Protein Science (1998), 7:2578–2586. Cambridge University Press. Printed in the USA.

Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions

O. KESKIN,[1] I. BAHAR,[1,2] A.Y. BADRETDINOV,[3,4] O.B. PTITSYN,[2,4] AND R.L. JERNIGAN[2]

[1]Chemical Engineering Department & Polymer Research Center, Bogazici University, and TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey
[2]Molecular Structure Section, Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-5677
[3]Laboratory of Molecular Biophysics, 1230 York Ave., Rockefeller University, New York, New York 10021-6399
[4]Institute of Protein Research, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia

(RECEIVED March 16, 1998; ACCEPTED July 20, 1998)

Abstract

Whether knowledge-based intra-molecular inter-residue potentials are valid to represent inter-molecular interactions taking place at protein-protein interfaces has been questioned in several studies. Differences in the chain connectivity effect and in residue packing geometry between interfaces and single chain monomers have been pointed out as possible sources of distinct energetics for the two cases. In the present study, the interfacial regions of protein-protein complexes are examined to extract inter-molecular inter-residue potentials, using the same statistical methods as those previously adopted for intra-molecular residue pairs. Two sets of energy parameters are derived, corresponding to solvent-mediation and "average residue" mediation. The former set is shown to be highly correlated (correlation coefficient 0.89) with that previously obtained for inter-residue interactions within single chain monomers, while the latter exhibits a weaker correlation (0.69) with its intra-molecular counterpart. In addition to the close similarity of intra- and inter-molecular solvent-mediated potentials, they are shown to be significantly more residue-specific and thereby discriminative compared to the residue-mediated ones, indicating that solvent-mediation plays a major role in controlling the effective inter-residue interactions, either at interfaces, or within single monomers. Based on this observation, a reduced set of energy parameters comprising 20 one-body and 3 two-body terms is proposed (as opposed to the 20 × 20 tables of inter-residue potentials), which reproduces the conventional 20 × 20 tables with a correlation coefficient of 0.99.

Keywords: knowledge-based potentials; protein interfaces; protein solvation

Coarse grained models of proteins, where only one or two points per residue are considered, have become popular because they simplify conformational considerations by avoiding the obfuscation of including all of the atoms. Knowledge-based residue-residue potentials are a necessary adjunct to the coarse grained models for protein simulations and for meaningful comparisons of different protein structures. The physical sense and possibilities were discussed by Finkelstein et al. (1995). In recent years, there have been several studies aimed at extracting potentials of mean force for inter-residue interactions from information available in protein structure databases, as described in several reviews (Sippl, 1995; Jernigan & Bahar, 1996; Jones & Thornton, 1996; Torda, 1997), and are exemplified by two recent studies (Huber & Torda, 1998; Zhang & Skolnick, 1998). While increasingly more detailed statistical methods have been utilized with the growing number of databank structures to obtain more accurate potentials, attention has also been paid to the limitations of these potentials, regarding the effects of chain connectivity and environment (Thomas & Dill, 1996), the reproducibility of the parameters (Zhang & Skolnick, 1998), and the various factors influencing their discriminatory abilities between correctly folded and misfolded structures (Kocher et al., 1994; Mirny & Shakhnovich, 1996; Park & Levitt, 1996).

From the physical point of view, inter-residue potentials should be potentials of mean force. These potentials represent a free energy change of the whole system (residues and surroundings) upon bringing together two residues from an infinitely large separation. A practical way to obtain these potentials is to extract them from frequencies of contacts between different residues in proteins with known three-dimensional (3D) structures (Miyazawa & Jernigan, 1985). The principal difficulty in this approach is, however, the connectivity of the protein chain. Since all residues are connected

Reprint Requests to: R.L. Jernigan, Molecular Structure Section, Laboratory of Mathematical Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-5677; e-mail: jernigan@structure.nci.nih.gov.

by chemical bonds, each globular structure leads to relatively small distances even for those pairs of residues that might repel each other. Therefore, the potentials of mean force obtained by this approach, also referred to as "solvent-mediated" potentials, appear to be attractive for all pairs of residues including, for example, even those with charges of the same sign (Miyazawa & Jernigan, 1985, 1996; Bahar & Jernigan, 1997; see also Table 1). This "connectivity effect" has been pointed out to introduce some biases in the knowledge-based inter-residue potentials of mean force.

One approach to avoid this bias toward compactness is to use along with solvent-mediated potentials, also "residue-mediated" potentials. These potentials represent the differences between the free energy of a residue in a given 3D structure (with the rather specific environment of each of the residues) and its free energy averaged over all globular structures, i.e., over all possible contacting residues (Miyazawa & Jernigan, 1985, 1996; Finkelstein et al., 1995; Bahar & Jernigan, 1997). As a result, the connectivity effect may be approximately canceled, when we compare the free energies of different globular conformations of the given amino acid sequence. The recent study of Skolnick et al. (1997) points out, in fact, that neglect of chain connectivity does not introduce errors in solvent mediated potentials. In an extremely crude way, we can postulate that solvent-mediated potentials are related to protein folding, i.e., its transition from more or less unfolded conformations into the native 3D structure. On the other hand, residue-mediated potentials are more appropriate for threading of protein cores, i.e., for comparing the free energies of buried residues in different folded conformations of a given amino acid sequence.

So far we have discussed the knowledge-based residue-residue potentials obtained from probabilities of different contacts inside one protein chain. An alternative way is to extract these potentials from probabilities of contacts between different protein monomers forming a quaternary structure. Although these inter-molecular potentials also can be influenced by chain connectivity, this influence should be smaller than in the case of intra-molecular potentials. On the other hand, the statistics for inter-molecular potentials is weaker because the dataset is smaller while the intra-molecular potentials are based on the tertiary structures of a large set of proteins; whereas inter-molecular ones can only be based on the more limited contact surfaces in quaternary structures of proteins consisting of two or more chains.

Intra-molecular residue-residue interactions have been investigated in detail in our previous papers (Miyazawa & Jernigan, 1985, 1996; Bahar & Jernigan, 1997). Here a set of quaternary 3D structures of proteins available in the Protein Data Bank (http://www.pdb.bnl.gov/) will be examined using the same model and methods, and the extracted residue-residue potentials will be compared with those obtained from tertiary 3D structures.

