Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions

YUEDONG YANG[1,2] AND YAOQI ZHOU[1,2]

[1]Indiana University School of Informatics, Indiana University–Purdue University, Indianapolis, Indiana 46202, USA
[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

(RECEIVED November 8, 2007; FINAL REVISION March 22, 2008; ACCEPTED April 2, 2008)

Abstract

One of the common methods for assessing energy functions of proteins is selection of native or near-native structures from decoys. This is an efficient but indirect test of the energy functions because decoy structures are typically generated either by sampling procedures or by a separate energy function. As a result, these decoys may not contain the global minimum structure that reflects the true folding accuracy of the energy functions. This paper proposes to assess energy functions by ab initio refolding of fully unfolded terminal segments with secondary structures while keeping the rest of the proteins fixed in their native conformations. Global energy minimization of these short unfolded segments, a challenging yet tractable problem, is a direct test of the energy functions. As an illustrative example, refolding terminal segments is employed to assess two closely related all-atom statistical energy functions, DFIRE (distance-scaled, finite, ideal-gas reference state) and DOPE (discrete optimized protein energy). We found that a simple sequence-position dependence contained in the DOPE energy function leads to an intrinsic bias toward the formation of helical structures. Meanwhile, a finer statistical treatment of short-range interactions yields a significant improvement in the accuracy of segment refolding by DFIRE. The updated DFIRE energy function yields success rates of 100% and 67%, respectively, for its ability to sample and fold fully unfolded terminal segments of 15 proteins to within 3.5 Å global root-mean-squared distance from the corresponding native structures. The updated DFIRE energy function is available as DFIRE 2.0 upon request.

Keywords: new methods; protein structure prediction; statistical energy function

Energy functions of proteins are developed to quantitatively capture the physical interaction that determines how proteins fold and interact with other biologically active molecules. Existing energy functions of proteins are obtained through a physical-based approach (Brooks et al. 1983; Weiner et al. 1986; Ponder and Richards 1987; Jorgensen et al. 1996; Scott et al. 1999), a statistical (knowledge-based) approach (Tanaka and Scheraga 1976), or their combination (empirical approach). A knowledge-based or statistical energy function is obtained directly from statistical analysis of known protein structures (Tanaka and Scheraga 1976).

Different statistical energy functions differ only in how their reference states are defined. We have introduced a physical reference state of uniformly distributed ideal gas points in finite protein-size spheres and assumed that the number of pairs of ideal gas points is proportional to the constant fractional power of the distance between

Reprint requests to: Yaoqi Zhou, Indiana University School of Informatics, Indiana University–Purdue University Indianapolis, and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Avenue, Walker Plaza Building Suite 319, Indianapolis, IN 46202, USA; e-mail: yqzhou@iupui.edu; fax: (317) 278-9201.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.033480.107.

two ideal gas points (r^{α}) with $\alpha < 2$ to account for the finite-size effect. This reference state together with a distance-scaling approximation yields the DFIRE (distance-scaled, finite, ideal-gas reference state) statistical energy function (Zhou and Zhou 2002). Recently, Shen and Sali introduced the theory of the discrete optimized protein energy (DOPE) based on the same finite ideal-gas reference state but with an analytical, protein-size dependent α parameter and a different scaling scheme (Shen and Sali 2006). A large-scale comparison among DFIRE, DOPE, and more than 20 other scoring functions (Eramian et al. 2006) suggests that DFIRE and DOPE perform superiorly over other scoring functions in near-native selections, whereas the difference between the two is small.

Decoy selection is only an indirect test of an energy function because decoys are often generated by a separate energy function or by sampling only. Here, we propose a more direct test of energy functions through ab initio refolding of completely unfolded segments with secondary structures while the rest of the protein remains folded. We hypothesize that a more direct test could detect the hidden difference between two closely related energy functions. In addition to assessing energy functions, terminal-segment refolding itself is biologically relevant. The folding of many proteins is assisted by a prefolded domain (pro-domain) (Llinas and Marqusee 1998; Atwell and Wells 1999).

Refolding the terminal segment while fixing the rest of a protein, however, is challenging because terminal regions are more flexible and often exposed (Jacob and Unger 2007) and, thus, searching backbone and side chain conformations at the same time is nontrivial. In fact, Zhu et al. (2006) showed that several physical-based and knowledge-based energy functions have difficulty folding even partially unfolded segments with secondary structures, and the best performance is given by the DFIRE energy function. Here, in order to separate testing energy functions from testing sampling techniques, we limit the maximum size of unfolded regions to less than 15 residues for a strand-containing segment, and less than 25 residues for a helix-containing segment.

