Specific interactions for ab initio folding of protein terminal regions with secondary structures

Yuedong Yang and Yaoqi Zhou*

Indiana University School of Informatics, Indiana University-Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202

ABSTRACT

Proteins fold into unique three-dimensional structures by specific, orientation-dependent interactions between amino acid residues. Here, we extract orientation-dependent interactions from protein structures by treating each polar atom as a dipole with a direction. The resulting statistical energy function successfully refolds 13 out of 16 fully unfolded secondary-structure terminal regions of 10–23 amino acid residues in 15 small proteins. Dissecting the orientation-dependent energy function reveals that the orientation preference between hydrogen-bonded atoms is not enough to account for the structural specificity of proteins. The result has significant implications on the theoretical and experimental searches for specific interactions involved in protein folding and molecular recognition between proteins and other biologically active molecules.

INTRODUCTION

The most well-studied specific interaction for protein folding is hydrogen bonding.[1] Little attention, however, has been paid to the orientation dependence of interactions between polar atoms that are not hydrogen bonded, despite evidence of their role in the formation of α-helices and β-sheets.[2,3] Moreover, the possible orientation dependence of interactions between polar and nonpolar atoms is ignored even though the hydrophobic effect is caused by the reorientation of water molecules near a hydrophobic surface.[4]

Recently, Zhu et al.[5] compared several statistical energy functions and physical-based energy functions and analyzed their respective abilities to refold partially unfolded helices or strands. They found that among the energy functions tested, the most effective one is an all-atom, distance-dependent, pairwise statistical energy function based on a Distance-scaled, Finite-Ideal gas REference (DFIRE) state.[6] In one test, more than 80% of conformations from 104 segments of 81 proteins (4 Å rmsd, in average from the native conformation) were refined to within 2 Å. This happened despite the lack of hydrogen bonding or any orientation-dependent term in the DFIRE energy function. However, success deteriorates significantly as the initial structures of the helical/strand segments deviate more from their respective native conformations.[5]

Here, we propose a "dipolar" DFIRE (dDFIRE) energy function based on the orientation angles involved in dipole–dipole interactions. This is done by treating each polar atom as a dipole. The orientation of the dipole is defined by the bond vectors that connect the polar atom with other heavy atoms. The dDFIRE energy function is then extracted from protein structures based on the distance between two atoms and the three angles involved in dipole–dipole interactions. This approach takes into account the hydrogen bonding interaction via the physical dipole–dipole interaction. More importantly, it provides a consistent treatment for the possible orientation-dependent interactions between polar and nonpolar atoms and between polar atoms that are non-hydrogen-bonded. Moreover, an integrated treatment of distance and angle dependence produces a parameter-free statistical energy function. Existing orientation-dependent knowledge-based energy functions are limited to either hydrogen bonding[7–9] or geometry-based orientation in coarse-grained models.[10–12]

This all-atom statistical energy function was employed to fold protein terminal regions with secondary structures. Folding completely unfolded terminal segments is challenging because it requires the restoration of both main-chain and side-chain conformations. Moreover, compared with internal regions, terminal regions are more flexi-

ble[13] and often exposed.[14] This test is necessary because native-like fragment structures are difficult to produce by contemporary energy functions,[15] and the prevailing structure-prediction techniques[16,17] are to mix and/or match known native structures either in whole (template-based modeling) or in part (fragment assembly). The ab initio refolding of a completely unfolded segment also has its own biological significance, as protein folding assisted by a prefolded domain (pro-domain) is common in many proteins.[18,19] For example, a small extra domain in the PDZ-3 domain construct (1BE9) can fold only after the PDZ-3 domain has folded.[20]

THEORY

The DFIRE potential

The DFIRE-based statistical energy function $[\bar{u}^{\mathrm{DFIRE}}(r_{ij})]$ is built on distance-scaling, a reference state of uniformly distributed ideal gas points, and the statistics of two atoms at a distance apart in known protein structures. The derivation of the equations, the method for extracting the DFIRE-based potential using a structure database, as well as the application of the resulting potential have been described previously.[6,21] Here, we give a brief summary below.

The atom–atom potential of mean force $\bar{u}^{\mathrm{DFIRE}}(r_{ij})$ between atom types i and j that are distance r apart is given by[6]

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

where R is the gas constant, $T = 300$ K, $\alpha = 1.61$, $N_{\mathrm{obs}}(i,j,r)$ is the number of (i,j) pairs within the spherical shell at distance r observed in a given structure database, residue specific atomic types are used (167 atomic types), $r_{\mathrm{cut}} = 14.5$ Å, and $\Delta r(\Delta r_{\mathrm{cut}})$ is the bin width at $r(r_{\mathrm{cut}})$. ($\Delta r = 2$ Å, for $r < 2$ Å; $\Delta r = 0.5$ Å for 2 Å $< r < 8$ Å; $\Delta r = 1$ Å for 8 Å $< r < 15$ Å). The number of observed atomic (i,j) pair with the spherical shell at distance $r[N_{\mathrm{obs}}(i,j,r)]$ is obtained from a structural database of 3574 nonredundant (less than 30% homology) high-resolution proteins (resolution < 2.0 Å and $R\%$ factor < 0.25) from Hobohm et al.[22] This is a larger database than 1011 proteins used originally.[6]

The value of α ($\alpha = 1.61$) was determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in 1011 finite protein-size spheres. Recently, Zhu et al.[5] found that $\alpha = 1.51$ improves the accuracy of restoring partially unfolded strands and hairpins for the proteins they studied. Shen and Sali[23] also introduced an analytical atom–atom distance and protein-size dependent α value. Here, we fixed α at 1.61 because the focus of this work is the effect of orientation dependence on fragment refolding.

The dDFIRE potential

The dipolar DFIRE potential adds orientation dependence by separating polar atoms from nonpolar atoms. Nonpolar atoms are carbon atoms, while polar atoms consist of nitrogen and oxygen atoms in all residues, and the sulfur atom in the residue Cys. Each polar atom possesses a reference direction that mimics the orientation of a dipole. The reference vector of a given polar atom, $\vec{r}_p^{\mathrm{ref}}$, is defined based on the sum of the bond vectors that chemically connect the polar atom to other heavy atoms ($\vec{r}_p^{\mathrm{ref}} \equiv \sum_i \vec{r}_{ip}$ with $\vec{r}_{ip} = \vec{r}_p - \vec{r}_i$). There is a special case where the sum of connected bond vectors yields approximately a zero vector. This happens when a polar atom bonded with three heavy atoms in a planar sp2-hybridization. In this case, the polar atom is treated as a nonpolar atom. Only the backbone nitrogen atom of residue Pro belongs to this category. Bonds to hydrogen atoms are ignored because we are developing a potential for heavy atoms only. We also do not separate double from single covalent bonds in defining the reference dipole direction, because they will be taken care by using different atom types.