Solvent-mediated and residue-mediated effective contact potentials

As mentioned above, it is possible to conceive of two reference states for the interaction of two residues: either solvent exposure, or a bath of residues packed in conformity with the packing characteristics of native folds. The fact that the inter-residue potentials of mean force could differ depending on the reference state is obvious, due to the basic derivation method and the essential physical meaning of potentials of mean force. We will designate the effective inter-residue potentials expressed with reference to these two media as e^0_{ij} and e^r_{ij}; the superscripts refer to the solvent and residue environments, respectively, and the subscripts describe the types i and j of the interacting amino acids. The single letter codes of amino acids will be conveniently substituted therein for referring to the potentials between specific pairs of amino acids, and the subscripts "0" and "r" will be adopted for representing the solvent and "average residue," respectively.

In addition, either one of the two reference state potentials coming from "0" or "r" may be experienced in different types of interactions, intra-molecular and inter-molecular; the first type of interaction, for occurrence of inter-residue contacts within a single chain or monomer, and the second at the interface between two molecules. The differences between the potentials of mean force for the intra-molecular case and the inter-molecular cases, shortly referred to as the intra- and inter-molecular potentials, will be explored in the present study with respect to both reference states, coming from solvent exposure and from exposure to an average residue.

The solvent- and residue-mediated effective contact potentials e^r_{ij} and e^0_{ij} are defined for intra-molecular interactions in monomers as (Miyazawa & Jernigan, 1985, 1996; Bahar & Jernigan, 1997)

$$e^0_{ij}(\text{intra}) = W_{ij}(\text{intra}) + W_{00}(\text{intra}) - W_{i0}(\text{intra}) - W_{j0}(\text{intra}) \tag{1}$$

and

$$e^r_{ij}(\text{intra}) = W_{ij}(\text{intra}) + W_{rr}(\text{intra}) - W_{ir}(\text{intra}) - W_{jr}(\text{intra}) \tag{2}$$

where W_{ij} is the database extracted potential of mean force corresponding to residues i and j, these being located within a distance r_c sufficiently close for contact; W_{00} and W_{i0} refer to the solvent-solvent and solvent-residue (i) pairs; W_{rr} is the average potential between all residue pairs (r-r), and W_{ir} that between a residue of a given type i and all residues (r) in folded structures. A cutoff distance of r_c = 6.5 Å suitably includes all sites within a first coordination shell in the neighborhood of a central interaction site (Miyazawa & Jernigan, 1985; Bahar & Jernigan, 1997). The argument (intra) in Equations 1 and 2 indicates that these potentials are derived by examining a dataset of *intra-molecular* interactions, and are thus representative of the effective potentials between residue pairs within a single chain or a monomer.

Similar expressions may also be used for defining the potentials operating at interfaces, i.e.,

$$e^0_{ij}(\text{inter}) = W_{ij}(\text{inter}) + W_{00}(\text{inter}) - W_{i0}(\text{inter}) - W_{j0}(\text{inter}) \tag{3}$$

and

$$e^r_{ij}(\text{inter}) = \mathrm{W_{ij}}(\text{inter}) + W_{rr}(\text{inter}) - W_{ir}(\text{inter}) - W_{jr}(\text{inter}). \tag{4}$$

Here, as indicated by the arguments, the potentials refer to inter-molecular interactions, the interacting residues belonging to different chains. Accordingly, the interface region in a dataset of

protein-protein complexes, or multimeric proteins is considered for extracting the potentials expressed by Equations 3 and 4.

The procedure for the evaluation of the potentials W_{ij} and W_{i0} for a given set of databank structures is summarized in Materials and methods. W_{ir} is the weighted average of W_{ij} values over the 20 different types of residue j, the contribution of each residue pair being weighted according to the number of occurrences of the particular contacts; and W_{rr} is found from the further averaging of W_{ir} over all residue types i.

The potentials W_{i0} and W_{ir} characterize the single-body behavior of the amino acid of type i in two different environments, i.e., with a solvent contact or participating in a contact with an average native-like residue. These terms play an important role in determining the effective inter-residue potentials, as may be deduced from Equations 1–4. In particular, the residue-solvent potentials of mean force, W_{i0}, will be distinguished by their strong and unique dependence on residue type, and by their insensitivity to chain connectivity and other effects associated with local packing geometry (which may differ between intra- and inter-molecular contacts). The robustness of these potentials will be evidenced by the reproducibility of results previously obtained for the intra-molecular cases, with the new dataset of interfacial residue pairs.

Results and discussion

Correlations between intra-molecular and inter-molecular potentials

The effective contact potentials between residues at inter-molecular interfaces are presented in Table 1. The upper diagonal and lower diagonal portions refer to solvent-mediated and residue-mediated potentials, e^0_{ij}(inter) and e^r_{ij}(inter), respectively, in RT units. The solvent mediated potentials for self-interactions e^0_{ii}(inter) are underlined for clarity. These will be compared with their counterparts obtained for single chain proteins, i.e., the intra-molecular potentials e^0_{ij}(intra) and e^r_{ij}(intra) presented in our previous studies (Miyazawa & Jernigan, 1985, 1996; Bahar & Jernigan, 1997).

A general feature apparent from the first examination of contact potentials is the significantly wider range of e^0_{ij} values compared to e^r_{ij} values. This immediately signals that the solvent plays a major role in inducing residue specificities. The dominant effect of residue-solvent interactions was also observed in the case of intra-molecular potentials.

In Figure 1, the 210 distinct potentials (20 self-interactions of type $[i,i]$ and 190 cross-interactions of type $[i,j]$, $i \neq j$) obtained for the pairs at interfaces are compared with those observed (Bahar & Jernigan, 1997) for the intra-molecular cases. Figures 1A and 1B display the potentials e^0_{ij} and e^r_{ij} associated with the two different reference states. The abscissa and ordinate correspond to intra- and inter-molecular potentials, in both parts. The best fitting lines obtained by linear regression, and the corresponding equations are shown. The solvent-mediated potentials (Fig. 1A) are not significantly different between the intra-molecular and inter-molecular cases, as evidenced by their relatively high (0.89) correlation coefficient. The residue-mediated ones, on the other hand, exhibit a lower correlation coefficient (0.69), indicating that the preferences of residues sequestered in protein interiors are not as strong as those on the surface, and therefore these are more sensitive to changes in environment.

