OPUS-Ca: A knowledge-based potential function requiring only Cα positions

YINGHAO WU,[1,4] MINGYANG LU,[2,4] MINGZHI CHEN,[3] JIALIN LI,[3] AND JIANPENG MA[1,2,3]

[1]Department of Bioengineering, Rice University, Houston, Texas 77005, USA
[2]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA
[3]Graduate Program of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas 77030, USA

(RECEIVED January 28, 2007; FINAL REVISION April 14, 2007; ACCEPTED April 16, 2007)

Abstract

In this paper, we report a knowledge-based potential function, named the OPUS-Ca potential, that requires only Cα positions as input. The contributions from other atomic positions were established from pseudo-positions artificially built from a Cα trace for auxiliary purposes. The potential function is formed based on seven major representative molecular interactions in proteins: distance-dependent pairwise energy with orientational preference, hydrogen bonding energy, short-range energy, packing energy, tri-peptide packing energy, three-body energy, and solvation energy. From the testing of decoy recognition on a number of commonly used decoy sets, it is shown that the new potential function outperforms all known Cα-based potentials and most other coarse-grained ones that require more information than Cα positions. We hope that this potential function adds a new tool for protein structural modeling.

Keywords: knowledge-based potential function; decoy recognition; structure prediction; protein folding

Protein folding is one of the most challenging problems in both computational and experimental biophysics (Dobson and Karplus 1999). The goal is to determine three-dimensional structures from one-dimensional amino acid sequences. In computational studies, a potential function plays a central role in accurately predicting the structures. There are two general types of potential functions: One is physics-based and another is knowledge-based. The physics-based potential functions are derived from quantum mechanical calculations, e.g., the CHARMM force field (MacKerell et al. 1998), the essence of which is molecular mechanics. The knowledge-based potential functions are derived from statistical analysis of known protein structures, the essence of which is the potential of mean force, or free energy. In many applications, it has been shown that the knowledge-based potential functions outperform the physics-based ones. There are many comprehensive reviews for various potential functions in the literature (Sippl 1995; Jernigan and Bahar 1996; Moult 1997; Lazaridis and Karplus 2000; Gohlke and Klebe 2001; Meller and Elber 2002; Russ and Ranganathan 2002; Buchete et al. 2004a; Poole and Ranganathan 2006; Skolnick 2006; Zhou et al. 2006).

The knowledge-based potential functions can usually be divided into two types: atomic level potentials (DeBolt and Skolnick 1996; Zhang et al. 1997; Melo and Feytmans 1998; Samudrala and Moult 1998; Gatchell et al. 2000; Lu and Skolnick 2001; Zhou and Zhou 2002; McConkey et al. 2003; Hubner et al. 2005; Qiu and Elber 2005; Shen and Sali 2006) and coarse-grained potentials (Tanaka and Scheraga 1976; Miyazawa and Jernigan 1985; Hendlich et al. 1990; Sippl 1990; Hinds and Levitt 1992; Jones et al. 1992; Godzik et al. 1995; Miyazawa and Jernigan 1996; Bahar and Jernigan 1997; Eisenberg et al. 1997; Betancourt

[4]These authors contributed equally to this work.

Reprint requests to: Jianpeng Ma, One Baylor Plaza, BCM-125, Baylor College of Medicine, Houston, TX 77030, USA; e-mail: jpma@bcm.tmc.edu; fax: (713) 796-9438.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.072796107.

and Thirumalai 1999; Liwo et al. 1999; Simons et al. 1999; Tobi and Elber 2000; Melo et al. 2002; Zhang et al. 2003, 2004, 2006; Buchete et al. 2004b; Loose et al. 2004; Colubri et al. 2006; Dehouck et al. 2006; Dong et al. 2006; Rajgaria et al. 2006). The latter have been demonstrated to be highly effective in reducing the computational cost in modeling native protein structures, although they are sometimes thought not to be physically rigorous enough to reflect the entire landscape of the potential surface (Thomas and Dill 1996; Skolnick 2006). The performance and applicability of coarse-grained potential functions are largely modulated by the choice of a coarse-graining scheme. In many applications, an ability to accurately calculate the potential energy solely based on Cα positions would certainly give one some advantages. Typical examples are recent studies on modeling protein chain topology based on low-resolution density maps (Wu et al. 2005a) and on a coarse-grained folding simulation based on a Cα model (Wu et al. 2005b).

In this study, we have developed a knowledge-based potential function, named the OPUS-Ca potential, that requires only the Cα positions as input. The potential function contains seven terms for representing typical molecular interactions in proteins. They are distance-dependent pairwise energy with orientational preference, hydrogen bonding energy, short-range energy, packing energy, tri-peptide packing energy, three-body energy, and solvation energy. It was tested against a number of commonly used decoy sets. The results show that the OPUS-Ca potential outperforms all known Cα-based potentials and most other coarse-grained ones that require more information than Cα positions. We hope that this potential function adds a new tool for protein structural modeling.

Results

Performance of individual terms

We first demonstrate the performance of five major individual energy terms in Equation 1 (see Materials and Methods) in terms of decoy recognition. They are distance-dependent pairwise energy with orientational preference $E_{pairwise}$, hydrogen bonding energy E_{Hbond}, short-range energy E_{short_range}, packing energy $E_{packing}$, and solvation energy $E_{solvation}$. There are two other terms: tri-peptide packing energy $E_{tri-peptide}$ and three-body energy E_{3-body}. Due to their relatively small contributions, their individual performance is not presented in detail here.

The decoy sets used in this study were from two collections. One was the so-called Decoys'R'Us collection, which included decoy sets 4state_reduced (seven proteins) (Park and Levitt 1996), fisa (four proteins) (Simons et al. 1997), fisa_casp3 (five proteins) (Simons et al. 1997), lattice_ssfit (eight proteins) (Samudrala et al. 1999; Xia et al. 2000), and lmds (eight proteins) (Keasar and Levitt 2003). In total, there are 32 proteins in the Decoys'R'Us collection. Another collection was the LKF decoy set (185 proteins) (Loose et al. 2004).

Table 1 gives the detailed ranking and Z-scores for individual proteins in the Decoys'R'Us collection. Note that only the results for 25 commonly used proteins in Decoys'R'Us (Tobi and Elber 2000; Dehouck et al. 2006) are listed.

Distance-dependent pairwise energy with orientational preference

For the distance-dependent pairwise energy term $E_{pairwise}$, the energy was calculated with respect to pseudo-Cβ atoms built from Cα atoms. In the literature (Zhang et al. 2004), it had been shown that pairwise energy based on Cβ atoms taken from X-ray structures was better than that based on Cα atoms because the distance between two Cβ atoms could better represent side chain packing than that between Cα atoms. This was confirmed in this study. Moreover, it has been shown that it is advantageous to include the orientational preference of residues (Buchete et al. 2004b; Miyazawa and Jernigan 2005). In this study, the pairwise energy in the OPUS-Ca potential took into account the relative orientation of two pairing residues. The comparison of decoy recognition for pairwise energy with and without orientation preference indicates that the energy with orientation preference could recognize the native conformation of more decoy sets than that without orientation preference. Also, the average Z-scores for the native structure in the two collections of decoy sets was observed to be 0.2–0.3 better in the case with the orientation preference than the case without. Figure 1 shows the performance on two decoy set collections, Decoys'R'Us (25 proteins) and LKF (185 proteins). The upper and middle panels give the percentage of proteins in the decoy sets whose native conformations were correctly ranked as the top 1 and within the top 10, respectively. It is clear that the trends in both decoy set collections are consistent; the performance of pairwise energy with pseudo-Cβ is better than the case without, and the performance of the energy with the orientational preference is better than the case without. Finally, the average Z-scores show exactly the same trend as well (Fig. 1, lower panels).

Hydrogen bonding energy

The hydrogen bonding energy term E_{Hbond} is required to build pseudo-backbone atoms from the original Cα atoms. They were the N and H atoms of amide groups and the C and O atoms of carbonyl groups. To compensate for error from building backbone atoms, the hydrogen bonding criteria were slightly modified. First, Cα-based hydrogen bonding energy was compared with all-atom-based hydrogen bonding energy, which directly used original backbone atoms and standard hydrogen bond criteria. By testing on

Table 1. *Performance of the OPUS-Ca potential on the Decoys'R'Us decoy set*

(A) Individual proteins

		Size	Pairwise		H-bond		Short-range		Packing		Tri-peptide		Three-body		Solvation	
4state_reduced																
1	1ctf	631	1[a]	−3.14[b]	4	−2.18	1	−2.56	9	−1.96	1	−2.66	2	−3.40	4	−2.89
2	1r69	676	1	−3.02	1	−2.68	1	−2.95	2	−3.43	1	−3.28	25	−1.71	60	−1.48
3	1sn3	661	1	−4.12	1	−4.29	1	−3.54	2	−6.78	1	−9.42	4	−2.87	2	−2.35
4	2cro	675	3	−2.63	2	−2.77	1	−2.98	6	−2.50	2	−3.10	8	−2.61	66	−1.43
5	4pti	688	1	−3.63	1	−3.35	1	−2.95	3	−6.34	1	−6.28	7	−2.62	19	−1.82
6	4rxn	678	1	−3.17	1	−3.91	1	−3.42	1	−10.33	16	−2.89	7	−3.13	13	−2.29
fisa																
7	1fc2	501	478	2.03	448	1.24	499	3.26	446	1.34	463	1.52	14	−1.99	1	−3.79
8	1hdd-C	501	1	−2.90	217	−0.20	373	0.60	6	−2.34	18	−1.95	1	−3.42	1	−5.17
9	2cro	501	1	−3.41	1	−3.01	37	−1.37	4	−2.66	4	−2.75	32	−1.8	9	−2.16
fisa_casp3																
10	1bg8-A	1201	1	−3.86	12	−2.10	135	−1.22	127	−1.26	96	−1.40	211	−0.95	248	−0.78
11	1bl0	972	6	−2.62	11	−2.38	98	−1.28	542	0.17	164	−0.92	762	0.79	777	0.84
12	1jwe	1408	1	−4.95	1	−5.60	2	−3.24	54	−1.88	2	−3.47	881	0.33	1	−2.65
lattice_ssfit																
13	1ctf	2001	1	−6.31	2	−3.90	1	−4.41	2	−4.08	4	−3.72	3	−3.89	3	−3.33
14	1dkt-A	2001	1	−4.92	1	−4.60	1	−4.53	1	−23.69	1	−8.31	27	−2.27	773	−0.29
15	1fca	2001	1	−5.74	1	−4.50	1	−4.09	4	−4.89	202	−1.06	701	−0.24	72	−1.87
16	1nkl	2001	1	−4.96	1	−3.33	40	−1.93	1	−5.24	1	−5.23	2	−5.58	1	−4.04
17	1pgb	2001	1	−6.24	1	−14.17	1	−10.74	1	−33.87	1	−25.06	12	−3.29	1	−4.73
18	1trl-A	2001	1	−3.00	1	−2.99	1	−3.92	1	−4.64	1	−6.34	468	−0.64	390	−0.82
lmds																
19	1ctf	498	1	−3.25	9	−1.91	9	−2.01	6	−2.52	26	−1.77	21	−1.80	4	−2.31
20	1dtk	216	6	−2.23	15	−1.52	20	−1.33	1	−3.00	10	−2.21	50	−0.71	7	−1.98
21	1fc2	501	359	0.59	18	−1.76	13	−2.01	20	−1.68	11	−2.32	15	−2.02	1	−3.23
22	1igd	501	1	−2.85	1	−6.88	1	−4.39	2	−2.98	60	−1.15	9	−1.98	4	−2.61
23	1shf-A	438	1	−5.33	1	−6.95	1	−3.57	22	−1.85	28	−1.68	140	−0.47	1	−3.36
24	2cro	501	2	−3.14	1	−3.99	2	−2.75	61	−1.19	1	−3.18	1	−5.07	1	−5.41
25	2ovo	348	61	−0.90	1	−3.96	1	−2.64	85	−0.18	113	−0.03	178	0.25	15	−1.80

