J. Mol. Biol. (1997) **266**, 195–214

JMB

# Inter-residue Potentials in Globular Proteins and the Dominance of Highly Specific Hydrophilic Interactions at Close Separation

**I. Bahar[1,2] and R. L. Jernigan[1*]**

[1]*Molecular Structure Section Laboratory of Mathematical Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, MSC 5677, Room B-116, Bldg. 12B, Bethesda MD 20892-5677, USA*

[2]*Chemical Engineering Department and Polymer Research Center, Bogazici University, and TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815 Istanbul, Turkey*

**Corresponding author*

Residue-specific potentials between pairs of side-chains and pairs of side-chain–backbone interaction sites have been generated by collecting radial distribution data for 302 protein structures. Multiple atomic interactions have been utilized to enhance the specificity and smooth the distance-dependence of the potentials. The potentials are demonstrated to successfully discriminate correct sequences in inverse folding experiments. Many specific effects are observable in the non-bonded potentials; grouping of residue types is inappropriate, since each residue type manifests some unique behavior. Only a weak dependence is seen on protein size and composition. Effective contact potentials operating in three different environments (self, solvent-exposed and residue-exposed) and over any distance range are presented. The effective contact potentials obtained from the integration of radial distributions over the distance interval $r \leqslant 6.4$ Å are in excellent agreement with published values. The hydrophobic interactions are verified to be dominantly strong in this range. Comparison of these with a newly derived set of effective contact potentials for closer inter-residue separations ($r \leqslant 4.0$ Å) demonstrates drastic changes in the most favorable interactions. In the closer approach case, where the number of pairs with a given residue is approximately one, the highly specific interactions between charged and polar side-chains predominate. These closer approach values could be utilized to select successively the relative positions and directions of residue side-chains in protein simulations, following a hierarchical algorithm optimizing side-chain–side-chain interactions over the two successively closer distance ranges. The homogeneous contribution to stability is stronger than the specific contribution by about a factor of 5. Overall, the total non-bonded interaction energy calculated for individual proteins follows a dependence on the number of residues of the form of $n^{1.28}$, indicating an enhanced stability for larger proteins.

*Keywords:* potentials of mean force; globular protein structures; radial distribution functions; homogeneous interactions; specific interactions

## Introduction

Among the interactions responsible for the stabilization of the native structures in globular poteins, those occurring between sequentially distant but spatially close amino acid residues are recognized to play a dominant role. These are referred to as non-bonded or long-range interactions, in the sense that they involve residue pairs that are not near-neighbors in the primary structure, but close to each other in the three-dimensional configuration. Short-range interactions, on the other hand, refer to those occurring between near-neighbor residues along the chain backbone. Studies based on short-range interactions only, or local propensities for secondary structures, succeed in predicting the secondary structure up to about 72% accuracy (Rost & Sander, 1993). However, it is clear that the tertiary interaction problem is more difficult. The native state is stabilized by various residue-specific, non-bonded interactions that hold a protein together in a compact form, in a delicate balance.

Because a large number of degrees of freedom must be optimized simultaneously to reach the most probable state, coarse-grained models, or so-called low-resolution approaches including fewer numbers of conformational variables (Jernigan, 1992) have been adopted in several studies (Crippen & Viswanadhan, 1985; Wilson & Doniach, 1989; Sippl, 1990; Casari & Sippl, 1992; Sippl *et al.*, 1992; Bryant & Lawrence, 1993; Sun *et al.*, 1992; Sun, 1993; Jones *et al.*, 1992; Wallqvist & Ullner, 1994). The original work of Levitt & Warshel (1975) and Levitt (1976) revealed that such simple models could capture the characteristics of the overall folds. The basic idea therein was to combine the atoms into unified groups, or effective interaction sites and connect them by virtual bonds. Threading of sequences through structures indicated that such simplified models together with database-extracted residue-residue contact potentials could effectively discriminate among alternative folding motifs (Covell & Jernigan, 1990; Hendlich *et al.*, 1990; Bowie *et al.*, 1991; Jones *et al.*, 1992; Casari & Sippl, 1992; Maiorov & Crippen, 1992; Bryant & Lawrence, 1993; Kocher *et al.*, 1994; Miyazawa & Jernigan, 1996).

Yet, it is recognized that for design purposes and dynamic simulations, more precise expressions including both the residue-specific distance-dependence of the potentials and the conformational preferences of the polypeptide backbone, should be developed. Incorporation of an artificial distance-dependence to Miyazawa-Jernigan (MJ) (1985) contact potentials, for example, was shown by Park & Levitt (1996) to increase the efficiency of the potentials in discriminating X-ray and near-native folds from amongst an ensemble of decoy conformations. In a recent comparison of detailed and reduced protein models tested for simulating the folding of proteins, Monge *et al.* (1995) emphasized the importance of improving the inter-residue potential functions. That is the purpose of this study. Long-range potentials are elaborated here, using 302 structures from Brookhaven Data Bank (PDB: Bernstein *et al.*, 1977; Abola *et al.*, 1987), which were recently used for characterizing the coordination geometries of non-bonded side-chains (Bahar & Jernigan, 1996). In a separate study, we demonstrated with a thorough analysis of short-range potentials, the importance of combining the two types of potentials (I. B., M. Kaplan & R. L. J., unpublished results).

One deficiency of the MJ contact potentials is their quite weak specificity. The radius for interaction (6.5 Å) used for their derivation is an obvious source of this non-specificity. On average, about six-non-bonded residues are found in the sphere centered on a buried side-chain. Probably not all of these are interacting directly with the central side-chain; consequently, this overcounting of interacting pairs leads to a smoothing of the specificity. Here, we are going to take a smaller radius where only one interacting pair is found on average and, as will be seen, these yield a significantly stronger specificity.

Each residue is represented here by two interaction sites, one on the backbone, and the second on the amino acid side-chain. The side-chain sites are selected on the basis of the specific structure and energy characteristics of the amino acid (Bahar & Jernigan, 1996). Distance-dependent side-chain–side-chain interaction potentials extracted at 0.4 Å resolution are used to estimate the effective contact potentials operating over different distance ranges. For example, we reproduce the MJ contact potentials that were recently updated (Miyazawa & Jernigan, 1996), as one case for integration of the present potentials over the range $r \leqslant 6.4$ Å. We generate effective inter-residue potentials operating in two distinct distance ranges, close ($r \leqslant 4.0$ Å) and distant ($4.0 < r \leqslant 6.4$ Å). It is generally accepted that hydrophobic interactions exert the strongest forces leading to the collapse to a compact globular shape, whereas hydrophilic residues are distinguished by their more specific but weaker effects (Miyazawa & Jernigan, 1985, 1996). That hydrophobicity is a major determinant of protein structure is supported by the success of models based on atomic solvation alone (Wang *et al.* 1995a,b). Although the predominant role of hydrophobic interactions in the range $r \leqslant 6.4$ Å is confirmed in the present study, a contrasting behavior will be revealed for the close distance regime: the hydrophilic interactions are found to gain importance over hydrophobic interactions, and to dominate the contact preferences at close separations. This suggests the successive use of two sets of effective inter-residue potentials corresponding to (1) the broad distance range $2.0 \leqslant r \leqslant 6.4$ Å, and (2) close distance $2.0 \leqslant r \leqslant 4.0$ Å, *via* a hierarchical simulation algorithm, for a more precise discrimination of the structural preferences of globular proteins in the native state.

Finally, we note that the accuracy of statistical potentials extracted from protein structures was recently questioned by Thomas & Dill (1996). In view of their arguments on the limits of applicability of Boltzmann statistics to Data Bank structures, and on the biases in extracted potentials arising from chain connectivity and excluded volume effects, a systematic analysis of effective contact energies as a function of the size and composition of the learning dataset of proteins has been performed here. The reproducibility of the results has been verified using two independent protein sets, Set I and II, presented on the Internet (Bahar & Jernigan, 1996), comprising each $\geqslant$150 proteins. In contrast to the results of Thomas & Dill (1996), no significant dependence on the size of the proteins or on the fraction of hydrophobic residues is observed. Such excluded volume and sequence effects were apparently magnified in their cases for two dimensions and with two-letter model chains of $n \leqslant 18$ monomers. Furthermore, the partition propensity, a property related to the effectiveness of the burial of hydrophobic residues in a given protein, is shown to have only a small perturbing effect on the effective contact potentials. These results, together with successful threading experiments, confirm the adequacy of knowledge-based potentials and low-resolution models as a first-order approach to elucidate sequence-structure relations in proteins.

## Theory

### Potentials of mean force

The potential of mean force between two residues' side-groups $A$ and $B$ expressed relative to the average potential $E_{XX}(r)$ is given (Sippl, 1990) by:

$$\begin{aligned}\Delta E_{AB}(r) &\equiv E_{AB}(r) - E_{XX}(r)\\ &\equiv -RT\ln[\bar{g}_{AB}(r_k)/\bar{g}_{XX}(r_k)]\end{aligned} \qquad (1)$$

where $\bar{g}_{AB}(r)$ is the normalized radial pair distribution function corresponding to the pair A and B located at a distance $r \pm \Delta r$ from each other (see Materials and Methods) and $\bar{g}_{XX}(r)$ is the mean pair radial distribution function averaged over all types ($N = 20$) of amino acid pairs as:

$$\bar{g}_{XX}(r) \equiv N^{-2} \sum_{A=1}^{N} \sum_{B=1}^{N} \bar{g}_{AB}(r) \qquad (2)$$

$E_{XX}(r)$ may be regarded as a homogeneous interaction energy existing between an average pair of amino acid residues in the native state of globular proteins, upon which particular attractive and/or repulsive preferences $\Delta E_{AB}(r)$ are superimposed, leading to the specific interaction potentials $E_{AB}(r)$. In analogy with equation (2), it proves useful to define for each type ($A$) of residue a pair correlation function, $\bar{g}_{AX}(r)$, averaged over all types of interacting partners, as:

$$\bar{g}_{AX}(r) \equiv N^{-1} \sum_{B=1}^{N} \bar{g}_{AB}(r) \qquad (3)$$

The summation in equation (3) includes all contacts [$A$,X] of a given amino acid $A$, with all other non-bonded residues (X) located at a distance $r \pm \Delta r$.

We note that the expression for the potential of mean force presented in equation (1) differs from that adopted in the original work by Sippl (1990), in that the radial distribution functions are normalized here with respect to (1) the total number of contacts observed for each pair of amino acid residues, and (2) the volume $4\pi r^2\Delta r$ of the spherical shell associated with a given inter-residue distance range $r \pm \Delta r$ (see equation (19) in Materials and Methods). The second normalization ensures the convergence of the potentials to vanishingly small values at separations $r \geqslant 12$ Å. Additionally, no dependence on the topological separation of residues $A$ and $B$ along the backbone is included here, all pairs of sites separated by five or more virtual bonds being treated equally. Nearer neighbors, on the other hand, will be subject to a fundamentally different treatment based on Markov statistics, their interactions being strongly constrained by chain connectivity (I. B., M. Kaplan & R. L. J., unpublished results).

### Effective inter-residue contact energies

Effective inter-residue contact energies between the pair of residues $A$ and $B$ located at a separation $r \leqslant r_c$ are given by (Miyazawa & Jernigan, 1985):

$$e_{AB}(r_c) = E_{AB}(r_c) + E_{00}(r_c) - E_{A0}(r_c) - E_{B0}(r_c) \qquad (4)$$

and

$$e_{AB}'(r_c) = E_{AB}(r_c) - (E_{AA}(r_c) + E_{BB}(r_c))/2 \qquad (5)$$

Here $E_{00}(r_c)$ is the solvent-solvent interaction energy representative of all groups of water molecules located at a distance of $r_c$ or closer to each other, $E_{A0}(r_c)$ and $E_{B0}(r_c)$ are the interaction energies between solvent and residues of type $A$ and $B$, respectively. A group of solvent molecules 0 comprises the number of water molecules that are collectively the size of an average residue. $e_{AB}(r_c)$ is the energy difference accompanying the formation of the contact pairs [$A$, $B$] and [0, 0] from the contact pairs [$A$, 0] and [$B$, 0], and $e_{AB}'(r_c)$ is the energy associated with the formation of the contact pair [$A$, $B$] at the expense of the pairs [$A$, $A$] and [$B$, $B$]. $e_{AB}(r_c)$ and $e_{AB}'(r_c)$ will be referred to as solvent-mediated effective contact potentials and effective self contact potentials, respectively, in the following. $e_{AB}'(r_c)$ may be expressed in terms of the radial distribution functions as:

$$\begin{aligned}e_{AB}'(r_c) = &-RT\ln\left[\int_0^{r_c}\bar{g}_{AB}(r)\,dr\right.\\ &\left.\left[\int_0^{r_c}\bar{g}_{AA}(r)\,dr \times \int_0^{r_c}\bar{g}_{BB}(r)\,dr\right]^{-1/2}\right]\end{aligned} \qquad (6)$$

which follows from equations (1) and (5). On the other hand, to evaluate $e_{AB}(r_c)$ is somewhat more complex; it may be conveniently estimated from:

$$\begin{aligned}e_{AB}(r_c) = &-RT\ln\left[q_{AX}(r_c)q_{BX}(r_c)\int_0^{r_c}\bar{g}_{AB}(r)\,dr\right]\\ &- E_{A0}(r_c) - E_{B0}(r_c) + E_{00}^{*}(r_c)\end{aligned} \qquad (7)$$

Here $q_{AX}(r_c)$ represents the average coordination number of residue $A$, on the basis of all inter-residue contacts [$A$, X] within a spherical volume of radius $r_c$, centered about residue $A$. The product $q_{AX}(r_c)$ $q_{BX}(r_c)$ takes into consideration the differences in the intramolecular coordination numbers of different residues. $E_{00}^{*}$ $(r_c)$ includes the solvent–solvent interaction and other non-specific contributions within a sphere of radius $r_c$. $E_{A0}(r_c)$ or $E_{B0}(r_c)$ is determined from:

$$E_{A0}(r_c) = -RT\ln\left[1 - \frac{N_{AX}(r_c)}{q_A(r_c)N_A}\right] + const. \qquad (8)$$

where $N_{AX}(r_c) = \Sigma_B N_{AB}(r_c)$ is the total number of contacts between all residues of type $A$ in the dataset and all other residues located within a distance $r \leqslant r_c$ from $A$, and $N_A$ is the total number of residues of type $A$. Equation (8) is based on a lattice description of protein structure following the Bethe approximation (Miyazawa & Jernigan, 1985, 1996). $q_A(r_c)$ is the total coordination number of residues type $A$. It includes both solvent molecules and all other side-chains located at $r \leqslant r_c$, such that the total number of contacts involving $A$ may be written as:

$$P$$
$$_B N_{AB}(r_c) + N_{A0}(r_c) = q_A(r_c)N_A \qquad (9)$$

This conservation equation simply states that, on average, coordination shells are completed by solvent molecules. Rearrangement yields:

$$\underset{B}{\overset{X}{}} \quad N_{AB}(r_c)/[q_A(r_c)N_A] + N_{A0}(r_c)/[q_A(r_c)N_A] = 1$$
$$(10)$$

We note that the internal coordination number may be expressed as $q_{AX}(r_c) = \Sigma_B N_{AB}(r_c)/N_A$. Equation (9) provides a means of estimating the number $N_{A0}(r_c)$ of effective contacts with solvent molecules. As may be inferred from equation (10), the term in square brackets in equation (8) is equal to the probability of observing an $[A,0]$ contact within $r \leqslant r_c$. In the calculations, $E_{A0}(r_c)$ and $E_{B0}(r_c)$ will be conveniently expressed relative to the solvent–glycine interaction $E_{G0}(r_c)$, and all constant contributions will be incorporated into $E^*_{00}(r_c)$.