The extent of correlation between the two sets of potentials, intra-molecular and inter-molecular, may be further analyzed by considering the behavior of individual residues. Plots similar to Figure 1, drawn separately for each type of amino acid, yield the correlation coefficients presented in Table 2. The two columns therein refer to the results found with reference to the two different environments, i.e., solvent and protein interior.

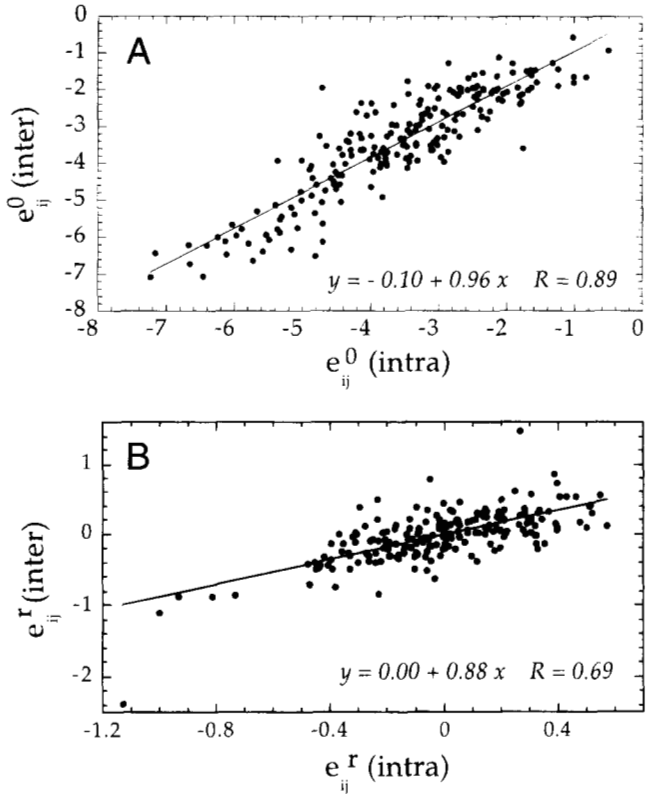

Fig. 1. Comparison of the effective inter-molecular and intra-molecular inter-residue contact potentials. The parameters obtained for interface regions of protein-protein complexes, or multimeric proteins (ordinate) are plotted against those extracted from single chain monomers (abscissa). (**A**) Solvent-mediated (e^0_{ij}) and (**B**) residue-mediated (e^r_{ij}) potentials shown in *RT* units. The best fitting line found by linear regression of the data for the 210 distinct pairs, and the corresponding equation and correlation coefficient (R) are displayed in each case. Solvent-mediated potentials are relatively insensitive to the choice of residue pairs, whether inter-molecular or intra-molecular, in their derivation, as illustrated in **A**. Residue-mediated potentials, however, do exhibit some dependence.

The results in Table 2 show that the solvent-mediated potentials determined for interfacial residues, and those for residues in a given monomer, generally exhibit high correlations as already suggested in Figure 1A. This indicates their robustness with respect to the choice of interacting residue pairs, whether the intra-molecular case or at inter-molecular interfaces, considered for their evaluations. Residue-mediated potentials, on the other hand, exhibit more variety. Some are almost uncorrelated (His, Tyr, Gln, Gly; correlation coefficient ≤ 0.35), others exhibit weak (≤ 0.70) correlations (Ala, Phe, Trp, Ser, Thr, Asn), with the remaining (hydrophobic-aliphatic and charged) being strongly correlated. Thus, the aromatic and polar residues are subject to distinctly different interaction energetics in the intra-molecular cases and at the inter-molecular interfaces, when the residue-mediated potentials are considered; whereas the aliphatic-hydrophobic and charged amino acids preserve the same characteristics, both for intra- and inter-molecular cases. The

Table 1. *Inter-residue contact potentials at interfaces (RT units)*[a]