(B) Average results

	# f proteins	Pairwise		H-bond		Short-range		Packing		Tri-peptide		Three-body		Solvation	
4state_reduced	6	5[c]/6[d]	−3.29[e]	4/6	−3.20	6/6	−3.07	1/6	−5.22	4/5	−4.61	0/5	−2.72	0/2	−2.04
fisa	3	2/2	−1.43	1/1	−0.66	0/0	0.83	0/2	−1.22	0/1	−1.06	1/1	−2.40	2/3	−3.71
fisa_casp3	3	2/3	−3.81	1/1	−3.36	0/1	−1.91	0/0	−0.99	0/1	−1.93	0/0	0.06	1/1	−0.86
lattice_ssfit	6	6/6	−5.20	5/6	−5.58	5/5	−4.94	4/6	−12.74	4/5	−8.29	0/2	−2.65	2/3	−2.51
lmds	7	3/5	−2.44	4/5	−3.85	3/5	−2.67	1/3	−1.91	1/2	−1.76	1/2	−1.69	3/6	−2.96
Summary	25	18/22	−3.35	15/19	−3.67	14/17	−2.80	6/17	−5.11	9/14	−3.95	2/10	−2.04	8/15	−2.47

[a] Ranking of the native conformation.
[b] Z-score of the native conformation.
[c] The number of native conformations ranked top-1.
[d] The number of native conformations ranked top-10.
[e] Average Z-score.

25 proteins in the Decoys'R'Us sets, it was found that Cα-based hydrogen bonding energy recognized more native conformations and had only a slightly lower Z-score than all-atom-based energy (Fig. 2). The Cα-based hydrogen bonding energy term also performed better than the all-atom-based calculation in the top-10 ranking.

Another feature of our hydrogen bonding energy was that an unfavorable energy barrier for hydrogen bond formation was eliminated by a constant energy shift. The occurrence of hydrogen bonds with a large CN distance and a large CON angle was rare. That caused hydrogen bonds to have higher energy at these regions than at the optimal hydrogen bonding region. The energy values sometimes could even be positive, so that hydrogen bond formation was not favorable. To better describe hydrogen bonding as an energetically favorable interaction, a constant energy shift was added to ensure that the energy was near zero when a hydrogen bond was about to form. Hence, hydrogen bonds could readily form without encountering an energy barrier when an amide group was close to a carbonyl group.

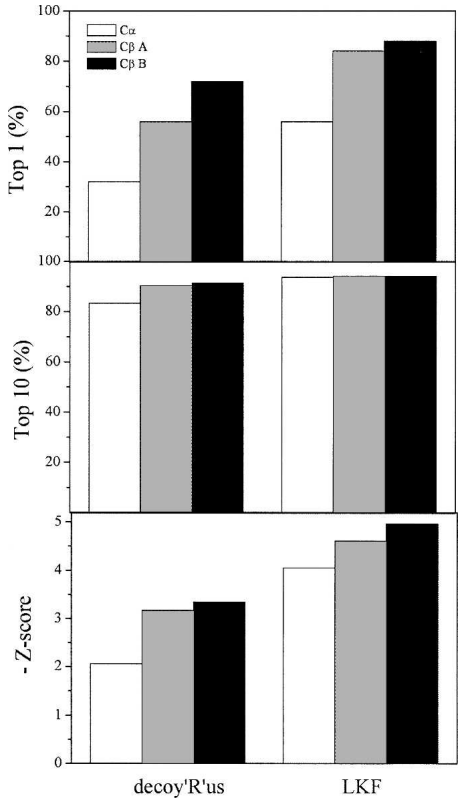

Figure 1. Performance of pairwise energy. (*Top* panel) Percentage of proteins in decoy sets whose native conformations were ranked top-1, (*middle* panel) percentage of proteins in decoy sets whose native conformations were ranked top-10, (*bottom* panel) the negative average Z-scores. (Cα) Pairwise energy based on Cα positions, (Cβ A) pairwise energy based on pseudo-Cβ positions built from Cα positions without orientation preference, (Cβ B) pairwise energy based on pseudo-Cβ positions built from Cα positions with orientation preference.

From Figure 2, one can see that, comparing with the case with a constant energy shift, the hydrogen bonding energy without a constant energy shift performed persistently worse in recognizing native conformation as top-1, top-10 ranking, and average Z-scores. It was also found that the ranking of the native conformations of three proteins (1nkl in lattice_ssfit, 1dtk in lmds, and 1fc2 in lmds) was dramatically worsened from within top-20 to below top-50. However, comparing with the case with an energy shift, it was found that our energy term performed worse in the lattice_ssfit and lmds decoy sets, while it performed better in the 4state_reduced and fisa decoy sets. So the effect of an energy shift was decoy-set-dependent, which was presumably related to how each decoy set was generated.

Short-range energy

For the short-range energy term E_{short_range}, different types of secondary structures were considered separately. This was because residues in different secondary structure types had different preferences for local conformations. From Table 1A, one can see that the short-range energy could perform quite well in all decoy sets except for fisa and fisa_casp3 decoy sets, in which case it couldn't recognize any native conformation and only one in the top 10 (PDB code: 1jwe). This was probably because the decoys in fisa and fisa_casp3 were generated by Rosetta based on native small fragments (Simons et al. 1997); thus, the native-like nature of short-range conformations caused insensitivity in the energy term.

Packing energy

The packing energy term $E_{packing}$ could be divided into seven smaller terms. They belong to three types: short-range packing that facilitated the formation of a single helix or single strand (E_{H_self}, E_{S_self}); long-range packing in paired strands that facilitated strand pairing ($E_{S\_pairing}$); and long-range packing between different helices or strands in stabilizing tertiary structure ($E_{H-H\_packing}$, $E_{H-S\_packing}$, $E_{S-S\_packing}$). Equal weight was used for all seven terms. At the $i,i + 3$ or $i,i + 4$ position in a single

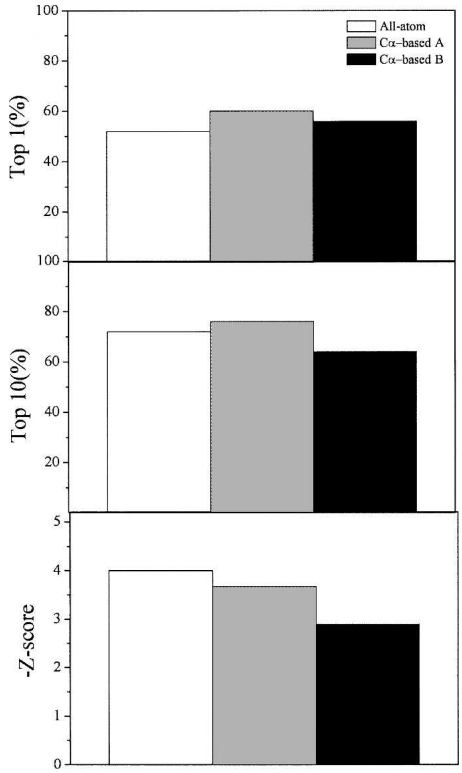

Figure 2. Performance of hydrogen bond energy. (*Top* panel) Percentage of proteins in decoy sets whose native conformations were ranked top-1, (*middle* panel) percentage of proteins in decoy sets whose native conformations were ranked top-10, (*bottom* panel) the negative average Z-scores. (All-atom) All-atom-based hydrogen bond energy, (Cα-based A) Cα-based hydrogen bond energy with an energy shift, (Cα-based B) Cα-based hydrogen bond energy without an energy shift. Results are shown for 25 proteins in the Decoys'R'Us collection.

helix, Pro and Gly were less likely to be involved, as the E_{H_self} was among the highest when packing pairs involved Pro and Gly. Ser, Thr, Asp, and Asn were the next four unfavorable residues. Also, Cys was less likely to pair with one of these four types of residues. In contrast, Ala was more likely to be involved in helices. It was also identified that Met–Met, Glu–Arg, and Glu–Lys pairs were favorable at both the $i,i + 3$ and $i,i + 4$ positions. A Met–Met pair had an E_{H_self} of -1.086 and -1.149 at the $i,i + 3$ and $i,i + 4$ positions, respectively; a Glu–Arg pair had an E_{H_self} of -1.333 and -1.275 at the $i,i + 3$ and $i,i + 4$ positions, respectively; and a Glu–Lys pair had an E_{H_self} of -1.177 and -1.248 at the $i,i + 3$ and $i,i + 4$ positions, respectively. For the $i,i + 2$ position in a strand, Pro was identified to be unfavorable, while hydrophobic residues preferred these positions. For example, a Val–Val pair had an E_{S_self} of -1.793, and a Val–Ile pair had an E_{S_self} of -1.677. For two paired strands, packing residues tended to have hydrogen bond and electrostatic interactions. Preferred contacting residues contained Cys–Cys, Glu/Asp–Arg/Lys, His–His, Ser–Asn/Gln, Trp–Trp pairs, etc. For example, $E_{S\_pairing}$ (averaged over seven cases) for Cys–Cys was -0.811; $E_{S\_pairing}$ (averaged over seven cases) for Glu–Arg was -0.894; $E_{S\_pairing}$ (averaged over seven cases) for His–His was -0.605. For long-range tertiary packing, it was found that hydrophobic and large aromatic residues were favorable. For example, Tyr–Trp had an $E_{S\_S\_packing}$ of -2.470, and Ile–Leu had an $E_{H\_H\_packing}$ of -1.687.

The overall performance of the packing energy term in Decoys'R'Us recognized six native conformations in the top-1 ranking and 17 in the top-10 ranking (Table 1).

Solvation energy

The solvation energy term $E_{solvation}$ was based on side chain solvent-accessible surfaces (SAS). An all-atom-based energy function was first established based on the SAS calculated from an all-atom model. Then, the SAS for the Cα model was estimated based on a coarse-grained method in which all parameters involved were systemically optimized from a structure database (see Materials and Methods). Using this estimated SAS as an approximate value, the solvation energy for the Cα model can be estimated from the all-atom-based energy function.