Finally, the residue-mediated effective contact energies (Miyazawa & Jernigan, 1985, 1996):

$$e_{AB}{}^{(r)}(r_c) \equiv E_{AB}(r_c) + E_{XX}(r_c) - E_{AX}(r_c) - E_{BX}(r_c) \qquad (11)$$

are found directly from the pair radial distribution functions, using the expression:

$$e_{AB}{}^{(r)}(r_c) = -RT\ln \quad \overset{\kappa Z_{r_c}}{\underset{0}{}} \bar{g}_{AB}(r)\,dr \quad \overset{Z_{r_c}}{\underset{0}{}} \bar{g}_{XX}(r)\,dr$$
$$\overset{\theta Z_{r_c}}{\underset{0}{}} \bar{g}_{AX}(r)\,dr \quad \overset{Z_{r_c}}{\underset{0}{}} \bar{g}_{BX}(r)\,dr \quad \iota_{-1}\lambda \qquad (12)$$

Equations (7) and (12) yield the effective contact energies for two different reference states, solvent-exposed and residue-exposed states. These energies may be evaluated for any distance range of interest upon suitable selection of the integration limits.

## Results and Discussion

### Side-chain pair correlation functions and hydrophobicity profiles

In order to gain an understanding of the general behavior of each type of residue, results have been consolidated into the integrated distributions $\bar{g}_{AX}(r)$ defined by equation (3). The results for hydrophobic and charged residues are illustrated in Figure 1(a) and (b), respectively. The continuous heavier curve is the mean distribution $\bar{g}_{XX}(r)$ defined by equation (2); it serves as a reference. The curves are drawn by interpolating results compiled at 0.4 Å intervals.

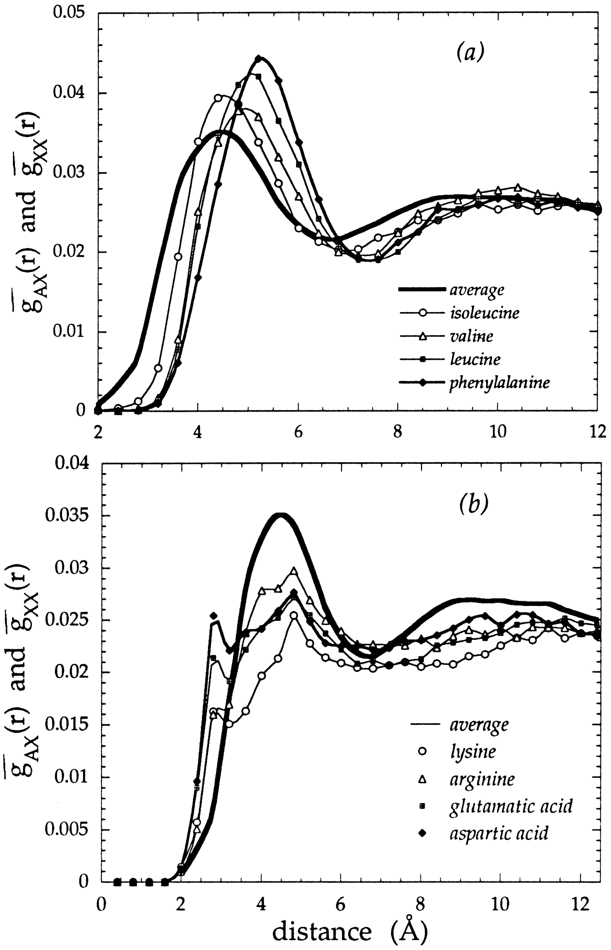

**Figure 1.** Normalized pair radial distribution functions $\bar{g}_{AX}(r)$ for contacts between pairs [$A$, X], for a particular residue type $A$ and all residues X in the same protein, defined by equation (3). The multiple atoms listed in Table 4 of Bahar & Jernigan (1996) characterize the interactions of each type of residue. The mean distribution $\bar{g}_{XX}(r)$ defined by equation (2) is shown by the continuous boldface curve. Curves in (a) are drawn for $A$ = Val, Leu, Ile and Phe; $\bar{g}_{AX}(r)$ for the charged residues $A$ = Arg, Lys, Asp and Glu are displayed in (b). $\bar{g}_{AX}(r)$ curves for the hydrophobic residues lie above the average $\bar{g}_{XX}(r)$, which is characteristic of residues preferentially located in the interior regions of proteins. In the case of charged residues (b) a new peak emerges near 2.5 Å that indicates a tendency for these residues to come into close proximity to particular neighbors. Erratum: in (b), for glutamatic read glutamic.

Two peaks located at about 4.5 Å and 9 Å are observed in the $\bar{g}_{XX}(r)$ curve. These peaks reflect the distance between a given side-chain and those in the first and second shell of neighbors, respectively. Examination of a particular radial distribution, on the other hand, reveals that the inter-residue spacing varies considerably with residue type. For the charged residues (Figure 1(b)), an additional peak appears at about 2.5 Å, indicating that those residues tend to come into closer proximity for some particular neighbors.

The radial distributions provide an indirect estimation of the location, on the surface or in the interior, of different residue types. In fact, the division by $4\pi r^2$ in evaluating the distributions at different separations (see Method) effectively diminishes the contribution of contacts occurring at larger $r$, and emphasizes the relatively shorter-range contacts. As a consequence, those residues buried in the globule, which experience a larger number of contacts at relatively shorter distances, exhibit higher $\bar{g}_{AX}(r)$ values on the observed scale, compared to residues on the surface. For example, the curves for the hydrophobic residues, in Figure 1(a), lie above the average $\bar{g}_{XX}(r)$. Met, Trp and Tyr exhibit (not shown) a similar trend. Charged and polar residues, on the other hand yield $\bar{g}_{AX}(r)$ curves that remain lower than $\bar{g}_{XX}(r)$, except for the extremely small $r$ region where strong specific interactions can occur. Among other residues not shown, Pro approximates the behavior of charged and polar residues; His, Ala and Gly do not show a strong preference for particular locations; Cys, and particularly those forming disulfide bridges, exhibit the strongest apparent hydrophobic behavior.

A quantitative assessment of the hydrophobicity profiles emerging from the present approach is made by evaluating standard deviation coefficients σ[*A*,*B*] for each residue pair as:

$$\sigma[A, B] = sign(\Delta\bar{g}_{AB})h(\bar{g}_{AX}(r) - \bar{g}_{BX}(r))^2 i \qquad (13)$$

Here the brackets represent the average over all separations and all types of residues X, and $sign(\Delta\bar{g}_{AB})$ refers to the sign of the mean deviation $\Delta\bar{g}_{AB} \equiv h\bar{g}_{AX}(r_k) - \bar{g}_{BX}(r_k)i$. A negative value for σ[*A*,*B*] means that the residue type *A* is more exposed to solvent compared to *B*. And the absolute value *j*σ[*A*,*B*]*j* provides a measure of the similarity between amino acid residues *A* and *B*, insofar as long-range interactions are concerned, with smaller *j*σ[*A*,*B*]*j* values corresponding to more similar pairs of residues. The complete list of σ[*A*,*B*] for all [*A*, *B*] is available upon request. The average of σ[*A*,*B*] over all *B* yields a hydrophobicity ranking in the order of Cys > Ile > Met > Phe > Val > Leu > Trp > Ala > Tyr > His > Pro > Gly > Thr > Ser > Asn > Gln > Arg > Asp > Glu > Lys. Hydrophobicity scales have been obtained in a large number of studies using various approaches (Nozaki & Tanford, 1971; Levitt, 1976; Meirovitch *et al.*, 1980; Wolfenden *et al.*, 1981; Miyazawa & Jernigan, 1985; Rose *et al.*, 1985; Cornette *et al.*, 1987). We note that the present ordering closely approximates that obtained by Rose *et al.* (1985) on the basis of mean-solvent accessibilities. In their study, the fractional area loss upon folding yields the quite similar rank Cys > Ile = Phe > Val > Met = Leu = Trp > His > Tyr > Ala > Gly > Thr > Ser > Pro = Arg > Asn > Gln = Asp = Glu > Lys.

## Homogeneous interactions

For estimating the mean potentials $E_{xx}(r)$ and $E_x(r)$ corresponding to the respective side-chain–side-chain (S-S) and side-chain–backbone (S-B) interactions averaged over all residue types, a limiting distance $r_{\mathrm{lim}}$ needs to be set, beyond which the non-bonded correlations become vanishingly small, and the interaction energies decay to zero. Accordingly, the homogeneous background potential between pairs of side-groups is expressed as:

$$E_{XX}(r) = -RT\ln[\bar{g}_{XX}(r)/\bar{g}_{XX}(r_{lim})] \qquad (14)$$

A similar expression holds for $E_x(r)$ in terms of the corresponding normalized pair radial distribution functions and cutoff distance.

Selection of the cutoff distance is made by analyzing the number of contacts taking place at separations $r\ \Sigma\ \Delta r$, within bins of size $\Delta r = 0.1$ Å. The results are shown in Figure 2, as a function of $r^2$. The S-B contacts are denoted as $N_{SB}(r\ \Sigma\ \Delta r)$ in (a), and the S-S contacts as $N_{SS}\ (r\ \Sigma\ \Delta r)$ in (b). At short separations, the preferences embodied in the radial distribution functions are effective; whereas at large separations, the numbers of contacts should increase linearly with $r^2$, conforming with a uniform density of particles. Such a region where the number of contacts scale linearly with $r^2$ may be delimited in the curve $N_{SB}(r\ \Sigma\ \Delta r)$ starting from $r\ \pi$ 11 Å and ending at $r\ \pi$ 15 Å. The slope of the tangent to this region, drawn by linear regression, reflects the uniform density of atoms in compact globular proteins. The decrease in the slope at larger separations reflects the limited sizes of the proteins. It is interesting to observe that the best fit line through the linear portion of the $N_{SS}\ (r\ \Sigma\ \Delta r)$ curve in Figure 2(b) also has the same slope as the $N_{SB}(r\ \Sigma\ \Delta r)$, confirming that the same uniform density region is approached regardless of the types of interacting atoms. It is clear from the $N_{SS}\ (r\ \Sigma\ \Delta r)$ curve in Figure 2(b) that a separation exceeding $\xi$ 12 Å would introduce biases from the finite sizes of the proteins. Likewise, $r = 15$ Å is the maximum separation at which the $\bar{g}_X(r)$ curves are likely to remain meaningful. In view of these considerations, the lower value $r_{\mathrm{lim}} = 12$ Å has been adopted.

The resulting potentials of mean force $E_{xx}(r)$ and $E_x(r)$ are presented in Figure 3 in dimensionless form. The absolute values of these curves at the minima are small. Yet, the overall contribution of homogeneous interactions to stability is large, as will be shown in Applications, in view of the fact that these are summed over all residues separated by $r \leqslant r_{\mathrm{lim}}$. The sum $E_{xx}(r) + E_x(r)$, which is the effective S-S and S-B homogeneous potential ex-

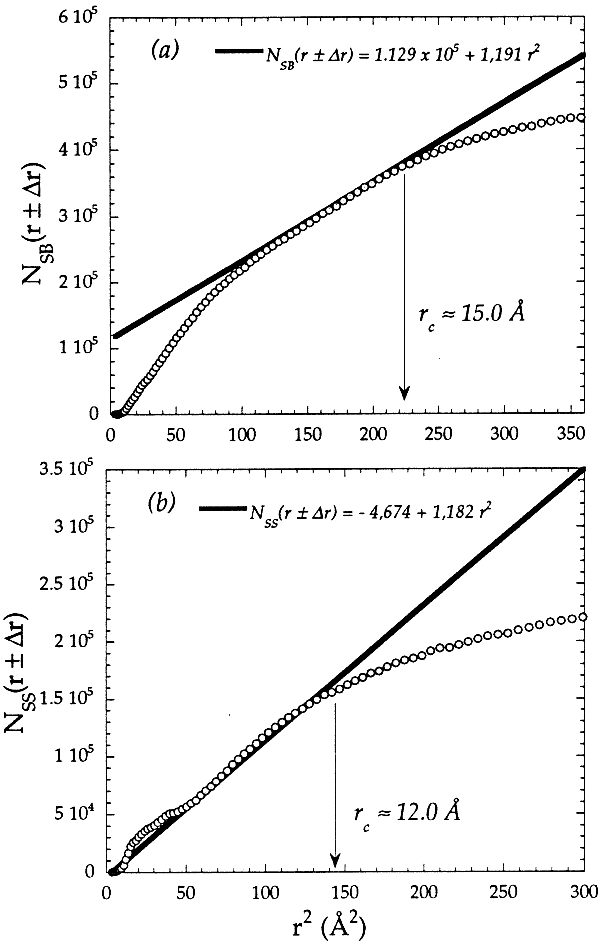

**Figure 2.** (a) Total number of contacts $N_{SB}(r\ \Sigma\ \Delta r)$ between side-groups (S) and backbone (B) atoms, taking place at separations $r\ \Sigma\ \Delta r$, in the 150 structures of Set I (Bahar & Jernigan, 1996). Results are given as a function of $r^2$, using bins of size $\Delta r = 0.1$ Å. $N_{SB}(r\ \Sigma\ \Delta r)$ scales linearly with $r^2$ in the range $11 \leqslant r \leqslant 15$ Å, approximately, in conformity with the uniform density of atoms at long separations. The tangent to this region is drawn by linear regression, and the corresponding equation is displayed. The decrease in the slope at $r \geqslant 15$ Å is due to the finite sizes of the proteins. Thus, $r_{\rm lim}\ \pi\ 15$ Å, indicated by the arrow, is the maximum separation at which $\bar{g}_X(r)$ curves can be reliably adopted in evaluating S-B potentials of mean force. (b) Total number of contacts between pairs of side-group atoms, $N_{SS}(r\ \Sigma\ \Delta r)$. The best fitting line in the range $r \leqslant 12$ Å exhibits the same slope as the tangent to $N_{SB}(r\ \Sigma\ \Delta r)$. This demonstrates the existence of a uniform distribution of atoms at long distances, regardless of the type of atoms. $r_{\rm lim}\ \pi\ 12$ Å is indicated as an upper limit beyond which the number of S-S contacts is biased by the finite size of the proteins.

perienced by a side-group, is also shown in Figure 3.