The above-defined reference direction approximately mimics the dipole direction of polar atoms. For example, the dipole direction of oxygen atoms in —C=O should be mostly determined by the partial charges in C and O. This will produce the dipole direction along the distance vector between atoms C and O. For backbone nitrogen atoms (—C$_2$N—H), the reference vector is the sum of $\vec{r}_{\mathrm{CN}}$ and $\vec{r}_{\mathrm{C_{\alpha}N}}$. This reference direction in the absence of polar hydrogen atoms in our model approximately captures the direction of $\vec{r}_{\mathrm{NH}}$, which is roughly the dipole direction for the backbone polar atom N because most of the partial charges are located in N and H atoms in a physical based force field such as CHARMM.[24]

The orientation dependence of polar interactions between polar atoms p and q is described by three angles in dipole–dipole interactions.[25] As illustrated in Figure 1, they are θ_{pq}, the angle between $\vec{r}_p^{\mathrm{ref}}$ and $\vec{r}_q^{\mathrm{ref}}$; θ_p, the
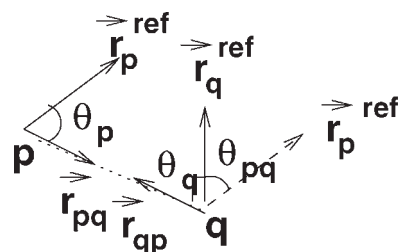
Figure 1

Definition of the orientation angles θ_p, θ_q, and θ_{pq}. Here, $\vec{r}_p^{\mathrm{ref}}$ and $\vec{r}_q^{\mathrm{ref}}$ are the reference directions for polar atoms p and q, respectively. $\vec{r}_p^{\mathrm{ref'}}$ is the parallel displacement of $\vec{r}_p^{\mathrm{ref}}$.

angle between $\vec{r}_p^{\text{ref}}$ and the distance vector $\vec{r}_{pq}$; and θ_q, the angle between $\vec{r}_q^{\text{ref}}$ and the distance vector $\vec{r}_{qp}$ ($\vec{r}_{qp} = -\vec{r}_{pq}$).

The above-defined reference directions and orientations have the following advantages: First, the positions of hydrogen atoms are not required. That is, one can develop an energy function for heavy atoms only. This is advantageous because only the positions of heavy atoms are known in X-ray structures. Second, the orientation angles involved in backbone hydrogen bonds between carboxy oxygen atoms and backbone nitrogen atoms can be described very well. The reference dipole direction for a nitrogen atom bonded with two carbon atoms is approximately the direction where the amide hydrogen atom is located. Hydrogen bonds are often described by the distance between a donor (D) and an acceptor (A), the AB-A⋯D angle (AB — the acceptor base) and the D-H⋯A angle.[26] These angles are closely related to θ_p and θ_q angles defined here (the minor difference is that a hydrogen bond is strongest when $\theta_{pq} \sim 180°$, $\theta_{p(q)} \sim 0°$, and AB-A⋯D and D-H⋯A angles are close to 180°). For a nitrogen atom bonded with one carbon atom and two hydrogen atoms (e.g., N_η of Arg), the reference "dipole" direction ($\vec{r}_{CN}$) will allow to measure the averaging effect of the two possible hydrogen bonds.

The equation for the dDFIRE potential can be obtained from a simple extension of Eq. (1) incorporating the above-mentioned three additional variables θ_p, θ_q, and θ_{pq}. Interaction between the two polar atoms p and q, $\bar{u}^{\text{dDFIRE}}(r_{pq}, \theta_p, \theta_q, \theta_{pq})$, is given by

$$\bar{u}^{\text{dDFIRE}}(r_{pq}, \theta_p, \theta_q, \theta_{pq})$$
$$= \begin{cases} -RT \ln \frac{N_{\text{obs}}(p,q,\theta_p,\theta_q,\theta_{pq},r)}{(\frac{r}{r_{\text{cut}}})^\alpha \frac{\Delta r}{\Delta r_{\text{cut}}} N_{\text{obs}}(p,q,\theta_p,\theta_q,\theta_{pq},r_{\text{cut}})}, & r < r_{\text{cut}}, \\ 0, & r \geq r_{\text{cut}}, \end{cases} \quad (2)$$

where $N_{\text{obs}}(p, q, \theta_p, \theta_q, \theta_{pq}, r)$ is the number of observed pairs of polar atoms p and q at distance r apart, with orientation angles θ_p, θ_q, and θ_{pq}.

To reduce the amount of structural data required to train the dDFIRE potential, we assume that the angle dependence on θ_p, θ_q, and θ_{pq} is independent of each other. A simple derivation leads to

$$\bar{u}^{\text{dDFIRE}}(r_{pq}, \theta_p, \theta_q, \theta_{pq}) = \bar{u}^{\text{DFIRE}}(r_{pq}) + \bar{u}(\theta_p | r_{pq})$$
$$+ \bar{u}(\theta_q | r_{pq}) + \bar{u}(\theta_{pq} | r_{pq}), \quad (3)$$

where $\bar{u}(\theta_p | r_{pq}) = -RT \ln \left[P_{pq}^{\text{obs}}(\theta_p | r) / P_{pq}^{\text{obs}}(\theta_p | r_{\text{cut}}) \right]$, $\bar{u}(\theta_q | r_{pq})$ $= -RT \ln \left[P_{pq}^{\text{obs}}(\theta_q | r) / P_{pq}^{\text{obs}}(\theta_q | r_{\text{cut}}) \right]$, and $\bar{u}^{\text{pp}}(\theta_{pq} | r_{pq}) = -RT \ln \left[P_{pq}^{\text{obs}}(\theta_{pq} | r) / P_{pq}^{\text{obs}}(\theta_{pq} | r_{\text{cut}}) \right]$. Here, $P_{pq}^{\text{obs}}(\theta_p | r)$ $[= N_{\text{obs}}(p,q,\theta_p | r) / N_{\text{obs}}(p,q,r)]$, $P_{pq}^{\text{obs}}(\theta_q | r)$ $[= N_{\text{obs}}(p,q,\theta_q | r) / N_{\text{obs}}(p,q,r)]$, and $P_{pq}^{\text{obs}}(\theta_{pq} | r)$ $[= N_{\text{obs}}(p, q, \theta_{pq} | r) / N_{\text{obs}}(p,q,r)]$ are conditional probabilities.

For the interaction between a polar atom p and a nonpolar atom n, there is no reference vector for atom n.

Thus, we have

$$\bar{u}^{\text{dDFIRE}}(r_{pn}, \theta_p) = \bar{u}^{\text{DFIRE}}(r_{pn}) + \bar{u}(\theta_p | r_{pn}). \quad (4)$$

For the interaction between the two nonpolar atoms n_1 and n_2, the dDFIRE potential is the same as the DFIRE potential. That is,

$$\bar{u}^{\text{dDFIRE}}(r_{n_1 n_2}) = \bar{u}^{\text{DFIRE}}(r_{n_1 n_2}). \quad (5)$$

Equations (3–5) show that the dDFIRE potential becomes the DFIRE potential when the orientation dependence is neglected. Therefore, the distance-bin procedure and atom types in the DFIRE potential are used to extract the dDFIRE potential. Additionally, all angles were divided into six bins based on $\cos(\theta)$. The ranges of $\cos(\theta)$ values for these six bins are $(-1, -2/3)$, $(-2/3, -1/3)$, $(-1/3, 0)$, $(0, 1/3)$, $(1/3, 2/3)$, and $(2/3, 1)$. Using six bins is a result of a balance between obtaining more information from the structural database (more bins) and having enough statistics from the available database. Because this study represents a preliminary assessment of the dDFIRE energy function, we did not test if a different number of angle bins would further improve the accuracy of the dDFIRE energy function.