As indicated in Figure 3, in the Decoys'R'Us test, solvation energy based on an all-atom SAS found native conformations of 11 decoy sets in the top-1 ranking, and 22 native conformations in the top-10 ranking. This implied reasonable accuracy of the solvation energy term when the SAS was obtained from the all-atom structure. For the Cα model, the energy function was not as good as its all-atom counterpart. However, it still recognized eight native conformations in the top-1 ranking and 15 in the top-10 ranking. The average Z-scores are also listed in Figure 3.

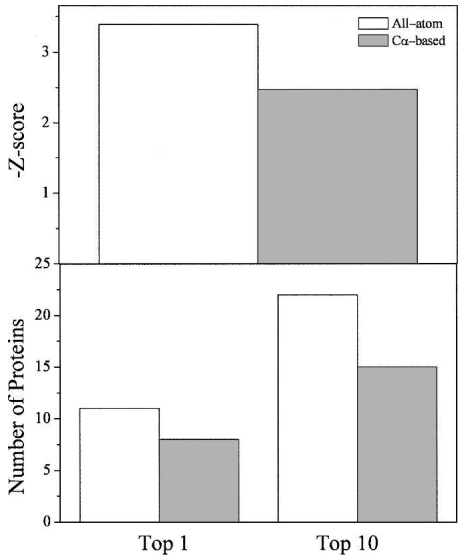

Figure 3. Performance of solvation energy. (*Top* panel) The negative average Z-scores, (*bottom* panel) the number of proteins in decoy sets whose native conformations were ranked top-1 (*left*) and top-10 (*right*). (All-atom) Solvation energy based on the solvent-accessible surface calculated from all atom positions, (Cα-based) solvation energy based on the solvent-accessible surface calculated from Cα positions only. Results are shown for 25 proteins in the Decoys'R'Us collection.

Performance of the overall energy function

To examine the performance of the overall energy function, weights had to be assigned to seven energy terms, a procedure that could sometimes be subjective. Two different ways of weight assigning were tried.

In the first way, all energy terms were calculated for all proteins in a non-homology database that had no chain break (a total of 1673 proteins [Wang and Dunbrack Jr. 2003]). The average energy was calculated for each term. Weights were assigned in such a way that they were anti-proportional to the average energy so as to make the numerical contribution from each term roughly equal. As indicated in Table 2, this scheme of weight assignment resulted in 18 out of 25 decoy sets in Decoys'R'Us with their native conformations correctly recognized as the lowest in energy (Subset 1 in Table 2). In the LKF decoy collection (Subset 2 in Table 2), it recognized 148 out of 151 decoy sets. As the tri-peptide and three-body energy terms could be regarded as higher order corrections of other terms, we also empirically lowered the magnitudes of these two weights to 0.1. It was found that the energy function with the modified weights could slightly improve the performance (19 out of 25 Decoys'R'Us decoy sets), indicating the less important nature of these two terms.

In the second way, all seven weights were optimized iteratively on three subsets of decoy sets (see Materials and Methods). Strikingly, it was found that the magnitudes of

Table 2. *Weights and performance in decoy set recognition*

	Type of weight	Equal contribution	Modified	Optimized
Weights	$w_{pairwise}$	0.3	0.3	0.3
	w_{Hbond}	0.3	0.3	0.3
	w_{short_range}	0.6	0.6	0.6
	$w_{packing}$	3.5	3.5	2.7
	$w_{tri\text{-}peptide}$	0.4	0.1	0.1
	$w_{3\text{-}body}$	5.6	0.1	0.1
	$w_{solvation}$	5.1	5.1	2.5
Performance	[a]Subset-1 (25[b])	18[c] (−4.11[d])	19 (−4.68)	21 (−4.81)
	[e]Subset-2 (151)	148 (−6.92)	148 (−7.13)	146 (−7.00)
	[f]Subset-3 (41)	37 (−6.40)	39 (−6.66)	39 (−6.51)

[a] A subset from the Decoys'R'Us collection.
[b] The number of proteins in the subset of the decoy collection.
[c] The number of native conformations with a rank of top-1.
[d] Average Z-score.
[e] A subset from the LKF decoy collection.
[f] The remaining proteins in the Decoys'R'Us and LKF decoy collections.

the optimized weights were very close to the modified weight mentioned above. With the optimized weights, the energy function could recognize 21 native conformations out of 25 decoy sets in Decoys'R'Us and 146 native conformations out of 151 LKF decoy sets. The performance in Subset-3 was also similar. Overall, the performance with the optimized weights was close to the case with the modified weight. This result indicates that the optimized weights are not very biased by the weight optimization procedure. We suggest using optimized weights in real applications as they have included the most diverse features of all decoy sets.

The correlation between the root mean square deviation (RMSD) of decoy conformations from the native conformation and the energy of decoy conformations was evaluated. As indicated in Figure 4, a most linear-like correlation between the RMSD and energy was observed for 4state_reduced. Decoy set LKF had reasonable correlations. However, in other decoy sets the correlations were not so good. This suggests that the correlation between RMSD and energy depended on how the decoy sets were generated.

The performance of the OPUS-Ca potential was also compared with that of other potentials. In the literature, there are a few energy functions solely based on Cα atoms (Loose et al. 2004; Rajgaria et al. 2006; Zhang et al. 2006). The results are listed in Table 3. The performance of the OPUS-Ca potential seems to be better in terms of decoy set recognition and Z-scores. It also outperformed many other coarse-grained potential functions that require more information than Cα positions (Hinds and Levitt 1992; Godzik et al. 1995; Miyazawa and Jernigan 1996; Bahar and Jernigan 1997; Betancourt and Thirumalai 1999; Tobi and Elber 2000; Zhang et al. 2004; Dong et al. 2006). In two cases (Zhou and Zhou 2002; Dehouck et al. 2006), the performance was similar.

Discussion

In this study, a knowledge-based potential function, named the OPUS-Ca potential, was developed. To evaluate the potential function, only Cα positions are needed as input. Since it is hard to establish a sensitive enough potential function based only on Cα positions, the contributions from other atomic positions were established from pseudo-positions artificially built from the Cα trace. The potential function was constructed based on seven major terms representing dominant molecular interactions in proteins. The seven terms are distance-dependent pairwise energy with orientational preference, hydrogen bonding energy, short-range energy, packing energy, tri-peptide packing energy, three-body energy, and solvation energy.

Decoy set recognition indicated that the overall potential function outperformed all known Cα-based potentials and most of the other coarse-grained ones that require more information than Cα positions. For the performance of individual terms, it was found that the distance-dependent pairwise energy with orientational preference performed the best, which could identify 18 native conformations alone (out of 25 proteins in the Decoys'R'Us collection). Hydrogen bonding and short-range energy could also identify 15 and 14 native conformations, respectively. If the top-10 was used for native conformation ranking, then five out of seven energy terms could identify >15 native conformations alone (especially, the distance-dependent pairwise energy with orientational preference could identify 22 native conformations). The performance of some individual terms could even perform better than some of the other potentials published in the literature. This highly optimized performance of individual terms is advantageous because, in certain situations, one may want to use the individual energy terms separately based on their physical nature.

An important and difficult issue in developing knowledge-based potential functions is the assignment of weight for each term (Feng et al. 2007). In general, each term represents one or more aspects of physical interactions, so the contribution of each term should be inherently determined by the physical features of protein structures. Ideally, the magnitudes of weights should be independent of their performance on decoy sets, and independent of the methods in generating decoy sets. However, there is no ab initio way to determine the contribution of each energy term. Besides, some energy terms, like pairwise, three-body, and tri-peptide terms, have a mixture of several basic physical interactions; i.e., they are not completely orthogonal to each other. That makes

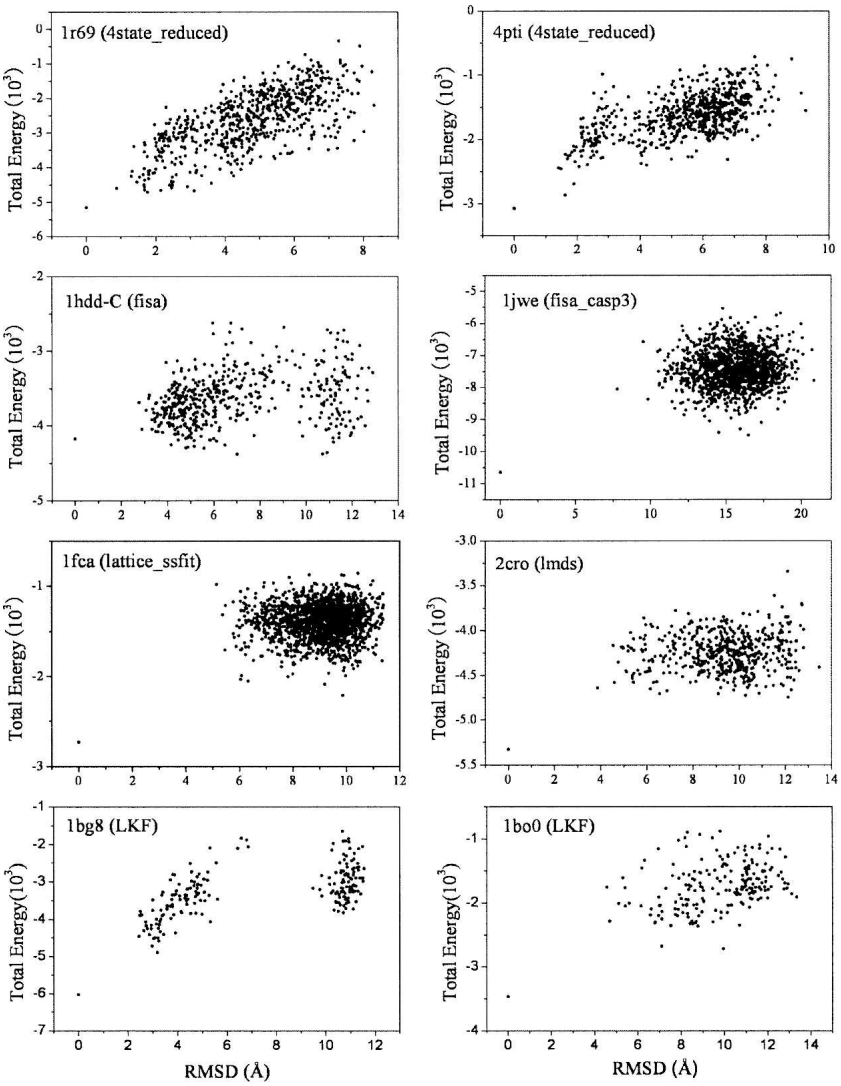

Figure 4. Scatter plots of total energy vs. RMSD of decoy from the native structure (based on Cα atoms). Results of eight proteins (1r69 and 4pti in the 4state_reduced decoy set, 1hdd-C in the fisa decoy set, 1jwe in the fisa_casp3 decoy set, 1fca in the lattice_ssfit decoy set, 2cro in the lmds decoy set, and 1bg8, 1bo0 in the LKF decoy set) are shown.

weight optimization even more subjective. This is why weight optimization by using a specific training set often introduces biases. In this study, several different ways of assigning weights were tried in order to minimize the bias.