		G	A	V	I	L	C	M	F	Y	W	S	T	D	N	E	Q	K	R	H	P
G	−0.27	−1.77	−2.24	−2.77	−3.50	−3.79	−3.43	−3.02	−3.76	−3.34	−3.77	−1.68	−1.81	−1.84	−1.63	−1.58	−3.02	−1.32	−2.71	−2.47	−1.01
A	0.08	−0.38	−3.51	−3.86	−4.45	−5.02	−3.99	−4.42	−4.43	−3.96	−4.66	−2.49	−2.38	−2.74	−2.06	−2.69	−3.65	−1.91	−3.26	−3.29	−1.78
V	0.17	−0.11	−0.49	−4.86	−5.31	−6.03	−4.82	−5.16	−5.33	−4.54	−4.70	−2.95	−2.87	−3.28	−2.76	−3.20	−4.29	−2.19	−3.70	−4.14	−2.43
I	0.05	−0.08	−0.33	−0.37	−5.97	−6.67	−5.53	−5.37	−6.11	−5.39	−5.79	−3.47	−3.61	−3.46	−3.42	−3.99	−4.65	−2.88	−4.29	−4.55	−3.09
L	0.33	−0.09	−0.48	−0.50	−0.42	−7.16	−6.13	−6.24	−6.65	−5.67	−6.40	−4.12	−4.28	−4.16	−3.94	−4.35	−5.36	−3.24	−5.01	−4.85	−3.75
C	−0.26	−0.00	−0.22	−0.30	−0.34	−2.38	−7.23	−5.07	−4.81	−4.71	−4.70	−3.41	−2.92	−3.56	−1.76	−3.02	−4.37	−2.28	−4.41	−4.79	−2.97
M	0.31	−0.27	−0.40	0.00	−0.30	−0.08	−0.74	−5.89	−5.58	−5.01	−5.49	−3.33	−3.33	−3.69	−3.05	−3.76	−4.46	−2.36	−4.13	−4.47	−3.11
F	−0.15	0.01	−0.27	−0.44	−0.40	0.49	−0.14	−0.71	−6.45	−5.33	−5.72	−4.01	−3.85	−3.52	−3.89	−4.10	−4.89	−2.70	−4.51	−3.82	−3.46
Y	−0.21	−0.00	0.03	−0.12	0.08	0.10	−0.05	−0.07	0.20	−4.58	−5.12	−3.43	−3.33	−3.63	−3.14	−3.62	−4.75	−2.64	−4.32	−3.78	−2.92
W	−0.31	−0.38	0.20	−0.26	−0.31	0.44	−0.20	−0.13	−0.02	0.27	−5.16	−2.86	−3.48	−3.81	−3.31	−3.84	−4.87	−3.12	−4.54	−4.50	−3.96
S	0.08	0.08	0.25	0.33	0.26	0.02	0.25	−0.14	−0.03	0.86	−0.14	−2.14	−2.22	−2.34	−2.13	−2.57	−3.06	−1.91	−2.78	−3.12	−1.56
T	−0.11	0.14	0.27	0.14	0.04	0.46	0.19	−0.02	−0.00	0.18	−0.27	−0.22	−2.12	−2.41	−2.05	−2.34	−3.44	−1.64	−2.43	−2.73	−1.24
D	*0.11*	*0.02*	*0.11*	*0.54*	*0.41*	*0.06*	*0.09*	*0.55*	−0.04	*0.11*	−0.13	−0.26	−0.07	−2.47	−2.50	−2.35	−3.15	−2.47	−3.92	−2.93	−1.24
N	−0.09	0.31	0.22	0.16	0.23	1.46	0.32	−0.22	0.04	0.20	−0.32	−0.30	−0.51	−0.40	−1.99	−2.43	−3.00	−1.63	−2.67	−3.05	−1.01
E	0.50	0.20	0.30	0.13	0.34	0.72	0.13	0.09	0.08	0.20	−0.25	−0.07	0.17	−0.30	−0.20	−2.85	−3.45	−2.83	−4.05	−3.15	−1.70
Q	−0.11	0.08	0.05	0.33	0.17	0.21	0.27	0.14	−0.21	0.01	0.10	−0.32	0.21	−0.05	0.03	−0.38	−4.71	−2.56	−4.13	−4.20	−2.67
K	−0.17	0.06	0.40	0.32	0.52	−0.50	0.61	0.57	0.14	1.00	−0.50	−0.29	−0.87	−0.44	−1.11	0.00	−0.22	−1.02	−2.11	−2.14	−0.50
R	−0.10	0.16	0.35	0.37	0.21	−0.12	0.30	0.22	−0.07	0.04	0.08	0.38	−0.87	−0.02	−0.87	−0.11	0.15	−0.26	−3.98	−3.24	−2.43
H	0.01	0.01	−0.23	−0.02	0.25	−0.64	−0.17	0.79	0.34	−0.05	−0.38	−0.05	0.00	−0.52	−0.10	−0.31	−0.01	0.35	0.38	−3.08	−1.80
P	*0.13*	*0.19*	*0.16*	*0.11*	*0.01*	*−0.15*	*−0.14*	*−0.19*	*−0.13*	*−0.85*	*−0.15*	*0.11*	0.36	0.19	0.01	*−0.11*	0.30	−0.17	0.33	−0.03	−0.83
	G	**A**	**V**	**I**	**L**	**C**	**M**	**F**	**Y**	**W**	**S**	**T**	**D**	**N**	**E**	**Q**	**K**	**R**	**H**	**P**	

[a]Upper diagonal parts are the solvent mediated potentials, e_{ij}^{0}(inter) for residue pairs indicated on the left column and first row; lower diagonal parts are the residue-mediated potentials, e_{ij}^{r}(inter), listed for the pairs indicated on the left column and last row.

Table 2. *Correlations between effective inter-residue contact potentials in the intra-molecular case and at inter-molecular interfaces*[a]

i	$C_i(e_{ij}^0)$	$C_i(e_{ij}^r)$
GLY	0.89	0.32
ALA	0.89	0.58
VAL	0.91	0.78
ILE	0.94	0.86
LEU	0.95	0.93
SER	0.84	0.69
THR	0.82	0.54
ASP	0.85	0.83
ASN	0.72	0.63
GLU	0.82	0.90
GLN	0.82	0.26
LYS	0.81	0.89
ARG	0.86	0.82
CYS	0.87	0.79
MET	0.91	0.76
PHE	0.90	0.50
TYR	0.89	0.00
TRP	0.88	0.59
HIS	0.83	0.00
PRO	0.92	0.75

[a] $C_i(e_{ij}^0)$ is the correlation coefficient obtained between e_{ij}^0 values obtained in the intra-molecular case and the inter-molecular case for inter-residue interactions involving residue i. $C_i(e_{ij}^r)$ is its counterpart for e_{ij}^r values.

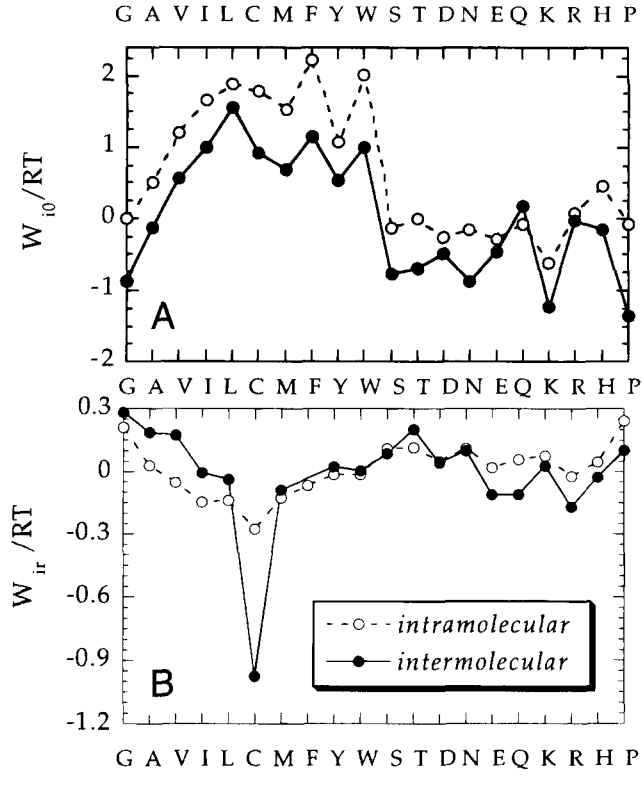

Fig. 2. Potentials of mean force W_{i0} and W_{ir} between (**A**) residue type i and solvent (0), and (**B**) residue type i and "average residue" (r) in folded structures, presented as a function of residue type (single letter amino acid code on the abscissa). The filled circles (and solid lines) refer to the inter-molecular inter-residue potentials; these are obtained using residue pairs located at protein-protein interfaces. The open circles (and dashed lines) are the intra-molecular inter-residue potentials, extracted from single chain proteins. No significant difference between the two sets is observed, apart from a slight enhancement of interactions at interfaces.

sensitivity of certain e_{ij}^r values to the choice of proteins (monomeric tertiary structures or interfacial regions of quaternary structures) used as the dataset is understandable in view of their weak (small in magnitude) preferences, when mediated by other residues.