Database dependence

A database of 3574 proteins was used to obtain $N_{\text{obs}}(p, q, \theta_p | r)$, $N_{\text{obs}}(p, q, \theta_q | r)$, and $N_{\text{obs}}(p, q, \theta_{pq} | r)$. We found that the average number of observed atomic pairs per angle bin was 737 from the database of 3574 nonredundant high-quality protein structures. We made no attempt to exclude the 16 proteins folded in this work from the 3574 proteins. The small number of proteins makes a negligible contribution to the statistics of such large database; as demonstrated in an early study, the inclusion or exclusion of a protein in generating the DFIRE energy function makes no difference in native-structure selection of that protein from decoys by the energy function.[6] To be sure, we also obtained a version of dDFIRE potential with 1011 proteins. This version of dDFIRE was tested to refold the segments of 1o82a and 1fltx. The resulting global rmsd values were 0.63 ± 0.03 Å for 1o82a and 1.86 ± 0.03 Å for 1fltx. These values are essentially the same as 0.8 ± 0.1 Å for 1o82a and 1.4 ± 0.1 Å for 1fltx, reported in Tables I and II.

Genetic algorithm for global minimization

The method used for this study is similar to the global minimization technique developed for a simple six-state model.[27]

Initial conformations

Each conformation is described by internal coordinates: the bond lengths, bond angles, planar torsion

Table I
Restoration of Unfolded Terminal Regions by a Genetic Algorithm with the DFIRE or dDFIRE Energy Function

PDB Id#[c]	# Res.[d]	Struc. Type[e]	Unfolded Range (#)[f]	Unfolded Type[g]	Local rmsd (Å)[a] Initial[h]	Local rmsd (Å)[a] DFIRE[i]	Local rmsd (Å)[a] dDFIRE[j]	Global rmsd (Å)[b] Initial[h]	Global rmsd (Å)[b] DFIRE[i]	Global rmsd (Å)[b] dDFIRE[j]
2guzb	65	5α	95–117 (23)	1α	7.0 ± 1.4	7.3 ± 1.1	0.54 ± 0.09	25 ± 9	9.7 ± 1.5	0.9 ± 0.1
1i2ta	61	4α	1051–69 (19)	1α	6.2 ± 1.2	6.4 ± 0.8	0.54 ± 0.10	20 ± 7	9.1 ± 0.6	0.9 ± 0.2
1u84a	81	4α	65–83 (19)	1α	6.1 ± 1.0	1.1 ± 0.6	0.33 ± 0.05	21 ± 6	1.7 ± 0.5	0.64 ± 0.03
1r690	61	5α	44–61 (18)	2α	6.2 ± 1.3	3.2 ± 0.6	0.67 ± 0.12	16 ± 4	6.5 ± 2.4	0.85 ± 0.01
1o82a	70	6α	51–70 (20)	2α	6.1 ± 1.0	3.9 ± 0.6	0.63 ± 0.04	20 ± 5	5.2 ± 1.0	0.8 ± 0.1
1opd0	85	3α4β	70–85 (16)	1α	5.4 ± 0.8	1.5 ± 0.4	1.19 ± 0.01	19 ± 6	2.0 ± 0.4	1.8 ± 0.04
2igd0	56	1α4β	52–61 (10)	1β	4.0 ± 1.2	0.67 ± 0.01	0.39 ± 0.03	18 ± 5	0.75 ± 0.03	0.51 ± 0.08
1vcc0	73	2α5β	63–73 (11)	1β	4.2 ± 1.0	3.3 ± 1.1	2.6 ± 0.2	17 ± 4	6.1 ± 2.1	4.1 ± 0.3
(Native bonds)[k]					(4.2 ± 1.0)	(2.7 ± 0.6)	(1.4 ± 0.5)	(18 ± 4)	(3.1 ± 0.3)	(1.7 ± 0.4)
2hsla	89	1α9β	82–93 (12)	1α1β	4.1 ± 0.8	1.6 ± 0.4	0.63 ± 0.07	17 ± 6	3.24 ± 0.03	1.2 ± 0.1
2cc6a	62	1α3β	52–62 (11)	1β	4.9 ± 1.7	0.44 ± 0.06	0.58 ± 0.01	18 ± 6	0.69 ± 0.1	0.69 ± 0.07
2ptl0	61	1α4β	68–78 (11)	1β	4.6 ± 1.5	2.3 ± 0.1	0.79 ± 0.08	22 ± 5	2.72 ± 0.02	1.03 ± 0.08
1csp0	64	5β	54–64 (11)	1β	4.7 ± 1.5	2.2 ± 0.7	0.36 ± 0.03	19 ± 5	2.4 ± 0.7	0.43 ± 0.04
1csp0[l]	64	5β	1–10 (10)	1β	4.1 ± 1.0	0.88 ± 0.44	0.46 ± 0.07	17 ± 4	1.0 ± 0.5	0.53 ± 0.09
1fltx	95	8β	214–26 (13)	1β	6.7 ± 1.9	4.71 ± 0.07	1.08 ± 0.11	22 ± 5	5.9 ± 0.2	1.4 ± 0.1
2ayda	66	5β	346–58 (13)	1β	4.7 ± 0.8	4.32 ± 0.02	4.46 ± 0.13	18 ± 5	7.82 ± 0.01	8.1 ± 0.1
(Native bonds)[k]					(4.8 ± 1.0)	(4.2 ± 0.2)	(1.0 ± 0.9)	(17 ± 4)	(8.07 ± 0.05)	(1.4 ± 1.0)
2extb	66	6β	61–72 (12)	2β	6.5 ± 1.5	4.56 ± 0.03	3.45 ± 0.02	19 ± 4	8.6 ± 0.1	7.755 ± 0.002
(Dimer)[m]					(5.8 ± 1.5)	(4.3 ± 1.8)	(0.58 ± 0.06)	(15 ± 5)	(6.7 ± 2.5)	(1.3 ± 0.2)
Average (median) (Å)					5.4 (5.1)	3.0 (2.7)	1.2 (0.63)	19 (19)	4.6 (4.2)	2.0 (0.86)

[a]Root-mean-squared distance (rmsd) of the global minimum structure of the refolded region from its native conformation.
[b]rmsd value of the global minimum structure of the entire protein from its native conformation.
[c]Protein Data Bank Identification number. The 4th digit is the chain ID.
[d]Number of residues in the native structure.
[e]The structural type represented by the number of α-helices and β-strands in the protein structure.
[f]Residue range of the unfolded regions (number of residues unfolded).
[g]The structural type of the unfolded region.
[h]The mean and standard deviation of rmsd values of initial 120 structures.
[i]The mean and standard deviation of the rmsd values from three independent global minimizations with the DFIRE energy function.
[j]The mean and standard deviation of the rmsd values from three independent minimizations with the dDFIRE energy function.
[k]The native improper dihedrals and bond lengths and angles are used for the unfolded segment, as in the rest of the protein.
[l]The N-terminal region is unfolded.
[m]The unfolded segment of a monomer is refolded in a dimeric structure.