Materials and Methods

The total energy function consists of seven terms,

$$E_{tot} = w_{pairwise}E_{pairwise} + w_{Hbond}E_{Hbond} + w_{short\_range}E_{short\_range} + w_{packing}E_{packing} + w_{tri-peptide}E_{tri-peptide} + w_{3-body}E_{3-body} + w_{solvation}E_{solvation}. \quad (1)$$

Here, w is the weight for that energy term. The statistical analysis of the knowledge-based potential function was performed over a nonhomologous structure database from the PISCES server by Dunbrack (Wang and Dunbrack Jr. 2003). Only X-ray structures were used. The percentage identity cutoff was 30%. The resolution cutoff was 1.8 Å. The R-factor cutoff was 0.25. The total number of chains was 2232.

Building pseudo-main chain atoms

Although the energy function only requires a Cα trace as input, to reliably build some of the terms in Equation (1), pseudo-main chain (including N, C, O, and H atoms) and Cβ conformations were established from a Cα trace. The procedure was based on the observation that main chain conformation can be mostly determined by local conformation of Cα atoms (Fidelis et al. 1994; Milik et al. 1997). In general, a main chain atom database was established, and it contained positional information of main chain atoms with respect to the local conformation of Cα atoms extracted from the nonhomologous structure database.

Table 3. *Comparison of performance between OPUS-Ca and other potential functions*

Energy functions		Decoys'R'Us (25[a])		LKF (151)		References
		Top-1 performance	Average Z-scores	Top-1 performance	Average Z-scores	
Cα-only potentials	OPUS-Ca	21	−4.81	146	−7.00	This study
	CALSP	10[b] (out of 18)	—	140	−6.42	Zhang et al. (2006)
	HR	—	—	86 (out of 110)	—	Rajgaria et al. (2006)
	LKF	—	—	93	−3.08	Loose et al. (2004)
Other coarse-grained potentials	HL	8	−2.67	—	—	Hinds and Levitt (1992)
	GKS	9	−2.36	—	—	Godzik et al. (1995)
	MJ	11	−2.82	—	—	Miyazawa and Jernigan (1996)
	BJ	15	−2.75	—	—	Bahar and Jernigan (1997)
	BT	9	−2.65	—	—	Betancourt and Thirumalai (1999)
	TE13	14	−3.53	64	−2.44	Tobi and Elber (2000)
	DFIRE-B	20	−4.22	—	—	Zhou and Zhou (2002)
	DFIRE-SCM	23[c] (out of 32)	−4.36	—	—	Zhang et al. (2004)
	DGR	21	−5.25	—	—	Dehouck et al. (2006)
	DWL	22[c] (out of 32)	−3.59	—	—	Dong et al. (2006)

[a] The total number of proteins in the decoy set collection.
[b] Two decoy sets, fisa and fisa_casp3, in Decoys'R'Us are not included; thus, there is a total of only 18 proteins. If compared with the performance on all 25 decoy sets used in Zhang et al. (2006) (seven in 4state_reduced, eight in lattice_ssfit, 10 in lmds), the OPUS-Ca potential had 22 Top-1s, and the Z-score was −5.33. The performance of CALSP in Zhang et al. (2006) was 15 Top-1s.
[c] It also included another seven Decoys'R'Us sets. The performance of the OPUS-Ca potential on the Top-1 ranking was 25 out of 32 decoy sets. Thus, it was better than DWL and DFIRE-SCM.

In detail, all main chain atoms of residue i were built from the position of four consecutive Cα atoms of residues $i - 1$, i, $i + 1$, and $i + 2$. The local Cα conformation was based on three parameters: Cα−Cα distance ($d_{i-1,i+1}$) between residues $i - 1$ and $i + 1$, Cα−Cα distance ($d_{i,i+2}$) between residues i and $i + 2$, and the dihedral angle (φ_i) formed by all four Cα atoms. Figure 5 schematically illustrates these parameters. Here, the distance was divided into 10 bins with a bin width of 0.3 Å. The range of the distance in the analysis was 4.6–7.6 Å. The dihedral angle was divided into 36 bins with a bin width of 10°. This led to a total of $10 \times 10 \times 36 = 3{,}600$ three-dimensional bins. To establish the main chain atom database, the positions of all main chain atoms (N, C, O, H) found in the structure database were averaged within each bin. To perform the averaging, a local reference frame was used. Its origin was set at the Cα atom of residue i, and Cartesian coordinate axes $\mathbf{v}_x$, $\mathbf{v}_y$, $\mathbf{v}_z$ were defined as:

$$\begin{aligned}\mathbf{v}_x &= \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}\\ \mathbf{v}_b &= \frac{\mathbf{r}_i - \mathbf{r}_{i-1}}{|\mathbf{r}_i - \mathbf{r}_{i-1}|},\\ \mathbf{v}_y &= \mathbf{v}_x \times \mathbf{v}_b\\ \mathbf{v}_z &= \mathbf{v}_x \times \mathbf{v}_y\end{aligned} \tag{2}$$

where $\mathbf{r}_i$ was the Cα positional vector of residue i and $\mathbf{v}_b$ was an auxiliary vector.

In addition, if the Cα distance between residues i and $i + 1$ was 2.7–3.3 Å, e.g., the *cis*-peptide bond in the case such as proline, statistical analysis of the histogram was performed separately. In the case where not enough statistical data were available for a particular bin, data from the most similar bins were assigned. For the first and last two residues, their local Cα conformations were assumed to be the same as those of the nearby residues so that their main chain atoms could be built as well (note, for most proteins, these residues were highly flexible). Also, to build the main chain atoms between the first Cα and second Cα atoms, local reference coordinate axes $\mathbf{v}'_x$, $\mathbf{v}'_y$, $\mathbf{v}'_z$ on the first Cα atom were defined as

$$\begin{aligned}\mathbf{v}'_b &= \frac{\mathbf{r}_3 - \mathbf{r}_2}{|\mathbf{r}_3 - \mathbf{r}_2|}\\ \mathbf{v}'_x &= \frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_2 - \mathbf{r}_1|}.\\ \mathbf{v}'_y &= \mathbf{v}'_x \times \mathbf{v}'_b\\ \mathbf{v}'_z &= \mathbf{v}'_x \times \mathbf{v}'_y\end{aligned} \tag{3}$$

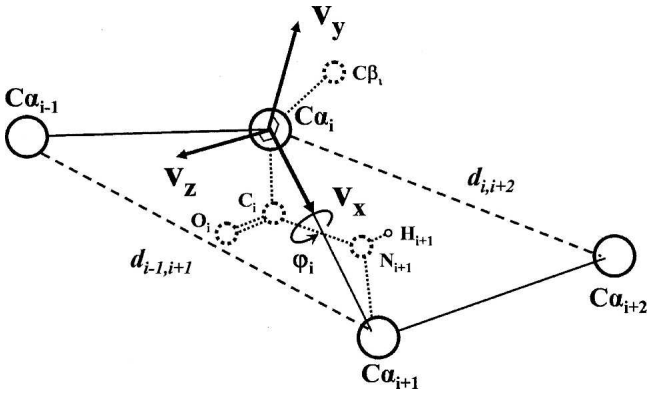

Figure 5. Schematic illustration of parameters used to build pseudo-main chain atoms and Cβ atoms from Cα positions. The three parameters are the Cα−Cα distance ($d_{i-1,i+1}$) between residues $i - 1$ and $i + 1$, the Cα−Cα distance ($d_{i,i+2}$) between residues i and $i + 2$, and the dihedral angle (φ_i) formed by all four Cα atoms. $\mathbf{v}_x$, $\mathbf{v}_y$, $\mathbf{v}_z$ are local Cartesian coordinate axes to build atoms.

In a real application, given the conformation of four consecutive Cα atoms, one would look up the corresponding bin based on the distances and dihedral parameters, and assign the main chain atoms coordinates extracted from the database. After establishing the positions of the main chain atoms, Cβ atoms could be built from the N, Cα, and C atoms according to the standard parameters: The bond length of the Cα−Cβ bond was 1.53Å, the bond angle of the N-Cα−Cβ angle was 110°, and the dihedral angle between plane N-Cα-C and plane Cα-C-Cβ was 124°.

Distance-dependent pairwise energy with orientational preference

The term $E_{pairwise}$ was the distance-dependent pairwise energy term. It had an orientational preference in such a way that cases in which the side chain of one residue points away from the partner residue and points toward the partner residue were distinguished. Specifically, a Cα to Cβ vector was used to represent the rough direction of the side chain. The distance-scaled finite ideal-gas reference state was used to normalize the statistical data (Zhou and Zhou 2002). For a pair of residues whose Cβ atoms were within the cutoff distance ($r_{cut} = 15Å$), the energy $E_{pairwise}$ for residues i with respect to residue j was given by

$$E_{pairwise}(A_i, A_j, O_{ij}, r_{ij}) = - RT \ln \frac{N_{obs}(A_i, A_j, O_{ij}, r_{ij})}{(r_{ij}/r_{cut})^{\alpha}(\Delta r_{ij}/\Delta r_{cut})N_{obs}(A_i, A_j, O_{ij}, r_{cut})}. \quad (4)$$

Here, A_i was the residue type, r_{ij} was the Cβ−Cβ distance between residues i and j, and Δr_{ij} and Δr_{cut} were the bin width at distance r_{ij} and r_{cut}. The constant R was the gas constant and T was temperature (both were set to 1 in practice). The total number of bins used in the study was 20. The bin width was 2 Å for $r_{ij} < 2$ Å, 0.5 Å for 2 Å $< r_{ij} <$ 8Å, and 1 Å for 8 Å $< r_{ij} <$ 15Å. The exponent α was 1.61. The term $N_{obs}(A_i, A_j, O_{ij}, r_{ij})$ gave the observed number of pairs of Cβ atoms at the designated distance in their respective orientation in the structural database. The symbol O_{ij} was expressed as

$$O_{ij} = \begin{cases} 1, & \mathbf{r}_{C_\alpha^i, C_\beta^i} \cdot \mathbf{r}_{C_\beta^i, C_\beta^j} > 0 \\ -1, & \mathbf{r}_{C_\alpha^i, C_\beta^i} \cdot \mathbf{r}_{C_\beta^i, C_\beta^j} < 0 \end{cases}, \quad (5)$$

where $\mathbf{r}_{\text{atom1,atom2}}$ was the displacement vector from atom 1 to atom 2. The symbol O_{ij} was used to distinguish the effect of the relative orientation of the two residues. If the value of O_{ij} was 1, residue i pointed toward residue j; if the value of O_{ij} was −1, residue i pointed away from residue j. Note this means that the case of i pointing toward j and the case of j pointing toward i can be different. Figure 6 schematically illustrates the two cases; in panel A, residue i points toward j, but residue j points away from i. In panel B, both residues point toward each other. Because of the normalization, the energy term $E_{pairwise}$ naturally decays to zero at cutoff distance r_{cut}. In the case of glycine, Cα atoms were used instead, and the effect of orientation was omitted.