Several variants of the DFIRE energy function are tested in this paper with a genetic algorithm for global energy minimization. We first update the DFIRE energy function with 30 distance bins that allow finer statistics to extract short-range repulsive and long-range interactions (uDFIRE). We then obtain an asymmetric DFIRE (aDFIRE) energy function that depends on the relative sequence positions of the interacting residues, an approach adopted in the DOPE energy function. Segment refolding by these energy functions suggests that both aDFIRE and DOPE have stronger local interactions and intrinsically bias toward the formation of longer helical segments than uDFIRE does, while uDFIRE appears to be more balanced in reproducing helical and strand segments. The most significant improvement, resulting from the change of the binning method from DFIRE to uDFIRE, highlights the importance of fine treatment of repulsive interactions for shape complementarity in segment refolding.

Results

Table 1 shows the refolding results of four single helices, two two-helix bundles, seven single strands, one mixed helix/strand segment, and one β hairpin that were unfolded and refolded in the presence of a fixed folded region. The accuracy of refolded segments is described by a local root-mean-squared distance (lrmsd) that is calculated by superposing the unfolded and native segments or by a global rmsd (grmsd) between the refolded and the native segment conformations that are calculated after the superposition of the fixed regions. The values of lrmsd and grmsd are calculated from the positions of C_{α} atoms. The former measures the accuracy of the segment structure only, while the latter describes both the structure and its orientation relative to the rest of the protein. We report the results of four energy functions: DFIRE based on the original 20 bins (2 Å for the first distance bin, 0.5 Å per bin up to 8 Å, and 1 Å per bin from 8 Å to 15 Å), updated DFIRE (uDFIRE) based on the 30 equal-distance bins (0.5 Å per bin for the same distance range), asymmetric DFIRE (aDFIRE), where the interaction between two atoms depends on the difference (positive or negative) in the two sequence positions of their respective residues, and DOPE, which is also an asymmetric energy function. Three independent global energy minimizations by a genetic algorithm (see Materials and Methods) are performed for each segment in the presence of the fixed, folded region, and only one global minimum structure from each minimization is employed to produce the results in Table 1.

It is clear from Table 1 that the uDFIRE energy makes a significant improvement over the DFIRE energy function with a smaller number of bins. Large reductions in rmsd values of refolded structures are observed for 1r690, 1o82a, and 2ayda, and smaller improvements are achieved for 1u84a, 1opd0, and 1csp0. Only three segments (1i2ta, 1vcc0, and 2extb) are refolded with significantly larger grmsd values by uDFIRE than by DFIRE (difference >0.5 Å). All three are misfolded by uDFIRE and DFIRE (grmsd > 6 Å). That is, the increases of grmsd values due to the use of uDFIRE for these segments are not that important. The overall success rates for refolding to within a 3.5 Å grmsd from the native conformation are 47% by DFIRE and 67% by uDFIRE, respectively, while aDFIRE and DOPE have the same success rate as uDFIRE.

While the overall success rates given by uDFIRE, aDFIRE, or DOPE are the same, there are some intrinsic

Table 1. *Restoration of unfolded terminal regions by a genetic algorithm with the original DFIRE, updated DFIRE (uDFIRE) with 30 bins, asymmetric DFIRE (aDFIRE) with a simple sequence dependence, and the DOPE energy function*