angles, backbone ϕ/ψ torsion angles, and side-chain χ angles. The bond lengths, bond angles, and planar torsion angles of a selected segment in a protein are fixed with standard values from the AMBER 99 force field.[28]

The backbone ϕ/ψ torsion angles of the residues in the segment are randomly assigned according to the observed residue-specific probability in the backbone-dependent rotamer library[29] (http://dunbrack.fccc.edu/bbdep, Ver-

Table II
The Accuracy of Refolding in Global rmsd Values (in Å) by Various Orientation-Dependent Components in the dDFIRE Energy Function

PDB ID[a]	Unfolded type[b]	DFIRE[c]	DFIRE + dDFIRE orientation components H-bond[d]	DFIRE + dDFIRE orientation components Polar–nonpolar[e]	DFIRE + dDFIRE orientation components Polar–Polar[f]	DFIRE + dDFIRE orientation components Full dDFIRE[g]
2guzb	α	9.7 ± 1.5[h]	0.8 ± 0.1	1.2 ± 0.6	1.3 ± 0.3	0.9 ± 0.1
1i2ta	α	9.1 ± 0.6	1.35 ± 0.03	0.97 ± 0.21	1.32 ± 0.05	0.9 ± 0.2
1o82a	2α	5.2 ± 1.0	2.0 ± 0.4	1.5 ± 0.6	1.7 ± 0.3	0.8 ± 0.1
1fltx	β	5.9 ± 0.2	5.8 ± 0.6	2.9 ± 1.7	1.8 ± 0.1	1.4 ± 0.1
2extb-dimer	2β	6.7 ± 2.5	4.7 ± 2.8	4.2 ± 3.2	4.1 ± 3.3	1.3 ± 0.2

[a]Protein Data Bank Identification number. The 4th digit is the chain ID.
[b]The structural type of the unfolded region.
[c]The DFIRE energy function.
[d]Only the orientation dependence between hydrogen-bonded atoms (e.g., O and N in the main-chains).
[e]Only the orientation dependence between polar and nonpolar atoms.
[f]Only the orientation dependence between polar atoms (including hydrogen-bonded atoms).
[g]The full dDFIRE energy function.
[h]The number in each cell is the global root-mean-squared distance and its standard deviation between the native protein and the global minimum of the segment-refolded protein (in Å). The standard deviation is calculated from three independent global minimizations.

sion of May 2002). The side-chain χ angles are then randomly assigned according to rotamer probability based on the previously assigned main-chain ϕ/ψ angles. In the backbone-dependent rotamer library, ϕ and ψ angles are divided into 36 bins. Each bin is $10°$. The exact values of ϕ/ψ angles with an angle bin are determined by adding a uniformly distributed random number between $-5°$ and $5°$ to the mean value of the angle bin. The exact values of side-chain χ angles are the average value plus a Gaussian-distributed random number according to the variance given by the rotamer library. A helical conformation has the highest probability in the rotamer library. To reduce the bias of generating too many helical residues in the initial conformation, the maximum probability of any ϕ/ψ angle bin is set to 10 times the average probability. Using observed probabilities in the protein structural database was to increase the efficiency of subsequent sampling. No specific native information of the protein is used at any stage of the global minimization. The initial conformations are generated 16 Å to 20 Å away from the native conformation in term of global rmsd values (See Table I).

Local minimization

For a given conformation, the new ϕ/ψ angle of a selected residue is randomly chosen from its own bin or its nearest neighboring 24 angle bins. We used only ϕ/ψ angle bins observed in the backbone-dependent rotamer library, however, the probabilities of each bin are not used in the selection of the new bin. Once the ϕ/ψ angle bin is selected, the side-chain χ angle bins and the actual values of ϕ, ψ, and χ angles are also determined by the method described in "Initial conformations." The new conformation is accepted if it has a lower energy than the current conformation and rejected if not. This procedure repeats until reaching either 100 successive rejections of new conformations or a total of 1000 attempted angle changes.

Fitness function

The fitness function of each conformation in generation l, relative to other conformations in the same generation, is calculated from

$$f_i^l = \frac{1}{\rho_i} \exp\left[-\frac{(e_i - e_{\min}^{l-1})}{(T_{\text{env}} \Delta e^{l-1})} \right], \tag{6}$$

where e_i is the energy of conformation i (included the fixed portion of the protein) based on either the DFIRE or dDFIRE energy function, $e_{\min}^{l-1}$ and Δe^{l-1} are the lowest energy and the root-mean-square deviation of the energies in the parent generation $l - 1$, respectively, with T_{env} set to 1.5. To increase the structural diversity of sampled conformations, the fitness function of a given

conformation is further normalized by its similarity density ρ_i. $\rho_i = \sum_{j,j \neq i} S_{ij}$, where S_{ij} is the number of residues in identical conformational states between conformations i and j. Two residues are considered in an identical conformation state if $|\phi_i - \phi_j| < 10°$, $|\psi_i - \psi_j| < 10°$, and $|\phi_i - \phi_j| + |\psi_i - \psi_j| < 15°$. This normalized fitness function reduces the probability of survival for a conformation in which many similar conformations exist.

Genetic algorithm

Initial conformations for a given terminal region of a protein are generated and locally minimized as described earlier. These conformations serve as the first parent generation. A population of 120 conformations (i.e., $N_c = 120$) was used in this study. The parent conformations are ranked with the density-normalized fitness function and the standard roulette-wheel selection procedure.[30] The roulette-wheel selection procedure is also called the fitness proportionate selection, in which a conformation is chosen from the conformation pool in random with a probability proportional to their fitness. The procedure continues until a preset number of conformations is selected. The sequentially ranked conformation pair are chosen to be the parent to breed two new conformations, first by two-point crossover and then by mutation operations. In the crossover operation, a randomly chosen fragment is selected and all the torsional angles in the fragment are exchanged between the two parents. To maintain the efficiency, the fragment size is chosen between 30 and 50% of the chain length of the unfolded protein segment. After crossover, each residue in the new child conformations is randomly subjected to a 2% probability for a mutation operation. A low probability of 2% is used to avoid a significant loss of the conformational identity of the parents. ϕ, ψ, and χ angles of a mutated residue are reselected according to the probabilities in the backbone-dependent rotamer library as described in generating initial conformations. These newly generated child conformations, together with the parent conformations, are subjected to structural filtering to maintain the conformational diversity. Two conformations are considered as structurally identical if they have more than 30% residues in the same ϕ/ψ states (defined earlier). The conformation with a higher energy is removed. New parent conformations are chosen from the surviving conformations according to the density-normalized fitness function and the standard roulette-wheel selection procedure.[30] This evolution process continues until the global minimum conformation is not changed for 100 successive generations or a total number of 400 generations is reached.

There are a few adjustable parameters in this algorithm (e.g., number of populations, cutoffs for fragment sizes, probablity of mutations, and the definition of conformational similarity). They were obtained by a few trials and

errors. These parameters, however, may not be optimal for effective sampling of near native structures. We defer any extensive investigation of these parameters to future publications, as the main purpose of this article is to evaluate the newly developed orientation-dependent energy function.