Importance of one-body potentials in different reference states

The above analysis suggests that the one-body potentials W_{i0} and W_{ir} play a dominant role in controlling the effective inter-residue potentials. These are displayed in Figures 2A and 2B, respectively. The results obtained for inter-molecular interfaces are shown by the filled circles connected by the solid lines to guide the eye, and those previously found (Bahar & Jernigan, 1997) for intra-molecular interactions by the open circles and dashed lines. Each point represents one type of amino acid, as indicated by the single letter code along the abscissa.

The results for intra-molecular and inter-molecular contacts exhibit a close similarity. In parallel with the e_{ij}^0 values, the residue-solvent potentials W_{i0} (intra- or inter-) exhibit a greater variability, covering a total range of about 3 RT, confirming that the broad range of e_{ij}^0 values originates in the contribution of relatively large residue one-body potentials W_{i0}. On the other hand, the average potentials between residue type i and folded residues in the surroundings, expressed by W_{ir} values, is about one order of magnitude smaller, if the outlier value of the potential W_{Cr}(inter) corresponding to cysteine is omitted.

An alternative comparison of the single-body terms occurring in intra-molecular cases and at inter-molecular interfaces is presented in Figure 3. Figure 3A compares the one-body potentials W_{i0}(intra) and W_{i0}(inter) for solvent exposure; and Figure 3B gives those, W_{ir}(intra) and W_{ir}(inter), for the "residue" environment. It is clear that the potentials W_{i0} are insensitive to the choice of dataset structures, values being comparable within monomers (intra-molecular) and at interfaces (inter-molecular), as indicated by the slope (0.99) and correlation coefficient (0.92) of the best fitting line in Figure 3A. W_{ir} values, on the other hand, are more environment dependent, which could be attributed to the slight changes in the packing characteristics of residues at inter-molecular interfaces, compared to those in the monomeric tertiary structures.

The fact that the residue-solvent interactions, W_{i0}, exhibit generally a pronounced and consistent specificity both at inter-molecular interfaces and in the intra-molecular cases, is a consequence of the strong solvent effect. This observation has the following important implications: Residue-mediated potentials e_{ij}^0 might be used with confidence for analyzing both monomeric proteins and protein-protein inter-molecular interfaces, and for providing guidance for structural preferences affecting folding and binding processes, insofar as specific residue contacts replace residue-solvent contacts.

We note that the close similarity between the inter-molecular and intra-molecular solvent-mediated potentials is consistent with a recent study of structural motifs at protein-protein interfaces in

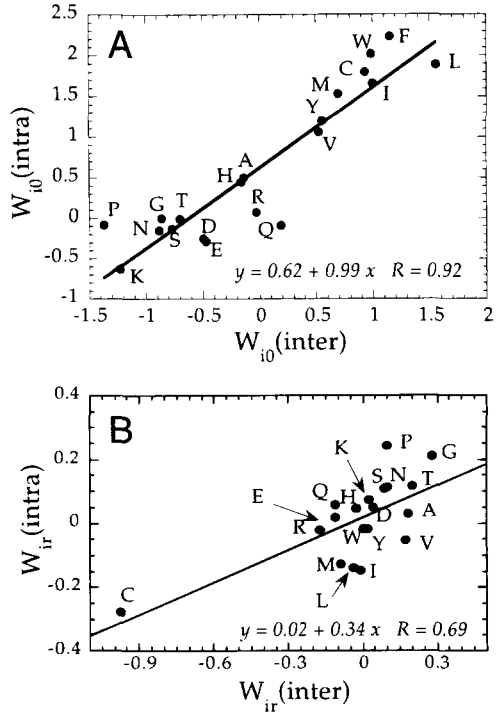

Fig. 3. An alternative method for comparing the inter-molecular and intra-molecular single-body potentials. Here the ordinates and abscissa refer to results obtained for the intra-molecular and inter-molecular cases, respectively; the (**A**) W_{i0} and (**B**) W_{ir} values are displayed. Residue-solvent potentials (W_{i0}) are more specific: they exhibit a strong dependence on residue type, which persists in both intra-molecular and inter-molecular cases (R = 0.92); whereas specificities are significantly weaker with reference to native-like packed residues, as evidenced by the clustering of most residues in **B**, and by the departure of the results derived for inter-molecular interfaces from those obtained for intra-molecular cases (R = 0.69). The differences between the single-body characteristics of amino acids for the two different reference states, solvent-exposure and residues-neighborhood, are also manifested in the ranges of the respective potentials, $3RT$ and $0.5RT$ (excluding Cys).

comparison to those occurring in protein cores, which showed that, despite the absence of chain connectivity, the global features of the architectural motifs, present in monomers, recur at the interfaces (Tsai et al., 1997; Tsai & Nussinov, 1997); the details of the motifs may vary, which could explain the small (in magnitude) differences in residue-mediated potentials.

Reduction of the set of empirical parameters

The robustness of residue-mediated potentials further suggests that, to a good approximation, these may be estimated by using a smaller number of parameters. For example, 20 energy parameters accounting for the single-body solvation or hydrophobicity characteristics of each of the different types of amino acids, and only a few (2–3) additional parameters accounting for particular two-body (residue-residue) interactions, which are more pronounced, could be determined. The suitability of such an approximation was supported by the recent analysis of Li et al. (1997). Therein, the 20 × 20 matrix of solvent-mediated potentials e_{ij}^0 derived by Miyazawa and Jernigan (1985, 1996) was shown to be well described by a total of 22 parameters, after eigenvalue decomposition of the original matrix, and retaining the dominant two eigenvectors (which are interdependent) and eigenvalues.

