Environment-dependent residue contact energies for proteins

Chao Zhang and Sung-Hou Kim*

Department of Chemistry and E. O. Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720

Contributed by Sung-Hou Kim, December 27, 1999

We examine the interactions between amino acid residues in the context of their secondary structural environments (helix, strand, and coil) in proteins. Effective contact energies for an expanded 60-residue alphabet (20 aa × three secondary structural states) are estimated from the residue–residue contacts observed in known protein structures. Similar to the prototypical contact energies for 20 aa, the newly derived energy parameters reflect mainly the hydrophobic interactions; however, the relative strength of such interactions shows a strong dependence on the secondary structural environment, with nonlocal interactions in β-sheet structures and α-helical structures dominating the energy table. Environment-dependent residue contact energies outperform existing residue pair potentials in both threading and three-dimensional contact prediction tests and should be generally applicable to protein structure prediction.

The main obstacle for protein structure prediction comes from the immense number of possible conformations accessible to a polypeptide chain and the complexity and variety of interactions involved in the folding process. This obstacle makes it a formidable task to carry out protein folding simulations using an all-atom representation of the polypeptide. Coarse-grained protein models, which treat amino acid residues as united interaction sites, offer a more practical approach to tackling the protein folding problem (1). Because such models omit many detailed features of atomic interactions, new energy parameters suitable for representing interactions at low level of resolution are required. These parameters often are derived empirically from the analysis of experimentally observed residue contact preferences in sets of known structures (2). Whether such knowledge-based potentials correctly reflect the actual physical forces stablizing the native structures of proteins remains a subject of debate (3–6). Nevertheless, structure-derived potentials have contributed substantially to the current theoretical studies of protein folding (7).

Over the years, many specifically designed residue pair potentials have been proposed (2, 7, 8), most of which use the 20-aa alphabet and assume a constant energy contribution for a given residue pair regardless of its local structural environment. However, it is known that the folded conformation of a protein is stabilized by a multitude of weak noncovalent interactions and the contribution of each of these interactions depends on its context within the folded structure (9–11). To derive better, more specific potential energy functions, one has to take into account the influence of varied structural circumstances on the specificity of inter-residue interactions. In this study, we examine pairwise amino acid interactions in the context of secondary structural environments (helix, strand, and coil) and report a set of residue contact energies for an expanded 60-residue alphabet (20 aa × three secondary structural states).

The discriminatory capability of the environment-dependent contact energies (ERCE) was tested first in the threading experiments where the secondary structural states of the template protein structures are experimentally determined. Though only the simple nongapped threading protocol was used, the improvement over existing residue pair potentials was already evident as ERCE correctly recognized the native structures for more than 97% of the testing proteins. The applicability of ERCE is greatly extended when combined with secondary structure prediction. The new generation of algorithms for secondary structure prediction benefits from using evolutionary information provided by multiple sequence alignment (12–14). Recently developed structure prediction methods, including both *ab initio* methods (15–17) and fold recognition-based methods (18–23), all used predicted secondary structures and have achieved an important degree of success. We compared ERCE with several existing residue pair potentials in predicting residue contacts in the native protein structure from the amino acid sequence. A test based on a large number of distinct protein domains showed that the use of ERCE in combination of predicted secondary structures led to a significant improvement in the accuracy of contact prediction. This result has broader implications for protein structure prediction, because an energy function that favors conformations with higher percentage of native contacts has a better chance to guide global optimization toward the native folded state (24–26).

Methods

Protein Structure Set. A set of representative protein domain structures were extracted from the Protein Date Bank (27) by referring to the SCOP database (version 1.37) (28). Specifically, we started with the $<40\%$ identity set built by the authors of SCOP and then performed additional sorting. First, we removed composite domains and domains with fewer than 50 residues or more than 300 residues. A single member then was selected from each SCOP family in the four structural classes, all-α, all-β, α/β, and $\alpha+\beta$; wherever applicable, structures that had been determined to the highest crystallographic resolution were chosen. Domains with only C^{α}-traces or containing large sequence gaps were excluded. The structures thus obtained were inspected at the SCOP superfamily level. If a superfamily had multiple representatives, those structures with resolution lower than 2.30 Å were deleted. Domains that do not form a compact globular structure (e.g., long α-helical coiled-coils) also were removed. The remaining protein domains were examined by pairwise sequence alignment; when the sequence identity between a pair was higher than 25%, only one structure, typically the one with higher resolution, was kept. The final data set contains 407 protein domains; of those, 80 are all-α, 105 are all-β, 91 are α/β, and 131 are $\alpha+\beta$ (the complete list is available from the authors on request).

The Definition of Secondary Structural States. The experimentally determined (real) secondary structural states (helix, strand, and

Abbreviations: ERCE, environment-dependent residue contact energies; MJ, Miyazawa–Jernigan procedure; JS, J. Skolnick *et al*. potential.

*To whom reprint requests should be addressed at: 220 Melvin Calvin Laboratory, University of California, Berkeley, CA 94720-5230. E-mail: SHKim@cchem.berkeley.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.040573597. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.040573597

coil) of residues in these structures were extracted from the DSSP database (29). A simple mapping scheme was used where only H states in DSSP were mapped to helix, E states mapped to strand, and all other states mapped to coil. The predicted secondary structural states were obtained by running the PSI-PRED program (14). This program uses the position-specific scoring matrices generated by PSI-BLAST (30) as input. To prepare for these matrices, we performed a PSI-BLAST search for each of the 407 protein domains against a nonredundant sequence database NRDB90 (31). The default E-value cutoff (0.0001) was used, and the maximum number of iterations was set to seven. The overall accuracy of secondary structure prediction for 407 protein domains is 79.5%.