PDB Id#[c]	Structure type[d]	Local rmsd (Å)[a]					Global rmsd (Å)[b]				
		Initial[e]	DFIRE[f]	uDFIRE[f]	aDFIRE[f]	DOPE[f]	Initial[e]	DFIRE[f]	uDFIRE[f]	aDFIRE[f]	DOPE[f]
2guzb	1α/5α	7.0 ± 1.4	7.3 ± 1.1	6.6 ± 0.5	2.6 ± 2.8	0.7 ± 0.2	25 ± 9	9.7 ± 1.5	9.4 ± 0.7	4.0 ± 4.0	1.4 ± 0.5
1i2ta	1α/4α	6.2 ± 1.2	6.4 ± 0.8	6.3 ± 1.0	0.50 ± 0.03	0.48 ± 0.03	20 ± 7	9.1 ± 0.6	9.8 ± 0.6	0.96 ± 0.1	0.71 ± 0.02
1u84a	1α/4α	6.1 ± 1.0	1.1 ± 0.6	0.55 ± 0.07	0.29 ± 0.02	0.27 ± 0.03	21 ± 6	1.7 ± 0.5	1.1 ± 0.05	0.65 ± 0.08	0.51 ± 0.06
1r690	2α/5α	6.2 ± 1.3	3.2 ± 0.6	0.66 ± 0.02	0.58 ± 0.02	0.43 ± 0.02	16 ± 4	6.5 ± 2.4	1.0 ± 0.1	0.82 ± 0.05	0.65 ± 0.08
1o82a	2α/6α	6.1 ± 1.0	3.9 ± 0.6	1.55 ± 0.08	2.4 ± 1.0	2.4 ± 1.4	20 ± 5	5.2 ± 1.0	2.0 ± 0.1	3.2 ± 1.6	3.1 ± 1.7
1opd0	1α/3α4β	5.4 ± 0.8	1.5 ± 0.4	0.94 ± 0.02	0.9 ± 0.1	1.06 ± 0.04	19 ± 6	2.0 ± 0.4	1.39 ± 0.06	1.3 ± 0.1	1.5 ± 0.1
2igd0	1β/1α4β	4.0 ± 1.2	0.67 ± 0.01	0.53 ± 0.05	0.49 ± 0.04	0.55 ± 0.02	18 ± 5	0.75 ± 0.03	0.62 ± 0.07	0.59 ± 0.04	0.73 ± 0.04
1vcc0	1β/2α5β	4.2 ± 1.0	3.3 ± 1.1	4.27 ± 0.08	3.8 ± 1.4	5.0 ± 0.3	17 ± 4	6.1 ± 2.1	8.0 ± 1.3	8.8 ± 5.5	6.8 ± 0.2
2hsla	1α1β/1α9β	4.1 ± 0.8	1.6 ± 0.4	2.13 ± 0.04	0.61 ± 0.03	0.61 ± 0.03	17 ± 6	3.24 ± 0.03	2.92 ± 0.07	1.51 ± 0.08	1.65 ± 0.09
2cc6a	1β/1α3β	4.9 ± 1.7	0.44 ± 0.06	0.6 ± 0.1	0.44 ± 0.03	0.5 ± 0.1	18 ± 6	0.69 ± 0.1	0.74 ± 0.08	0.62 ± 0.04	0.7 ± 0.1
2ptl0	1β/1α4β	4.6 ± 1.5	2.3 ± 0.1	2.27 ± 0.06	2.16 ± 0.05	2.5 ± 0.2	22 ± 5	2.72 ± 0.02	3.03 ± 0.08	2.80 ± 0.09	3.1 ± 0.3
1csp0	1β/5β	4.7 ± 1.5	2.2 ± 0.7	0.80 ± 0.05	2.3 ± 0.5	3.3 ± 0.9	19 ± 5	2.4 ± 0.7	1.14 ± 0.08	3.2 ± 0.5	5.0 ± 2.0
1fltx	1β/8β	6.7 ± 1.9	4.71 ± 0.07	4.48 ± 0.01	4.4 ± 0.1	7.2 ± 0.1	22 ± 5	5.9 ± 0.2	6.14 ± 0.07	6.1 ± 0.2	15.1 ± 0.1
2ayda	1β/5β	4.7 ± 0.8	4.32 ± 0.02	2.5 ± 0.2	3.6 ± 0.4	4.8 ± 0.2	18 ± 5	7.82 ± 0.01	2.9 ± 0.1	5.9 ± 0.8	16 ± 2
2extb	2β/6β	5.8 ± 1.5	4.3 ± 1.8	5.5 ± 1.1	4.7 ± 0.3	5.2 ± 0.9	15 ± 5	6.7 ± 2.5	8.4 ± 0.2	8.8 ± 0.1	8.6 ± 0.1
Success rate (<3.5 Å)[g]			9/15 (60%)	10/15 (67%)	11/15 (73%)	11/15 (73%)		7/15 (47%)	10/15 (67%)	10/15 (67%)	10/15 (67%)

[a] lrmsd is obtained by minimizing root-mean-squared distance through rotation of the refolded segment conformation with respect to the native segment conformation.
[b] grmsd between the refolded and native segments is calculated after the superposition of the fixed native regions in the two structures.
[c] Protein Data Bank identification number. The fourth digit is the chain ID.
[d] Unfolded-segment and whole-protein structural types represented by the number of α helices and β strands in the protein structure, respectively.
[e] Mean and standard deviation of lrmsd or grmsd values of the initial 120 structures.
[f] Mean and standard deviation of the lrmsd or grmsd values of the three global minimum structures from three independent global minimizations with the DFIRE, uDFIRE, aDFIRE, or DOPE energy function, respectively.
[g] Total number (percentage) of refolded segments whose lrmsd or grmsd values are <3.5 Å.

differences between the three energy functions. For example, the uDFIRE energy function fails to fold the terminal helix of 2guzb and 1i2ta (grmsd > 9 Å) while DOPE misfolds the terminal strand in 1csp0 (grmsd = 5 Å) and 2ayda (grmsd = 16 Å). A close examination of 2guzb and 1i2ta indicates that their C-terminal helices are long and partially exposed (1i2ta is shown in Fig. 1A). uDFIRE does not make an accurate prediction of the C-terminal segments of the two proteins because it attempts to tightly pack the terminal segment with the rest of the protein. Both aDFIRE and DOPE must have stronger local interactions than uDFIRE so that the partially exposed helix can be stabilized in the absence of strong interaction with the rest of proteins. We further discovered that slightly more accurate refolding of mixed α-helical and β-strand terminal regions of 2hsla by aDFIRE and DOPE than by uDFIRE is due to a more accurate prediction of the helical portion of the refolded segment. However, bias toward a long helical structure has a negative impact on folding of some other segments. For example, DOPE predicts helical segments for the terminal strands of 1csp0, 1fltx, and 2ayda (1csp0 is shown in Fig. 1C). Similar behavior is observed for the terminal segments in 1csp0 and 2ayda predicted by aDFIRE. Stronger local interactions are likely the source for the misfolded single helix by DOPE and aDFIRE rather than native two helices in the two-helix terminal segment of 1o82a (Fig. 1B).