To find a representative reduced set of energy parameters (e_{ij}^*), describing the solvent-mediated contact potentials, an optimization scheme, based on a minimization of the difference between database extracted values, and newly estimated values, was adopted. The calculation procedure is described in Materials and methods. The results are presented in Table 3. Therein, the parameter W_i^* refers to the single-body potential characteristic of residue-type i irrespective of its occurrence in intra-molecular or inter-molecular regions. $W_i^*$is used in the expression

$$e_{ij}^* = W_i^* + W_j^* + \Delta W_{ij}^* + W_{00}^* \qquad (5)$$

for evaluating the new set of energy parameters. Here, ΔW_{ij}^* is a two-body term that is set to zero except for the following pairs with distinctive attractive interactions: pairs of hydrophobic residues $[H,H]$, oppositely charged amino acids $[+,-]$, and disulfide bridges $[C,C]$. The respective ΔW_{ij}^* values are -0.3, -0.8, and -1.1. W_{00}^* is the newly optimized solvent-solvent interaction parameter, the absolute value of which may be readily adjusted upon comparison of the results (e_{ij}^*) with known e_{ij}^0 values. This simple rule yields contact potentials that reproduce almost exactly the well-established e_{ij}^0 values obtained by Miyazawa and Jernigan (1985, 1996). It is interesting to note that the correlation coefficients between the two sets of parameters is of the order of 0.99, for each of the 20 types of amino acids, as presented in the third column of Table 3. Thus, 23 parameters, comprised of 20 single-

Table 3. *Reduced set of energy parameters for evaluating inter-residue interactions for intra-molecular and inter-molecular cases*[a]

i	W_i^*	Correlation coefficient[b]
GLY	−0.845	0.992
ALA	−0.531	0.996
VAL	0.633	0.995
ILE	1.087	0.996
LEU	1.502	0.996
SER	−1.076	0.992
THR	−0.828	0.993
ASP	−1.302	0.992
ASN	−1.104	0.989
GLU	−1.334	0.973
GLN	−1.038	0.991
LYS	−1.648	0.973
ARG	−1.043	0.983
CYS	0.246	0.996
MET	0.707	0.996
PHE	1.512	0.997
TYR	0.355	0.991
TRP	0.656	0.990
HIS	−0.429	0.982
PRO	−0.907	0.995

[a]See Equation 5 for the use of the tabulated single-body potentials in evaluating inter-residue potentials

[b]Correlation coefficients refer to those between e_{ij}^* values presently calculated using Equation 5, and those e_{ij}^0(inter) reported by Miyazawa and Jernigan (1996).

body (W_i^*) and 3 distinct two-body (ΔW_{ij}^*) terms, suffice to describe the solvent-mediated inter-residue interactions, presently shown to operate both in the intra-molecular cases and at inter-molecular interfaces.

The single-body potentials W_i^* depart only slightly from W_{i0} values. In view of the small magnitude of W_{ir} values, it is natural to expect that the potentials W_{i0} play a major role in determining W_i^* values. A strong correlation between these two sets is indeed observed by plotting the W_i^* values against the W_{i0}, as illustrated in Figure 4.

Classification of residues on the basis of their one-body energetics; clustering of polar and hydrophobic residues

An interesting feature observed in Figure 4, which was also discernible in Figure 3A, is the occurrence of a separation between hydrophobic and polar residues, as indicated by the broken line dividing different residue types. In particular, polar residues are clustered in a more restricted region of the figure, certain residue pairs or triplets such as (Thr, Gly), (Ser, Asn, Gln) being almost indistinguishable. Here it is clear that the term "polar residues" is employed in a relaxed sense, to include all residues other than those (Leu, Phe, Ile, Met, Val, Trp, Tyr) usually classified as hydrophobic. Not surprisingly, Ala and His on the one side, and Tyr on the other, are located nearer the boundary between the two regions. Hydrophobic residue values are more broadly spread in general, which may be attributed to their stronger preferences, both aversion to solvent-exposure and propensity for forming a core, compared to other residues.

The possibility of reducing the amino acid types to two broad classes when their single-body energetics, which predominantly controls their folding and binding preferences, are taken into consideration justifies to some extent the simple H/P model chains (Dill et al., 1995). Notably, however, the individual residue type variability within each of the two groups is substantially larger than the gap between the two classes, as can readily be seen in Figure 4. Such clear behavior dictates against the use of the H/P conceptual model to represent real proteins.

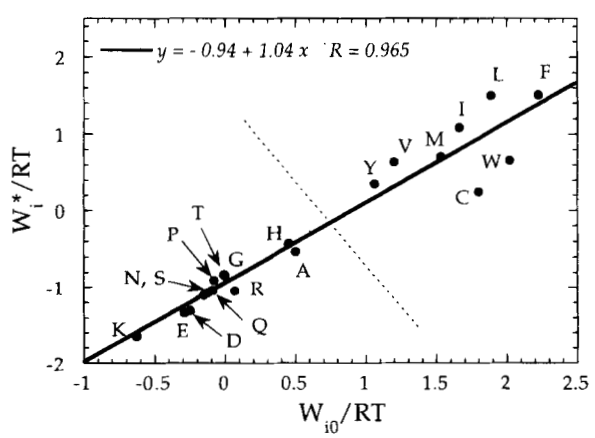

Fig. 4. Reduced set of parameters W_i^* representative of single-body preferences of individual amino acids, plotted against the intra-molecular solvent-residue potentials obtained (Bahar & Jernigan, 1997) for intra-molecular cases. The best fitting line and the corresponding equation are displayed. We note the clustering of hydrophobic and polar residues into two groups, delimited by the dashed line.

Conclusion

There are two important conclusions drawn from the present study:

1. The discriminatory ability of residue-residue interactions is strong and unequivocal only in the presence of solvent. In other words, inter-residue contact preferences are more selective, more pronounced, only when mediated by water. For a reference state of other packed native-like residues, the preferences are much weaker and dependent on the environment.

2. Inter-residue interaction potentials at inter-molecular interfaces bear a close resemblance to those operating intra-molecularly, provided that the solvent-mediated effective contact potentials (e_{ij}^0) are being considered. Perturbations in residue-specific interactions induced by the absence of chain connectivity and by the slight differences in inter-molecular structural motifs are minor.