Segment Refolding

We randomly selected 15 small globular proteins (<100 residues) with diverse structural topologies, both as a whole and in their terminal regions (Table I). Sixteen terminal regions of these proteins are unfolded with a random assignment of ϕ, ψ, and χ dihedral angles for the main-chain and side-chains according to their probabilities observed in a database of protein structures. We refolded the unfolded segments using ideal bond lengths and angles and improper dihedral angles, while the rest of the proteins was fixed in the native conformations. The above-described genetic algorithm was used to search for the global minimum given by a DFIRE (or dDFIRE) energy function. Each refolding was performed three times, with different random initial conformations for the unfolded segments. Multiple global minimizations were used to check the robustness of the final folded structures. Because this article focuses on evaluating the proposed energy function rather than testing sampling techniques, we limited the maximum size of unfolded regions to be less than 15 residues for a strand-containing segment, and less than 25 residues for a helix-containing segment. Each refolding event takes about 85 CPU h for the dDFIRE energy and 40 h for the DFIRE energy function on a single CPU in AMD Dual-Core Opteron Processor (2.4 GHz). The combined computational time for all tests performed exceeds 1 year on a single CPU.

RESULTS

Figure 2 and Table I summarize the refolding results of four single helices, two two-helix bundles, eight single strands, one mixed helix/strand segment, and one β-hairpin. The structural accuracy of a segment is described by a local rmsd between the refolded conformation and the native conformation of the segment and a global rmsd between the refolded and native conformations of the entire protein (calculated based on main-chain C_α atoms). The former measures the restoration of a segment structure and latter indicates the restoration of the segment structure and its orientation relative to the rest of the protein. Figure 2 shows that the global rmsd values of initial conformations range from 16 to 25 Å. The DFIRE energy function successfully folds three segments within 1 Å rmsd and five proteins within 2 Å rmsd. By comparison, the dDFIRE energy function restores nine segments within 1 Å rmsd and 13 segments within 2 Å
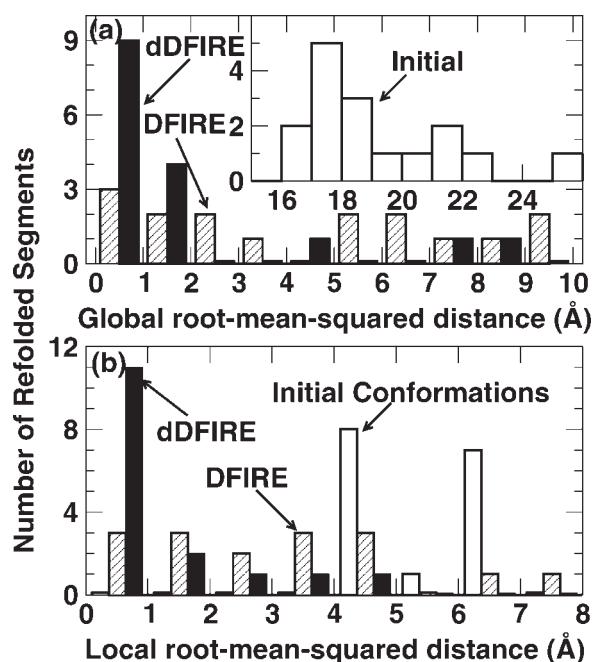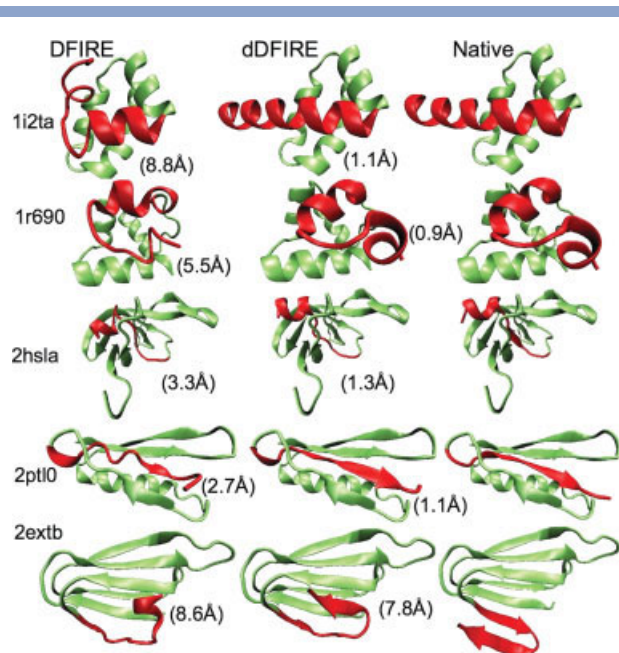

Figure 2

*The population distributions of global (**a**) and local (**b**) root-mean-squared distances (rmsd) between the native structure of a protein and the corresponding refolded structure. The initial structures of unfolded segments and structures refolded by DFIRE and dDFIRE are as labeled. A significantly high population at low local and global rmsd values indicates the successful restoration of segment structures and their orientations relative to the rest of the proteins.*

rmsd. The dDFIRE energy function improves over the DFIRE energy function even more in local rmsd values. There are 11 segments folded within 1 Å rmsd by the former and only three by the latter. Thus, the dDFIRE energy produces not only more accurate structures of unfolded segments, but also more accurate orientations of the segments relative to the rest of proteins than the DFIRE energy function. In addition, a factor of six reduction is observed for the standard deviations of the rmsd values of the global-minimum structures from three independent global-energy minimizations. Greater structural similarities among independent minimizations confirm that the dDFIRE energy function is more specific than the DFIRE energy function.

As illustrative examples, Figure 3 compares five native structures to structures whose fragments in five different structural elements are refolded by DFIRE and by dDFIRE, respectively. There is a clear difference between the structures refolded by dDFIRE and those by DFIRE. For example, dDFIRE can refold the C-terminal single helix segment of 1i2ta very well, while DFIRE breaks it into two segments. A similar phenomenon is observed for 2guzb and 1u84. In addition, unlike dDFIRE, DFIRE fails to yield two helices in 1r690 (as shown) and 1o82a. Moreover, for single strand, DFIRE produces either a

Figure 3

The segment structures (in red) refolded by DFIRE (left) and dDFIRE (center) for five proteins as labeled are compared with their respective native conformations (right). The fixed portion of each protein is colored in light green.

ture of another monomer that interacts with the C-terminal β-hairpin (i.e., the use of a dimer), the C-terminal β-hairpin can be refolded very accurately (1.3 Å in average global rmsd and 0.58 Å in average local rmsd) by dDFIRE. By comparison, the DFIRE energy function continues to misfold the β-hairpin with an average global rmsd of 6.7 Å (See Table I).

The other two poorly folded proteins are 1vcc0 (4 Å in global rmsd) and 2ayda (8 Å in global rmsd). Both refolded structures of 1vcc0 and 2ayda display a shortened loop and a shifted strand. The examination of global rmsd values of sampled conformations reveals a cluster of structures more native-like than the global minimum structure. Because standard bond lengths and bond angles are used for refolded segments and native bond values are employed for the rest of the proteins, this may have caused the shift of the refolded strand. Thus, we performed the global minimizations with native bond values for the unfolded segments. Both can now be refolded to within 2 Å in global rmsd by dDFIRE (See Table I). Further studies are needed to understand why these two segments are more sensitive than others to the difference between the native and standard values of bond lengths and angles.