The residue-specific potentials, $\Delta E_{AB}(r)$, differ from a typical 6-12 Lennard-Jones atomic potential function in the existence of multiple minima. This is a characteristic property of potentials of mean

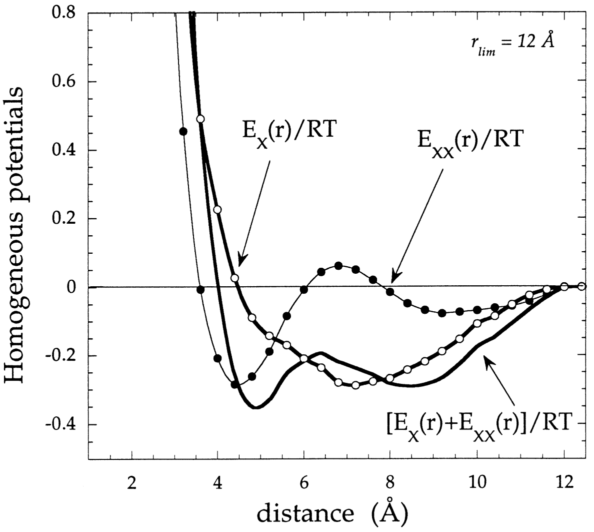

**Figure 3.** The homogeneous potentials $E_{XX}(r)$ and $E_X(r)$ obtained with $r_{\rm lim} = 12$ Å for all S-S and S-B contacts, respectively, regardless of the type of residue. The sum $E_{XX}(r) + E_X(r)$ is given as the heavy line.

force is dense systems. Another feature of potentials of mean force is their long correlation lengths. Cutoff separations in the range $10 \leqslant r_{\rm lim} \leqslant 15$ Å have been considered in previous studies (Wilson & Doniach, 1989; Sun, 1993; Jones *et al.*, 1992; Bryant & Lawrence, 1993). This range may be compared with $r_{\rm lim}\ \pi\ 7.0$ Å commonly used with Lennard-Jones potentials. In fact, information about supersecondary structures and overall topology is included in the potentials of mean force. For example, the residues belonging to parallel α-helices are separated by 8 to 10 Å (Wilson & Doniach, 1989). The range of $5.5 < r < 6.5$ Å is typical of the pairs $[C_i^\alpha, C_{i+4}^\alpha]$ in α-helices; whereas the same atoms are separated by $11 \leqslant r \leqslant 14$ Å in extended β-strands.

## Specific interactions between side-chains

Due to the large number, 210, of distinct side-chain pair types (S-S) for the 20 types of amino acids, we limit our presentation to a few illustrative examples. The complete set of interaction energies is available upon request.

The interaction potentials between hydrophobic residues and other groups of residues are illustrated in Figure 4(a). The curve labeled [Hϕ, Hϕ] represents the potentials of mean force $\Delta E_{AB}(r)$ averaged over the residues Hϕ ≡ Ile, Leu, Val, Phe and Met. [Hϕ, polar] refers to the average interaction between Hϕ and polar side-chains Asn, Gln, Ser, Thr and His. [Hϕ, Lys] illustrates the interaction between Hϕ and charged residues; the interactions with Glu and Asp closely approximate that of [Hϕ, Lys], while that of the pair [Hϕ, Arg] is slightly more favorable. Tyr and Trp are combined in the last curve, as their interactions with the hydrophobic residues exhibit comparable

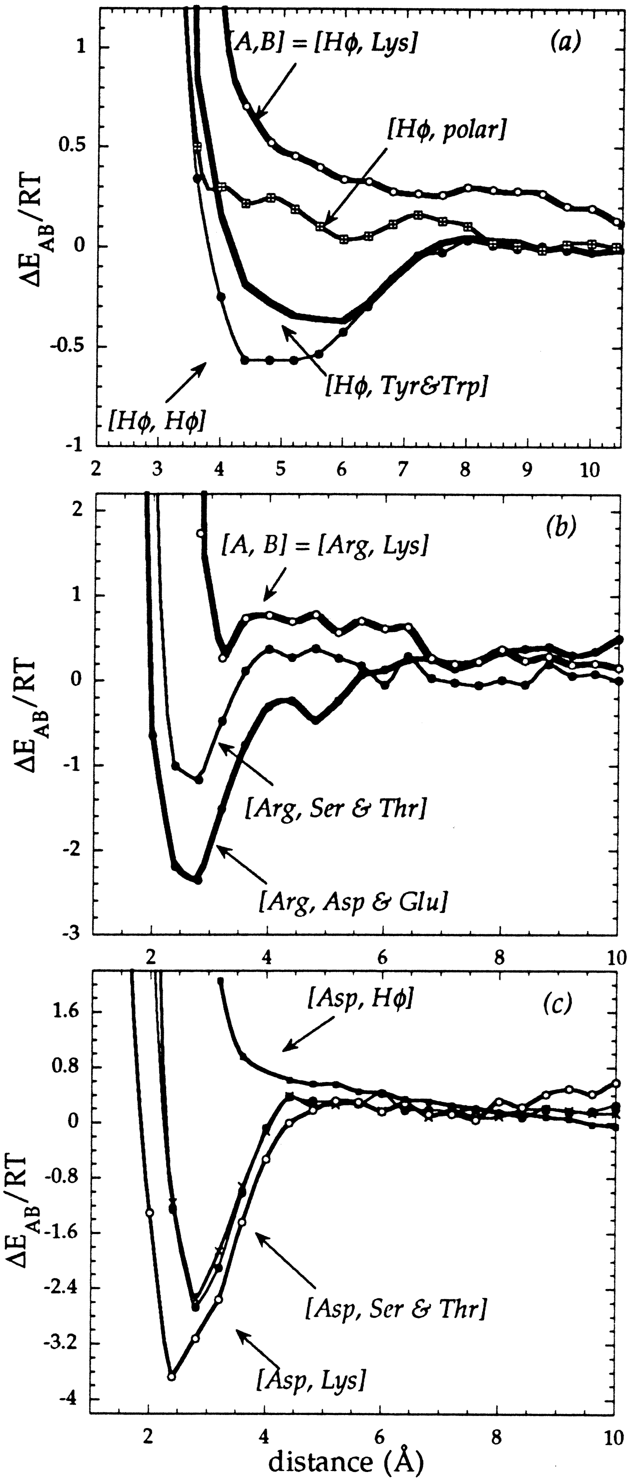

**Figure 4.** The potential of mean force $\Delta E_{AB}(r)$, the excess over the homogeneous interaction $E_{XX}(r)$, calculated with equation (1) for [$A$, $B$] = [Hϕ, Hϕ], [Hϕ, Lys], [Hϕ, polar] and [Hϕ, Tyr & Trp] in (a), [Arg, Lys], [Arg, Ser & Thr] and [Arg, Asp & Glu] in (b), and [Asp, Ser], [Asp, Thr], [Asp, Hϕ] and [Asp, Lys] in (c). The group Hϕ includes the hydrophobic residues Hϕ = Ile, Leu, Val, Phe and Met. The [Hϕ, polar] curve describes the potential of mean force averaged over the group Hϕ and the polar residues Asn, Gln, Ser, Thr and His. Tyr and Trp are also combined, their interactions with the hydrophobic residues being similar. The interactions of Ser and Thr with Arg are quite similar, as well as those of Asp and Glu with Arg, hence their representation by single representative curves in (b). Results are obtained at 0.4 Å intervals, starting from 2.0 Å, and connected by smooth curves for visualization. The strong tendency of Lys, Ser and Thr to cluster together with Asp is illustrated in (c).

trends. In many instances, Trp exhibits behavior quite similar to that of the hydrophobic residues, but it is definitely distinguished from them by its favorable interaction with Arg. This is an interaction upon which others have remarked (Levitt & Perutz, 1988; Flocco & Mowbray, 1994).

Figure 4(b) and (c) typify the interactions between charged and polar residues. Comparison with Figure 4(a) shows that the energy wells are now narrower and shifted to closer separations. We obtain an attractive potential of ξ − 2.5 $RT$ for oppositely charged residues at a separation of about 3 Å; whereas the same type of interaction was reported to be subject to an attraction of only ξ − 1.5 $RT$ at ξ 4 Å in the work of Wilson & Doniach (1989), or only ξ − 0.75 $RT$ at $0 \leqslant r \leqslant 5$ Å in work of Bryant & Lawrence (1993). The choice of side-group interaction centers at specific atoms in the present work (see Table 4 of Bahar & Jernigan, 1996) leads to stronger interactions at close distances.

Another observation is that the clusters of amino-aromatic or pairs of aromatic residues are favored, which was pointed out by Burley & Petsko (1988) to play a considerable role in stabilizing native structures. We have also observed in other studies (Wallqvist *et al.*, 1995) a segregation between aliphatic and aromatic carbon atoms. Stacking of $sp^2$ hybridized nitrogen atoms above aromatic rings such as Phe and Tyr has been pointed out by Mitchell *et al.* (1994) to be more frequent than the formation of hydrogen bonds between those groups. A relatively high frequency of Arg side-group nitrogen atoms approaching Tyr and Trp is found here; whereas a tendency for close approach of Arg to Phe is not observed.

## Effective inter-residue contact energies

In Table 1, we present the effective inter-residue contact energies $e_{AB}(r_c)$ and $e^{o}_{AB}(r_c)$ obtained from the integrations of the radial distribution functions in the range $2.0 \leqslant r \leqslant r_c = 6.4$ Å. The upper limit is chosen as 6.4 Å so as to be able to compare with the contact energies derived by Miyazawa & Jernigan (1985, 1996). The range $r \leqslant 2.0$ Å is excluded in all integrations, inasmuch as no contact takes place below this lower limit. Values of $e_{AB}(r_c)$ are given on the diagonal and in the upper triangular elements of the 20 Θ 20 matrix. The diagonal elements are underlined for clarity. The lower triangular elements represent the $e_{AB}{}^{o}(r_c)$ values. The residue types are listed in the first column and first row of the Table. The corresponding

**Table 1.** Effective broad contact energies in *RT* units for inter-residue distances $r \leqslant r_c = 6.4$ Å: $e_{AB}(r_c)$ for the upper triangular half and diagonal, $e_{AB}^{0}(r_c)$ for the lower half

| | Gly | Ala | Val | Ile | Leu | Ser | Thr | Asp | Asn | Glu | Gln | Lys | Arg | Cys | Met | Phe | Tyr | Trp | His | Pro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gly | −2.13 | −2.55 | −3.10 | −3.37 | −3.55 | −2.03 | −2.16 | −1.90 | −1.88 | −1.44 | −1.85 | −1.25 | −2.12 | −3.27 | −3.37 | −3.65 | −2.95 | −3.43 | −2.12 | −1.77 |
| Ala | 0.14 | −3.24 | −4.02 | −4.38 | −4.52 | −2.42 | −2.50 | −2.02 | −2.21 | −1.86 | −2.34 | −1.49 | −2.20 | −3.96 | −4.14 | −4.59 | −3.53 | −4.13 | −2.84 | −2.26 |
| Val | 0.37 | 0.01 | −4.81 | −5.27 | −5.42 | −2.95 | −3.06 | −2.33 | −2.74 | −2.59 | −3.13 | −1.97 | −3.10 | −4.61 | −5.01 | −5.44 | −4.36 | −5.04 | −3.54 | −3.03 |
| Ile | 0.54 | 0.08 | −0.03 | −5.69 | −5.86 | −3.42 | −3.62 | −2.93 | −3.00 | −3.09 | −3.36 | −2.54 | −3.47 | −5.20 | −5.51 | −5.93 | −4.82 | −5.48 | −3.83 | −3.47 |
| Leu | 0.53 | 0.11 | −0.01 | −0.01 | −6.02 | −3.43 | −3.58 | −2.95 | −3.30 | −3.15 | −3.68 | −2.61 | −3.67 | −5.24 | −5.66 | −6.11 | −4.92 | −5.47 | −4.13 | −3.47 |
| Ser | 0.04 | 0.21 | 0.46 | 0.43 | 0.59 | −2.01 | −2.09 | −1.97 | −1.92 | −1.99 | −1.94 | −1.29 | −2.05 | −3.13 | −3.33 | −3.49 | −2.77 | −3.14 | −2.64 | −1.77 |
| Thr | 0.04 | 0.25 | 0.48 | 0.35 | 0.56 | 0.05 | −2.26 | −2.06 | −2.06 | −1.95 | −2.10 | −1.46 | −2.23 | −3.29 | −3.38 | −3.60 | −2.81 | −3.32 | −2.60 | −1.86 |
| Asp | −0.08 | 0.36 | 0.82 | 0.66 | 0.81 | −0.21 | −0.17 | −1.50 | −1.85 | −1.47 | −1.82 | −1.90 | −2.54 | −2.38 | −2.75 | −3.04 | −2.75 | −2.97 | −2.51 | −1.37 |
| Asn | 0.16 | 0.39 | 0.65 | 0.83 | 0.69 | 0.07 | 0.05 | −0.12 | −1.96 | −1.81 | −2.15 | −1.54 | −1.98 | −3.02 | −3.20 | −3.38 | −2.84 | −3.32 | −2.39 | −1.56 |
| Glu | 0.24 | 0.37 | 0.43 | 0.36 | 0.47 | −0.37 | −0.20 | −0.11 | −0.21 | −1.23 | −1.72 | −1.98 | −2.64 | −2.59 | −3.09 | −3.20 | −2.76 | −3.10 | −2.45 | −1.45 |
| Gln | 0.15 | 0.23 | 0.22 | 0.42 | 0.27 | 0.00 | −0.03 | −0.13 | −0.23 | −0.17 | −1.88 | −1.64 | −2.30 | −3.00 | −3.42 | −3.73 | −3.05 | −3.49 | −2.51 | −1.90 |
| Lys | 0.09 | 0.41 | 0.70 | 0.57 | 0.68 | −0.01 | −0.06 | −0.88 | −0.28 | −1.09 | −0.43 | −0.54 | −1.08 | −2.02 | −2.58 | −2.85 | −2.50 | −2.96 | −1.56 | −0.90 |
| Arg | 0.10 | 0.57 | 0.46 | 0.53 | 0.49 | 0.11 | 0.05 | −0.63 | 0.16 | −0.87 | −0.21 | 0.34 | −2.31 | −2.94 | −3.15 | −3.80 | −3.22 | −3.90 | −2.59 | −2.07 |
| Cys | 0.60 | 0.47 | 0.60 | 0.45 | 0.57 | 0.68 | 0.65 | 1.17 | 0.76 | 0.83 | 0.75 | 1.06 | 1.02 | −5.61 | −4.99 | −5.49 | −4.26 | −4.61 | −3.96 | −3.16 |
| Met | 0.44 | 0.23 | 0.14 | 0.08 | 0.10 | 0.42 | 0.50 | 0.75 | 0.53 | 0.27 | 0.27 | 0.44 | 0.76 | 0.57 | −5.50 | −5.84 | −4.69 | −5.37 | −4.11 | −3.48 |
| Phe | 0.54 | 0.15 | 0.09 | 0.03 | 0.02 | 0.64 | 0.65 | 0.84 | 0.72 | 0.54 | 0.33 | 0.54 | 0.48 | 0.44 | 0.03 | −6.24 | −4.94 | −5.68 | −4.43 | −3.69 |
| Tyr | 0.15 | 0.13 | 0.08 | 0.06 | 0.13 | 0.27 | 0.36 | 0.03 | 0.17 | −0.11 | −0.08 | −0.19 | −0.03 | 0.59 | 0.10 | 0.21 | −4.07 | −4.64 | −3.62 | −3.09 |
| Trp | 0.24 | 0.10 | −0.03 | −0.03 | 0.15 | 0.48 | 0.42 | 0.39 | 0.27 | 0.12 | 0.06 | −0.08 | −0.14 | 0.81 | −0.01 | 0.05 | 0.00 | −5.22 | −4.07 | −3.65 |
| His | 0.59 | 0.43 | 0.52 | 0.66 | 0.52 | 0.01 | 0.18 | −0.11 | 0.24 | −0.18 | 0.07 | 0.36 | 0.21 | 0.49 | 0.29 | 0.34 | 0.06 | 0.18 | −3.30 | −2.30 |
| Pro | 0.09 | 0.16 | 0.17 | 0.17 | 0.34 | 0.03 | 0.07 | 0.18 | 0.22 | −0.04 | −0.16 | 0.17 | −0.12 | 0.44 | 0.07 | 0.23 | −0.26 | −0.24 | 0.15 | −1.59 |
| SLV | 0.00 | 0.28 | 0.75 | 1.13 | 1.33 | −0.08 | −0.01 | −0.08 | −0.04 | −0.09 | −0.03 | −0.24 | 0.06 | 0.25 | 1.03 | 1.67 | 0.70 | 1.53 | 0.27 | 0.01 |
| corr(*e*) | 0.98 | 0.99 | 0.98 | 0.99 | 0.99 | 0.97 | 0.97 | 0.97 | 0.96 | 0.97 | 0.98 | 0.95 | 0.97 | 0.98 | 0.99 | 0.99 | 0.98 | 0.98 | 0.97 | 0.98 |
| corr($e^0$) | 0.95 | 0.79 | 0.88 | 0.93 | 0.93 | 0.91 | 0.95 | 0.97 | 0.98 | 0.98 | 0.93 | 0.98 | 0.96 | 0.78 | 0.87 | 0.92 | 0.87 | 0.82 | 0.91 | 0.89 |