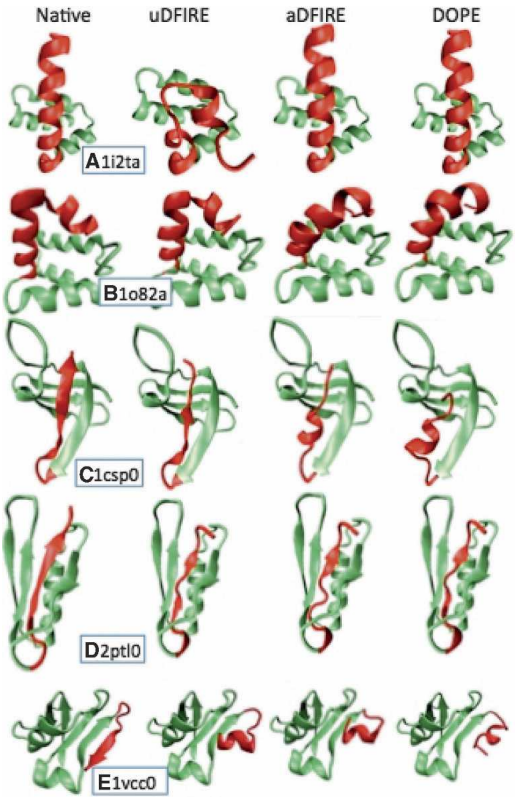

Figure 1. The segment structures (red) refolded by updated DFIRE (uDFIRE, *center left*), asymmetric DFIRE (aDFIRE, *center right*), DOPE (*right*) for five proteins—(*A*) 1i2ta, (*B*) 1o82a, (*C*) 1csp0, (*D*) 2ptl0, and (*E*) 1vcc0—are compared with their respective native conformations (*left*). The fixed portion of each protein is colored in light green. For segments whose three independent runs yield quite different structures, only the lowest energy structures are shown. In the first three cases, aDFIRE and DOPE have a stronger ability to form helices correctly (1i2ta) or incorrectly (1o82a and 1csp0) than uDFIRE.

While the difference between the three energy functions (uDFIRE, aDFIRE, and DOPE) is obvious for the seven above-mentioned segments, their similarity prevails for the eight remaining segments. All make high-resolution predictions of the terminal segments of 1u84a (1α), 1r690 (2α), 1opd0 (1α), 2igd0 (1β), and 2cc6a (1β) to within 1.5 Å grmsd. They also refold the terminal strand of 2ptl0 at 3 Å grmsd. An example (2ptl0) is shown in Figure 1D. All three predict a slightly curved strand, likely due to the lack of orientation-dependent polar interactions. The successful folding of these segments suggests that shape complementary is sufficient for accurate refolding of some segments in the presence of a prefolded region. All three energy functions, however, misfold the strands in 1vcc0 and 2extb. An example (1vcc0) is shown in Figure 1E. All three energy functions yield a helical structure because there is space available for helical formation and a helix has a better interaction with the rest of protein. Thus, a strand conformation, if not confined by shape complementary, requires an orientation-dependent interstrand interaction for its stabilization. This interaction is absent in all three energy functions.

The analysis above is limited to the average grmsd values of the global minimum structures. However, in some cases, very different global minimum structures are obtained from independent runs. This is reflected from several large standard deviations in grmsd values of the three global minimum structures from three independent runs (Table 1). Interestingly, only one standard deviation of three independent runs is >1 Å for uDFIRE. The corresponding number is three for either DOPE or aDFIRE. More tests are required to be certain whether or not uDFIRE is more specific than DOPE or aDFIRE in folding.

Different structures from independent runs indicate the limitation of the global minimization technique employed in this study. This raises the question of whether misfolding some segments is caused by the failure of sampling. Figure 2 shows two segment conformations (1o82a and 1vcc0, respectively) sampled by DOPE, aDFIRE, and uDFIRE. In the first case, uDFIRE is more specific (producing single stable structure). In the second case (1vcc0), different energy functions have different abilities to sample near-native structures. uDFIRE and aDFIRE can sample near-native conformations of <2 Å grmsd while DOPE cannot, although all three make wrong predictions. Thus, near-native structures can be successfully sampled even when their energy values are not competitive.