Extracting Contact Energies from Known Protein Structures. The combination of 20 aa and three secondary structural states gives rise to 60 residue types. We applied the Miyazawa–Jernigan procedure (MJ) (32) with a different definition of the reference state (33) to derive contact energies for this expanded residue alphabet from the 407 protein domain structures. Specifically, amino acid residues were represented by the centroids of their side chains, and two residues were considered to be in contact if the distance between their centroids fell within $R_C = 6.5$ Å. The numbers of contacts formed between two residues, i and j, and between them and the solvent molecules (represented by 0) were related to the contact energy by a hypothetical chemical reaction

$$(i - 0) + (j - 0) \rightleftharpoons (i - j) + (0 - 0)\,. \qquad [1]$$

The effective contact energy (e_{ij}) is defined as the negative logarithm of the equilibrium constant of the reaction (32)

$$e_{ij} = -\ln\left(\frac{\bar{n}_{ij}\bar{n}_{00}}{\bar{n}_{i0}\bar{n}_{j0}}\right), \qquad [2]$$

where $\bar{n}_{ij}$, $2\bar{n}_{i0}$, $2\bar{n}_{j0}$, and $\bar{n}_{00}$ are the number of residue i-residue j contacts, the number of residue i-solvent contacts, the number of residue j-solvent contacts, and the number of solvent-solvent contacts, respectively. In practice, e_{ij} was estimated by

$$e_{ij} = -\ln\left(\frac{N_{ij}N_{00}}{N_{i0}N_{j0}}\frac{C_{i0}C_{j0}}{C_{ij}C_{00}}\right) \qquad [3]$$

to minimize the bias from amino acid compositional heterogeneity and polypeptide chain connectivity, where N_{ij}, N_{i0}, N_{j0}, and N_{00} are the contact numbers observed in known structures, and C_{ij}, C_{i0}, C_{j0}, and C_{00} are the corresponding quantities expected in a reference state.

The calculations of $N_{ij} = \Sigma_p n_{ij;p}$ and $N_{i0} = \Sigma_p n_{i0;p}$ require the knowledge of the contact numbers $n_{ij;p}$ and $n_{i0;p}$ in individual proteins (represented by p). Whereas $n_{ij;p}$ was counted directly from the structure, $n_{i0;p}$ was derived from

$$n_{i0;p} = q_i n_{i;p}/2 - \sum_{j=1}^{18} n_{ij;p}\,, \qquad [4]$$

where q_i is the precalculated coordination number of residue i (Table 1) and $n_{i;p}$ is the number of residue i in protein p.

We assume that the reference state of a protein exhibits the same compactness as the native folded state but with randomly arranged residue-residue and residue-solvent contacts (for details see ref. 33). Under this assumption, we have

$$C_{ij} = \sum_p n_{rr;p}\frac{q_i n_{i;p} q_j n_{j;p}}{(\sum_{k=1} q_k n_{k;p})^2} \qquad [5]$$

Table 1. Coordination numbers of amino acids in different secondary structural environments for $R_C = 6.5$ Å

Amino acid	α-Helix (α)	β-Strand (β)	Coil (C)
ALA	6.59 (1.38)	6.16 (1.26)	6.18 (1.31)
ARG	6.07 (1.34)	6.04 (1.33)	6.13 (1.25)
ASN	6.35 (1.40)	6.26 (1.10)	6.13 (1.34)
ASP	6.33 (1.22)	6.16 (1.28)	6.12 (1.28)
CYS	6.51 (1.22)	6.26 (1.24)	6.26 (1.35)
GLN	6.17 (1.26)	6.30 (1.28)	6.19 (1.23)
GLU	6.22 (1.17)	6.24 (1.22)	6.18 (1.26)
GLY	6.41 (1.46)	5.90 (1.34)	5.80 (1.30)
HIS	6.15 (1.24)	6.33 (1.26)	6.01 (1.26)
ILE	6.38 (1.35)	5.99 (1.31)	5.97 (1.26)
LEU	6.48 (1.37)	6.27 (1.24)	6.27 (1.28)
LYS	6.07 (1.17)	6.07 (1.27)	6.02 (1.50)
MET	6.31 (1.28)	6.12 (1.19)	6.21 (1.25)
PHE	6.13 (1.32)	6.15 (1.27)	6.07 (1.37)
PRO	5.95 (1.15)	5.93 (1.27)	5.84 (1.32)
SER	6.39 (1.33)	6.05 (1.39)	5.98 (1.24)
THR	6.58 (1.36)	6.05 (1.20)	6.17 (1.27)
TRP	5.72 (1.26)	5.76 (1.38)	5.61 (1.28)
TYR	6.17 (1.29)	6.15 (1.24)	6.05 (1.28)
VAL	6.59 (1.38)	6.15 (1.28)	6.19 (1.30)

The coordinate number q_i (i = 1, 60) is calculated as the number of non-nearest neighbor contacts formed by a buried residue of type i averaged over 407 protein domain structures. A residue is considered buried if more than 95% of its surface area becomes inaccessible to solvent when the protein is folded. Values in parentheses are the SDs.

$$C_{i0} = \sum_p n_{r0;p}\frac{q_i n_{i;p}}{\sum_{k=1} q_k n_{k;p}} \qquad [6]$$

$$C_{00} = N_{00}\,, \qquad [7]$$

where $n_{rr;p} = \Sigma_i \Sigma_j n_{ij;p}$ and $n_{r0;p} = \Sigma_i n_{i0;p}$.

Testing the Derived Energies. To assess whether ERCE offer better discriminatory power than the existing residue pair potentials, we first tested these potentials by using a nongapped threading protocol (34–36). The sequences of proteins with 200 or fewer residues in the data set were threaded through the structures of all proteins of the same or larger size at all possible positions. When the energy of the native structure of a protein is lower than that of any threaded model, we consider that the sequence of the protein successfully recognizes its structure.

