhou Y, An accurate residue-level pair potential of mean force for folding and binding based on the distance-scaled ideal-gas reference state, *Protein Sci* **13**:400–411, 2004.

41. Jones D, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol* **292**:195–202, 1999.

42. Chakravarty S, Varadarajan R, Residue depth: a novel parameter for the analysis of protein structure and stability, *Struct Fold Des* **15**:723–732, 1999.

Hongzhi Li obtained his B.Sc. and Ph.D. in Chemistry from the University of Science and Technology of China in 1991 and the Institute of Chemistry, Chinese Academy of Sciences in 1996 respectively. He received his M.Sc. in Computer Science at Queen's University, Canada in 2003 and has been a Post-Doctoral Research Associate in the Department of Physiology and Biophysics at the State University of New York at Buffalo, USA, since 2003.

Yaoqi Zhou obtained his B.Sc. and Ph.D. in Chemical Physics from the University of Science and Technology of China in 1984 and the State University of New York in Stony Brook, USA, in 1990 respectively. After a brief stint in the industry, he became a post-doctoral fellow in 1994 at the State University of New York in Stony Brook, USA, and worked with George Stell and Harold Friedman. From 1994 to 1995, he was a post-doctoral fellow with Carolina State University and worked with Carol Hall, and from 1995 to 2000, he was at Harvard University and worked with Martin Karplus. He became an Assistant Professor in the Department of Physiology and Biophysics at the State University of New York in Buffalo, USA, in 2000 and has been an Associate Professor since 2004.