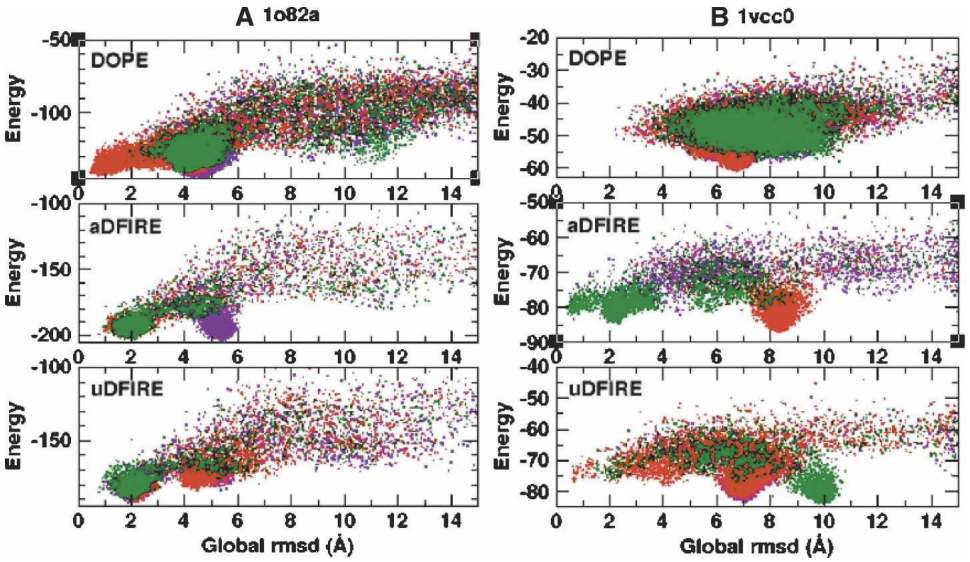

Figure 2. The energies of the segment conformations are sampled for proteins 1o82a (*A*) and 1vcc0 (*B*) by the genetic algorithm coupled with either DOPE (*top*), aDFIRE (*middle*), or uDFIRE (*bottom*). They are plotted as a function of their global rmsd values from their respective native structures. The results of three independent global minimizations are shown in three different colors.

To further examine the ability of different energy functions to sample near-native conformations, Table 2 shows the lowest global rmsd values (average over three independent runs) sampled by DFIRE, uDFIRE, aDFIRE, and DOPE. Although the success rate of folding to within 3.5 Å grmsd is only 67%, the success rate of reaching (sampling) near-native structures (within 3.5 Å grmsd) is significantly higher. uDFIRE has the highest success rate of 100%, followed by aDFIRE and DFIRE (93%) and DOPE (87%). However, the actual difference among the four is small (two segments). Thus, more studies are needed to be certain if the uDFIRE energy function is best suitable for sampling both helical- and strand-containing fragments.

Discussion

In this paper, we demonstrate that terminal-segment refolding allows the revelation of the fine difference between the energy functions that are otherwise difficult to detect by commonly used decoy selections. We find that DOPE and aDFIRE, because of their dependence of relative sequence positions, favor helices over sheets (2ayda, 1fltx, and 1csp0) and longer helices over short ones (2guzb, 1i2ta, and 1o82a). This hidden bias produces more accurate structures for some segments (2guzb, 1i2ta, 2hsla) while it misfolds others (2ayda, 1csp0, 1o82a). Stronger local interactions for helical formation likely result from the simple two-state dependence on the sequence positions of two residues I and J ($I > J$ or $J < I$), which overestimates the effect of the dependence on sequence positions. It overestimates because the dependence on sequence positions should be limited to neighboring residues ($|I - J|$ within a small number), rather than for all residues. More sophisticated sequence-position dependence has been proposed (Melo et al. 2002; Zhang and Skolnick 2004; Bradley et al. 2005; Ferrada and Melo 2007). We are currently testing various approaches to account for covalent bonding (Melo et al. 2002; Cheng et al. 2007; Ferrada and Melo 2007).

The effect caused by the protein-size-dependent α (the difference between DOPE and aDFIRE) is small. This was observed previously in decoy selections (Eramian et al. 2006; Shen and Sali 2006). For example, the relative occurrence of the most accurate 20% models among the 20% best scoring models compared with that for the entire decoy set is 3.85 for DFIRE and 3.92 for DOPE (Shen and Sali 2006) in the moulder decoy set of 20 proteins. For the same decoy set, the success rate of selecting best near-native structures is 25.4% by DFIRE and 24.7% by DOPE (Eramian et al. 2006). We compare the energy parameters of DOPE and aDFIRE by calculating the correlation coefficients for the two sets of parameters for all atomic pairs at each distance bin. The correlation coefficients are all >0.9 from the bin of 0.5–1 Å to the bin of 9.5–10 Å. For example, at the distance bin of 4–4.5 Å, the correlation coefficient between the two sets of data is 0.9888 with a slope of 0.996. The difference between the two sets of parameters is even smaller for attractive interactions (negative values) and larger for repulsive interactions. Repulsive interactions typically involve fewer occurrences, and, thus, the larger difference