To assess the potential in a more realistic setting where only predicted secondary structure information can be obtained, we applied ERCE to the prediction of the three-dimensional contacts in protein structures. In particular, we concentrated on nonlocal contacts, i.e., contacts formed by residues not closely associated by the polypeptide chain. For contacts between on-chain neighbors, short-range interactions are probably more important. A minimum of four residue separations were required for a contact to be considered. For the purpose of comparing energy functions, we used a simple method for contact prediction where any residue pair that had energy lower than a specified cutoff was predicted to be in contact. By varying the cutoff value, the tradeoff between the completeness (or coverage) of the prediction (how many native contacts are predicted) and the accuracy of the prediction (how many predicted contacts correspond to actually observed contacts) can be examined. The actual three-dimensional contacts were identified from protein structures by using the criterion that two residues in contact must form at least four atom–atom contacts (two atoms are considered in contact if the distance between them is

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	
	-2.52	-1.24	-1.22	-1.22	-3.40	-1.17	-0.77	-2.84	-1.69	-3.27	-3.29	-0.70	-2.63	-2.99	-1.64	-1.74	-1.87	-2.20	-2.57	-3.07	ALA
		0.03	-0.80	-1.51	-0.78	-0.68	-1.24	-1.25	-0.94	-1.41	-1.18	0.28	-0.98	-1.44	-0.84	-0.81	-0.89	-1.39	-1.57	-1.26	ARG
ALA	-1.65		-0.48	-1.17	-1.72	-0.74	-0.67	-1.69	-0.67	-1.41	-1.45	-0.35	-1.32	-1.41	-0.78	-1.16	-0.90	-1.13	-1.47	-1.33	ASN
ARG	0.02	1.08		-0.07	-1.34	-0.71	-0.18	-1.64	-1.44	-1.28	-1.26	-1.00	-1.04	-1.43	-0.27	-1.09	-1.00	-1.00	-1.50	-1.08	ASP
ASN	-0.25	0.25	0.14		-3.74	-1.97	-0.83	-3.17	-2.31	-3.73	-3.63	-0.74	-3.48	-3.61	-1.92	-2.23	-2.39	-3.03	-2.92	-3.33	CYS
ASP	-0.06	-0.16	0.09	0.53		-0.10	-0.47	-1.60	-1.04	-1.45	-1.38	-0.42	-1.66	-1.59	-0.58	-1.10	-1.03	-1.72	-1.31	-1.33	GLN
CYS	-2.04	-0.54	-0.95	-1.04	-2.81		0.72	-1.07	-1.17	-1.48	-1.23	-1.06	-0.64	-1.22	0.38	-0.85	-1.17	-1.21	-1.27	-1.07	GLU
GLN	0.00	0.41	0.20	0.15	-0.41	0.59		-2.60	-2.08	-2.93	-2.98	-0.83	-2.48	-3.01	-1.85	-2.03	-1.84	-2.58	-2.69	-2.78	GLY
GLU	0.29	-0.11	0.52	1.32	-0.38	0.47	1.24		-1.69	-1.86	-1.86	-0.61	-1.55	-1.83	-0.81	-1.53	-1.29	-1.50	-1.74	-1.70	HIS
GLY	-1.74	-0.11	-0.62	-0.67	-2.35	-0.01	0.37	-1.68		-3.52	-3.72	-1.32	-3.11	-3.48	-1.59	-1.62	-2.00	-2.86	-2.86	-3.46	ILE
HIS	-0.48	0.34	0.10	-0.09	-1.21	0.43	-0.09	-0.42	-0.40		-3.55	-1.16	-3.14	-3.50	-1.93	-1.77	-1.82	-2.83	-2.72	-3.46	LEU
ILE	-2.09	-0.77	-0.84	-0.65	-2.99	-1.00	-0.52	-2.15	-1.18	-2.72		0.47	-1.17	-0.92	-0.41	-0.83	-1.12	-1.55	-1.42	-1.21	LYS
LEU	-1.89	-0.63	-0.63	-0.31	-2.74	-0.64	-0.22	-1.87	-0.79	-2.93	-2.69		-1.95	-3.13	-1.43	-1.72	-1.40	-2.31	-2.26	-2.96	MET
LYS	0.08	1.21	0.32	0.09	-0.46	0.53	0.09	-0.16	0.40	-0.56	-0.20	1.52		-3.10	-1.75	-1.77	-1.52	-2.81	-2.73	-3.21	PHE
MET	-1.28	-0.10	-0.46	0.14	-2.17	-0.23	0.27	-1.50	-1.00	-2.44	-2.37	0.20	-1.85		0.20	-0.48	-0.72	-2.05	-1.69	-1.53	PRO
PHE	-1.70	-0.39	-0.43	-0.09	-2.87	-0.48	-0.06	-1.63	-1.21	-2.46	-2.58	-0.21	-2.25	-2.45		-1.03	-1.31	-1.50	-1.61	-1.83	SER
PRO	-0.26	0.45	0.59	0.44	-1.12	0.94	0.87	-0.45	0.31	-1.14	-0.69	0.64	0.07	-0.42	1.19		-1.29	-1.00	-1.46	-1.82	THR
SER	-0.83	0.24	-0.25	0.11	-1.36	0.24	0.50	-1.11	-0.08	-1.35	-1.24	0.37	-0.57	-1.13	0.32	-0.36		-1.61	-1.95	-2.42	TRP
THR	-0.78	0.09	-0.23	0.08	-1.60	-0.03	0.31	-1.23	-0.40	-1.52	-1.48	0.09	-1.12	-1.08	0.00	-0.40	-0.37		-1.59	-2.40	TYR
TRP	-1.40	-0.65	-0.48	-0.04	-2.02	-0.88	-0.43	-1.58	-0.78	-2.40	-2.27	-0.55	-2.20	-2.40	-0.75	-0.72	-1.22	-1.21		-3.13	VAL
TYR	-1.10	-0.49	-0.28	-0.22	-2.05	-0.33	-0.12	-1.05	-0.59	-2.01	-1.93	-0.41	-1.69	-2.05	-0.41	-0.62	-0.74	-1.82	-1.13		
VAL	-1.84	-0.47	-0.71	-0.37	-2.72	-0.32	-0.23	-1.98	-0.88	-2.66	-2.57	-0.37	-2.05	-2.39	-0.97	-1.19	-1.38	-1.93	-1.76	-2.35	
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	-0.94	1.26	0.55	0.76	-1.54	1.14	1.57	-0.78	0.44	-1.59	-1.64	1.91	-0.90	-1.49	0.28	0.20	-0.04	-0.92	-0.75	-1.45
ARG	0.56	1.79	2.31	0.79	-0.67	2.54	0.72	1.09	0.94	-0.01	0.01	3.68	0.89	-0.05	1.37	0.83	1.35	0.00	0.33	0.44
ASN	0.59	2.21	1.82	0.77	-0.90	0.46	3.06	-0.16	0.63	-0.33	0.20	2.43	0.99	0.63	0.54	0.24	0.63	0.11	-0.19	0.23
ASP	0.66	0.76	0.76	1.19	-0.21	1.66	2.22	0.29	0.57	0.59	0.79	1.13	1.41	0.49	1.70	1.03	1.19	1.85	0.18	0.86
CYS	-1.75	0.78	-1.00	0.32	-3.64	0.48	0.87	-1.67	-0.62	-2.77	-2.32	0.19	-1.22	-2.67	-1.62	-0.83	-1.14	-0.52	-1.94	-2.35
GLN	0.33	2.15	1.22	1.26	1.37	1.17	2.56	0.92	1.02	0.11	0.00	2.58	0.79	-0.26	0.53	1.19	1.11	0.21	0.39	0.15
GLU	0.82	1.05	2.18	2.11	0.01	2.42	2.58	1.15	0.97	0.20	0.31	1.31	1.25	0.12	2.00	1.09	1.13	0.58	0.31	0.39
GLY	-0.40	0.95	0.03	0.14	-1.00	0.34	0.99	-1.32	0.13	-1.40	-1.36	1.59	-0.90	-1.41	0.82	-0.27	0.21	-0.59	-1.27	-1.09
HIS	-0.75	2.19	0.13	0.68	-1.37	1.98	1.13	0.01	1.52	-0.83	-0.58	2.26	-0.82	-1.01	0.53	-0.17	0.02	-0.49	-0.61	-0.56
ILE	-1.99	0.25	-0.20	1.00	-2.44	-0.12	0.88	-1.54	-0.05	-2.64	-2.33	0.75	-1.85	-2.46	-1.06	-0.59	-0.65	-1.82	-1.88	-2.45
LEU	-2.02	0.34	-0.04	0.13	-2.29	0.24	0.73	-1.27	-0.46	-2.53	-2.44	0.67	-1.80	-2.28	-1.29	-0.40	-0.34	-1.76	-1.66	-2.26
LYS	0.60	3.11	2.23	1.06	0.50	1.80	1.65	0.82	1.25	0.10	0.34	3.51	0.98	-0.21	1.15	2.09	1.30	-0.14	0.28	0.13
MET	-1.54	-0.06	-0.63	1.76	-2.51	0.14	0.72	-1.74	0.07	-2.27	-2.22	1.27	-1.77	-1.87	0.34	-0.02	-0.21	-0.93	-1.54	-1.81
PHE	-2.12	0.33	-0.70	0.17	-2.30	-0.59	0.26	-1.60	-0.88	-2.53	-2.44	-0.42	-1.83	-2.68	-1.40	-0.82	-0.61	-1.63	-1.83	-2.25
PRO	0.63	2.43	-0.19	1.31	-1.63	1.46	1.91	0.08	1.11	-0.20	0.47	1.94	-0.34	0.15	0.57	0.00	1.15	0.06	0.26	-0.06
SER	-0.41	0.88	1.02	1.04	-0.21	1.27	0.94	0.04	0.75	-0.48	-0.67	2.28	0.45	-0.92	0.75	0.50	0.96	0.22	-0.19	-0.54
THR	-0.32	1.48	0.35	0.43	-1.44	0.38	1.36	-0.38	0.20	-1.14	-1.00	1.38	-0.35	-0.97	-0.05	-0.16	0.29	-0.53	-0.76	-0.73
TRP	-1.85	0.45	-0.03	0.80	-1.64	-0.23	0.11	-0.95	0.67	-1.58	-2.13	0.61	-1.75	-1.59	-1.07	-0.34	-0.40	-1.29	-1.27	-1.79
TYR	-0.88	-0.20	-0.29	0.14	-1.31	0.09	0.71	-0.56	-0.57	-1.66	-1.38	1.40	-1.60	-1.97	-0.73	-0.32	-0.37	-1.40	-0.96	-1.38
VAL	-1.74	0.85	0.24	0.72	-2.25	0.45	0.81	-1.29	-0.24	-2.46	-2.38	0.37	-1.21	-2.16	-1.00	-0.10	-0.57	-1.34	-1.52	-2.31