corr(*e*) and corr($e^0$) are correlation coefficients between present results and those of Miyazawa & Jernigan (1996) calculated for $e_{AB}$ and $e_{AB}^{0}$, respectively. SLV refers to solvent, and the corresponding row represents the interaction potentials $E_{A0}(r_c)$ between solvent and residue type *A*, normalized with respect to that between solvent and glycine.

residue-solvent (SLV) interaction energies $E_{A0}(r_c)$ normalized with respect to that of glycine, and the correlation coefficients between the present results for $e_{AB}(r_c)$ and $e_{AB}^{0}(r_c)$ and those of the updated MJ potentials (Miyazawa & Jernigan, 1996) are listed in the last three rows, respectively. Here $E^{*}_{00}(r_c)$ has been simply taken as −3.3 *RT*, the value leading to the highest correlation between the two sets of results.

We note that Miyazawa & Jernigan (1996) used a similar distance integration ($r_c = 6.5$ Å), but different centers for interaction. They used side-chain centers and here we use selected atom pairs. Overall, excellent agreement between the two sets of data is observed. The pair of residues subject to the most favorable intramolecular interaction $e_{AB}^{0}(r_c)$ is [Lys, Glu] in both cases; whereas [Asn, Cys] is subject to the strongest repulsion. On the basis of the solvent-mediated contact energies $e_{AB}(r_c)$, on the other hand, [Leu, Leu] and [Lys, Lys] pairs exhibit the extreme behaviors of opposite character, in conformity with the MJ results; the respective contact energies are −6.02 *RT* and −0.54 *RT*. The agreements are excellent for the $e_{AB}(r_c)$ values, on average, better than 0.98; for the $e_{AB}^{0}(r_c)$ values, correlation coefficients are slightly lower, averaging 0.94. This agreement verifies the consistency of the present approach with previous work. The data and formalism presented here permit, on the other hand, the derivation of effective inter-residue contact potentials for any distance range, which may be conveniently used in on and off-lattice simulations of proteins.

We note from Table 1 that the strongest attractions are those between pairs of hydrophobic side-chains; whereas interactions between hydrophilic groups are relatively weaker. However, from the distance-dependence of the particular potentials of mean force $\Delta E_{AB}(r)$, illustrated in Figure 4, it is not hard to realize that the character of the effective contact potentials should be considerably altered at shorter separations. In fact, most of the attractive potentials between hydrophobic groups occur in the range $4 \leqslant r \leqslant 6$ Å (Figure 4(a)); whereas pairs of polar and charged groups experience the strong attractions in the interval $2 \leqslant r \leqslant 4$ Å, approximately (Figure 4(b) and (c)).

In order to quantify these arguments, calculations are repeated for $r_c = 4.0$ Å. The coordination numbers of the different residue types excluding first neighbors along the chains are given in Table 2, for both $r_c = 4.0$ and 6.4 Å. These are obtained by applying the approach of Miyazawa & Jernigan (1985) to buried residues (located within a distance of 7.0 Å of the center of protein). The proteins with size $n \geqslant 100$ in our dataset (Bahar & Jernigan, 1996) were considered in evaluating $q(r_c)$. Also shown are the total numbers of contacts $N_{AX}(r_c)$ observed between *A* and all other side-chain groups located within a distance $r_c$ of *A*. It can be seen that the average coordination number of $6.1\,\Sigma\,0.9$ for $r_c = 6.4$ Å is now reduced for $r_c = 4.0$ Å to $0.9\,\Sigma\,0.6$. This indicates that highly specific, individual pairs are now to be observed within this narrower distance range of $r_c \leqslant 4.0$ Å.

The results are presented in Table 3, in the same format as Table 1. In order to facilitate the com-

**Table 2.** Coordination numbers, $q_A(r_c)$, for $r_c = 4.0$ and 6.4 Å

| $A$ | $r_c = 4.0$ Å | | $r_c = 6.4$ Å | |
|---|---|---|---|---|
| | $N_{AX}(r_c)^a$ | $q_A(r_c)$ | $N_{AX}(r_c)^a$ | $q_A(r_c)$ |
| Gly | 750 | 1.48 | 7095 | 6.61 |
| Ala | 880 | 1.45 | 8963 | 6.26 |
| Val | 439 | 0.88 | 8986 | 6.17 |
| Ile | 410 | 0.75 | 7717 | 5.92 |
| Leu | 480 | 0.71 | 12,080 | 5.90 |
| Ser | 888 | 1.26 | 5148 | 6.98 |
| Thr | 767 | 1.15 | 5131 | 6.75 |
| Asp | 742 | 1.05 | 4308 | 6.26 |
| Asn | 439 | 1.03 | 3360 | 6.31 |
| Glu | 679 | 0.80 | 4046 | 6.17 |
| Gln | 307 | 0.92 | 2804 | 6.54 |
| Lys | 638 | 0.76 | 3654 | 6.80 |
| Arg | 462 | 0.49 | 3223 | 6.23 |
| Cys | 895 | 1.29 | 3149 | 6.59 |
| Met | 156 | 0.72 | 2702 | 5.96 |
| Phe | 188 | 0.54 | 5839 | 5.58 |
| Tyr | 211 | 0.65 | 4278 | 5.72 |
| Trp | 113 | 0.26 | 1856 | 5.30 |
| His | 199 | 0.69 | 2232 | 6.03 |
| Pro | 159 | 1.31 | 3327 | 5.91 |

[a] Total number of inter-residue contacts occurring between residue $A$ and any other residue at $r \leqslant r_c$.

parison, $e_{AB}(r_c)$ values for $A = B = \text{Gly}$ in the two Tables are set equal, by choosing $E^*_{00}(r_c = 4\text{ Å})$ as $-5.6$ $RT$. No normalization is required for $e_{AB}{}^0(r_c)$ values. Values missing in Table 3 correspond to pairs of residues not observed at separations $r \leqslant 4$ Å, mainly [Trp, Trp], [Trp, Tyr], [Trp, Glu], [Trp, Met], [Met, Asp] and [Trp, SLV]. The most striking observations in Table 3, in comparison to Table 1 are:

(1) In Table 1, hydrophobic pairs, and in particular those involving Phe and Leu, experience the strongest attractions, of the order of $-5$ $RT$, insofar as the solvent-mediated effective contact potentials $e_{AB}(r_c)$ are concerned; whereas these interactions do not appear to be as favorable at close separations. In fact, an average value of only about $-1.5$ $RT$ for the $e_{AB}(r_c)$ values of hydrophobic pairs follows from Table 3.

(2) For charged residues the situation is reversed: even the [Lys, Lys] pair, which was subject to the weakest $e_{AB}(r_c)$ among all pairs at $r \leqslant 6.4$ Å, is found to be much more favorable ($-3.70$ $RT$), when the range $r \leqslant 4.0$ Å is considered. Oppositely charged side-chains are now subject to the strongest interactions, as expected, their interaction being of the order of $-7.0$ $RT$. The strongest attraction occurs between Arg and Glu.

(3) Interactions between polar and charged side-chains also emerge as an important group of attractive potentials at $r \leqslant 4.0$ Å. The corresponding effective contact potential is about $-5$ $RT$, on the average, and varies from $-3.7$ $RT$ for [Asp, Gln] to $-6.6$ $RT$ for [Arg, Thr]. These may be compared with the respective values $-1.72$ $RT$ and $-2.23$ $RT$ in Table 1. We also note an enhancement in the attractive interaction of the pair [His, Arg].

(4) In both Tables, pairs involving Cys exhibit highly favorable interactions.

(5) From the comparison of the effective self contact potentials $e_{AB}{}^0(r_c)$ in the lower triangular part of Tables 1 and 3, the first observation is the broadening of the range of effective potentials from 2.3 $RT$ at $r_c = 6.4$ Å to 4.2 $RT$ at $r_c = 4.0$ Å. This indicates a significant increase in specificity for the narrower distance range. The pair [Lys, Glu] is subject to the strongest attraction, $-1.1$ $RT$ and $-2.3$ $RT$ in the respective Tables 1 and 3.

(6) Some amino-aromatic pairs such as [Phe, Arg], [Tyr, Arg], [Pro, Asp], [Tyr, Lys] and [His, Asp] exhibit an enhancement to the number of close contacts, leading to more favorable contact poten-

**Table 3.** Effective close contact energies in $RT$ units for inter-residue distances $r \leqslant r_c = 4.0$ Å: $e_{AB}(r_c)$ for the upper half and diagonal, $e_{AB}{}^0(r_c)$ for the lower half

| | Gly | Ala | Val | Ile | Leu | Ser | Thr | Asp | Asn | Glu | Gln | Lys | Arg | Cys | Met | Phe | Tyr | Trp | His | Pro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gly | −2.13 | −2.72 | −1.31 | −1.63 | −1.42 | −3.90 | −3.92 | −3.89 | −3.28 | −4.04 | −3.05 | −4.92 | −4.92 | −5.92 | −3.33 | −2.44 | −2.67 | – | −3.05 | −0.34 |
| Ala | 0.05 | −3.44 | −2.41 | −2.84 | −2.23 | −3.86 | −3.79 | −3.78 | −3.33 | −3.41 | −3.16 | −2.60 | −5.18 | −6.14 | −3.17 | −2.33 | −2.72 | – | −3.16 | −0.83 |
| Val | 0.47 | 0.00 | −1.41 | −1.73 | −1.15 | −2.66 | −2.69 | −5.44 | −1.70 | −2.18 | −2.22 | −2.77 | −4.20 | −5.38 | −2.13 | −1.17 | −1.73 | – | −2.39 | 0.75 |
| Ile | 0.55 | −0.05 | 0.07 | −2.15 | −1.73 | −2.92 | −2.78 | −1.82 | −1.96 | −2.43 | −2.05 | −1.82 | −4.91 | −5.44 | −2.42 | −1.80 | −2.08 | – | −4.73 | −3.20 |
| Leu | 0.09 | −0.06 | −0.02 | −0.20 | −0.91 | −2.26 | −2.44 | −1.56 | −1.84 | −2.29 | −2.28 | −1.33 | −3.81 | −5.23 | −3.08 | −1.92 | −1.51 | – | −1.95 | 0.74 |
| Ser | −0.43 | 0.25 | 0.44 | 0.59 | 0.58 | −4.81 | −4.86 | −5.38 | −4.41 | −5.21 | −4.26 | −4.43 | −6.28 | −6.74 | −3.15 | −1.63 | −2.08 | – | −4.70 | −2.28 |
| Thr | −0.26 | 0.53 | 0.62 | 0.87 | 0.32 | 0.06 | −5.19 | −5.22 | −4.55 | −5.19 | −4.40 | −4.55 | −6.56 | −6.71 | −3.11 | −1.88 | −2.36 | – | −4.38 | −2.31 |
| Asp | −0.85 | −0.04 | 1.25 | 1.28 | 0.89 | −1.89 | −0.62 | −4.03 | −4.01 | −4.25 | −3.71 | −5.96 | −7.58 | −6.23 | −1.88 | −0.56 | −1.09 | – | −6.43 | −1.39 |
| Asn | −0.22 | 0.37 | 0.95 | 0.33 | 0.59 | 0.01 | 0.02 | −0.02 | −4.06 | −4.66 | −3.74 | −4.29 | −5.37 | −5.30 | −2.54 | −1.70 | −1.98 | – | −3.42 | −0.89 |
| Glu | −0.51 | 0.46 | 0.52 | 0.66 | 0.31 | −0.85 | −0.47 | −0.23 | 0.35 | −4.31 | −3.85 | −5.99 | −7.75 | −6.11 | −2.61 | −0.91 | −2.34 | – | −4.49 | −1.52 |
| Gln | −0.17 | −1.27 | 0.32 | 0.89 | −0.02 | −0.02 | 0.09 | −0.06 | 0.08 | 0.09 | −3.74 | −4.30 | −5.84 | −5.88 | −2.71 | −0.40 | −1.85 | – | −3.18 | −1.57 |
| Lys | 0.02 | 0.96 | 1.66 | 1.06 | 0.77 | −0.14 | −0.14 | −2.07 | −1.16 | −2.27 | −0.68 | −3.70 | −6.77 | −7.42 | −2.25 | −1.29 | −3.05 | – | −3.90 | −0.86 |
| Arg | 0.22 | 0.73 | 0.55 | 0.31 | 0.68 | 0.16 | 0.26 | −1.33 | 0.65 | −1.93 | −0.03 | 0.54 | −7.65 | −7.26 | −5.44 | −4.65 | −4.90 | – | −5.95 | −3.00 |
| Cys | 0.90 | 1.37 | 1.23 | 1.43 | 0.95 | 1.50 | 1.71 | 1.75 | 1.63 | 1.76 | 1.62 | 1.52 | 1.98 | −10.64 | −6.48 | −4.77 | −5.20 | – | −6.65 | −5.39 |
| Met | −0.64 | −0.38 | −0.25 | −0.48 | −0.24 | 0.03 | 0.57 | – | 1.29 | 0.20 | −0.07 | 0.13 | 0.93 | 0.59 | −2.07 | 0.34 | −1.92 | – | −3.01 | −1.19 |
| Phe | −0.61 | −0.52 | −0.36 | −0.74 | −0.40 | 0.82 | 0.79 | 1.52 | 0.23 | 1.25 | 1.66 | 0.63 | −1.18 | 1.17 | 1.88 | −0.42 | 1.06 | – | −2.38 | −0.12 |
| Tyr | −1.52 | −0.67 | −1.34 | −1.41 | −0.81 | 0.70 | 0.38 | 0.84 | 0.20 | 1.65 | −0.07 | −1.38 | −0.94 | 0.72 | −1.05 | 1.44 | −0.24 | – | −3.60 | −1.05 |
| Trp | −1.01 | 0.36 | −0.32 | −0.04 | −0.61 | 0.65 | 1.01 | 1.91 | −0.02 | – | 1.71 | 0.25 | −0.34 | 1.47 | – | 0.71 | – | – | – | – |
| His | −0.24 | 0.12 | −0.06 | −0.10 | 0.00 | −0.68 | −0.12 | −0.96 | 0.22 | −0.56 | 0.37 | −0.20 | −0.25 | 1.27 | 0.37 | −0.64 | −0.75 | 0.30 | −3.30 | −1.57 |
| Pro | 0.10 | 0.23 | −1.01 | 0.95 | 0.56 | −0.41 | −0.37 | −1.42 | 0.35 | −0.05 | −0.35 | 0.36 | 0.30 | 0.93 | −0.15 | −0.75 | −1.14 | −1.23 | −0.52 | 1.06 |
| SLV | 0.00 | 0.19 | −0.27 | −0.03 | −0.34 | 0.58 | 0.66 | 0.70 | 0.22 | 0.79 | 0.29 | 0.57 | 2.56 | 3.04 | 0.04 | −0.31 | −0.09 | – | 0.65 | −0.98 |