It is important to learn which orientation-dependent interaction is responsible for the success of the dDFIRE

strand that is coil-like (2ptl, as shown, 1fltx, 1csp) or even a helix (2extb, as shown), while dDFIRE produces strands that have a more normal structural pattern. Thus, there is a marked difference in the quality of the secondary-structure segments refolded by the two energy functions as indicated by the local rmsd values (Fig. 2 and Table I).
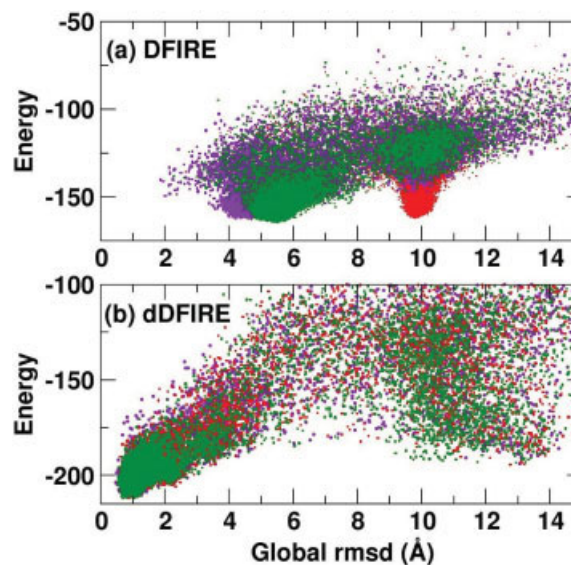
Figure 4 displays the energies of all conformations sampled for protein 1r69 as a function of their rmsd values from the native structure. Three independent global minimizations with the DFIRE energy function produced very different structures (about 5 and 10 Å in global rmsd, separately). In contrast, essentially the same structure is obtained in three minimizations with the dDFIRE energy function. This confirms that the orientation-dependent energy function produces more specific structures.

Only 3 out of 16 segments refolded by dDFIRE have a global rmsd value >2 Å. One of them is 2extb. As seen in Figure 3, the dDFIRE correctly produces a C-terminal β-hairpin, but the last β-strand interacts with a fixed strand in the protein rather than the other strand in the terminal β-hairpin. This "misfolded" β-hairpin obviously has a much stronger interaction with the rest of the protein than the native β-hairpin does, suggesting that the native C-terminal β-hairpin is stable only in a multimeric conformation. Indeed, the PDB 2ext file contains an engineered 12-member ring from the cyclic 11-mer TRAP.[31] If the fixed structure includes the native struc-

Figure 4

*The energies of the conformations are sampled for protein 1r69 by the genetic algorithm coupled with either the DFIRE (**a**) or the dDFIRE energy function (**b**). They are plotted as a function of their global rmsd values from the native 1r69 structure. The results of three independent global minimizations are shown in three different colors. The DFIRE energy samples different minimums across three independent minimizations (one around 10 Å and the other around 5 Å). The dDFIRE energy samples essentially have the same global minimum state, around 0.8 Å in global rmsd value, in all three minimizations.*

energy function in segment refolding. Three orientation components of the dDFIRE energy function are employed to refold five terminal regions (two single-helix segments, one two-helix bundle, one strand, and one β-hairpin of five separate proteins). The three dDFIRE components are the orientation dependence involving hydrogen-bonded polar atoms, polar–nonpolar atoms, and polar atoms only (Note that the last one includes hydrogen-bonded atoms). The results are shown in Table II. The three individual orientation components can restore single helix in 2guzb and 1i2ta as accurately as the full dDFIRE energy function. However, they produced slightly less accurate structures (1.5–1.7 Å in global rmsd) than the dDFIRE (0.8 Å) for the terminal two-helix bundle in 1o82a. While every single orientation component can refold helix-containing segments with reasonable accuracy, they cannot restore the structures of strand-containing segments well. The orientation components between hydrogen-bonded atoms and between polar and nonpolar atoms failed to fold the C-terminal β-strand of 1fltx within 2 Å global rmsd. Additionally, none of the three individual components can refold the C-terminal β-hairpin of 2extb (in a dimeric form) to within 2 Å in global rmsd. Thus, orientation-dependent interactions between polar and nonpolar atoms and between hydrogen-bonded and nonhydrogen-bonded polar atoms are all essential for restoring secondary-structure containing segments, most particularly for β-strands.

The dDFIRE energy function takes into account the hydrogen bonding interaction. As one example, the distance dependence of the interactions between atom $O_{\varepsilon 1}$ of Glu and atom $N_{\eta 1}$ of Arg is shown in Figure 5 for three representative bins for the angle component θ_{pq}. These results indicate that the orientation dependence is very strong at short distance and disappears at large distance ($r > 8$ Å) as expected. A strong orientation dependence at short distance confirms the need for an orientation-dependent energy function. Moreover, a θ_{pq} of around 180° (the angle between $\vec{r}_{C_\delta O_{\varepsilon 1}}$ and $\vec{r}_{C_\zeta N_{\eta 1}}$) is the most favorable angle as expected from the hydrogen bonding requirement.

More importantly, we are interested in what causes the orientation dependence for the interactions between non-hydrogen-bonded polar atoms and between a polar atom and a nonpolar atom. Figure 6 shows two examples. An orientation-dependent component in three angle bins between two nonhydrogen-bonded polar atoms ($O_{\varepsilon 1}$ of Glu and $O_{\delta 1}$ of ASP) is shown as a function of distance in Figure 6(a). The optimal angle at a short atomic distance (indicated by a low energy value) is 90°–180° for the angle between $\vec{r}_{C_\delta O_{\varepsilon 1}}^{\text{Glu}}$ and $\vec{r}_{C_\gamma O_{\delta 1}}^{\text{Asp}}$. This means that the two oxygen atoms are facing each other with carbon atoms hidden behind [see Fig. 6(a)]. This orientation happens frequently because the two oxygen atoms are linked by a hydrogen atom [e.g., from residue Arg in Fig.
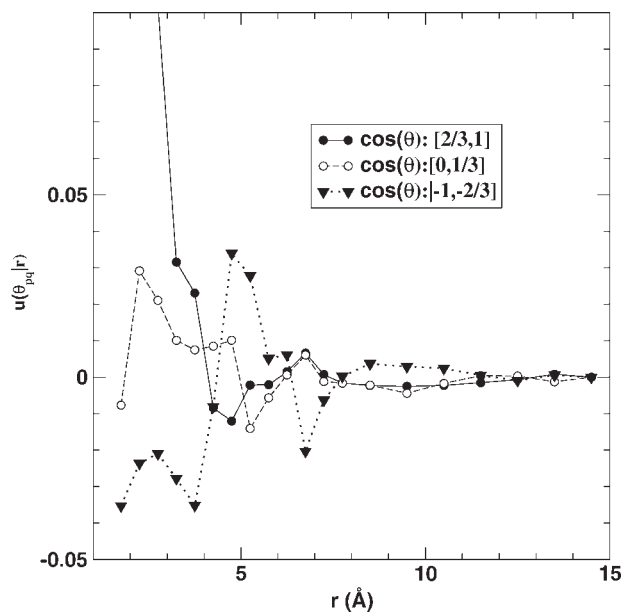

Figure 5

One angular component of the dDFIRE energy function between $O_{\varepsilon 1}$ of Glu and $N_{\eta 1}$ of Arg. The distance dependence of three angle bins based on $\cos(\theta)$ is shown as labeled.