A corollary to these conclusions is that the dominant effect of hydrophobicity, or solvation effect, since the "solvent-mediation" is manifested principally in the preferential clustering of hydrophobic residues in the protein interior, or at buried regions of interfaces between complexes, and by the solvation of hydrophilic groups on the solvent-exposed regions of either monomers or interfaces. In particular, the drive for burial of hydrophobic patches originally on the surface of monomers appears to be an important factor guiding multimeric complexation, as can also be inferred from other studies (Young et al., 1994; Lijnzaad & Argos, 1997). Not surprisingly, highly simplified sets of energy parameters based on solvation energetics, and mainly reflecting the hydrophobicity scale of individual residues, have been successfully utilized in previous evaluations of conformational preferences. The presently proposed reduced set of parameters is another example of such a scale, consistent with the dominant effect of (specific) residue-solvent interactions. An important conclusion is that the same set of energy parameters is valid, to a good approximation, for both the inter-molecular and intra-molecular regimes, a result that removes long-standing reservations about adopting the same force fields for binding and for folding. These solvent-mediated potentials are more appropriate for folding considerations, and only following the achievement of a sufficiently compact, closed structure would residue mediation become more operative and the specific effects related to those particular packing characteristics come into play.

Materials and methods

Materials

The Protein Data Bank (PDB) (Bernstein et al., 1977) structures considered in evaluating the inter-molecular inter-residue potentials are listed in Table 4. These are complexes, or multimers, the structures of which were determined by X-ray crystallography at a resolution of 3.0 Å or better. Structures determined by NMR spectroscopy were not included in the list. To eliminate homologous sequences, PDB files with similar COMPND statements were excluded. The homologous sequences were further filtered out by pairwise sequence alignments using the program CLUSTALW (Thompson et al., 1994).

The amino acid composition at the inter-molecular interface was observed to be similar to that of the intra-molecular cases, as

Table 4. *PDB codes of the structures used in examining residue pairs at inter-molecular interfaces*

148l, 1abj, 1abr, 1acb, 1acm, 1acy, 1ahg, 1apt, 1atn, 1atp, 1ayb, 1bbr, 1bbt, 1bgs, 1bll, 1bmv, 1bn2, 1bql, 1cau, 1cdm, 1cgi, 1cgl, 1cgs, 1cka, 1coh, 1cpc, 1ctp, 1dbj, 1dfb, 1dhl, 1dlh, 1eap, 1eed, 1epl, 1ept, 1ets, 1fcd, 1fdl, 1fgv, 1fiv, 1fpt, 1fru, 1glb, 1gmd, 1gss, 1hag, 1hai, 1hcd, 1hcg, 1hcn, 1hco, 1hiv, 1hle, 1hoc, 1hpb, 1hpg, 1hsb, 1htl, 1htr, 1huc, 1ice, 1igi, 1ind, 1laf, 1lgc, 1lpb, 1lpr, 1ltb, 1mam, 1mda, 1mdy, 1mee, 1min, 1mmo, 1mrd, 1myp, 1nih, 1ola, 1pad, 1pcr, 1pll, 1prt, 1psd, 1pso, 1pts, 1pya, 1rlc, 1rmf, 1rmu, 1scm, 1scu, 1sha, 1shd, 1smr, 1sps, 1srn, 1ssa, 1tab, 1tci, 1tec, 1tmc, 1tme, 1tmn, 1tnr, 1tpa, 1tro, 1tya, 1tyt, 1vab, 1wat, 1wsy, 2aai, 2ach, 2bbv, 2bpa, 2btf, 2cwg, 2fb4, 2fbj, 2hmg, 2hnt, 2ins, 2lal, 2lgs, 2ltn, 2mev, 2mth, 2pcc, 2pcd, 2pka, 2plv, 2ptc, 2rmb, 2tpi, 2wrp, 3aah, 3cpa, 3eca, 3gct, 3hvt, 3ins, 3mon, 3rub, 3sc2, 3sga, 3sic, 4fab, 4ins, 4ts1, 6rlx, 7api, 7est, 7hvp

illustrated in Figure 5. Here, all residue pairs, belonging to two distinct monomers, and located within a side-chain to side-chain interaction distance of 6.4 Å or less, are included. The total number of interfacial residues presently examined was 17,102, Cys344 and Trp360 being the fewest. The curves display the normalized probability distributions of amino acids of different types (abscissa), calculated for inter-molecular interfaces (filled circles and solid lines), and for intra-molecular cases (open circles and dashed lines). It should be noted that the results for certain pairs are subject to statistical errors. Examples are Trp-Trp, Trp-Cys, Met-Met, Cys-Gln, and Lys-Cys, whose numbers of observation were lower than 20.

The two-site per residue virtual bond model and the statistical method used in our previous study (Bahar & Jernigan, 1997) for extracting inter-residue potentials for monomeric globular proteins is adopted here for calculating *inter-molecular* potentials. Accordingly, each residue is represented by two sites, one on the backbone, and the other on the amino acid side chain. C^{α} atoms are chosen as the backbone sites, and side-chain centroids or centers of functional groups, are selected as the second site, as previously described (Bahar & Jernigan, 1996).

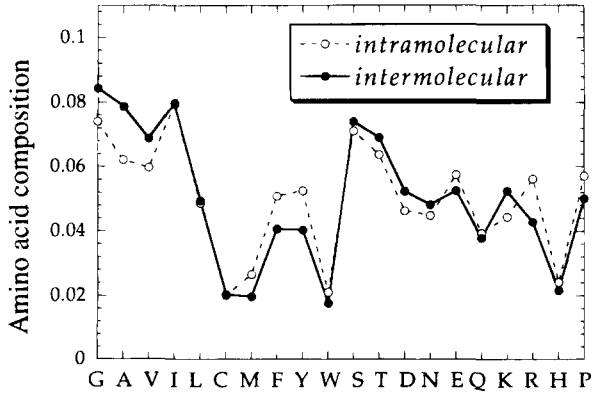

Fig. 5. Amino acid composition at inter-molecular interfaces (filled circles and solid lines) and intra-molecular cases (open circles and dashed lines). The distribution curves are normalized. The types of amino acids are indicated along the abscissa. Clearly the differences are relatively small.