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.12	1.17	0.84	0.90	-0.81	1.16	1.44	0.10	0.69	-0.81	-0.78	1.16	-0.22	-0.67	0.61	0.47	0.36	-0.72	-0.37	-0.43
ARG	0.98	1.65	1.16	0.60	-0.21	1.26	1.12	1.09	1.16	-0.04	-0.09	2.37	0.47	-0.04	1.22	1.05	0.92	-0.09	0.06	0.32
ASN	0.69	1.16	1.16	1.22	-0.06	1.23	1.45	0.96	0.88	0.26	0.12	1.48	0.32	0.03	1.14	0.73	0.62	0.62	0.53	0.23
ASP	0.90	0.40	1.06	1.45	0.58	1.88	2.18	1.13	0.69	0.43	0.65	0.95	0.75	0.33	1.41	0.39	0.54	-0.10	0.12	0.77
CYS	-0.83	0.10	0.40	0.12	-2.65	-0.24	0.96	-0.26	-0.26	-1.61	-1.77	0.80	-1.02	-1.47	-0.31	-0.31	-0.49	-1.30	-0.98	-1.62
GLN	1.13	1.10	1.28	1.37	0.14	1.62	1.84	1.29	1.31	0.05	-0.05	1.50	0.41	0.20	1.14	0.86	0.62	0.45	0.31	0.48
GLU	1.33	0.91	1.33	1.60	0.31	1.60	1.93	1.62	1.01	0.33	0.38	1.12	0.82	0.55	1.54	0.78	0.54	0.23	0.52	0.86
GLY	-0.22	0.72	0.27	0.47	-0.95	0.42	1.39	-0.23	0.40	-0.48	-0.81	1.04	-0.62	-0.36	0.41	0.23	-0.04	-0.71	0.08	-0.35
HIS	0.47	0.81	0.95	0.51	-1.56	0.90	0.89	0.86	0.20	-0.43	-0.48	1.31	-0.63	-0.41	0.56	0.40	0.28	-0.20	-0.22	-0.21
ILE	-0.58	0.17	0.61	0.46	-1.17	0.24	0.80	0.04	-0.16	-1.64	-1.66	0.87	-0.89	-1.56	-0.27	0.02	-0.32	-1.40	-1.13	-1.36
LEU	-0.44	0.20	0.50	0.71	-1.56	0.11	0.82	0.28	-0.15	-1.67	-1.62	0.72	-0.96	-1.55	0.02	0.19	-0.09	-1.46	-0.95	-1.32
LYS	1.07	2.48	1.75	0.98	0.42	1.68	1.04	1.31	1.39	0.41	0.29	2.95	0.98	0.27	1.63	1.51	1.48	0.32	0.60	0.64
MET	-0.22	0.65	0.76	0.88	-0.95	0.68	1.92	0.27	0.31	-1.32	-1.04	1.02	-0.57	-1.60	0.07	0.47	0.04	-1.29	-0.85	-0.82
PHE	-0.33	-0.06	0.42	0.42	-1.90	0.25	0.64	0.12	-0.01	-1.64	-1.50	0.58	-1.36	-1.77	-0.30	0.02	0.04	-1.41	-1.36	-1.34
PRO	0.78	1.30	1.31	1.27	-0.04	1.44	1.71	0.69	0.84	0.05	0.15	1.68	0.38	0.27	1.05	1.19	0.83	-0.24	0.23	0.12
SER	0.46	1.07	1.04	0.73	-0.31	1.47	1.23	0.57	0.58	-0.11	-0.24	1.37	0.08	-0.34	0.76	0.51	0.48	-0.04	0.47	0.18
THR	0.50	0.90	0.75	0.91	-0.26	1.03	1.25	0.55	0.55	-0.20	-0.26	1.42	0.50	-0.22	0.88	0.69	0.56	0.41	0.11	-0.15
TRP	-0.41	-0.06	-0.19	0.32	-0.79	-0.14	0.58	0.07	-0.62	-1.58	-1.16	0.18	-1.03	-1.33	-0.56	0.15	-0.19	-1.83	-0.67	-0.92
TYR	-0.22	-0.07	0.52	0.46	-0.87	0.38	0.59	0.40	-0.17	-1.29	-1.15	0.83	-0.98	-1.16	-0.16	0.34	-0.12	-0.79	-0.77	-0.78
VAL	-0.51	0.49	0.48	0.67	-1.40	0.66	0.63	-0.06	0.28	-1.25	-1.50	1.14	-0.93	-1.36	-0.04	0.10	-0.01	-1.11	-0.82	-1.14