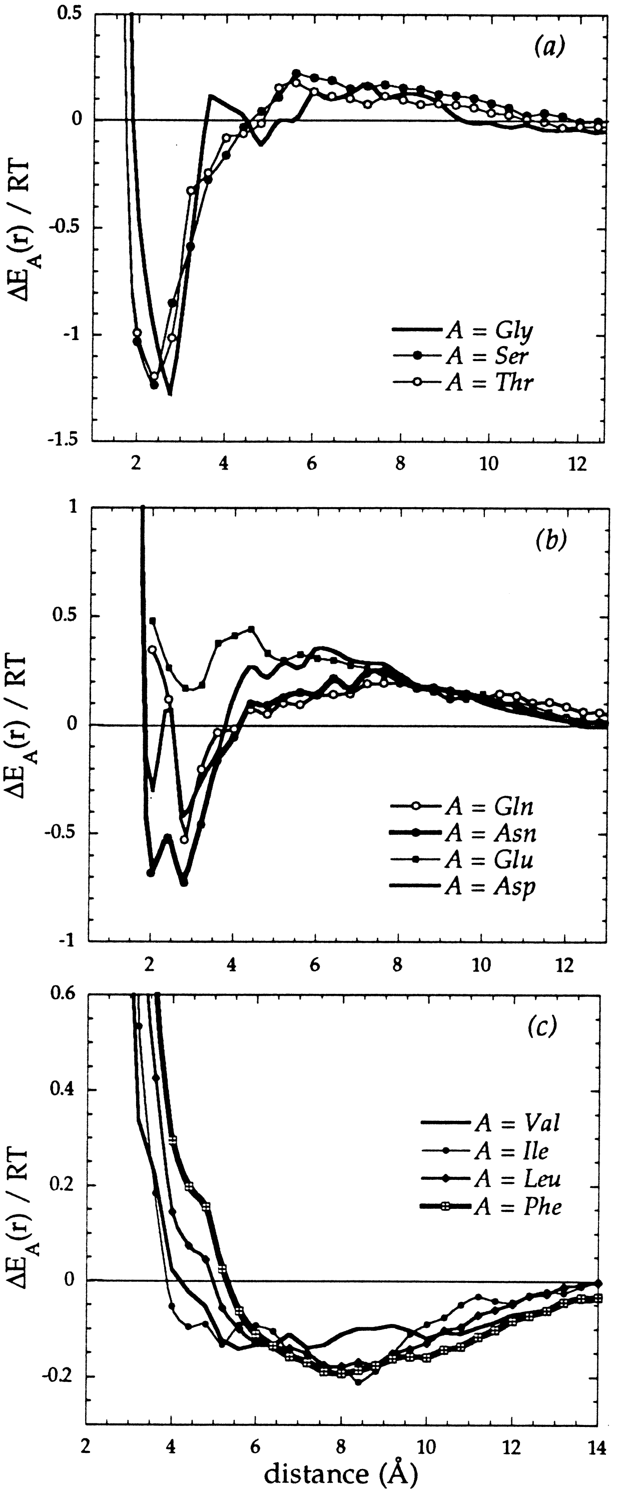

**Figure 5.** Potential of mean force $\Delta E_{A}(r) \equiv E_{A}(r) - E_{x}(r)$ for the S-B interaction between the side-group of residue $A$ and backbone atoms separated by at least four intervening residues, shown for (a) $A$ = Ser, Thr and Gly, (b) $A$ = Asn, Gln, Asp and Glu, and (c) A = Leu, Ile, Val and Phe. The three residues in (a) exhibit the strongest affinity for backbone atoms; $\Delta E_{A}(r)$ is calculated by using the counterpart of equation (1), with the S-B pair radial distribution functions $\bar{g}_{A}(r)$ and $\bar{g}_{X}(r)$.

tials $e_{AB}^{0}(r_{c})$ at $r \leqslant 4.0$ Å, compared with those at $r \leqslant 6.4$ Å. [Tyr, Gly] and [Asp, Gly] are also distinguished by their enhanced attraction at close distances.

In summary, the range $r \leqslant 6.4$ Å may be viewed as consisting of two different regimes of effective contact potentials: at short distances, $r \leqslant 4$ Å, highly specific interactions, predominantly attractive, occur between pairs of hydrophilic residues, while pairs of hydrophobic side-chains exhibit relatively weaker contact interactions. This behavior is reversed at longer separations, i.e. $4.0 \leqslant r \leqslant 6.5$ Å.

### Side-chain–backbone (S-B) interactions

The side-chains that display the strongest affinity for backbone are Gly, Ser and Thr, as illustrated in Figure 5(a). The departure from the homogeneous S-B interaction $E_{X}(r)$ is designated here as $\Delta E_{A}(r)$ for residue type $A$ interacting with the backbone. The residues that exhibit the next most favorable interactions are Lys > Arg $\pi$ Asn > Gln $\pi$ Asp > Glu, in order of decreasing strength of attraction. Some of these are illustrated in Figure 5(b). We note that Asn and Asp display two minima, which may be attributed to the specific interactions of the two terminal atoms. The behavior of Glu is quite distinct from the chemically comparable residue Asp, which brings attention to the effect of side-chain flexibility on S-B interactions. Leu, Ile, Val, Phe, Met, Cys, Tyr and Trp have qualitatively different and rather weak, but favorable, interactions with the backbone, centered near 7.5 Å as illustrated in Figure 5(c).

## Applications

### Relative contributions of homogeneous and specific interactions to stability

The overall non-bonded interaction energy between side-groups in a given protein, hereafter referred to as $E_{\text{tot}}$(S-S), is evaluated by summation of the potentials $E_{AB}(r)$ over all S-S pairs separated by at least five intervening virtual bonds. Nearer pairs along the chain are not included in this category, since their interactions are likely to be better accounted for by short-range energies. Each interaction may be expressed as the sum of two terms, $E_{XX}(r)$ and $\Delta E_{AB}(r)$, representing the homogeneous and specific contributions. We will compare the strengths of the two contributions. The overall specific interaction potential in a given protein will be designated with the superscript $s$, as $E_{\text{tot}}^{s}$(S-S).

Results are displayed in Figure 6, as a function of the number $n$ of residues in each of the proteins given in Set I on the Internet (Bahar & Jernigan, 1996). The filled circles represent the total S-S energies in dimensionless form, $E_{\text{tot}}(\text{S-S})/RT$. The open circles show the contribution of specific inter-

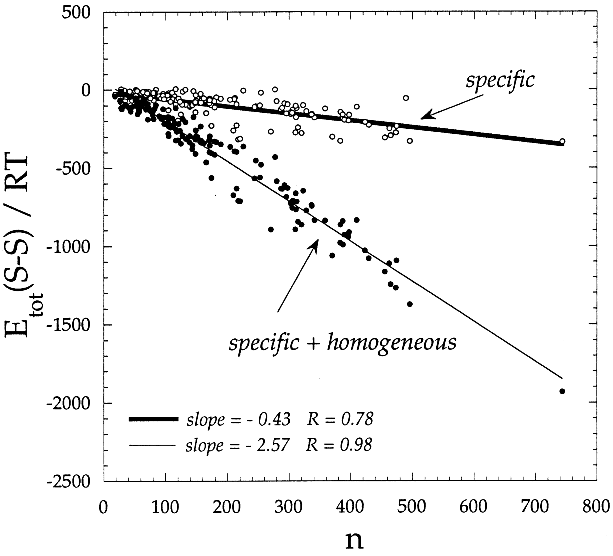

**Figure 6.** Total interaction energy $E_{tot}$(S-S) between the side-chain groups of 150 PDB structures of Set I (Bahar & Jernigan, 1996), as a function of the protein size. The abscissa gives the number of residues in each protein. The filled circles are obtained from the summation of homogeneous ($E_{XX}$) and specific ($\Delta E_{AB}$) interactions between all side-chain groups separated by $r \leqslant 12$ Å, excluding the nearest neighbors; the open circles represent the contribution of specific interactions to the total energy. Lines obtained by least-squares fitting indicate an average total of −2.57 *RT* per residue stabilizing globular proteins, and, of this, −0.43 *RT* comes from specific effects.

actions. From the slopes of the best fit lines obtained by linear regression, the average S-S potential of mean force per residue is estimated to be −2.57 *RT*. Of this, only −0.43 *RT* is from specific interactions. A similar analysis for S-B interactions yields the respective average values of −0.13 *RT* and −1.07 *RT* for the specific and homogeneous contributions to S-B potentials. These results show that the native structures are stabilized mostly by homogeneous attractions rather than by specific interactions. However, we note that the homogeneous contribution is invariant to primary structure; whereas the specific contribution, alone, discriminates between different amino acid sequences for a given folded state.

## Detection of non-native-like structures

The S-S and S-B potentials of some proteins are found to be relatively weak compared to other native structures. The Brookhaven Protein Data Bank (PDB) names of these proteins, their number of residues, the resolution of the X-ray coordinates, if applicable, and the corresponding total interaction potentials are listed in Table 4. The results are divided into two groups. The upper part of the Table lists relatively small proteins ($n < 120$), in which both the S-B and S-S interactions per residue are found to be weaker than −1.0 *RT*. The lower part shows those proteins with $n > 150$, whose non-bonded interaction energies are weakest among the 302 PDB structures

**Table 4.** Proteins with weak non-bonded potentials

| PDB code | $n^a$ | Resolution (Å) | $E_{tot}$(S-S)/$n$ | $E_{tot}$(S-B)/$n$ | Others[b] |
|---|---|---|---|---|---|
| A. *Small* | | | | | |
| 9ins | 20 | 1.70 | −0.76 | −0.79 | |
| 2zta[c] | 31 | 1.80 | −0.41 | −0.57 | |
| 1ctaa[c] | 33 | NMR | −0.73 | −0.54 | $NH_2$, $Ca^{2+}$ |
| 1tabi | 34 | 2.30 | −0.64 | −0.10 | |
| 2bpa3[c] | 35 | 3.40 | −0.26 | −0.13 | |
| 1r094[c] | 39 | 2.90 | −0.43 | −0.30 | JEN, DMS |
| 1ltsc[c] | 40 | 1.95 | −0.24 | −0.37 | |
| 2madl[c] | 50 | 2.25 | −0.79 | −0.49 | |
| 1aaf | 54 | NMR | −0.96 | −0.51 | $Zn^{2+}$ |
| 2mev4[c] | 57 | 3.00 | −0.30 | −0.19 | $PO_4^{3-}$ |
| 1pi2 | 60 | 2.50 | −0.35 | −0.74 | |
| 1nxb[c] | 61 | 1.38 | −0.65 | −0.70 | $SO_4^{2-}$ |
| 3ebx | 61 | 1.40 | −0.46 | −0.97 | $SO_4^{2-}$ |
| 2cdv[c] | 106 | 1.80 | −0.96 | −0.98 | Heme |
| 1cy3[c] | 117 | 2.50 | −0.59 | −0.64 | Heme |
| B. *Large* | | | | | |
| 7wga | 170 | 2.00 | −1.08 | −1.28 | |
| 1hgeb[c] | 174 | 2.60 | −1.13 | −0.96 | |
| 3pgm[c] | 229 | 2.80 | −1.48 | −1.33 | $SO_4^{2-}$, MP3 |
| 4rcrh[c] | 280 | 3.00 | −1.26 | −1.36 | |
| 2plv1 | 296 | 2.88 | −1.45 | −1.22 | MYR, SPH |
| 3pgk | 414 | 2.50 | −1.67 | −1.60 | ATP, $Mg^{2+}$, MP3 |
| 2dpv[c] | 547 | 3.25 | −1.66 | −1.49 | |

Energies are given in *RT* units.

[a] Number of residues whose coordinates are reported in the PDB.

[b] Heterogeneous group reported in PDB. JEN, $C_{16}H_{20}N_4O$; DMS, $C_2H_6O$ S; MP3, 3-phosphoglycerate; MYR, $C_{14}H_{26}O_2$; SPH, $C_{18}H_{35}NO_2$.