Hydrogen bonding energy

The term E_{Hbond} was the main chain hydrogen bonding energy. It was developed first via statistical analysis of those residue pairs in the nonhomologous structure database based on an all-atom structure model. Then, for Cα models in real applications, the energy was computed based on the constructed pseudo-backbone atom positions.

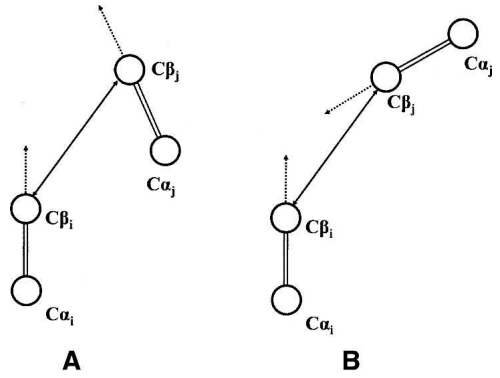

Figure 6. Schematic illustration of the different orientations of interacting Cβ pairs. (*A*) Residue i points toward j, but residue j points away from i. (*B*) Both residues point toward each other.

The hydrogen bonding criterion based on the all-atom structure model was from Fabiola et al. (2002),

$$r_{O,H} \le 2.7\,\text{Å},\ r_{O,N} \le 3.5\,\text{Å},\ \angle CON \ge 90°,\ \angle NHO \ge 90°. \quad (6)$$

Figure 7 illustrates the parameters. For a particular pair distance $r_{ij} = r_{C,N}$ and interaction angle $\theta_{ij} = \angle CON$, the total number of main chain hydrogen bonds was counted as $N(r_{ij},\theta_{ij})$ in a space region defined from (r_{ij},θ_{ij}) to $(r_{ij} + \Delta r_{ij},\ \theta_{ij} + \Delta\theta_{ij})$ with volume $V(r_{ij},\theta_{ij})$ (Δr_{ij} was 0.1 Å and $\Delta\theta_{ij}$ was $\pi/36$). This region was cylindrically symmetric with respect to the main chain carbonyl bond (it was assumed that the nitrogen atoms in hydrogen bonding interactions in this region were uniformly distributed). Then, the hydrogen bonding energy E_{Hbond} as a function of (r_{ij},θ_{ij}) was given by

$$E_{Hbond}(r_{ij}, \theta_{ij}) = - RT \ln \frac{V_{Total} N(r_{ij}, \theta_{ij})}{V(r_{ij}, \theta_{ij}) \sum_{r_{ij},\theta_{ij}} N(r_{ij}, \theta_{ij})} + RT \ln \frac{V_{Total} \times 1}{V\left(r_{ij}^{\max}, \pi/2\right) \sum_{r_{ij},\theta_{ij}} N(r_{ij}, \theta_{ij})}. \quad (7)$$

Here, V_{Total} was the volume of the entire search space defined as the spherical shell with r_{ij} in the range of 1.8–3.3 Å, $r_{ij}^{\max}$ was 4.8 Å, and θ_{ij} was in the range of $[0,\pi]$. The sum in the denominator gave the total number of hydrogen bonding pairs in the research region. The counting only applied to residues that were at least two residues apart in sequence. The second part of the energy term was a constant energy shift to eliminate the energy barrier during hydrogen bond formation. Note that in this study, proline was never considered a donor, and chain C termini were never considered as acceptors.

In real applications, main chain atoms (including hydrogen atoms) were built from Cα atoms first. It was found that the N and C atoms built from Cα atoms had ~0.1 Å RMSD from native positions, while O atoms had ~0.3–0.4 Å. In order to

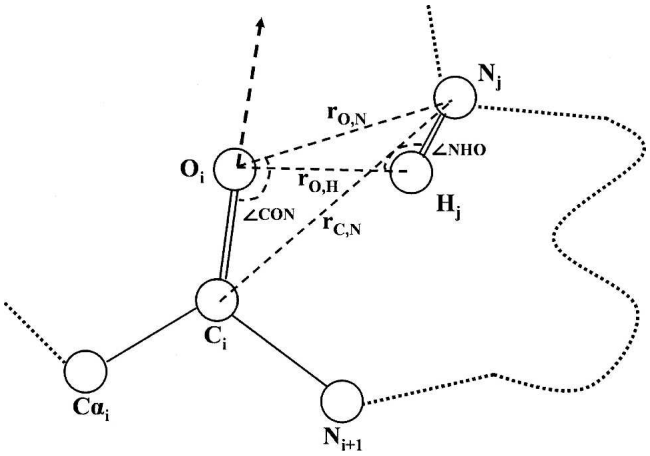

Figure 7. Schematic illustration of the parameters used in H-bond energy. $\angle CON$, $\angle NHO$, $r_{O,N}$, $r_{O,H}$ are used in hydrogen bonding criterion. $\angle CON$, $r_{C,N}$ are used as energy parameters.

avoid the wrong assignment of hydrogen bonds owning to the error of the estimated main chain, a modified criterion was used:

$$r_{O,H} \leq 2.8\,\text{Å},\ r_{O,N} \leq 3.7\,\text{Å},\ \angle CON \geq 90°,\ \angle NHO \geq 90°. \quad (8)$$

It was found that by using this criterion, ~91% of the hydrogen bonds identified by the old criterion in Equation 6 were found by the new criterion in Equation 8. Thus, the current only-Cα-based method could provide a reasonably close energy value to the all-atom main chain hydrogen bond energy.

Cα-based secondary structure assignment

Several terms in the energy function required the secondary structure assignment. The Cα trace alone does not allow one to accurately identify the secondary structure by methods such as the DSSP algorithm (Kabsch and Sander 1983), and the positions of main chain atoms built were pseudo-positions for auxiliary purposes; i.e., they were not accurate enough for regular secondary structure assignment. As indicated above, a modified definition of hydrogen bonds was used for DSSP analysis. Since, with the new definition, only 9% of hydrogen bonds were missed, it was expected that the accuracy of the secondary structure assignment on the sole Cα level would be reasonable. Only three types of secondary structure elements were used: α-helix, 10–3 helix, and π-helix were categorized as helix; extended sheet and β-bridge were categorized as sheet; others, such as loop and bend, were categorized as loop.

Short-range term

The term E_{short_range} is a short-range energy term. The conformation of each pentapeptide fragment was divided into discrete bins, associated with the sequence information. The correlation between the sequence and local secondary structure for each pentapeptide fragment was constructed and transferred into energy functions based on the statistical distribution extracted from the nonhomologous structure database. This short-range energy term presents the structural preference of local fragments.

The conformation of the Cα trace for a protein of N residues was thus defined by $3N - 6$ parameters: $N - 1$ pseudo-bonds connecting two neighboring Cα atoms, $N - 2$ pseudo-bond angles (θ) formed by three Cα atoms, and $N - 3$ pseudo-dihedral angles (φ) formed by four Cα atoms. All the degrees of freedom are illustrated in Figure 8. The energy function was expressed as:

$$E_{short\_range}(A_i;\theta_i,\varphi_{i-1},\varphi_i,S_{2nd}) = -RT\ln\frac{N_{obs}(A_i;\theta_i,\varphi_{i-1},\varphi_i,S_{2nd})/\sin\theta_i}{\left(\sum_{\theta_i,\varphi_{i-1},\varphi_i} N_{obs}(A_i;\theta_i,\varphi_{i-1},\varphi_i,S_{2nd})\right)\Big/\sum_{\theta_i}\sin\theta_i}. \quad (9)$$

Here, A_i was the residue type of the central residue in the pentapeptide (20-letter code) and S_{2nd} was the secondary structure type. The bond angle, which was from 0°–180°, was divided into six bins. The dihedral angle, which was from −180° to 180°, was divided into 24 bins.

Packing term

The term $E_{packing}$ was for pairwise packing energy related to the side chain orientation, residue type, and secondary structure. The packing energy can be expressed as a sum of six terms,

$$E_{packing} = E_{H\_self} + E_{S\_self} + E_{S\_pairing} + E_{H-H\_packing} + E_{H-S\_packing} + E_{S-S\_packing}. \quad (10)$$

The first term, E_{H_self}, was the helix self-packing energy. Side chain interactions within a helix have been analyzed previously (Stapley and Doig 1997; Adamian and Liang 2001; Andrew et al. 2001; Shi et al. 2002), and (i,i + 3), (i,i + 4) residue pairs play a significant role in stabilizing helix structure. Hence, the sequence propensity of such residue pairs in a helix was statistically analyzed in the structure database. The term E_{H_self} can be expressed as:

$$E_{H\_self}(A_i,A_j) = \begin{cases} -RT\ln\left(\dfrac{N_3^h(A_i,A_j)\sum_{A_i,A_j}N_3^{nh}(A_i,A_j)}{N_3^{nh}(A_i,A_j)\sum_{A_i,A_j}N_3^h(A_i,A_j)}\right), j=i+3 \\ -RT\ln\left(\dfrac{N_4^h(A_i,A_j)\sum_{A_i,A_j}N_4^{nh}(A_i,A_j)}{N_4^{nh}(A_i,A_j)\sum_{A_i,A_j}N_4^h(A_i,A_j)}\right), j=i+4 \end{cases}, \quad (11)$$

where $N_3^h(A_i, A_j)$ was the number of cases in which residue i of type A_i was three residues ahead in sequence of residue j of type

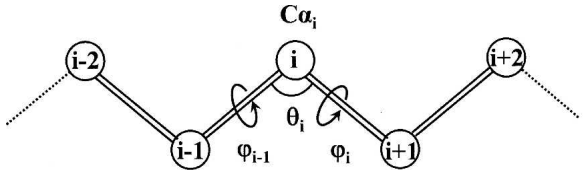

Figure 8. Schematic illustration of short-range parameters. θ is the pseudo-bond angle formed by three consecutive Cα atoms; φ is pseudo-dihedral angles formed by four consecutive Cα atoms.

A_j on the helix, $N_3^{nh}(A_i, A_j)$ was for the cases in which both residues were not in the helix, $N_4^{h}(A_i, A_j)$ was the number of cases in which residue i of type A_i was four residues ahead in sequence of residue j of type A_j in the helix, and $N_4^{nh}(A_i, A_j)$ was for the cases in which both residues were not on the helix.

The second term, E_{S_self}, is sheet self-packing energy, very similar to the first term. The sequence propensity of $(i,i+2)$ residue pairs in a sheet was statistically analyzed in the structure database. So,

$$E_{S\_self}(A_i, A_j) = -RT \ln \left(\frac{N_2^s(A_i, A_j) \sum_{A_i, A_j} N_2^{ns}(A_i, A_j)}{N_2^{ns}(A_i, A_j) \sum_{A_i, A_j} N_2^s(A_i, A_j)} \right), \; j = i+2, \tag{12}$$

where $N_2^s(A_i, A_j)$ was the number of cases in which residue i of type A_i was two residues ahead in sequence of residue j of type A_j on the strand, and $N_2^{ns}(A_i, A_j)$ was for the cases in which both residues were not on the strand.