6(a)]. The three-body effect provides a necessary orientation-dependent correction to the repulsive DFIRE energy function around 4 Å [shown as a black solid line with open circles in Fig. 6(a)].

Figure 6(b), on the other hand, shows an optimal angle of 90° at an atomic distance of 4 Å for the angle between $\vec{r}_{C_\zeta N_{\eta 1}}^{\text{Arg}}$ and $\vec{r}_{N_{\eta 1}^{\text{Arg}}, C_{\delta 1}^{\text{Leu}}}$. This orientation preference between polar atom $N_{\eta 1}$ of Arg and nonpolar atom $C_{\delta 1}$ of Leu at a short atomic distance results from the tendency of the two terminal N atoms of the Arg side-chain to be on top of the two terminal C atoms of the Leu side-chain for an optimal van der Waals interaction, as illustrated in Figure 6(b). Such an orientation preference reduces the overall repulsion given by DFIRE in certain orientations between the hydrophobic and hydrophilic atoms. Thus, the orientation-dependent dDFIRE function provides a fine tuning of the DFIRE energy function by incorporating the orientation dependence resulting from all possible interactions, including polar interactions, geometric compatibility, and multibody effects.

DISCUSSION

We demonstrated that the dDFIRE energy function significantly improves over DFIRE in segment refolding. The latter was shown[5] to refold partially unfolded segments with secondary structures more accurately than the RAPDF energy function, another all-atom knowl-
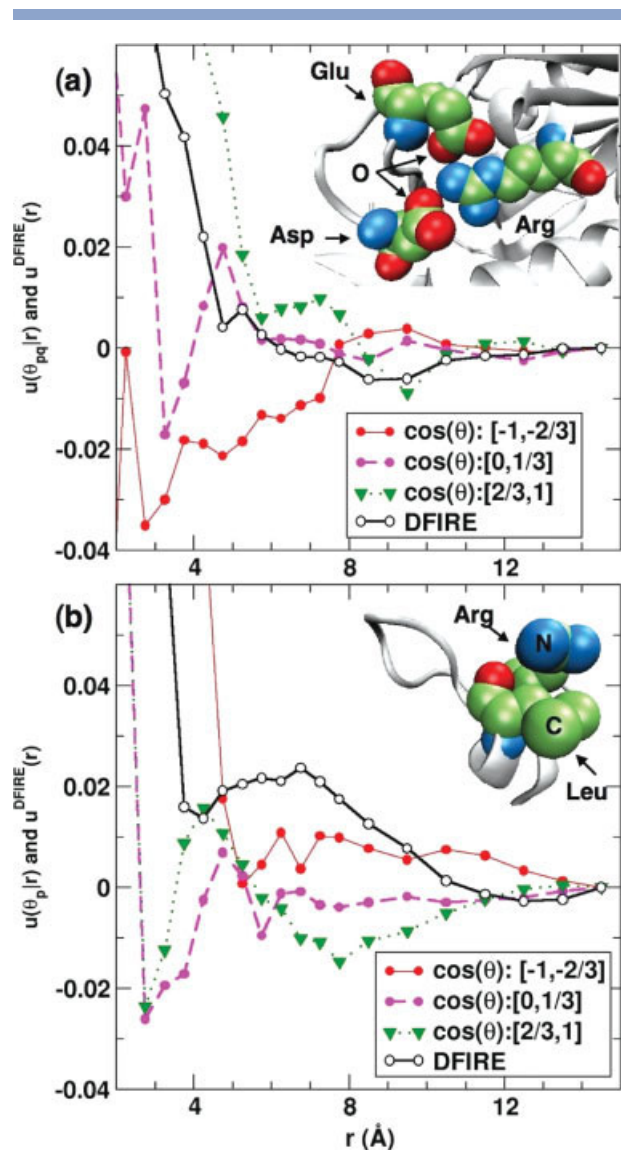
Figure 6

(*a*) *The distance-dependence of the angular component of the dDFIRE energy function between $O_{\varepsilon 1}$ of Glu and $O_{\delta 1}$ of ASP for the angle between $\vec{r}^{Glu}_{C_{\delta}O_1}$ and $\vec{r}^{Asp}_{C_{\gamma}O_{\delta 1}}$ (θ_{pq}). Three angle bins (in color) based on the range of cos(θ) are shown as labeled along with the orientation-independent DFIRE energy function (in black). A θ_{pq} of around 180° (curve in red) is the most favorable angle at 4 Å. This angle corresponds to the two oxygen atoms facing each other and stabilized by a hydrogen atom (from Arg in illustration). (*b*) The angular component of the dDFIRE energy function between polar atom $N_{\eta 1}$ of Arg and nonpolar atom $C_{\delta 1}$ of Leu for the angle between $\vec{r}^{Arg}_{C_{\zeta}N_{\eta 1}}$ and $\vec{r}_{N^{Arg}_{\eta 1},C^{Leu}_{\delta 1}}$ (θ_p). A θ_p at 90° (in violet) is the most favorable angle at 4 Å. This reflects the preference of the two terminal N atoms of Arg side-chain to be on the top of the two terminal C atoms of Leu side-chain for an optimal van der Waals interaction as demonstrated.*

edge-based energy function,[32] and molecular-mechanics-based refinement by energy minimization (the all-atom OPLS force field[33] and the generalized-Born solvent model[34]) and molecular dynamics simulations (the GROMOS96 43a1 force field and explicit SPC water

model[35]). We also implemented the Rosetta all-atom energy function[36] in our genetic algorithm. The Rosetta all-atom energy function (Score # 12 for protein design and refinement) was obtained from the software code of Rosetta ab initio (version 2.1.1) from the David Baker's group (http://www.bakerlab.org/). We found that the energy function is unable to fold four terminal segments of 1fltx, 1opd0, 2guzb, and 2hsla. All the four segments remain solvent exposed after several hundreds of generations. The final structures of the four segments also contain helical structure including the β-terminal region of 1fltx. This is perhaps because that the Rosetta all-atom energy function was optimized for refining (or relaxing) over-packed coarse-grained models with prebuilt native segments from different proteins. The result further confirms that native-like fragment structures are difficult to produce by contemporary energy functions.[15]

The results reported here underline the importance of orientation-dependent interactions, in addition to the well-studied hydrogen-bonding interaction, for the successful restoration of specific structural segments of proteins. The absence of orientation dependence leads to short helices or coils rather than secondary-structure elements. These results confirm the importance of orientation preference between nonhydrogen-bonded atoms in the formation of secondary structures of proteins[2,3] (for a recent review, see Ref. 37). Additionally, the results call for the attention to the relative orientation between polar and nonpolar atoms. So far, orientation-dependent interactions other than hydrogen bonding have been ignored in constructing all-atom knowledge-based or empirical energy functions.[38,39] This explains why contemporary energy functions are difficult to produce native-like fragment structures.[15] Thus, this work has significant implications for developing more specific energy function for folding and molecular recognition. For example, it is straightforward to extend the dDFIRE energy function to protein–ligand and protein–DNA interactions, considering DFIRE's reasonable success in binding affinity prediction.[40]