Table 2. *The lowest global rmsd (grmsd) values for the structures sampled by the original DFIRE, updated DFIRE (uDFIRE) with 30 bins, asymmetric DFIRE (aDFIRE) with a simple sequence dependence, and the DOPE energy function*

	Lowest grmsd (Å)			
PDB Id#[a]	DFIRE[b]	uDFIRE[b]	aDFIRE[b]	DOPE[b]
2guzb	3.5 ± 0.9	1.9 ± 0.5	0.52 ± 0.09	0.7 ± 0.3
1i2ta	3.8 ± 1.0	2.3 ± 1.9	0.39 ± 0.03	0.36 ± 0.02
1u84a	1.0 ± 0.7	0.52 ± 0.04	0.47 ± 0.05	0.35 ± 0.02
1r690	2.3 ± 0.3	0.64 ± 0.04	0.62 ± 0.03	0.47 ± 0.04
1o82a	1.9 ± 0.4	0.85 ± 0.07	1.1 ± 0.4	1.0 ± 0.5
1opd0	0.80 ± 0.01	0.84 ± 0.05	0.85 ± 0.05	0.73 ± 0.04
2igd0	0.38 ± 0.02	0.39 ± 0.02	0.370 ± 0.003	0.42 ± 0.02
1vcc0	2.10 ± 0.08	1.1 ± 0.7	1.1 ± 1.0	2.4 ± 0.2
2hsla	0.86 ± 0.07	0.83 ± 0.03	0.61 ± 0.02	0.50 ± 0.04
2cc6a	0.48 ± 0.03	0.37 ± 0.07	0.370 ± 0.006	0.43 ± 0.09
2ptl0	0.62 ± 0.03	0.78 ± 0.07	0.75 ± 0.05	1.2 ± 0.4
1csp0	0.74 ± 0.04	0.38 ± 0.03	2.0 ± 0.2	1.7 ± 0.9
1fltx	1.5 ± 0.3	1.1 ± 0.2	1.7 ± 0.5	4.3 ± 1.3
2ayda	2.8 ± 0.8	1.5 ± 0.9	5.0 ± 1.2	4.2 ± 1.7
2extb	2.1 ± 0.7	1.8 ± 0.3	1.9 ± 0.2	1.8 ± 0.6
Success rate (<3.5 Å)[c]	14/15 (93%)	15/15 (100%)	14/15 (93%)	13/15 (87%)

[a] Protein Data Bank identification number. The fourth digit is the chain ID.
[b] The mean and standard deviation of the grmsd values of three best near-native structures from three independent global minimizations with the DFIRE, uDFIRE, aDFIRE, and DOPE energy function, respectively.
[c] The total number (percentage) of refolded segments whose lowest grmsd values <3.5 Å.

for repulsive interactions is in part due to different statistical information from different databases for DOPE (1472 proteins) and aDFIRE (3572 proteins). Using the same database might further increase the similarity between aDFIRE and DOPE. For segment refolding, aDFIRE and DOPE yield similarly accurate structures for 10 out of 15 segments (grmsd difference <0.5 Å). aDFIRE produces a more accurate terminal β segment in 1csp0, while DOPE folds the terminal helix of 2guzb more accurately. We found that DOPE is more likely than aDFIRE to form helical structures.

It is somewhat surprising that uDFIRE makes a significant improvement over DFIRE. uDFIRE produces more accurate structures for six segments and similarly accurate structures for all other proteins (either the grmsd difference is <0.5 Å or both grmsd values are ≥6 Å, i.e., misfolded structures). The only major difference between uDFIRE and DFIRE is that the former has more bins within the same interaction range of 15 Å. We believe that the improvement is mostly contributed by a fine treatment of short-range interactions from a single 2 Å bin to four 0.5 Å bins, rather than a fine treatment of long-range interactions from seven 1 Å bins to 14 0.5 Å bins between 8 and 15 Å. Indeed, changing four 0.5 Å bins back to a single 2 Å bin makes uDFIRE fold 1o82a at the same level of accuracy (4.5 Å) as DFIRE does (5.2 Å). This highlights the importance of an accurate description of repulsive interactions for induced fitting in segment refolding. It is interesting to note that DFIRE and DOPE are able to form near perfect helices and reasonable strands in the absence of hydrogen-bonding interactions. This is mostly due to induced fitting of the unfolded segment to prefolded domains. It remains to be seen if DFIRE or any other distance-dependent only potential can make an ab initio folding for some helical proteins or even β proteins. We have implemented orientation-dependent dipolar interactions into the DFIRE energy functions. A test of refolding by the new dipolar DFIRE energy function to the same 15 terminal fragments yields a folding rate of 87% (13/15) to within 2 Å grmsd from the corresponding native conformations (Yang and Zhou 2008). The successful refolding (including the β hairpin in 2extb) highlights the importance of orientation-dependent polar interactions in the formation of secondary and tertiary structures of proteins.