The third term, $E_{S\_pairing}$, was the intrasheet strand–strand pairing energy. In this term, the sequence propensity of any interacting residue pair in both antiparallel and parallel cases was analyzed. Comparing with what has been reported in the literature (Hutchinson et al. 1998; Steward and Thornton 2002), a more complete set of interacting types for residue pairs was included. This energy term is very useful in determining the sequence register of pairing β-strands. Let T_{ij} denote the type of interacting residue pairs i, j. There were in total four types of interacting residue pairs for the antiparallel sheet (schematically shown in Fig. 9A): a hydrogen-bond-involving pair (type AA, $T_{ij} = (0, 0)$), a non-hydrogen-bond-involving pair (type aa, $T_{ij} = (1, 1)$), a hydrogen-bond-involving residue interacting with the next hydrogen-bond-involving residue on the opposite strand [type AB, $T_{ij} = (0,2)$], and a non-hydrogen-bond-involving residue interacting with the next non-hydrogen-bond-involving residue on the opposite strand [type ab, $T_{ij} = (1,3)$]. Similarly, there were three types of interacting residue pairs for the parallel sheet (schematically shown in Fig. 9B): a hydrogen-bond-involving residue interacting with a non-hydrogen-bond-involving residue [type Aa, $T_{ij} = (0,1)$], a hydrogen-bond-involving residue interacting with the next non-hydrogen-bond-involving residue on the opposite strand toward the C terminus [type Ab, $T_{ij} = (0,3)$], and a non-hydrogen-bond-involving residue interacting with a hydrogen-bond-involving residue on the opposite strand toward the C terminus [type aB, $T_{ij} = (3,0)$]. Note that the four types in the antiparallel sheet were symmetric, while the three types in the parallel sheet were asymmetric with respect to the direction of the polypeptide chain. The term $E_{S\_pairing}$ could be expressed as:

$$E_{S\_pairing}(A_i, A_j, T_{ij}) = -RT \ln \frac{N_{obs}(A_i, A_j, T_{ij}) \sum_{A_i, A_j} N_{obs}(A_i, A_j, T_{ij})}{\sum_{A_j} N_{obs}(A_i, A_j, T_{ij}) \sum_{A_i} N_{obs}(A_i, A_j, T_{ij})}, \tag{13}$$

where $N_{obs}(A_i, A_j, T_{ij})$ was the observed number of one specific interacting residue pair of type A_i and A_j.

The fourth term, $E_{H\_H\_packing}$, was the interhelix packing energy. The sequence propensity of the packing residue pairs in different helices was analyzed. The strategy was to define a Cα-based condition for interhelix packing, then to develop an energy term based on that. First, a residue-type-dependent cutoff distance, $d_{cut}^{hh}(A_i, A_j)$, was defined as the distance between the Cβ (or Cα in the case of glycine) of two residues for an interhelix interaction. To determine $d_{cut}^{hh}(A_i, A_j)$, the distances between the Cβ atoms (or Cα in the case of glycine) of two residues in different helices whose side chains had contacts was analyzed in the structure database. Two side chains were considered to have

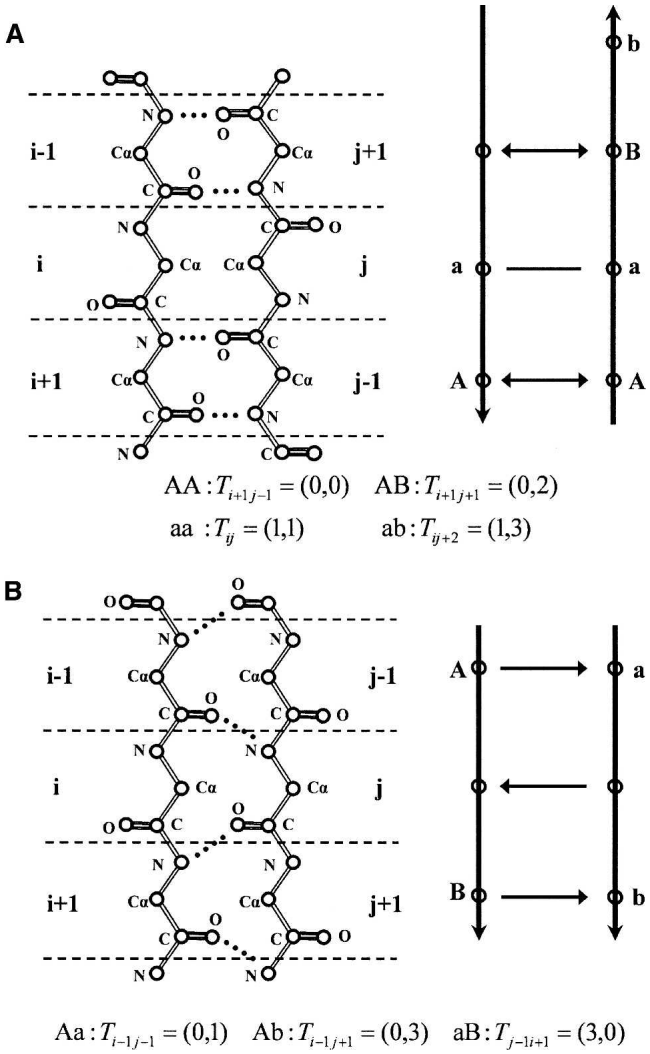

Figure 9. Seven types of interacting residue pairs in two pairing β-strands. (*A*) Four types of interacting residue pairs in antiparallel β-strands. (AA) A hydrogen-bond-involving pair [$T_{i+1j-1} = (0,0)$; note for illustration purposes that the subscripts of T_{pq} are based on the diagram in the figure]; (aa) a non-hydrogen-bond-involving pair [$T_{ij} = (1,1)$]; (AB) a hydrogen-bond-involving residue interacting with the next hydrogen-bond-involving residue on the opposite strand [$T_{i+1j+1} = (0,2)$]; (ab) a non-hydrogen-bond-involving residue interacting with the next non-hydrogen-bond-involving residue on the opposite strand [$T_{ij+2} = (1,3)$]. (*B*) Three types of interacting residue pairs in parallel β-strands. (Aa) A hydrogen-bond-involving residue interacting with a non-hydrogen-bond-involving residue [$T_{i-1j-1} = (0,1)$]; (Ab) a hydrogen-bond-involving residue interacting with the next non-hydrogen-bond-involving residue on the opposite strand toward the C terminus [$T_{i-1j+1} = (0,3)$]; (aB) a non-hydrogen-bond-involving residue interacting with the next hydrogen-bond-involving residue on the opposite strand toward the C terminus [$T_{j-1i+1} = (3,0)$].

contacts if the distance between two atoms from each side chain was <5 Å. The cutoff distance $d_{cut}^{hh}(A_i, A_j)$ was chosen to include most of the contacting residue pairs while having reasonably low false positives, and its value was kept in a lookup table. The interhelix packing criterion based on the pseudo-Cβ position was:

$$\left|\mathbf{r}_{C_\beta^i, C_\beta^j}\right| \le 6\text{Å} \quad \text{or} \quad \left|\mathbf{r}_{C_\beta^i, C_\beta^j}\right| \le d_{cut}^{hh}(A_i, A_j) \quad \text{and} \quad \mathbf{r}_{C_\alpha^i, C_\beta^i} \cdot \mathbf{r}_{C_\beta^i, C_\beta^j} \ge 0. \tag{14}$$

The energy term can be expressed as:

$$E_{H-H\_packing}(A_i, A_j) = -RT \ln \frac{N_{obs}(A_i, A_j) \sum_{A_i} N^h(A_i) \sum_{A_j} N^h(A_j)}{N^h(A_i) N^h(A_j) \sum_{A_i, A_j} N_{obs}(A_i, A_j)}, \tag{15}$$

where $N_{obs}(A_i, A_j)$ was the observed number of interacting packing pairs in a helix of residues of type A_i and A_j, and $N^h(A_i)$ was the total number of residues of type A_i in the helix.

The fifth term, $E_{H-S\_packing}$, was the helix-strand packing energy. The helix-sheet packing criterion was almost the same as the interhelix packing criterion in Equation 14, except that the cutoff distance $d_{cut}^{hs}(A_i, A_j)$ for helix-strand packing was extracted from the structure database. The energy term can be expressed as:

$$E_{H-S\_packing}(A_i, A_j) = -RT \ln \frac{N_{obs}(A_i, A_j) \sum_{A_i} N^h(A_i) \sum_{A_j} N^s(A_j)}{N^h(A_i) N^s(A_j) \sum_{A_i, A_j} N_{obs}(A_i, A_j)}, \tag{16}$$

where $N_{obs}(A_i, A_j)$ was the observed number of interacting packing pairs for two residues of type A_i in a helix and A_j in a sheet, $N^h(A_i)$ was the total number of residue of type A_i in a helix, and $N^s(A_j)$ was the total number of residues of type A_j in a sheet.

The sixth term, $E_{S-S\_packing}$, was the intersheet strand–strand packing energy. Differing from the third term, this term represented the sequence propensity of packing residue pairs in different β-sheets. The criterion of packing residue pairs in strand–strand packing was:

$$\mathbf{r}_{C_\alpha^i, C_\beta^i} \cdot \mathbf{r}_{C_\alpha^i, C_\beta^j} \le 0 \quad \text{and} \quad \left|\mathbf{r}_{C_\beta^i, C_\beta^j}\right| \le d_{cut}^{ss}(A_i, A_j) \tag{17}$$

where $d_{cut}^{ss}(A_i, A_j)$ was the cutoff distance extracted from the structure database. Note that the residue pairs belonging to two contacting strands in the same sheet were excluded. The energy term can be expressed as:

$$E_{S-S\_packing}(A_i, A_j) = -RT \ln \frac{N_{obs}(A_i, A_j) \sum_{A_i} N^s(A_i) \sum_{A_j} N^s(A_j)}{N^s(A_i) N^s(A_j) \sum_{A_i, A_j} N_{obs}(A_i, A_j)}, \tag{18}$$

where $N_{obs}(A_i, A_j)$ was the observed number of interacting packing pairs in a sheet for two residues of type A_i, A_j, and $N^s(A_i)$ was the total number of residues of type A_i in a sheet.