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	-0.57	0.47	0.30	0.62	-1.60	0.45	0.61	-0.24	0.07	-1.64	-1.63	0.62	-1.03	-1.55	-0.11	-0.10	-0.34	-1.44	-0.39	-1.55
ARG	0.23	0.79	0.76	0.39	-0.41	0.92	0.76	0.52	0.51	-0.30	0.13	1.58	0.88	-0.07	0.60	0.65	0.37	0.14	0.32	0.17
ASN	-0.28	0.74	0.57	0.87	-0.68	0.52	1.00	-0.07	0.32	-0.31	-0.08	0.87	0.29	-0.17	0.57	0.11	0.19	0.04	0.24	-0.23
ASP	0.15	-0.25	0.46	0.69	-0.46	0.41	1.34	0.56	-0.51	-0.23	0.27	0.59	0.60	-0.38	1.02	0.08	0.05	-0.48	0.02	0.34
CYS	-1.19	-0.46	0.21	0.51	-3.30	0.26	0.20	-1.03	-0.72	-1.55	-1.71	0.27	-1.24	-1.70	-0.50	-0.55	-0.97	-0.67	-1.26	-1.62
GLN	0.63	1.18	0.92	1.37	-0.30	0.93	1.27	0.56	0.91	-0.28	-0.11	0.98	0.15	-0.30	0.64	0.88	0.68	-0.44	0.66	0.15
GLU	0.97	0.89	1.37	1.89	0.30	1.25	2.34	0.98	0.58	0.20	0.50	0.67	1.23	0.58	1.26	0.95	1.06	0.04	0.87	0.48
GLY	-0.64	0.12	0.27	0.31	-1.37	0.38	0.98	-0.40	-0.12	-1.58	-1.40	0.78	-0.46	-1.38	-0.21	0.05	-0.26	-1.41	-0.61	-1.13
HIS	-0.02	0.75	0.68	0.14	-0.58	0.73	0.84	0.41	-0.64	-0.75	0.03	1.46	-0.16	-0.49	0.52	0.31	-0.11	-1.00	-0.58	0.03
ILE	-0.94	-0.14	0.31	0.26	-1.70	0.07	0.46	-0.37	-0.50	-1.88	-1.79	0.84	-0.99	-1.82	-0.47	-0.05	-0.54	-1.65	-1.09	-1.64
LEU	-0.76	0.32	0.43	0.25	-1.63	0.22	0.68	-0.17	-0.40	-1.84	-1.70	0.47	-1.06	-1.76	-0.39	0.09	-0.42	-1.81	-1.15	-1.64
LYS	1.02	1.99	1.18	0.59	0.08	1.10	0.60	0.61	0.95	0.24	0.34	2.69	0.97	-0.03	1.23	1.07	0.83	0.00	0.26	0.36
MET	-0.16	0.83	0.47	0.92	-1.63	0.36	0.71	-0.20	0.90	-1.00	-1.12	1.55	-0.31	-1.35	-0.01	0.34	0.20	-1.70	-0.60	-0.79
PHE	-0.70	0.03	0.63	0.15	-1.26	0.29	0.35	-0.11	-0.36	-1.73	-1.55	0.71	-0.97	-1.55	-0.28	-0.09	-0.32	-1.23	-0.91	-1.30
PRO	0.17	0.50	0.60	0.67	-1.31	0.50	0.94	0.02	-0.45	-1.26	-0.91	1.08	0.83	-0.87	0.63	0.31	0.26	-0.50	-0.55	-0.79
SER	-0.06	0.99	0.73	0.86	-0.89	0.85	0.67	0.08	0.06	-0.22	-0.29	0.94	-0.08	-0.41	0.67	0.33	0.13	-1.01	0.13	-0.24
THR	0.26	0.93	0.70	0.87	-0.78	0.58	1.20	0.12	0.52	-0.30	-0.24	1.11	0.01	-0.08	0.65	0.47	0.41	-0.31	0.12	-0.32
TRP	-0.03	-0.11	0.27	0.66	-1.50	0.65	0.50	-0.12	-0.32	-1.13	-1.01	0.52	-1.08	-1.04	-0.32	-0.03	-0.10	-0.67	-0.73	-0.64
TYR	-0.44	0.20	0.20	0.20	-1.26	0.16	0.10	-0.21	-0.52	-1.26	-1.30	0.60	-0.76	-1.17	-0.42	0.05	-0.27	-1.20	-0.75	-0.84
VAL	-0.83	0.20	0.48	0.62	-1.44	0.17	0.73	-0.12	-0.26	-1.64	-1.59	0.52	-0.70	-1.55	-0.28	0.12	-0.17	-1.16	-0.85	-1.42