[c] Proteins whose specific S-S or S-B interactions are also unusually weak ($\leqslant -0.15$ *RT* per residue) or repulsive, in addition to their homogeneous S-S and S-B interactions.

considered in the present study, the complete list of which is available on the Internet (Bahar & Jernigan, 1996).

The structures subject to weak $E_{\mathrm{tot}}$(S-S) and $E_{\mathrm{tot}}$(S-B) may be susceptible to unfolding, unless the total energy is supplemented by other effects not included here. In this connection, we note that these non-bonded potentials do not include the enhanced stability from (1) covalent bonds between the disulfide bridge-forming cysteine residues, (2) coordination with a metal ion or other prosthetic group buried in the protein, (3) interactions with small molecules if present or the other monomers in multimeric proteins, (4) strong interactions with immobile structural water molecules or (5) the environment difference for membrane proteins. Some of these effects are present among the proteins listed in Table 4. For example, 7wga, 2madl, 1nxb and 3ebx have 16, seven, four and four disulfide bridges, respectively, which might offset the weak non-bonded attractions to stabilize the structure. Other proteins, such as 1tabi, 2plv1, 3pgm, 1cy3, 2cdv and 1r094, form complexes with oligomeric or prosthetic compounds, and their weak energies suggest that these structures might be unfolded, except for the stability imparted by their complex formation. Examination of the specific contribution to $E_{\mathrm{tot}}$(S-S) and $E_{\mathrm{tot}}$(S-B) reveals that many of the proteins listed in Table 4 (marked [c]) have relatively unfavorable ($\leqslant -0.15$ $RT$ per residue) specific interaction potentials. Some counter-examples, i.e. proteins whose specific interactions are much more favorable than the average native structure, are 2mrt, 1mhu, 1cbh, 1atf, 4rxn, 3sici, 2ssi, 1tho, 1ppfe, 1sgt, 1tpa, 1tgs, 1s01, 8adh and 1ace.

Smaller proteins ($n \leqslant 60$) are generally subject to relatively weaker non-bonded interactions. This feature may be confirmed by plotting the logarithm of [$E_{\mathrm{tot}}$(S-S) + $E_{\mathrm{tot}}$(S-B)] against log $n$ (Figure 7) for the 302 proteins of Sets I and II. A power law of the form $E_{\mathrm{tot}}$ ξ $n^{1.28}$ is found from the slope of the best fit line from linear regression giving a strong correlation coefficient, 0.987. The proteins in Table 4, whose non-bonded energetics are relatively weak, are shown by the plus signs on the figure, some of which are also labeled; whereas the results for the remaining 279 proteins are indicated by the open squares.

The observed departure from a linear dependence on $n$ is evidence of the enhanced stability of larger proteins. Smaller proteins might owe a considerable part of their stability to disulfide bridge-forming cysteine residues, in a rather natural way to compensate for their weaker internal non-bonded energies. In fact, the average number of cysteine residues forming disulfide bridges decreases with increasing size of the proteins. The 46 proteins of our dataset (Bahar & Jernigan, 1996) having $n \leqslant 60$ exhibit, on average, 1.65 disulfide-bonded Cys per protein; whereas this number drops to 0.87 in the case of the 65 proteins with $60 \leqslant n \leqslant 120$, and to 0.49 for the remaining 187 structures with $n > 120$.

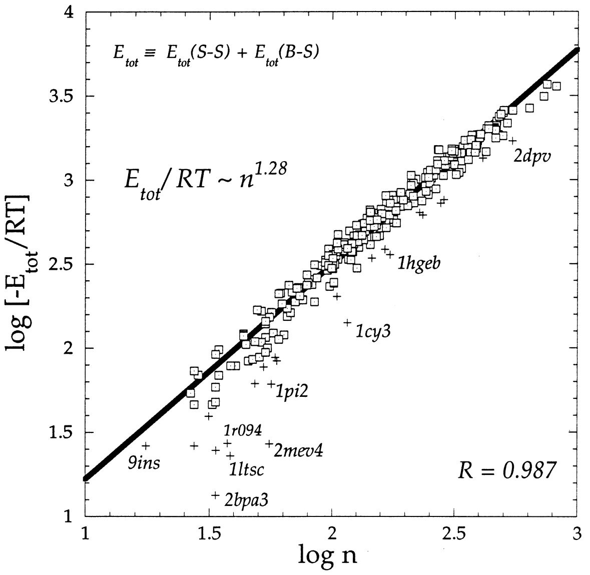

**Figure 7.** Dependence of the overall non-bonded energies $E_{\mathrm{tot}}$ of databank protein structures on the number of residues. Results are plotted for 302 known structures given in Sets I and II on the Internet (Bahar & Jernigan, 1996). The logarithm of $E_{\mathrm{tot}}/RT$ is plotted against log $n$. This yields a power law of the form $E_{\mathrm{tot}}$ ξ $n^{1.28}$ with a correlation coefficient of 0.987. The proteins listed in Table 4, which are subject to unusually weak interaction potentials, are shown by + signs; these are excluded from the least-squares calculation of the best fit line. The results for the remaining 279 proteins are indicated by squares.

## Threading experiments

Threading or inverse folding experiments are performed without gaps or insertions using a structure-recognizes-sequence protocol (Hendlich *et al.*, 1990; Kocher *et al.*, 1994) for 62 PDB structures of size $29 \leqslant n \leqslant 292$, determined to a resolution better than 2.8 Å. Sequences of 32 additional proteins with $185 \leqslant n \leqslant 488$ are utilized so as to increase the number of variants (sequence fragments of all larger size proteins, obtained by advancing one residue at a time) threaded through the 62 target structures, in parallel with the approach of Hendlich *et al.* (1990). The results are presented in Table 5. The first two columns give the PDB name and the size of the known structure. The number of sequence variants threaded onto each structure is given in the third column. The rank of the native sequence amongst all variants, classified on the basis of total energy $E_{\mathrm{tot}} = E_{\mathrm{tot}}$(S-S) + $E_{\mathrm{tot}}$(S-B), is given in the fourth column. Of 62 structures 52 correctly recognize the native sequence. The fifth column represents the energy difference $\Delta E_{\mathrm{tot}}$ per residue between the native sequence and the variant yielding the lowest energy. The ten cases that yield a positive $\Delta E_{\mathrm{tot}}$ and therefore fail to identify

**Table 5.** Results of threading experiments based on residue-specific S-S and S-B potentials

| PDB name | $n$ | Number of variants | Rank (long-range) | $\Delta E/nRT$ (long-range) | Lowest energy variant |
|---|---|---|---|---|---|
| 1rhd | 292 | 1548 | 1 | −5.25 | |
| 1pyp | 280 | 1873 | 1 | −1.36 | |
| 1dri | 270 | 2155 | 1 | −3.51 | |
| 1aai | 261 | 2419 | 1 | −2.08 | |
| 1dnk | 259 | 2452 | 1 | −3.80 | |
| 1caj | 258 | 2570 | 1 | −4.50 | |
| 2ca2 | 256 | 2581 | 1 | −5.62 | |
| 1baa | 242 | 3057 | 1 | −1.95 | |
| 3pgm | 229 | 3540 | 1 | −0.51 | |
| 2cla | 213 | 4032 | 1 | −0.02 | |
| 1bbt2 | 209 | 4341 | 1 | −1.38 | |
| 1abm | 197 | 4871 | 1 | −1.27 | |
| 3adk | 193 | 5008 | 1 | −0.78 | |
| 1gky | 185 | 5429 | 1 | −0.75 | |
| 1cpc | 173 | 6011 | 1 | −0.03 | |
| 1cpcl | 173 | 6011 | 1 | −0.80 | |
| 1cd4 | 172 | 6063 | 1 | −1.58 | |
| 2fcr | 172 | 6063 | 1 | −0.55 | |
| 5p21 | 165 | 6433 | 1 | −1.07 | |
| 1l84 | 161 | 6665 | 1 | −1.20 | |
| 3dfr | 161 | 6665 | 1 | −0.96 | |
| 5tnc | 160 | 6711 | 1 | −0.09 | |
| 1mbn | 152 | 7172 | 1 | −0.53 | |
| 1lh3 | 152 | 7172 | 1 | −0.70 | |
| 1f3g | 149 | 7333 | 1 | −0.81 | |
| 1aak | 149 | 7333 | 1 | −1.32 | |
| 4cln | 147 | 7475 | 4 | +0.11 | 1po4, 1 |
| 1mba | 146 | 7537 | 1 | −0.95 | |
| 1fx1 | 146 | 7537 | 1 | −1.48 | |
| 1bab | 145 | 7601 | 1 | −1.52 | |
| 1bar | 137 | 8116 | 7 | +0.50 | 1po4, 10 |
| 1end | 136 | 8181 | 1 | −0.11 | |
| 1eco | 135 | 8250 | 1 | −0.85 | |
| 2snm | 134 | 8245 | 1 | −0.46 | |
| 1bbh | 130 | 8595 | 1 | −0.82 | |
| 1ifb | 130 | 8595 | 1 | −0.08 | |
| 1lhm | 129 | 8666 | 1 | −0.78 | |
| 1bw4 | 124 | 9024 | 1 | −0.55 | |
| 4p2p | 123 | 9097 | 1 | −0.48 | |
| 1alc | 121 | 9245 | 1 | −0.43 | |
| 1paz | 119 | 9391 | 1 | −0.64 | |
| 1cy3 | 117 | 9541 | 22 | +0.32 | 1brd, 102 |
| 1cd8 | 113 | 9844 | 1 | −0.76 | |
| 2ssi | 106 | 10,381 | 1 | −1.31 | |
| 1acx | 106 | 10,381 | 1 | −0.88 | |
| 1fkf | 106 | 10,381 | 1 | −0.92 | |
| 1fdd | 105 | 10,468 | 1 | −0.62 | |
| 1aps | 97 | 11,142 | 1 | −0.10 | |
| 1ten | 89 | 11,711 | 1 | −0.06 | |
| 2gn5 | 86 | 12,603 | 1 | −0.08 | |
| 1c5a | 65 | 13,736 | 65 | +0.73 | 1po4, 21 |
| 1nxb | 61 | 14,066 | 348 | +1.15 | 1acx, 34 |
| 1aaf | 54 | 14,649 | 1 | −0.01 | |
| 1egf | 52 | 14,819 | 19 | +0.45 | 8adh, 174 |
| 1gps | 46 | 15,331 | 1 | −0.06 | |
| 1atx | 45 | 15,418 | 1 | −3.25 | |
| 1cbn | 45 | 15,418 | 1 | −0.26 | |
| 1pdc | 41 | 15,774 | 1 | −0.12 | |
| 2bpa | 35 | 16,311 | 4684 | +0.29 | 1aai, 121 |
| 1bba | 35 | 16,311 | 178 | +0.59 | 1po4, 64 |
| 2mrt | 29 | 16,864 | 11 | +0.48 | 1brd, 166 |
| 1mhu | 29 | 16,864 | 4 | +0.16 | 1po4, 56 |

The rank refers to the position of the native sequence in the energy-sorted list obtained for all variants (column 3) threaded onto the reference structure (first column). The evaluation function is $E^{\mathrm{s}}_{\mathrm{tot}}$(S-S) + $E^{\mathrm{s}}_{\mathrm{tot}}$(S-B). The fifth column is the excess energy of the native sequence relative to the lowest energy variant. It assumes positive values if a non-native sequence (column 6) yields a lower energy than the native sequence.

The PDB name of the 32 additional proteins used to enlarge the set of sequences to be threaded onto the tabulated structures are 1lap, 1gly, 3ts1, 6icd, 3pgk, 3cp4, 1etu, 1efm, 8adh, 1atna, 1ald, 2reb, 2liv, 5ldh, 1abh, 1avha, 4tms, 4tln, 1ads, 1fnr, 2gbp, 3cpa, 3ccp, 1pyp, 3blm, 1brd, 1sgt, 2act, 1bbt1, 1po4, 1abma and 4sgb.

the correct sequence-structure match are generally smaller proteins, as expected. The PDB name and the index of its starting residue for these best ranking variants are given in the last column.

We have found that including the additional contributions of short-range interactions together with those of the non-bonded potentials in evaluating threaded sequences improves the recognition of correct sequence-structure pairs due to a compensating effect between short-range and long-range interactions (I. B., M. Kaplan & R. L. J., unpublished results).

## Discussion and Conclusion

### How accurate are contact potentials?

Thomas & Dill (1996) have drawn attention to the limits of applicability of knowledge-based potentials, and to possible systematic errors arising from the neglect of chain connectivity and excluded volume. Most of the observed features in databank-extracted potentials are pointed out to be biased by hydrophobic (H) interactions. For example, charged or polar residues (P) are said to be driven to the protein surface by the non-polar attractions of other amino acid residues, rather than their favorable interaction with the solvent. As a consequence, the contact potentials are pointed out to be strongly dependent on the size and composition of the proteins, on the surface-to-volume ratio, and particularly, on the extent of burial of hydrophobic residues in the protein interior. The partition propensity $\pi = 2n_c/(n_H q_H)$ of the protein, for example, is presented as an important parameter controlling the values extracted for the effective contact potentials. Here $n_c$ and $n_H$ are the total number of contacts and the number of hydrophobic residues in a given protein, and $q_H$ is the average coordination number of hydrophobic residues.

The basic assumptions and approximations adopted in the extraction of empirical potentials from known structures were recently discussed in some detail (Jernigan & Bahar, 1996), and will not be elaborated in depth here. The reader is referred to this review for a more thorough description of the limitations and/or achievements of knowledge-based potentials. However, in view of the recent issues raised by Thomas & Dill (1996), we perform here a rigorous analysis of the dependence of the effective H-H, H-P and P-P contact potentials on the size, the fraction of hydrophobic residues, and the partition propensities of the proteins included in the learning dataset. Also a few remarks on the adoption of Boltzmann statistics and on the effect of environment on extracted potentials are presented.

We note that Thomas & Dill (1996) base most of their remarks on the results obtained with short model chains ($n \leqslant 18$ monomers) of two types (H and P) of residues on a two-dimensional lattice. As also recognized by those authors, excluded volume is a more stringent constraint in two dimensions than in three, size and sequence effects may be more pronounced in their short-chain models and the effect of dominant interactions may be overestimated in chains composed of only two types of monomers, the real energetics of the proteins undoubtedly being more complex. The results from calculations presented below aim at giving an estimate of the actual errors incurred by neglecting size and composition effects.