Tri-peptide packing term

The term $E_{tri-lpeptide}$ was for the tri-peptide energy, defined as the contact energy of two specific tri-peptides with corresponding secondary structure types. The amino acids were grouped into four categories based on their physicochemical properties and sizes: (Asp, Glu, Lys, Arg, His), (Ser, Thr, Asn, Gln), (Gly, Ala, Val, Cys, Met), and (Ile, Leu, Pro, Phe, Tyr, Trp). Three types of secondary structure, α-helix, β-strand, and loop, were used. Therefore, there was a total of $64 \times 3 = 192$ different types of tri-peptides, in which $64 = 4 \times 4 \times 4$ was for the coarse-grained residue types, and three was for the secondary structure types. The tertiary packing potential was given by

$$E_{tri-peptide}(T_i, S_i; T_j, S_j) = -RT \ln \frac{N_{obs}(T_i, S_i; T_j, S_j)}{N_{obs}(S_i; S_j) \times \chi(T_i) \times \chi(T_j)}. \tag{19}$$

Here, T_i was for the ith tri-peptide, S_i was for the secondary structure type of that tri-peptide, and $\chi(T_i)$ was the mole fraction of tri-peptide i extracted from the structural database. Also, N_{obs} was the observed number of contact pairs in the structural database: $N_{obs}(T_i,S_i;T_j,S_j)$ was for the contacts between tri-peptides and $N_{obs}(S_i;S_j)$ was for the contacts between two secondary structural elements defined as a pair of secondary structural elements with at least one pair of tri-peptide contacts. To define a contact between two tri-peptides, a 3×3 distance matrix D_{ij} was constructed for the pair, in which the element $d_{ki,lj}$ of the matrix gave the distance between the kth residue in tri-peptide i and the lth residue in tri-peptide j. Two tri-peptides were regarded as being in contact if more than five elements of their 3×3 distance matrix were within the cutoff distance, which was set to 5 Å for a strand–strand contact, 10 Å for a helix–helix contact, and 12 Å for all other contacts.

Three-body term

The term $E_{3\text{-}body}$ was a three-body energy for including the multi-body effect. A triplet of residues was defined as three residues (not nearest neighbor in sequence) with their Cβ atoms in long-range contact (defined as all three pair distances smaller than a cut-off distance $r_c < 7.5$ Å). All the triplets in the nonredundant structure database were recorded. The energy term for a residue triplet (type A_i, A_j, and A_k) was given by

$$E_{3-body}(A_i, A_j, A_k) = \begin{cases} -RT \ln \dfrac{N_{obs}(A_i, A_j, A_k)}{(\sum_{i,j,k} N_{obs}(A_i, A_j, A_k)) \chi(A_i) \chi(A_j) \chi(A_k) C}, & r_{ij}, r_{jk}, r_{ik} \le r_c \\ 0, & otherwise \end{cases}, \tag{20}$$

where $N_{obs}(A_i, A_j, A_k)$ was the number of triplets of type (A_i, A_j, A_k) extracted from the database, and $\chi(A_i)$ was the mole fraction of the residue type A_i. C was a factor defined as:

$$C = \frac{3!}{\prod_{v=1}^{\eta} t_v!}. \quad (21)$$

Here, η was the number of distinct residue types in the triplet ($1 \le \eta \le 3$), and t_v was the number of residues of type v in the triplet.

Solvation energy based on the solvent-accessible surface

The term $E_{solvation}$ was for the solvation energy based on the solvent-accessible surface (SAS). It was developed via statistical analysis of the side chain solvent-accessible surface in the nonhomologous structure database based on the all-atom structure model. For the Cα model, an approximate method for calculating the side chain SAS was developed, and the solvation energy was evaluated accordingly.

An approximate method was developed to calculate the side chain SAS from Cα atoms. Here, $r_{res}(i)$ was defined as the effective radius of the whole residue i, and $S_{SC}(i)$ was defined as the effective total solvent-accessible surface for the side chain of residue i. $S_{SC}(i)$ could be expressed by $S_{SC}(i) = 4\pi(r_{SC}(i) + r_{H_2O})^2$; $r_{SC}(i)$ was the effective radius for the side chain of residue i, and r_{H_2O} was the radius of a water molecule (set to 1.4 Å). Also, $d_{SC}(i)$ was defined as the distance between the Cα atom and effective side chain center of residue i, and the positional vector of the side chain center was defined as $\mathbf{R}_{SC}(i)$. It was assumed that the effective side chain center was always along the Cα to Cβ direction (Fig. 10). So, $\mathbf{R}_{SC}(i) = \mathbf{R}_{C\alpha}(i) + d_{SC}(i)\hat{\mathbf{r}}_{C_\alpha^i, C_\beta^i}$, where $\mathbf{R}_{C\alpha}(i)$ was the Cα positional vector of residue i, and $\hat{\mathbf{r}}_{C_\alpha^i, C_\beta^i}$ was the unit vector along the Cα to Cβ direction. The side chain SAS for residue i could then be calculated by:

$$\begin{aligned}
SAS(i) &= A_i - B_i \qquad (SAS(i) = 0 \quad \text{if } A_i < B_i) \\
A_i &= S_{SC}(i) \prod_j \left(1 - \frac{b_i^j - b_i^{'j}}{S_{SC}(i)}\right) \\
B_i &= \sum_j b_i^{'j} \\
r_{i,j}^{cut} &= r_{res}(j) + r_{SC}(i) + 2r_{H_20} \\
D_{i,j} &= \|\mathbf{R}_{SC}(i) - \mathbf{R}_{C\alpha}(j)\|
\end{aligned}$$

$$b_i^j = \begin{cases} 0, & D_{i,j} > r_{i,j}^{cut} \\ \pi \cdot (r_{SC}(i) + r_{H_20})(r_{i,j}^{cut} - D_{i,j})\left(1 + \dfrac{r_{res}(j) - r_{SC}(i)}{D_{i,j}}\right), & D_{i,j} < r_{i,j}^{cut} \end{cases}$$

$$b_i^{'j} = \begin{cases} 0, & D_{i,j} > r_{i,j}^{cut} \\ \max\left(0, \pi \cdot (r_{SC}(i) + r_{H_20})(r_{i,j}^{cut} - D_{i,j} - s)\left(1 + \dfrac{r_{res}(j) - r_{SC}(i) - s}{D_{i,j}}\right)\right), & D_{i,j} < r_{i,j}^{cut}, \end{cases} \quad (22)$$

where $s = 2.5$ Å according to the literature (Wodak and Janin 1980). Here, there were 20 of $r_{res}(i)$, 19 of $S_{SC}(i)$, and 19 of $d_{SC}(i)$ as all the parameters in the SAS calculation.

In order to obtain the side chain SAS accurately for the coarse-grained model, all 58 parameters were trained against the atomic side chain SAS. The atomic side chain SAS was calculated based on look-up table methods (Bystroff 2002) and they were regarded as expected values. Then, a simulated annealing Monte Carlo simulation was used to optimize parameters according to the following target function on a set of 392 protein chains, which were selected from the nonhomologous structure database whose total number of residues range from 60 to 150 and there were no chain break, heteroatoms,, and missing atoms. The target function was:

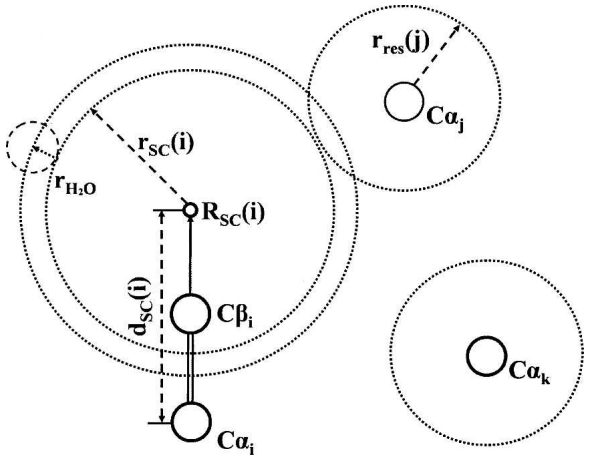

Figure 10. Schematic illustration of the parameters of the side-chain SAS from the Cα position. For residue i, $r_{res}(i)$ is the effective radius of the whole residue, $r_{SC}(i)$ is the effective radius for the side chain, $d_{SC}(i)$ is the distance between the Cα atom and effective side chain center, and $\mathbf{R}_{SC}(i)$ is the positional vector of the side chain center. r_{H_2O} is the radius of a water molecule (1.4 Å).

$$\Delta = \frac{\sum_{training\ set} \sum_{i=1}^{N_{tot}} |\zeta(i) - \zeta^{\exp}(i)|}{\sum_{training\ set} N_{tot}}, \quad (23)$$

where N_{tot} was the total number of residues of a protein chain, and $\zeta(i)$, $\zeta^{\exp}(i)$ were the coarse-grained and expected fraction of solvent-accessible surface, respectively, which can be defined as the ratio between the side chain solvent-accessible surface and the total side chain surface area of that residue in isolation with the same configuration; i.e.,

$$\begin{aligned}
\zeta(i) &= \frac{SAS(i)}{S_{SC}(i)} \\
\zeta^{\exp}(i) &= \frac{SAS^{\exp}(i)}{S_{SC}^{\exp}(i)}
\end{aligned} \quad (24)$$

The optimized parameters can be found in Table 4, and the best target function value after optimization was 0.0917, indicating the existence of small error.

Finally, the energy term was related to the fraction of solvent-accessible surface ζ, which had a value between [0,1], and was uniformly divided into $n_{bins} = 20$ bins. It was given by

$$E_{solvent}(A_i, \zeta^{\exp}(i)) = -RT \ln \frac{n_{bins} N(A_i, \zeta^{\exp}(i))}{\sum_{bins} N(A_i, \zeta^{\exp}(i))}. \quad (25)$$

Here, A_i was the residue type of the target residue, and $N(A_i, \zeta^{\exp}(i))$ was the observed number of occurrences of residue type A_i.

Table 4. *Values of 58 optimized parameters for determining the SAS of side chains*

	Residue type	r_{res} (Å)	S_{SC} (Å^2)	d_{SC} (Å)
1	GLY	1.900	—	—
2	ALA	2.108	330.853	0.892
3	VAL	2.871	484.497	0.655
4	ILE	2.910	548.097	0.650
5	LEU	2.785	540.512	0.680
6	SER	2.252	302.490	0.807
7	THR	2.594	369.292	0.759
8	ASP	2.619	421.927	0.784
9	ASN	2.482	372.825	0.815
10	GLU	2.657	523.778	0.846
11	GLN	2.587	459.227	0.869
12	LYS	2.612	637.273	0.907
13	ARG	2.813	565.030	0.889
14	CYS	2.521	360.000	0.512
15	MET	2.691	512.042	0.726
16	PHE	3.112	560.370	0.574
17	TYR	3.072	445.473	0.793
18	TRP	3.370	678.563	0.496
19	HIS	2.723	433.573	0.810
20	PRO	2.800	422.424	0.668

When using the solvent energy term, ζ from the coarse-grained side chain SAS was used as the approximation of atomic value.

Weight optimization

Weights were optimized against all proteins in the LKF and Decoys'R'Us decoy set collections. To ease the decoy set dependence of weight optimization, these decoy sets were regrouped into three subsets based on the literature: The first subset consisted of 25 proteins in the Decoys'R'Us sets (Tobi and Elber 2000) (Subset-1 in Table 2), the second group consisted of 151 proteins in the LKF set (Loose et al. 2004; Zhang et al. 2006) (Subset-2 in Table 2), the third group consisted of the remaining 34 proteins in the LKF set and seven proteins in the Decoys'R'Us sets (Subset-3 in Table 2). The seven proteins from Decoys'R'Us in Subset-3 were 3icb in the 4state_reduced decoy set, 4icb in the fisa decoy set, 1eh2 and smd3 in the fisa_casp3 decoy set, 1beo and 4icb in the lattice_ssfit decoy set, and 4pti in the lmds decoy set.