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.12																			
ARG	0.56	1.18																		
ASN	0.59	0.83	0.83																	
ASP	0.65	0.33	0.61	1.15																
CYS	-0.74	-0.25	-0.07	-0.12	-2.42															
GLN	0.70	0.94	0.87	0.96	-0.05	1.22														
GLU	1.32	0.54	1.09	1.48	0.20	1.11	2.18													
GLY	0.11	0.65	0.79	0.75	-0.56	0.73	1.16	0.35												
HIS	0.35	0.47	0.91	0.29	-0.83	0.83	0.74	0.53	0.28											
ILE	-0.45	0.05	0.40	0.47	-1.27	0.24	0.70	0.01	-0.21	-1.04										
LEU	-0.25	0.26	0.37	0.57	-1.25	0.38	0.82	0.13	-0.03	-1.15	-1.04									
LYS	1.06	1.65	1.11	0.58	0.76	1.29	0.93	1.08	1.23	0.48	0.78	2.23								
MET	0.29	0.58	0.68	0.75	-0.74	0.65	1.08	0.54	0.19	-0.52	-0.68	1.15	-0.12							
PHE	-0.41	0.24	0.46	0.38	-1.44	0.39	0.55	0.09	-0.28	-1.09	-1.05	0.74	-0.60	-1.09						
PRO	0.48	0.82	1.09	1.26	-0.23	1.09	1.25	0.74	0.59	0.14	0.10	1.59	0.54	-0.04	1.11					
SER	0.45	0.55	0.81	0.54	-0.47	0.75	0.87	0.61	0.46	0.09	0.01	1.24	0.56	0.17	0.94	0.87				
THR	0.30	0.80	0.54	0.51	-0.37	0.81	0.81	0.38	0.31	-0.28	-0.05	1.10	0.46	0.08	0.67	0.54	0.69			
TRP	-0.28	-0.04	0.13	0.43	-0.63	-0.01	0.44	-0.17	-0.44	-1.28	-0.98	0.23	-0.50	-1.02	-0.33	0.14	-0.19	-0.46		
TYR	-0.07	0.22	0.40	0.38	-0.58	0.09	0.47	0.28	-0.18	-0.74	-0.56	0.62	-0.26	-0.63	0.04	0.23	0.27	-0.87	-0.18	
VAL	-0.18	0.43	0.54	0.59	-1.22	0.23	0.74	0.04	0.09	-0.90	-0.93	0.81	-0.31	-0.77	0.17	0.20	-0.10	-0.57	-0.38	-0.31
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL

Fig. 1. Secondary structural ERCE (in *RT* units). The energy parameters are divided into six groups, α-α, β-β, α-β, α-*C*, β-*C*, and *C*-*C*, where α, β, and *C* represent helix, strand, and coil states, respectively.

Table 2. Comparisons between energy parameters

		Pairwise correlation coefficients									
	Average energy*	MJ	α-α	β-β	α-$\beta^{\dagger}$	α-$\beta^{\ddagger}$	α-$C^{\dagger}$	α-$C^{\ddagger}$	β-$C^{\dagger}$	β-$C^{\ddagger}$	*C*-*C*
MJ	−2.93 (1.45)	—									
α-α	−0.74 (0.99)	0.86	—								
β-β	−1.66 (0.93)	0.80	0.94	—							
α-$\beta^{\dagger}$	−0.10 (1.33)	0.83	0.92	0.89	—						
α-$\beta^{\ddagger}$	−0.04 (1.33)	0.84	0.92	0.89	0.88	—					
α-$C^{\dagger}$	0.23 (0.93)	0.89	0.92	0.87	0.89	0.90	—				
α-$C^{\ddagger}$	0.18 (0.94)	0.91	0.92	0.89	0.90	0.92	0.90	—			
β-$C^{\dagger}$	0.01 (0.87)	0.80	0.89	0.87	0.86	0.90	0.91	0.90	—		
β-$C^{\ddagger}$	−0.14 (0.88)	0.81	0.91	0.90	0.89	0.89	0.89	0.91	0.83	—	
C-*C*	0.25 (0.70)	0.86	0.91	0.88	0.87	0.89	0.95	0.93	0.92	0.91	—

The three asymmetric energy groups, α-β, β-*C*, and β-*C*, are subdivided into lower and upper off-diagonal triangular components and represented by † and ‡, respectively. MJ, ref. 32.
*In *RT* units. Values in parentheses are the SDs.

less than 6.0 Å). The results were essentially the same when a more (or less) stringent definition of contact was used.

Results

The Influence of Secondary Structural Environments on Inter-Residue Interactions. The calculated contact energies are given in Fig. 1. The 60-by-60 ERCE parameters are separated into six groups with respect to the secondary structural environments of the interacting pair: α-α, β-β, α-β, β-*C*, β-*C*, and *C*-*C* (where α, β, and *C* represent helix, strand, and coil, respectively). Similar to the contact energies determined by MJ (32), the energies derived here are dominated by hydrophobic interactions. The correlation coefficients of the contact energies from different groups with each other and with those determined by MJ (32) are summarized in Table 2.