Method

The potential of mean force ΔW_{ij} between residues i and j, expressed relative to the average potential W_{rr} between all residue pairs, is given by Bahar and Jernigan (1997) as

$$\Delta W_{ij} = W_{ij} - W_{rr} = -kT \ln\left[\int_0^{rc} g_{ij}(s)\, ds \Big/ \int_0^{rc} g_{rr}(s)\, ds \right] \tag{6}$$

where k is the Boltzmann constant, T is the absolute temperature, and $g_{ij}(s)$ is the normalized radial distribution function. $g_{ij}(s)$ is expressed in terms of the number $N_{ij}(s)$ of pairs of type i and j (or r) located at a distance $s \pm \Delta s$ from each other, as

$$g_{ij}(s) = (4\pi s^2)^{-1} N_{ij}(s) \Big/ \int_0^{\infty} (4\pi s^2)^{-1} N_{ij}(s)\, ds. \tag{7}$$

The interaction sites of the residues are taken at the side-chain centroid or functional group, depending on the type of amino acid, as described in our previous work (Bahar & Jernigan, 1996). The integrations in Equations 6 and 7 are conveniently replaced by summations over shells of thickness Δs = 0.4 Å, and the upper limit of integration in Equation 6 is taken as 13 Å, a separation beyond which no statistically significant data are obtainable (Bahar & Jernigan, 1997). The resulting intra-molecular potentials were shown in previous threading calculations to successfully recognize the correct sequence-structure matches in a series of test proteins (Bahar et al., 1997). We note that Equation 2 may be rewritten as

$$e_{ij}^{r} = \Delta W_{ij} - \Delta W_{ir} - \Delta W_{jr} \tag{8}$$

using Equation 3. The solvent-residue potentials W_{i0}, on the other hand, are estimated here using the relationship

$$W_{i0} = -kT \ln[P_i(\text{solvent-exposed})/P_i(\text{core})] \tag{9}$$

where P_i(solvent-exposed) is the probability of being solvent-exposed for a residue of type i with respect to the set of all solvent-exposed residues, and P_i(core) is that of being buried in the core. A given residue i is accepted to be solvent-exposed if its coordination number z_i within a sphere of radius 6.4 Å is below or equal to four, and to be in the core if $z_i \geq 7$. Side-chain centers were considered for determining the coordination numbers. The threshold values for classifying the two types of environment were determined by examining packing densities, or the coordination number distributions of amino acids of different types in the folded structures. This method of estimating the solvent-residue potentials W_{i0} was verified to reproduce, with a correlation coefficient above 0.96, the potentials obtained (for the intra-molecular cases) using the somewhat more complex quasi-chemical approximation, and was therefore adopted for estimating the solvent-residue potentials for residues located at inter-molecular interfaces. Finally, W_{00} at inter-molecular interfaces is adjusted (as W_{00}^{*}) so as to equate the average $\langle e_{ij}^{0} \rangle$ over all solvent-mediated inter-residue contacts observed at inter-molecular interfaces to that of the intra-molecular case.

Optimization scheme for determining a reduced set of energy parameters

At the first step of our optimization scheme, the potentials of mean force W_{ij} are approximated as

$$W_{ij} = [W_{ir} + W_{jr}]/2 + \Delta W_{ij}^*. \quad (10)$$

Here ΔW_{ij}^* is the second order correction term, which may be taken as zero, as a starting point. The latter was, in fact, constrained to be zero for all pairs, except for a few cases. Equation 10 permits us to express the newly generated solvent-mediated contact potentials as

$$e_{ij}^* = (W_{ir}/2 - W_{i0}) + (W_{jr}/2 - W_{j0}) + \Delta W_{ij}^* + W_{00}^*. \quad (11)$$

The one-body terms in parentheses are optimized by an iterative scheme minimizing the mean-square deviation between the estimated e_{ij}^* values and those directly found (e_{ij}^0) from statistical examination of databank structures. The resulting optimized values $[W_{ir}/2 - W_{i0}]^*$ are simply designated as W_i^*. The value $W_{00}^* = -3.645RT$ is adopted, which ensures that the mean value of the set e_{ij}^* matches that of known potentials e_{ij}^0.

References

Bahar I, Jernigan RL. 1996. Coordination geometry of non-bonded residues in globular proteins. *Fold Des 1*:357–370.

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

Bahar I, Kaplan M, Jernigan RL. 1997. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins 29*:292–308.

Bernstein F, Koetzle T, Williams G, Meyer E, Brice M, Rodgers J, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein databank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. 1995. Principles of protein folding. A perspective from simple exact models. *Protein Sci 4*:562–602.

Finkelstein AV, Badretdinou AY, Gutin AM. 1995. Why do protein architectures have a Boltzmann-like statistics? *Proteins 23*:142–150.

Huber T, Torda AE. 1998. Protein fold recognition without Boltzmann statistics or explicit physical basis. *Protein Sci 7*:142–149.

Jernigan RL, Bahar I. 1996. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol 6*:195–209.

Jones DT, Thornton JM. 1996. Potential energy functions for threading. *Curr Opin Struct Biol 6*:210–216.

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

Li H, Tang C, Wingreen NS. 1997. Nature of driving force for protein folding: A result from analyzing the statistical potential. *Phys Rev Lett 79*:765–768.

Lijnzaad P, Argos P. 1997. Hydrophobic patches on protein subunit interfaces: Characteristics and prediction. *Proteins 28*:333–343.

Mirny L, Shakhnovich E. 1996. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol 204*:1164–1179.

Miyazawa S, Jernigan RL. 1985. Estimation of effective inter-residue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

Miyazawa S, Jernigan RL. 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.

Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol 258*:367–392.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Skolnick J, Jaroszewski L, Kolinski A, Godzik A. 1997. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci 6*:676–688.

Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol 257*:457–469.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res 22*:4673–4680.

Torda AE. 1997. Perspectives in protein fold recognition. *Curr Opin Struct Biol 7*:200–205.

Tsai C, Nussinov R. 1997. Hydrophobic folding units at protein-protein interfaces: Implications to protein folding and protein-protein association. *Protein Sci 6*:1426–1437.

Tsai C, Xu D, Nussinov R. 1997. Structural motifs at protein-protein interfaces: Protein cores versus two-state and three-state model complexes. *Protein Sci 6*:1793–1805.

Young L, Jernigan RL, Covell DG. 1994. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci 3*:717–729.

Zhang L, Skolnick J. 1998. How do potentials derived from structural databases relate to "true" potentials? *Protein Sci 7*:112–122.