#### Dependence of effective contact potentials on the size of the proteins

We have calculated the solvent-mediated effective contact potentials for subsets of proteins of different sizes selected from our original dataset of 302 structures. Results are displayed in Figure 8 for the case $r_c = 6.4$ Å. The filled circles refer to subsets of proteins with $n$ ranking in the successive intervals $0 \leqslant n \leqslant 150$, $50 \leqslant n \leqslant 200$, $\ldots$, $350 \leqslant n \leqslant 500$. Such overlapping ranges are selected so as to ensure that each subset comprises at least 50 protein structures. The abscissa represents the average number of residues of the proteins in each subset, and the ordinate represents the solvent-mediated effective contact energies averaged over different categories of residue pairs, mainly H-H, H-P, P-P and C-C. In the group H, we included the residues Ala, Val, Ile, Leu, Met, Phe, Trp, Tyr and Cys, following Thomas & Dill (1996); all other residues belong to the group P. C-C refers to oppositely charged residue pairs. The dependence of the effective contact potentials on the size of the proteins is found to be negligibly small. Calculations repeated for the effective self contact potentials $e_{AB}{}^0$ and for the close distance interval $r \leqslant r_c = 4.0$ Å confirm that the extracted potentials calculated according to equations (6) and (7) are practically independent of the size of the proteins included in the learning set. Results for the extremely short model chains used by Thomas & Dill (1996) showed a stronger dependence on chain size.

#### Dependence on the fraction of hydrophobic residues

Examination of the compositions of known proteins demonstrates that the fraction $x_H$ of hydrophobic residues is rather narrowly distributed. Precisely, 248 out of the complete set of 302 proteins have the fraction of hydrophobic residues in the range $0.36 \leqslant x_H \leqslant 0.46$. Evidently, within such a narrow range, the changes in effective contact potentials are well bounded. Calculations performed for subsets of at least 50 proteins lying in successive $x_H$ ranges showed that the variations in the effective H-H, H-P, P-P and C-C contact potentials from their mean values remain lower than $\Sigma 5\%$, in general (results not shown).

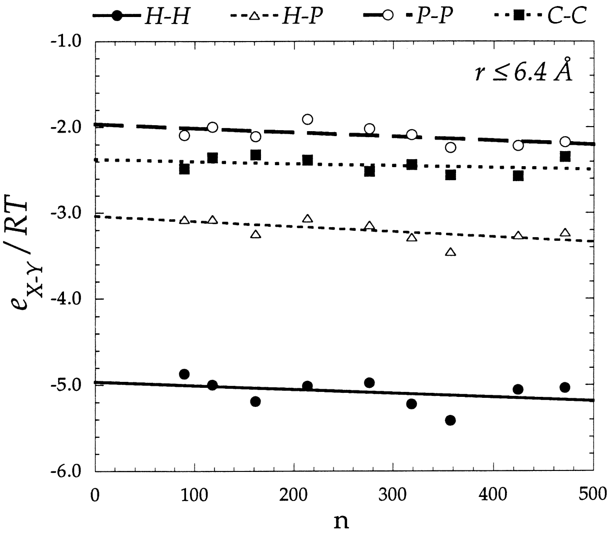

**Figure 8.** The weak dependence of solvent-mediated contact energies on the size (number of residues) of the proteins included in the learning dataset. Results are shown for H-H, H-P, P-P and C-C pairs, for subsets of $\geqslant 50$ proteins lying in a given size interval. The results are the averages over the individual contact potentials $e_{AB}(r_c)$ obtained using equation (7) for all pairs of side-chains $A$ and $B$ belonging to the groups H and P. C-C represents oppositely charged side-chains.

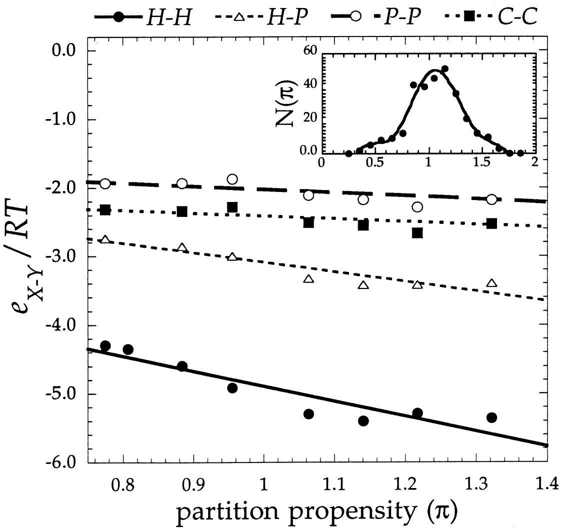

**Figure 9.** Dependence of solvent-mediated contact energies on partition propensity, π. The inset displays the distribution of partition propensity in terms of the number of proteins belonging to successive π intervals of width $\Delta\pi = 0.1$. The effective H-H, H-P, P-P and C-C contact potentials are shown for subsets of given π range comprising $\geqslant 50$ proteins.

### Dependence of the effective contact potentials on partition propensities

Results of calculations performed for $r_c \leqslant 6.4$ Å are displayed in Figure 9. The inset shows the distribution of partition propensities, i.e. the numbers of proteins lying in successive overlapping π intervals, each comprising $\geqslant$50 proteins. The solvent-mediated contact potentials appear to be more favorable as the partition propensities of the proteins increase, i.e. as the surface-to-volume ratio and composition of the protein permit a more efficient burial of the hydrophobic residues in the core, in qualitative agreement with the results of Thomas & Dill (1996). Yet the deviations from the mean contact potentials are confined to the range Σ10% in general, which may be viewed as a secondary effect. Even the H-H contacts, which are those most sensitive to π among all observed categories, exhibit a moderate departure (Σ12%) from their mean values over the range $0.76 \leqslant \pi \leqslant 1.36$. This range includes 240 proteins out of 302. This result differs significantly from the much stronger effect shown in their Figure 8 by Thomas & Dill (1996). The curves displayed in our Figures 8 and 9 provide a measure of the errors brought about by the adoption of the mean effective contact energies extracted from the complete dataset of proteins, without considering size and composition effects. If one so desired, these curves could be utilized to correct for these minor effects.

### On the validity of a thermodynamic equilibrium and the applicability of Boltzmann statistics

A basic postulate underlying the present approach is that the native state is the most thermodynamically stable form. However, the possibility remains that the native state may be at a non-global energy minimum attained as a result of a particular folding pathway favored by lower-energy barriers in the conformational space accessible to a given protein. Such effects would of course hamper the computational determination of the protein's native fold on the basis of knowledge-based mean fields. The more detailed semi-empirical potentials commonly used in atomic simulations would also suffer from the same limitation. Furthermore, the Boltzmann statistics applies, in a strict sense, to a single system that can visit several configurational states at a given temperature; whereas in knowledge-based approaches an ensemble of systems (proteins) having each a fixed (native) configuration is analyzed on the basis of the assumption that the ensemble average represents the equilibrium populations of different microstates. The validity of an effective Boltzmann temperature, roughly equal to room temperature, has been discussed elsewhere, although this is still an open issue (Thomas & Dill, 1996). Here, we have chosen to test the reproducibility of the results by performing the calculations for two independent sets of proteins, comprising each $\geqslant$150 proteins, and repeating the calculations for subsets of proteins of given sizes, compositions and partition propensities. In addition, threading experiments have been performed here and elsewhere (I. B., M. Kaplan &

R. L. J., unpublished results) which confirm the success of the potentials in recognizing correct structure-sequence matches; experimental hydrophobicity scales have been satisfactorily reproduced, and widely tested inter-residue effective contact potentials, which were recently validated for 1661 protein subunits (Miyazawa & Jernigan, 1996), have been obtained as a special case of the present potentials of mean force. All these results lend support to the applicability and utility of knowledge-based approaches, as an excellent first-order approximation to a complex problem.

### *Effect of environment*

It is intersting to note that similar knowledge-based interaction parameters derived from high-resolution X-ray structures and from low-resolution X-ray structures agree with a correlation of 0.91 (Godzik *et al.*, 1995), whereas the parameters extracted from NMR structures show a significantly weaker correlation (0.46) with those extracted from X-ray structures. This invites attention to the role of environment in determining the effective interactions, and the necessity of properly defining the reference state in simulations. As we discussed in a recent review (Jernigan & Bahar, 1996), a mixture of two reference states, for solvent exposure and residue exposure, might be appropriate in folding simulations, and either solvent-mediated or residue-mediated effective inter-residue contact energies (equations (7) or (12)) should be utilized depending on the local environment of a given residue.

## Major findings and their implications

### *Hydrophilic interactions are more important than hydrophobic ones for close inter-residue distances*

The connection between the knowledge-based pair radial distribution functions and the effective inter-residue contact energies operative over any distance range in different environments is established through equations (6), (7) and (12). Application of these expressions to two different distance ranges, reveals that at ''close'' distances, i.e. $r \leqslant 4$ Å, specific interactions between pairs of hydrophilic residues are predominantly important; whereas at longer separations hydrophobic interactions supersede in this role. The latter are much stronger and dominate the apparent behavior over the ''broad'' range $r \leqslant 6.4$ Å. These observations have important implications insofar as the simulations of low-resolution models are concerned. Broad distance potentials (Miyazawa & Jernigan, 1985, 1996), which are closely reproduced in Table 1, have proven to be useful in numerous studies for recognizing native-like folds (Jernigan & Bahar, 1996). However, a finer level of description may now be achieved by adopting a hierarchical approach, mainly using the close distance effective potentials (Table 3) after native-like structures have been attained with the broad distance potentials.

### *Stability is predominantly imparted by homogeneous interactions, the contribution of specific interactions being about five times weaker*

Two types of non-bonded interactions are distinguished here: specific interactions that discriminate between the different sequences for a given fold, and homogeneous interactions that stabilize a given fold regardless of the identity of the amino acid residues. The latter is found from the average of all S-S and S-B interactions. Homogeneous S-S interactions contribute by $-2.1$ *RT* per residue to the stability of native structures, and are stronger than the specific interactions by a factor of approximately 5. Thus, the drive for forming a compact structure appears to be much stronger than the tendency to select a particular conformation satisfying individual specific interactions. The latter type of interaction gains importance at close inter-residue separations only, i.e. after a certain degree of collapse is achieved. These two stages could conceivably correspond to the hydrophobic collapse manifested in the molten globule intermediate and the subsequent folding to native form.

### *The total non-bonded energy in globular proteins scales with the number* n *of residues as* $E \xi n^{1.28}$

This relationship is obeyed with a correlation coefficient of 0.987 by the enlarged ensemble of 302 proteins available on the Internet (Bahar & Jernigan, 1996). The departure from a linear dependence evidences the enhanced stability of larger proteins. The fact that the observed probability of disulfide bridges in small proteins is more than twice that for large proteins, suggests that smaller proteins may owe a considerable part of their stability to disulfide bridge-forming cysteine residues, in a rather natural way to compensate for their weaker internal non-bonded energies.

### *It is inappropriate to combine residues into a reduced set of representative groups*

Apart from hydrophobic residues whose non-bonded interactions with the surroundings are similar, the potentials of mean force between individual pairs of amino acid residues are strongly residue-specific and cannot be satisfactorily accounted for in terms of unified groups. And even the residues classified within the hydrophobic group exhibit some unique characteristics, imparted by size and shape effects. These distinctions become more pronounced upon examination of the short-range effects and geometric preferences on a local scale (I. B., M. Kaplan & R. L. J., unpublished results).

### Future directions

The precise evaluation of specific interactions takes on a critical importance for protein design and engineering applications. One suggested improvement to these non-bonded interactions is to incorporate effects of directionality in interactions (Bahar & Jernigan, 1996). Inasmuch as the protein interior is composed of a network of intramolecular hydrogen bonds, each residue comprising at least two polar bonds, the specificity imparted by the directional effects for interaction may be critically important in reducing the search space for the selection of native folds. One particularly interesting feature revealed in the present analysis is the characteristic difference in behavior, over distance, of hydrophobic and hydrophilic pairs. This characteristic is true for both S-S and S-B pairs. For hydrophobic pairs, the minimum in energy occurs at larger distances: in the range of 4 to 6 Å for most S-S pairs, and 6 to 8 Å for these residues with backbone atoms. The minima for the polar pairs always occur at smaller separations, usually in the range of 2 to 4 Å, as may be seen in Figures 4 and 5. This distance corresponds to an average of only one residue interaction per central residue. The behavior of the residue pairs in the closer range reflects the same overall intermolecular characteristics recently reported in an atomic surface interaction investigation (Wallqvist *et al.*, 1995), which also corresponds to a close approach. Perhaps the values in Table 3 for close approach can serve to span the gap between atomic and single-point residue models.

## Materials and Methods

In the low-resolution model adopted here, each residue is represented by two interaction sites, one on the backbone, and the second on the amino acid side-group. $\alpha$-Carbon atoms are conveniently used for the backbone structure points. Such a simple representation of the polypeptide backbone finds its roots in the pioneering work of Brant & Flory (1965). Side-group interaction centers are determined on the basis of a selection of their atoms that are subject to the most distinctive interactions (Table 4 in Bahar & Jernigan, 1996).

Similar simplified representations of protein side-groups have been adopted in a number of studies (Wilson & Doniach, 1989; Sun, 1993; Hendlich *et al.*, 1990; Sippl *et al.*, 1992; Bryant & Lawrence, 1993; Maiorov & Crippen, 1992). The model used by Wallqvist & Ullner (1994) provides a relatively more refined description of side-groups by allowing for more than a single interaction site for flexible and bulky amino acid residues. Here, the potential of average force assigned to each pair of residues is evaluated on the basis of multiple interactions taking place between the selected atoms along the side-chain, as opposed to a consideration of a single centroid point. The advantage of adopting such multiple site correlations is twofold. First, strong interactions incidentally involving particular atoms of the side-groups are explicitly taken at their actual locations. This can yield a more specific set of residue-residue interactions. We note that an extreme approach in this direction was that of Godzik & Skolnick (1992), in which residue contacts were identified on the basis of the close approach of any two atoms. Secondly, the use of multiple interactions enlarges the sample size, and consequently enhances the statistical accuracy and the smoothness of the data. These multiple interactions are of course correlated due to the connectivity of the side-group atoms, and consequently do not provide as much information as would an independent sample of the same size. Yet, their explicit consideration provides a more accurate mean field assessment of residue interactions. In another recent study of protein-ligand interactions we have developed atomic surface area based interaction energies (Wallqvist *et al.*, 1995); that approach provides another way to obtain more specific interactions.

### Materials. Overall characteristics

The proteins examined in the present study consist of two sets, I and II, available on the Internet (Bahar & Jernigan, 1996). The structures in Set I are selected from the set of proteins studied by Brauer & Beyer (1994) in generating pair potentials based on mutation data matrices. The original set was derived by Hobohm *et al.* (1992) and comprises 185 high-resolution structures, chosen to avoid homologous proteins. Protein structures in Set II are utilized for veryifying the reproducibility of the results obtained from Set I, and for further applications such as threading experiments, detection of structures whose energetics deviate from native-like behavior. These are taken from the enlarged set of 305 non-homologous structures presented by Hobohm & Sander (1994). We exclude those crystal or NMR structures whose side-group coordinates are not reported, and of course those already included in Set I.