The high success rate in refolding fully unfolded fragments by the all-atom statistical energy function highlights its potential for structural refinement. By comparison, it is known that molecular dynamics simulations often fail to refine homology (Baker and Sali 2001; Summa and Levitt 2007) models or partially unfolded structures (Zhu et al. 2006). One possible cause is that the entropy of packing is not included in a typical physical-based force field and it cannot be described adequately by short nanosecond dynamics simulations. This is supported by the fact that a more accurate description of the free energy of packing leads to 20% improvement in the success rate of fragment refolding from DFIRE to uDFIRE. Certainly, one cannot discount other equally plausible causes such as inadequate conformational sampling and shortcomings of the force field and/or solvent model.

Materials and Methods

The genetic algorithm and the sampling method are the same as those described elsewhere (Yang and Zhou 2008). We include them here briefly for completeness.

Genetic algorithm for global minimization of unfolded terminal segments

The selected terminal segment of a given protein is described by internal coordinates: the bond lengths, bond angles, planar torsion angles, backbone ϕ/ψ torsion angles, and side-chain χ angles. The bond lengths, bond angles, and planar torsion angles of the segment are fixed with standard values from the AMBER 99 force field (Weiner et al. 1986). The initial conformation of the segment is obtained by randomly assigned backbone ϕ/ψ torsion angles according to the observed residue-specific probability in the backbone-dependent rotamer library (Dunbrack Jr. and Karplus 1993) (http://www.fccc.edu/research/labs/dunbrack/bbdep.html; version as 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. The initial conformations are generated 16 Å–20 Å away from the native conformation in terms of global rmsd values (see Table 1). A given conformation is locally minimized by randomly choosing a new ϕ/ψ angle of a selected residue from its own bin or its nearest-neighboring 24 angle bins. The side-chain χ angles are then chosen based on the new backbone angles as above. 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.

The minimized initial conformations serve as the first parent generation. The parent conformations are ranked with the normalized fitness function and the standard fitness proportionate selection procedure (Goldberg and Smith 1987). We define the fitness function of each conformation in generation l, relative to other conformations in the same generation, as

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

where the conformation density $\rho_i = \sum_{j,j \neq i} S_{ij}$, S_{ij} is the number of residues in identical conformational states between conformations i and j, e_i is the energy of conformation i (including the fixed portion of the protein) based on either the DFIRE or DOPE energy function, $e_{\min}^{l-1}$ and $\triangle e^{l-1}$ are the lowest energy and the rmsd of the energies in the parent generation $l - 1$, respectively, with T_{env} set to 1.5. Two residues are considered in an identical conformation state if $|\phi_i - \phi_j| < 10°$, $|\psi_i - \psi_j| < 10°$, and $|\phi_i - \phi_j| + |\psi_i - \psi_j| < 15°$. The sequentially ranked conformation pair is chosen to be the parent to breed two new conformations, first by a two-point crossover and then by mutation operations. 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. A population of 120 conformations (i.e., $N_c = 120$) was used in this study.

Segment refolding

Table 3 lists 15 small globular proteins (<100 residues) with diverse structural topologies, both as a whole and in their terminal regions. 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 paper focuses on evaluating the proposed energy function rather than testing sampling techniques, we limited the maximum size of unfolded regions to be <15 residues for a strand-containing segment, and <25 residues for a helix-containing segment. Each refolding event takes ~40 h for the DFIRE energy function on a single CPU in an AMD Dual-Core Opteron Processor (2.4 GHz).

The DOPE, aDFIRE, and uDFIRE energy functions

All DFIRE variants were extracted from a database of 3574 nonredundant (≤30% homology) high-resolution proteins (resolution <2.0 Å and R % factor <0.25) from Hobohm et al. (1992).

Table 3. *Fifteen unfolded terminal regions*

			Unfolded	
PDB Id#[a]	Residue #[b]	Structure type[c]	Range (#)[d]	Type[e]
2guzb	65	5α	95–117 (23)	1α
1i2ta	61	4α	1051–69 (19)	1α
1u84a	81	4α	65–83 (19)	1α
1r690	61	5α	44–61 (18)	2α
1o82a	70	6α	51–70 (20)	2α
1opd0	85	3α4β	70–85 (16)	1α
2igd0	56	1α4β	52–61 (10)	1β
1vcc0	73	2α5β	63–73 (11)	1β
2hsla	89	1α9β	82–93 (12)	1α1β
2cc6a	62	1α3β	52–62 (11)	1β
2ptl0	61	1α4β	68–78 (11)	1β
1csp0	64	5β	54–64 (11)	1β
1fltx	95	8β	214–226 (13)	1β
2ayda	66	5β	346–358 (13)	1β
2extb	66	6β	61–72 (12)	2β