Despite the high correlations, the average values of contact energies from different groups differ significantly (Table 2). The α-α and β-β contacts are generally more favorable than other types of interactions. This result is understandable considering that the cores of most proteins consist of closely packed regular secondary structural elements where helices tend to gather into bundles and β-strands often appear in assembled β-sheets. The β-β contact energies are on average nearly one *RT* unit lower than the α-α contact energies. Because most of the tertiary contacts formed by β residues are cross-β-bridge contacts, this result suggests that there may be a fundamental difference between the structural and energetic principles governing β-strand register and α-helix packing. Among the energy groups, the α-β contact energies show the largest variations, and therefore, highest specificity, consistent with the unique geometric feature of α-β packing observed in protein structures (37, 38).

Threading Experiments. There are 316 protein domains in our data set that contain 200 or fewer residues; the number of conformations generated by threading ranges from 4,000 (for a 200-residue protein) to 40,000 (for a 50-residue protein). As shown in Table 3, ERCE consistently outperformed MJ in the treading tests in all four structural categories, with an overall 6% improvement in success rate. For 88 all-β proteins and 49 α/β proteins, perfect recognition was achieved. In seven of the nine cases where ERCE failed to identify the native structures, the native structures were among the five most favorable conformations. MJ failed three times more; in half of those cases, the native structures were ranked below five.

Table 3. Results of the threading experiments

	all-α (71)	all-β (88)	α/β (49)	$\alpha+\beta$ (108)	Overall (316)
MJ	87.3%	94.3%	93.9%	90.7%	91.4%
ERCE	90.1%	100%	100%	98.1%	97.2%

The numbers in parentheses are the number of proteins in each category. MJ, ref. 32.

Contact Prediction. Fig. 2 shows the performance of ERCE in contact prediction for 407 proteins. Both predicted and real secondary structures were used, and the prediction based on real secondary structures in conjunction with ERCE (ERCE-Real) is noticeably better. However, a significant improvement over MJ

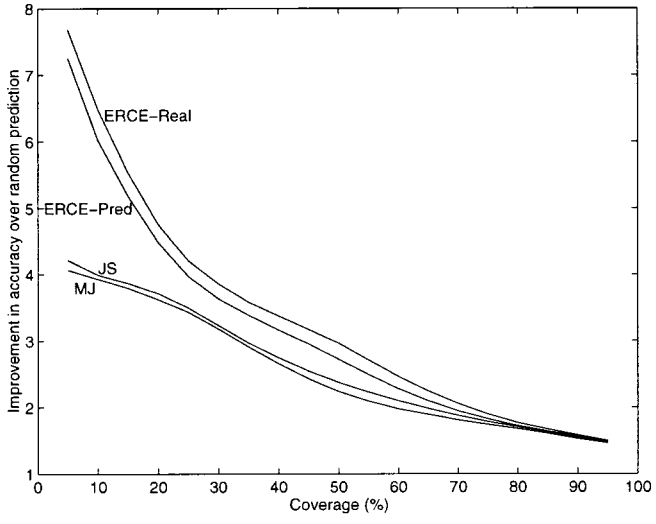

Fig. 2. Contact prediction using three different potentials: MJ, JS, and ERCE. ERCE predictions based on predicted (ERCE-Pred) and real secondary structures (ERCE-Real) both are shown. The *x* axis indicates the fraction of experimentally determined contacts that are correctly predicted (defined here as the coverage). Each coverage value *x* corresponds to an energy cutoff such that the percentage of native contacts whose energies are lower than the cutoff happens to be *x*. There are other residue pairs that do not form contacts in the native protein structure but also have energies lower than the cutoff. The fraction of predicted contacts that correspond to actually observed contacts define the accuracy. Here we compare different prediction methods with a common standard: the random prediction. In theory, the accuracy of a random prediction averaged over many trials equals the ratio of native contacts to the number of all possible residue pairs in a protein. This value varies with protein size and contact density. For each method, the *y* axis indicates how much more accurate the prediction is relative to that from a random method.

BIOPHYSICS

Table 4. Comparisons of the accuracy of contact prediction by different methods

Coverage*	Method/potential†	Size categories			Structural classes				Overall
		50–114	115–179	180–299	All-α	All-β	α/β	$\alpha+\beta$	
	Random prediction	3.72	2.97	2.36	2.49	4.00	2.21	3.30	3.08
20%	MJ	15.03	8.87	5.41	10.05	12.28	7.00	10.78	10.18
	JS	15.43	9.11	5.55	10.31	12.73	7.15	11.01	10.45
	ERCE-Pred	18.60	11.32	6.58	9.91	15.42	9.83	14.19	12.69
	ERCE-Real	19.61	12.07	7.20	10.17	16.87	10.10	15.20	13.50
50%	MJ	9.17	5.47	3.39	6.00	7.50	4.45	6.68	6.26
	JS	9.66	5.85	3.66	6.31	8.15	4.65	7.04	6.65
	ERCE-Pred	11.56	6.82	4.01	6.13	9.81	5.36	8.86	7.79
	ERCE-Real	12.75	7.49	4.32	6.06	11.16	5.70	9.95	8.55

The accuracy is defined as the percentage of predicted contacts that correspond to actually observed contacts in protein structures.
*See Fig. 2 for definition.
†Abbreviations are explained in Fig. 2.

in terms of prediction accuracy was obtained by both methods. The encouraging result achieved by ERCE in combination with predicted secondary structures (ERCE-Pred) adds a practical value to ERCE. We also have applied other published residue pair potentials for the same task; of those, the only potential that outperformed MJ was one recently reported by Skolnick and coworkers (JS) (5). Interestingly, a recent update to MJ performed slightly worse than the original energy table (data not shown), although the updated energy table was derived from a much larger set of protein structures (36).

More detailed comparisons of different prediction methods are shown in Table 4. The 407 testing proteins were divided into three size categories and four structural classes. Because the fraction of residue pairs that form tertiary contacts in a larger protein is less than that in a smaller protein, the level of difficulty in predicting contact rises as the size of the protein increases. This trend is clearly seen in the case of a random prediction (Table 4). Although the use of potential energy function improves the chance of detecting native contacts, the basic trend of prediction accuracy versus protein size persists. ERCE-based predictions are consistently better than other residue pair potentials for all three size categories.