The dDFIRE energy function repeatedly folds the majority of unfolded terminal regions within 2 Å global rmsd from their respective native structures. This suggests that the dDFIRE energy function is highly specific, at least for folding short segments. Using short segments was an attempt to separate the effect of a sampling technique from that of an energy function. These two effects are difficult to separate in some cases. This is not the case in this article, because an identical sampling technique produced dramatically diffent results for different fitness functions. It remains to be seen if this success is reproducible for larger segments or even entire proteins because the conformational sampling will be significantly more challenging, as the size and complexity of protein structures increase.[16] The current sampling technique sometimes yields different structures with different ran-

dom initial conformations [Fig. 4(a)]. This suggests that the method is not yet a true global minimization technique and can lead to a local minimum. A more efficient sampling such as fragment assembly[16,17] may be necessary for using dDFIRE to fold larger segments or proteins. Recently, Sancho and Rey[41] successfully folded several helix-bundle proteins up to five helices and 155 residues with a rigid-fragment assembly guided by a coarse-grained DFIRE energy function.[42] Fragment assembly coupled with the dDFIRE energy function is in progress.

In this study, we have focused on a specific test of the dDFIRE energy function in segment refolding. More "traditional" tests including decoy discriminations, near native selections, loop predictions, mutation-induced changes in stablility, docking decoy selections, and ab initio protein folding are the subject of our on-going research.

ACKNOWLEDGMENTS

REFERENCES

1. Haber E, Anfinsen CB. Regeneration of enzyme activity by air oxidation of reduced subtilisin-modified ribonuclease. J Biol Chem 1961;236:422–424.
2. Maccallum PH, Poet R, Milner-White EJ. Coulombic interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of α-helices and anti-parallel β-sheet. A new method for analysing the forces between hydrogen bonding groups in proteins includes all the coulombic interactions. J Mol Biol 1995;248:361–373.
3. Deane CM, Allen FH, Taylor R, Blundell TL. Carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. Protein Eng 1999;12:1025–1028.
4. Blokzijl W, Engberts JBFN. Hydrophobic effects—opinions and facts. Angew Chem Int Ed Engl 1993;32:1545–1579.
5. Zhu J, Xie L, Honig B. Structural refinement of protein segments containing secondary structure elements: local sampling, knowledge-based potentials and clustering. Proteins 2006;65:463–479.
6. 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 2002;11:2714–2726.
7. 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 2003;326:1239–1259.
8. Scherga HA, Liwo A, Oldziej S, Czaplewski C, Pillardy J, Ripoll DR, Vila JA, Kazmierkiewicz R, Saunders JA, Arnautova YA, Jagielska A, Chinchio M, Nanias M. The protein folding problem: global optimization of the force fields. Front Biosci 2004;9:3296–3323.
9. Grzybowski B, Ishchenko A, DeWitte R, Whitesides G, Shakhnovich E. Development of a knowledge-based potential for crystals of small organic molecules: Calculation of energy surfaces for C=O···H—N hydrogen bonds. J Phys Chem B 2000;104:7293–7298.
10. Miyazawa S, Jernigan RL. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? J Chem Phys 2005;122:024901.
11. Hoppe C, Schomburg D. Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. Protein Sci 2005;14:2682–2692.
12. Buchete N-V, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improve native fold recognition. Protein Sci 2004;13:862–874.
13. Pandey BP, Zhang C, Yuan X, Zi J, Zhou Y. Protein flexibility prediction by an all-atom mean-field statistical theory. Protein Sci 2005;14:1772–1777.
14. Jacob E, Unger R. A tale of two tails: why are terminal residues of proteins exposed? Bioinformatics 2007;2:E225–E230.
15. Bujnicki JM. Protein-structure prediction by recombination of fragments. ChemBioChem 2006;7:19–27.
16. Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
17. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
18. Llinas M, Marqusee S. Subdomain interactions as a determinant in the folding and stability of T4 lysozyme. Protein Sci 1998;7:96–104.
19. Atwell S, Wells JA. Selection for improved subtiligases by phage display. Proc Natl Acad Sci USA 1999;96:9497–9502.
20. Feng H, Vu ND, Bai Y. Detection of a hidden folding intermediate of the third domain of PDZ. J Mol Biol 2005;346:345–353.
21. Zhou Y, Zhou H, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys 2006;46:165–174.
22. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
23. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
24. Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular systems. J Chem Phys 1996;105:1902–1921.
25. Jackson JD. Classical electrodynamics. 2nd ed. New York: Wiley; 1975.
26. Fabiola F, Bertram R, Korostelev A, Chapman MS. An improved hydrogen bond potential: impact on medium resolution protein structures. Protein Sci 2002;11:1415–1423.
27. Yang Y, Liu H. Genetic algorithms for protein conformation sampling and optimization in a discrete backbone dihedral angle space. J Comput Chem 2006;27:1593–1602.
28. Weiner SJ, Kollman P, Nguyen D, Case D. An all atom force field for simulations of proteins and nucleic acids. J Comput Chem 1986; 7:230–252.
29. Dunbrack RL, Jr, Karplus M. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. J Mol Biol 1993; 230:543–574.
30. Goldberg DE, Smith RE. Nonstationary function optimization using genetic algorithm with dominance and diploidy. In Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application (Cambridge, MA). Grefenstette JJ, editor. New Jersey: Lawrence Erlbaum Associates; 1987. pp 59–68.
31. Heddle J, Yokoyama T, Yamashita I, Park S, Tame J. Rounding up: engineering 12-membered rings from the cyclic 11-mer TRAP. Structure 2006;14:925–933.
32. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
33. Jorgensen WL, JulianTirado-Rives DSM. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
34. Zhu J, Alexov E, Honig B. Comparative study of generalized Born models: Born radii and peptide folding. J Phys Chem B 2005;109: 3008–3022.

35. vanGunsteren WFB, Eising AA, Hunenberger PH, Kruger P, Mark AE, Scott WRP, Tironi IG. Groningen molecular simulation (GROMOS) system. Groningen, The Netherlands: University of Groningen 1996.

36. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322:65–78.

37. Paulini R, Muller K, Diederich F. Orthogonal multipolar interactions in structural chemistry and biology. Angew Chem Int Ed Engl 2005;44:1788–1805.

38. Skolnick J. In quest of an empirical potential for protein structure prediction. Curr Opin Struc Biol 2006;16:166–171.

39. Gohlke H, Klebe G. Statistical potentials and scoring functions applied to protein-ligand binding. Curr Opin Struct Biol 2001;11: 231–235.

40. Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein and protein-DNA complexes. J Med Chem 2005;48:2325–2335.

41. deSancho D, Rey A. Assessment of protein folding potentials with an evolutionary method. J Chem Phys 2006;125:014904.

42. 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 2004;13:400–411.