The radii of gyration $s$ of the proteins in Set I are calculated to depend on the number $n$ of residues as $\log s^2 = (2/3) \log n + 0.92$, with a correlation coefficient of 0.92. A power law of the form $s^2 \xi\ n^{2/3}$ is in fact anticipated for spherical compact globular shapes; this indicates a direct proportionality between the volumes occupied by the chains and their chain lengths. This relationship supports the use of the radius of gyration, within a reasonable tolerance, as a property constraining the structure of a globular protein of known size, $n$ (Hao *et al.*, 1992). We note that Maiorov & Crippen (1992) presented a similar equation, $s_{\min} = -1.26 + 2.79\ n^{1/3}$, as a lower bound $s_{\min}$ for the radius of gyration of native proteins.

Analyses of the locations of different types of residues with respect to the centers of mass of the proteins yielded the probability distributions displayed in Figure 10. The frequencies of occurrences of residues within various concentric zones from the centroid of the protein reflect the spatial preferences of the residues (Prabhakaran & Ponnuswamy, 1980). The curves are shown here for (a) the group of hydrophobic residues Val, Ile, Leu, Phe, (b) the charged residue group Lys, Arg, Glu, Asp, and (c) all residues. All cases are normalized for comparison. The variable $(r - r_{\mathrm{cm}})$ of the abscissa denotes the distance of a given residue from the center of mass of the protein. It is normalized with respect to the radius of gyration $s$ of the protein, so as to eliminate biases due to size effects. The distribution of hydrophobic (or charged) residues indicates their tendency to be placed at inner (or outer) positions, as expected. It is interesting to observe that these distri-

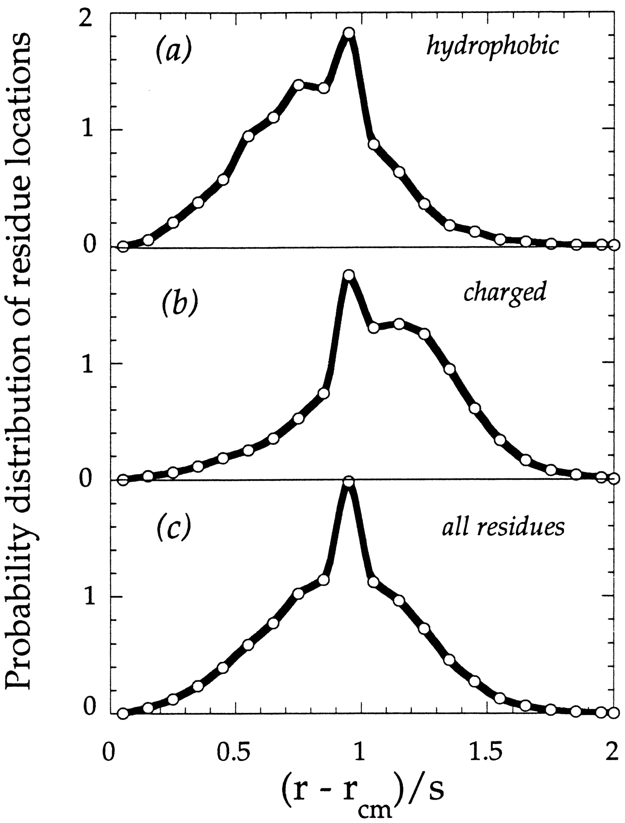

**Figure 10.** Normalized probability distributions of the radial location for different types of residues with respect to the centers of mass of proteins. The curves in (a), (b) and (c) represent the behavior of hydrophobic, charged and all residues, respectively. These are given for the groups (a) Val, Ile, Leu, Phe, (b) Lys, Arg, Glu, Asp, and (c) all residues. $(r - r_{cm})/s$ is the distance of a given residue from the center of mass of the protein, divided by the radius of gyration $s$ of the protein.

butions are non-Gaussian and exhibit a rather bimodal character. The highest peak in all cases is closely coincident with the root-mean-square radius of gyration. Calculations indicate that the second peaks arise from the contribution of particular residues such as Phe and Val in Figure 10(a), and Lys and Glu in Figure 10(b). Thus, residues presumed to be comparable in character, insofar as the interaction energetics are concerned, exhibit different preferences for location in globular structures.

### Method

The potential of mean force between two particles $A$ and $B$ located at spartial positions $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively, may be written as (Ben-Naim, 1992):

$$E_{AB}(\mathbf{r}_1, r_2) = -RT \ln g_{AB}(\mathbf{r}_1, \mathbf{r}_2) \qquad (15)$$

where $g_{AB}$ ($\mathbf{r}_1$, $\mathbf{r}_2$) is the pair correlation function, also called the radial distribution function, $R$ is the gas constant and $T$ is the absolute temperature. $g_{AB}$ ($\mathbf{r}_1$, $\mathbf{r}_2$) is proportional to the joint probability $P_{AB}$($\mathbf{r}_1$, $\mathbf{r}_2$) d$\mathbf{r}_1$ d$\mathbf{r}_2$ of finding particle $A$ at a position $\mathbf{r}_1$, and $B$ at $\mathbf{r}_2$, within the differential volume elements d$\mathbf{r}_1$ and d$\mathbf{r}_2$, respectively. In the case of a uniform distribution of independent particles in a multicomponent system, this joint probability reduces to the product of the respective mole fractions $x_A$ and $x_B$; whereas in the presence of concentration fluctuations imposed by specific interactions, a relationship of the form:

$$P_{AB}(\mathbf{r}_1, \mathbf{r}_2) = x_A x_B g_{AB}(\mathbf{r}_1, \mathbf{r}_2) = (f\rho_A f\rho_B/\rho^2)g_{AB}(\mathbf{r}_1, \mathbf{r}_2) \quad (16)$$

holds. Here ρ is the mean number density of the system, and $\rho_A \equiv x_A\rho$ and $\rho_B \equiv x_B\rho$ are the densities for components $A$ and $B$. The product $x_A$ $x_B$ gives the *a priori* probability of having a residue pair or a contact of type [$A$,$B$] in the ensemble, and $g_{AB}$ ($\mathbf{r}_1$,$\mathbf{r}_2$) describes the further conditional probability that this residue pair will be located at ($\mathbf{r}_1$, $\mathbf{r}_2$), in the volume elements d$\mathbf{r}_1$ and d$\mathbf{r}_2$ respectively, given that we are considering the subset of residue pairs [$A$,$B$].

For the spherically symmetric interactions adopted here, $g_{AB}$ ($\mathbf{r}_1$, $\mathbf{r}_2$) is a function of the scalar distance $j\mathbf{r}_1 - \mathbf{r}_2 j \equiv$ r only, and is proportional to the number $N_{BjA}(r)$ of particles of type $B$ within a differential spherical shell of volume d$\mathbf{r} = 4\pi r^2$ d$r$, at a distance $r$ from a central particle $A$. The number $N_{AB}(r)$ of pairs [$A$,$B$] at a separation $r$ from each other is:

$$N_{AB}(r) = N_{BjA}(r)N_A = N_A\rho_B g_{AB}(r)4\pi r^2\, dr \qquad (17)$$

In the present analysis of protein crystal structures, $N_{AB}(r)$ is simply evaluated by counting pairs at a given separation as:

$$N_{AB}(r) = \underset{i}{\mathrm{X}}\;\underset{j}{\mathrm{X}}\;\delta(j\mathbf{r}_{Ai} - \mathbf{r}_{Bj}j - r) \qquad (18)$$

Where $\mathbf{r}_{Ai}$ indicates the position vector of the $i$th residue of type $A$, and $\mathbf{r}_{Bj}$ that of the $j$th residue of type $B$, δ($x$) is the Kronecker delta, which is equal to 1 if its argument is equal to zero, and zero otherwise. The summation in equation (18) is performed over all occurrences of residues $i$ and $j$, of types $A$ and $B$, respectively, in the set of protein structures, sequentially separated by five or more virtual bonds. Bins of thickness $\Delta r = 0.4$ Å are used for collecting the values of $N_{AB}(r)$. From equation (17), we obtain the normalized radial pair distribution function in a different form:

$$\bar{g}_{AB}(r) \equiv \frac{\cancel{P}g_{AB}(r)}{{}_r g_{AB}(r)} = \cancel{P}\frac{N_{AB}(r)/(4\pi r^2)}{{}_r[N_{AB}(r)/(4\pi r^2)]} \qquad (19)$$

$\bar{g}_{AB}(r)$ reflects the probability of contact for a given pair, as a function of distance, unbiased by the frequencies of the particular residues in the database. The overbar in $\bar{g}_{AB}(r)$ indicates that $g_{AB}$ is normalized for the subset of pairs [$A$,$B$]. The corresponding potential of mean force is:

$$E_{AB}(r) = -RT \ln \overset{\kappa}{\phantom{x}}(4\pi r^2)^{-1}N_{AB}(r)\,\overset{\delta}{\phantom{x}}\underset{r}{\mathrm{X}}\,(4\pi r^2)^{-1}N_{AB}(r)\overset{\lambda}{\phantom{x}} \qquad (20)$$

The summations in equations (18) to (20) have been performed over all contacts between side-chains separated by five or more virtual bonds. The contribution of the terms for larger $r$ ($r \geqslant 16$ Å for instance) is vanishingly small, and the exact choice of the upper limit for the above summations has only a negligible effect on the results. The contacts with separations $r < 2.0$ Å, which are presumably due to the inaccuracies in the crystallographic measurements and modeling, are approximated here by $r = 2$ Å.

## Acknowledgement

Partial support by NATO CRG Project no. 951240 is gratefully acknowledged.

## References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) . The protein databank. In *Crystallographic Databases-Information Content Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Sievers, R., ed.) p. 107, Data Commission of the International Union of Crystallography, Bonn, Cambridge and Chester.

Bahar, I. & Jernigan, R. L. (1996). Coordination geometry of non-bonded residues in globular proteins. *Folding Des.* **1**, 357–370. Supplementary material and paper available on the WWW at http://BioMedNet.Com/cbiology/fad.htm.

Ben-Naim, A. (1992). *Statistical Thermodynamics for Chemists and Biochemists*, Plenum Press, New York.

Bernstein, F., Koetzel, 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.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known 3-dimensional structure. *Science,* **253**, 164–170.

Brant, D. A. & Flory, P. J. (1965). The configuration of random polypeptide theory. *J. Am. Chem. Soc.* **87**, 2791–2800.

Brauer, A. & Beyer, A. (1994). An improved pair potential to recognize native protein folds. *Proteins: Struct. Funct. Genet.* **18**, 254–261.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.

Burley, S. K. & Petsko, G. A. (1988). Weakly polar interactions in proteins. *Advan. Protein Chem.* **39**, 125–189.

Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **244**, 725–732.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. & Delisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195**, 659–685.

Covell, D. G. & Jernigan, R. L. (1990). Conformations of folded proteins in restricted spaces. *Biochemistry,* **29**, 3287–3294.

Crippen, G. M. & Viswanadhan, N. V. (1985). Sidechain and backbone potential functions for conformational analysis of proteins. *Int. J. Pept. Protein Res.* **25**, 487–509.

Flocco, M. M. & Mowbray, S. L. (1994). Planar stacking interactions of arginine and aromatic side-chains in proteins. *J. Mol. Biol.* **235**, 709–717.

Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins. Applications to supersecondary and tertiary structure determination. *Proc. Natl Acad. Sci. USA,* **89**, 98–102.

Godzik, A., Kolinski, A. & Skolnick, J. (1995). Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* **4**, 2107–2117.

Hao, M., Rackovsky, S., Liwo, A., Pincus, M. R. & Scheraga, H. A. (1992). Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc. Natl Acad. Sci. USA,* **89**, 6614–6618.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & 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.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Exhaustive matching of representative protein data sets. *Protein Sci.* **1**, 409–417.

Jernigan, R. L. (1992). Protein folds. *Curr. Opin. Struct. Biol.* **2**, 248–256.

Jernigan, R. L. & Bahar, I. (1996). Simple empirical potentials derived from structures and their use in protein simulations. *Curr. Opin. Struct. Biol.* **6**, 195–209.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86–89.

Kocher, J.-P. A., Rooman, M. J. & Wodak, S. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598–1613.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Levitt, M. & Perutz, M. F. (1988). Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol.* **201**, 751–754.

Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature,* **253**, 694–698.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

Meirovitch, H., Rackovsky, S. & Scheraga, H A. (1980). Empirical studies of hydrophobicity. 1. Effect of protein size and the hydrophobic behavior of amino acids. *Macromolecules,* **13**, 1398–1405.

Mitchell, J. B. O., Nandi, C. L., McDonald, I. K., Thornton, J. M. & Price, S. (1994). Amino-aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? *J. Mol. Biol.* **239**, 315–331.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules,* **18**, 534–552.

Miyazawa, S. & 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.

Monge, A., Lathrop, E. J., Gunn, J. R., Shenkin, R. S. & Friesner, R. A. (1995). Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995–1012.

Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine peptide in aqueous ethanol and dioxane solutions. Establishment of hydrophobicity scales. *J. Biol. Chem.* **246**, 2211.

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

Prabhakaran, M. & Ponnuswamy, P. K. (1980). Spatial assignment of amino acid residues in globular proteins: an approach from information theory. *J. Theor. Biol.* **87**, 623–637.

Rose, G., Geselowitz, A., Lesser, G., Lee, R. & Zehfus, M. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science,* **229**, 834–838.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–589.

Sippl, M. (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., Hendlich, M. & Lackner, P. (1992). Assembly of polypeptide and protein backbone conformation from low energy ensembles of short fragments: development of strategies and construction of models for myoglobin, lysozyme, and thymosin $b_4$. *Protein Sci.* **1**, 625–640.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Sun, S., Luo, N., Ornstein, R. & Rein, R. (1992). Protein structure prediction based on statistical potential. *Biophys. J.* **62**, 104–106.

Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457–469.

Wallqvist, A. & Ullner, M. (1994). A simplified amino acid potential for use in structure predictions of proteins. *Proteins: Struct. Funct. Genet.* **18**, 267–280.

Wallqvist, A., Jernigan, R. L. & Covell, D. G. (1995). A preference-based free energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci.* **4**, 1881–1903.

Wang, Y., Zhang, H., Li, W. & Scott, R. A. (1995a). Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl Acad. Sci. USA,* **92**, 709–713.

Wang, Y., Zhang, H. & Scott, R. A. (1995b). A new computational model for protein folding based on atomic solvation. *Protein Sci.* **4**, 1402–1411.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins: Struct. Funct. Genet.* **6**, 193–209.

Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. (1981). Affinities of amino acid sidechains for solvent water. *Biochemistry,* **20**, 849–855.

***Edited by B. Honig***

*(Received 20 September 1996; received in revised form 26 October 1996; accepted 26 October 1996)*