[a] Protein Data Bank identification number. The fourth digit is the chain ID.
[b] Number of residues in the native structure.
[c] The structural type represented by the number of α helices and β strands in the protein structure.
[d] Residue range of the unfolded regions (number of residues unfolded).
[e] The structural type of the unfolded region.

uDFIRE is an updated DFIRE energy function by increasing the number of bins from 20 to 30 without changing the cutoff distance of 15 Å. uDFIRE uses a constant 0.5 Å for each bin and a total of 30 bins, 158 residue-specific atom types with identical atom types merged (e.g., NH1 and NH2 in Arg), while the original DFIRE used a variable bin width (2 Å for the first 2 Å, 0.5 Å from 2 Å and 8 Å, and 1 Å from 8 to 15 Å) and 167 residue-specific atom types (Zhou and Zhou 2002). Moreover, uDFIRE is a continuous potential based on a linear interpolation. No interpolation was used in DFIRE (Zhou and Zhou 2002). The same bin procedure and linear interpolation are used in DOPE.

We further introduce aDFIRE for examining the effect of the dependence on relative sequence positions adopted in DOPE. The pair interaction energy between two atoms i and j of residues I and J, u_{ij}, is different from u_{ji}. The former is exacted from the database when the sequence position of residue I is greater than that of J, and the latter is used when the opposite is true.

The latest version of DOPE was kindly provided by the authors. It uses 30 distance bins with 0.5 Å per bin, asymmetric pairwise interaction, and linear interpolation between bins. We have integrated their energy function into our genetic algorithm for global energy minimization.

Acknowledgment

This work was supported by the NIH (R01 GM 966049 and R01 GM 068530).

References

Atwell, S. and Wells, J.A. 1999. Selection for improved subtiligases by phage display. *Proc. Natl. Acad. Sci.* **96:** 9497–9502.

Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93–96.

Bradley, P., Misura, K.M.S., and Baker, D. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science* **309:** 1868–1871.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Cheng, J., Pei, J., and Lai, L. 2007. A free-rotating and self-avoiding chain model for deriving statistical potentials based on protein structures. *Biophys. J.* **92:** 3868–3877.

Dunbrack Jr., R.L. and Karplus, M. 1993. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.* **230:** 543–574.

Eramian, D., Shen, M., Devos, D., Melo, F., Sali, A., and Marti-Renom, M.A. 2006. A composite score for predicting errors in protein structure models. *Protein Sci.* **15:** 1653–1666.

Ferrada, E. and Melo, F. 2007. Nonbonded terms extrapolated from nonlocal knowledge-based energy functions improve error detection in near-native protein structure models. *Protein Sci.* **16:** 1410–1421.

Goldberg, D.E. and Smith, R.E. 1987. Nonstationary function optimization using genetic algorithm with dominance and diploidy. In *Proceedings of the second international conference on genetic algorithms and their application* (ed. J.J. Grefenstette), pp. 59–68. Lawrence Erlbaum Associates, Inc., Cambridge, MA, USA.

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. 1992. Selection of representative protein data sets. *Protein Sci.* **1:** 409–417.

Jacob, E. and Unger, R. 2007. A tale of two tails: Why are terminal residues of proteins exposed? *Bioinformatics* **2:** E225–E230. doi: 10.1093/bioinformatics/btl318.

Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118:** 11225–11236.

Llinas, M. and Marqusee, S. 1998. Subdomain interactions as a determinant in the folding and stability of T4 lysozyme. *Protein Sci.* **7:** 96–104.

Melo, F., Sanchez, R., and Sali, A. 2002. Statistical potentials for fold assessment. *Protein Sci.* **430:** 430–448.

Ponder, J.W. and Richards, F.M. 1987. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.* **8:** 1016–1026.

Scott, W.R.P., Hünenberger, P.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., Krüger, P., and van Gunsteren, W.F. 1999. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A* **103:** 3596–3607.

Shen, M. and Sali, A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15:** 2507–2524.

Summa, C.M. and Levitt, M. 2007. Near-native structure refinement using in vacuo energy minimization. *Proc. Natl. Acad. Sci.* **104:** 3177–3182.

Tanaka, S. and Scheraga, H.A. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9:** 945–950.

Weiner, S.J., Kollman, P., Nguyen, D., and Case, D. 1986. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **7:** 230–252.

Yang, Y. and Zhou, Y. 2008. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* (in press). doi: 10.1002/prot.21968.

Zhang, Y. and Skolnick, J. 2004. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci.* **101:** 7594–7599.

Zhou, H. and Zhou, Y. 2002. 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.

Zhu, J., Xie, L., and Honig, B. 2006. Structural refinement of protein segments containing secondary structure elements: Local sampling, knowledge-based potentials and clustering. *Proteins* **65:** 463–479.