In this study, an iterative protocol of Monte Carlo-simulated annealing was used on the three subsets of decoy collections. The cost function for optimization was:

$$F = \bar{z} + 1.5N_{\text{missing}}, \tag{26}$$

where $\bar{z}$ was the average Z-score for all proteins in the group and N_{missing} was the number of proteins whose native structures failed to be ranked first in energy. The Z-score of the native structure was defined as:

$$z = \frac{E_{tot}^{native} - \bar{E}_{tot}}{\sigma(E_{tot})}, \tag{27}$$

where E_{tot}^{native} and E_{tot} were the energy of the native and decoy structures for a particular protein, respectively, $\bar{E}_{tot}$ and $\sigma(E_{tot})$ were the average and standard deviation of energy of all decoys for a particular protein. The temperature factor k_BT in the Monte Carlo simulation was decreased gradually during simulated annealing from 1.0 to 0.01 in 19,800 steps. Then, k_BT was set to zero in Metropolis sampling, and the scoring function in Equation 26 was minimized for another 200 steps. The simulation started with predefined initial weights. Then, a randomly selected weight was increased or decreased by 0.1 in each Monte Carlo move, if the weight was within the predefined allowed range (see below for more details).

In detail, simulated annealing optimization was first performed on a randomly picked decoy subset with randomly assigned weights to obtain a set of optimized weights. Then, those weights were set as the new initial values for another round of simulated annealing optimization on a different decoy subset picked randomly. Optimizations were repeated among three decoy subsets 100 times. To make the simulation converge, the percentage changes of each weight with respect to the initial weight were restricted. In each round of simulation, the allowed percentage changes gradually decreased from 300% to a minimal 20%. However, to prevent the weights from being trapped at zero, the absolute allowed changes were no less than 0.5. According to weight optimization, weights finally converged. But the overall performance was not necessarily the best for the weight in the last step of annealing. So, we selected the best performing weights from the last few annealing steps.

Acknowledgments

M.C. and M.L. are partially supported by a predoctoral fellowship from the W.M. Keck Foundation of the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology. Y.W. is partially supported by a grant from the Doer Foundation. J.M. acknowledges support from a grant from the National Institutes of Health (R01-GM067801).

References

Adamian, L. and Liang, J. 2001. Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.* **311:** 891–907.

Andrew, C.D., Penel, S., Jones, G.R., and Doig, A.J. 2001. Stabilizing nonpolar/polar side-chain interactions in the α-helix. *Proteins* **45:** 449–455.

Bahar, I. and Jernigan, R.L. 1997. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* **266:** 195–214.

Betancourt, M.R. and Thirumalai, D. 1999. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **8:** 361–369.

Buchete, N.V., Straub, J.E., and Thirumalai, D. 2004a. Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.* **14:** 225–232.

Buchete, N.V., Straub, J.E., and Thirumalai, D. 2004b. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.* **13:** 862–874.

Bystroff, C. 2002. MASKER: Improved solvent-excluded molecular surface area estimations using Boolean masks. *Protein Eng.* **15:** 959–965.

Colubri, A., Jha, A.K., Shen, M.Y., Sali, A., Berry, R.S., Sosnick, T.R., and Freed, K.F. 2006. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. *J. Mol. Biol.* **363:** 835–857.

DeBolt, S.E. and Skolnick, J. 1996. Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: Atomic burial position and pairwise non-bonded interactions. *Protein Eng.* **9:** 637–655.

Dehouck, Y., Gilis, D., and Rooman, M. 2006. A new generation of statistical potentials for proteins. *Biophys. J.* **90:** 4010–4017.

Dobson, C.M. and Karplus, M. 1999. The fundamentals of protein folding: Bringing together theory and experiment. *Curr. Opin. Struct. Biol.* **9:** 92–101.

Dong, Q., Wang, X., and Lin, L. 2006. Novel knowledge-based mean force potential at the profile level. *BMC Bioinformatics* **7:** 324.

Eisenberg, D., Luthy, R., and Bowie, J.U. 1997. VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods Enzymol.* **277:** 396–404.
Fabiola, F., Bertram, R., Korostelev, A., and Chapman, M.S. 2002. An improved hydrogen bond potential: Impact on medium resolution protein structures. *Protein Sci.* **11:** 1415–1423.
Feng, Y., Kloczkowski, A., and Jernigan, R.L. 2007. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins* doi: 10.1002/prot.21362.
Fidelis, K., Stern, P.S., Bacon, D., and Moult, J. 1994. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* **7:** 953–960.
Gatchell, D.W., Dennis, S., and Vajda, S. 2000. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* **41:** 518–534.
Godzik, A., Kolinski, A., and Skolnick, J. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* **4:** 2107–2117.
Gohlke, H. and Klebe, G. 2001. Statistical potentials and scoring functions applied to protein–ligand binding. *Curr. Opin. Struct. Biol.* **11:** 231–235.
Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M.J. 1990. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216:** 167–180.
Hinds, D.A. and Levitt, M. 1992. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci.* **89:** 2536–2540.
Hubner, I.A., Deeds, E.J., and Shakhnovich, E.I. 2005. High-resolution protein folding with a transferable potential. *Proc. Natl. Acad. Sci.* **102:** 18914–18919.
Hutchinson, E.G., Sessions, R.B., Thornton, J.M., and Woolfson, D.N. 1998. Determinants of strand register in antiparallel β-sheets of proteins. *Protein Sci.* **7:** 2287–2300.
Jernigan, R.L. and Bahar, I. 1996. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6:** 195–209.
Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. A new approach to protein fold recognition. *Nature* **358:** 86–89.
Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure—Pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577–2637.
Keasar, C. and Levitt, M. 2003. A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* **329:** 159–174.
Lazaridis, T. and Karplus, M. 2000. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10:** 139–145.
Liwo, A., Lee, J., Ripoll, D.R., Pillardy, J., and Scheraga, H.A. 1999. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci.* **96:** 5482–5485.
Loose, C., Klepeis, J.L., and Floudas, C.A. 2004. A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins* **54:** 303–314.
Lu, H. and Skolnick, J. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **44:** 223–232.
MacKerell, A.D., Bashford Jr., D., Bellott, M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem.* **B102:** 3586–3616.
McConkey, B.J., Sobolev, V., and Edelman, M. 2003. Discrimination of native protein structures using atom–atom contact scoring. *Proc. Natl. Acad. Sci.* **100:** 3215–3220.
Meller, J. and Elber, R. 2002. Protein recognition by sequence-to-structure fitness: Bridging efficiency and capacity of threading models. *Adv. Chem. Phys.* **120:** 77–130.
Melo, F. and Feytmans, E. 1998. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **277:** 1141–1152.
Melo, F., Sanchez, R., and Sali, A. 2002. Statistical potentials for fold assessment. *Protein Sci.* **11:** 430–448.
Milik, M., Kolinski, A., and Skolnick, J. 1997. Algorithm for rapid reconstruction of protein backbone from α carbon coordinates. *J. Comput. Chem.* **18:** 80–85.
Miyazawa, S. and Jernigan, R.L. 1985. Estimation of effective interresidue contact energies from protein crystal-structures—Quasi-chemical approximation. *Macromolecules* **18:** 534–552.
Miyazawa, S. and Jernigan, R.L. 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256:** 623–644.
Miyazawa, S. and Jernigan, R.L. 2005. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? *J. Chem. Phys.* doi: 10.1063/1.1824012.
Moult, J. 1997. Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **7:** 194–199.
Park, B. and Levitt, M. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258:** 367–392.
Poole, A.M. and Ranganathan, R. 2006. Knowledge-based potentials in protein design. *Curr. Opin. Struct. Biol.* **16:** 508–513.
Qiu, J. and Elber, R. 2005. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins* **61:** 44–55.
Rajgaria, R., McAllister, S.R., and Floudas, C.A. 2006. A novel high resolution Cα–Cα distance dependent force field based on a high quality decoy set. *Proteins* **65:** 726–741.
Russ, W.P. and Ranganathan, R. 2002. Knowledge-based potential functions in protein design. *Curr. Opin. Struct. Biol.* **12:** 447–452.
Samudrala, R. and Moult, J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275:** 895–916.
Samudrala, R., Xia, Y., Levitt, M., and Huang, E.S. 1999. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pacific Symposium on Biocomputing* **4:** 505–516.
Shen, M.Y. and Sali, A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15:** 2507–2524.
Shi, Z., Olson, C.A., and Kallenbach, N.R. 2002. Cation-π interaction in model α-helical peptides. *J. Am. Chem. Soc.* **124:** 3284–3291.
Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268:** 209–225.
Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34:** 82–95.
Sippl, M.J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213:** 859–883.
Sippl, M.J. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5:** 229–235.
Skolnick, J. 2006. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **16:** 166–171.
Stapley, B.J. and Doig, A.J. 1997. Hydrogen bonding interactions between glutamine and asparagine in α-helical peptides. *J. Mol. Biol.* **272:** 465–473.
Steward, R.E. and Thornton, J.M. 2002. Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. *Proteins* **48:** 178–191.
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.
Thomas, P.D. and Dill, K.A. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **257:** 457–469.
Tobi, D. and Elber, R. 2000. Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins* **41:** 40–46.
Wang, G. and Dunbrack Jr., R.L. 2003. PISCES: A protein sequence culling server. *Bioinformatics* **19:** 1589–1591.
Wodak, S.J. and Janin, J. 1980. Analytical approximation to the accessible surface-area of proteins. *Proc. Natl. Acad. Sci. USA* **77:** 1736–1740.
Wu, Y., Chen, M., Lu, M., Wang, Q., and Ma, J. 2005a. Determining protein topology from skeletons of secondary structures. *J. Mol. Biol.* **350:** 571–586.
Wu, Y., Tian, X., Lu, M., Chen, M., Wang, Q., and Ma, J. 2005b. Folding of small helical proteins assisted by small-angle X-ray scattering profiles. *Structure* **13:** 1587–1597.
Xia, Y., Huang, E.S., Levitt, M., and Samudrala, R. 2000. Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* **300:** 171–185.
Zhang, C., Vasmatzis, G., Cornette, J.L., and DeLisi, C. 1997. Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.* **267:** 707–726.
Zhang, Y., Kolinski, A., and Skolnick, J. 2003. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.* **85:** 1145–1164.
Zhang, C., Liu, S., Zhou, H., and Zhou, Y. 2004. 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.
Zhang, J., Chen, R., and Liang, J. 2006. Empirical potential function for simplified protein models: Combining contact and local sequence-structure descriptors. *Proteins* **63:** 949–960.
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.
Zhou, Y., Zhou, H., Zhang, C., and Liu, S. 2006. What is a desirable statistical energy function for proteins and how can it be obtained? *Cell Biochem. Biophys.* **46:** 165–174.