The analysis of prediction performance with respect to structural classes is more revealing. Compared with the random prediction, the use of residue pair potentials such as MJ and JS yields the maximal gain in contact prediction for all-α proteins. The added value of using ERCE is that it also improves predictions on other three structural classes (all-β, α/β, and $\alpha+\beta$) such that the overall improvement ratios relative to a random prediction are comparable for all four structural classes.

Discussion

Long-range interactions play an important role in determining the tertiary structures of proteins. However, computational simulations of such interactions have encountered great difficulties in the past. Coarse-grained reside pair potentials attempt to strike a balance between the accuracy of energy representation and the computational expediency by harnessing the rich amount of information about intramolecular interactions provided by experimentally solved structures. In this study, we extended the contact energy formulation (32) to include the influence of secondary structural environments on the specificity of residue interactions, taking advantage of the amount and quality of structural data currently available. Other ways to divide up the data to extract features usable in various categories of protein simulations also have been reported (10, 39–41).

The added value of ERCE over existing residue pair potentials was explored in both threading and three-dimensional contact prediction experiments. The merit of using contact prediction to test energy functions is that it is independent of the conformational search algorithms and can be readily applied to a large set of protein structures. Because the objectives of two-dimensional contact map prediction and three-dimensional structure prediction are essentially the same (16, 42–44), the combined power of ERCE and secondary structure prediction illustrated here should be of general interest to protein structure prediction.

It should be noted that we used the same set of 407 structures to derive and test ERCE. It has been noticed before that contact energies derived by the MJ formulation is relatively insensitive to the changes in the database size and content (36, 40). Because our data set contains a large number of nonredundant structures, a Jackknife test produces essentially the same results (data not shown). In fact, we also used an independent set of 91 protein chains (33) to derive an ERCE table, and the parameters thereof were highly correlated with those given in Fig. 1 and achieved comparable accuracy in threading and contact prediction experiments.

We thank Inna Dubchak and Bob Jernigan for critical reading of the paper. We also thank David Jones for communicating the PSI-PRED program. This work was supported by grants from the Department of Energy (DE-AC03-76SF00098), the National Science Foundation (97–23352), and the National Institutes of Health (CA78406).

1. Levitt, M. (1976) *J. Mol. Biol.* **104,** 59–107.
2. Jernigan, R. L. & Bahar, I. (1996) *Curr. Opin. Struct. Biol.* **6,** 195–209.
3. Godzik, A., Kolinski, A. & Skolnick, J. (1995) *Protein Sci.* **4,** 2107–2117.
4. Thomas, P. D. & Dill, K. A. (1996) *J. Mol. Biol.* **257,** 457–469.
5. Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997) *Protein Sci.* **6,** 676–688.
6. Zhang, C. (1998) *Proteins Struct. Funct. Genet.* **31,** 299–308.
7. Hao, M.-H. & Scheraga, H. A. (1999) *Curr. Opin. Struct. Biol.* **9,** 184–188.
8. Rooman, M. & Gilis, D. (1998) *Eur. J. Biochem.* **254,** 135–143.
9. Minor, D. L. J. & Kim, P. S. (1994) *Nature (London)* **371,** 264–267.
10. Cootes, A. P., Curmi, P. M. G., Cunningham, R., Donnelly, C. & Torda, A. E. (1998) *Proteins Struct. Funct. Genet.* **32,** 175–189.
11. Hutchinson, E. G., Sessions, R. B., Thornton, J. M. & Woolfson, D. N. (1998) *Protein Sci.* **7,** 2287–2300.
12. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
13. Cuff, J. A. & Barton, G. J. (1999) *Proteins Struct. Funct. Genet.* **34,** 508–519.
14. Jones, D. T. (1999) *J. Mol. Biol.* **292,** 195–202.
15. Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997) *J. Mol. Biol.* **265,** 217–241.
16. Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1020–1025.
17. Huang, E. S., Samudrala, R. & Ponder, J. W. (1999) *J. Mol. Biol.* **290,** 267–281.
18. Russel, R. B., Copley, R. R. & Barton, G. J. (1996) *J. Mol. Biol.* **259,** 349–365.
19. Di Francesco, V., Garnier, J. & Munson, P. J. (1997) *J. Mol. Biol.* **267,** 446–463.
20. Rice, D. W. & Eisenberg, D. (1997) *J. Mol. Biol.* **267,** 1026–1038.

21. Rost, B., Schneider, R. & Sander, C. (1997) *J. Mol. Biol.* **270,** 471–480.
22. Aurora, R. & Rose, G. D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 2818–2823.
23. Grigoriev, I. V. & Kim, S.-H. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 14318–14323.
24. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem.* **48,** 545–600.
25. Dill, K. A. & Chan, H. S. (1997) *Nat. Struct. Biol.* **4,** 10–19.
26. Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 5482–5485.
27. Berstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
28. Murzin, A., Brenner, S. E., Hubband, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
29. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
30. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
31. Holm, L. & Sander, C. (1998) *Bioinformatics* **14,** 423–429.
32. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18,** 534–552.
33. Zhang, C., Vasmatzis, G., Cornette, J. L. & DeLisi, C. (1997) *J. Mol. Biol.* **267,** 707–726.
34. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990) *J. Mol. Biol.* **216,** 167–180.
35. Kocher, J. P., Rooman, M. J. & Wodak, S. J. (1994) *J. Mol. Biol.* **235,** 1598–1613.
36. Miyazawa, S. & Jernigan, R. L. (1996) *J. Mol. Biol* **256,** 623–644.
37. Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. (1982) *J. Mol. Biol.* **156,** 821–862.
38. Reddy, B. V. B., Nagarajaram, H. A. & Blundell, T. L. (1999) *Protein Sci.* **8,** 573–586.
39. Bahar, I. & Jernigan, R. L. (1996) *Folding Des.* **1,** 357–370.
40. Bahar, I. & Jernigan, R. L. (1997) *J. Mol. Biol.* **266,** 195–214.
41. Miyazawa, S. & Jernigan, R. L. (1999) *Proteins Struct. Funct. Genet.* **15,** 347–356.
42. Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994) *Proteins Struct. Funct. Genet.* **18,** 309–317.
43. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994) *Protein Eng.* **7,** 349–358.
44. Olmea, O., Rost, B. & Valencia, A. (1999) *J. Mol. Biol.* **293,** 1221–1239.