# A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. I. Functional Forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data

**A. LIWO,[1,2] S. OŁDZIEJ,[1] M. R. PINCUS,[3] R. J. WAWAK,[2] S. RACKOVSKY,[4] H. A. SCHERAGA[2]**

[1]*Department of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland*
[2]*Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301*
[3]*Department of Pathology, Brooklyn Veterans Administration Medical Center, Brooklyn, New York 11209 and State University of New York, Health Science Center, Brooklyn, New York 11203*
[4]*Department of Biomathematical Sciences, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, New York 10029*

*Received 7 June 1996; accepted 11 September 1996*

**ABSTRACT:** A two-stage procedure for the determination of a united-residue potential designed for protein simulations is outlined. In the first stage, the long-range and local-interaction energy terms of the total energy of a polypeptide chain are determined by analyzing protein–crystal data and averaging the all-atom energy surfaces. In the second stage (described in the accompanying article), the relative weights of the energy terms are optimized so as to locate the native structures of selected test proteins as the lowest energy structures. The

*Correspondence to*: H. A. Scheraga; e-mail: has5@cornell.edu

This article includes Supplementary Material available from the authors upon request or via the Internet at ftp.wiley.com/public/journals/jcc/suppmat/18/849 or http://www.journals.wiley.com/jcc

Contract grant sponsors: Polish State Committee for Scientific Research, contract grant number: PB 190/T09/96/10; National Institute on Aging, contract grant number: AG 00322; National Institute of General Medical Sciences, contract grant number: GM 14312; National Science Foundation, contract grant number: MCB95-13167; National Cancer Institute, contract grant number: CA 42500



CCC 0192-8651 / 97 / 070849-25

goal of the work in the present study is to parameterize physically reasonable functional forms of the potentials of mean force for side-chain interactions. The potentials are of both radial and anisotropic type. Radial potentials include the Lennard–Jones and the shifted Lennard–Jones potential (with the shift parameter independent of orientation). To treat the angular dependence of side-chain interactions, three functional forms of the potential that were designed previously to describe anisotropic systems are evaluated: Berne–Pechukas (dilated Lennard–Jones); Gay–Berne (shifted Lennard–Jones with orientation-dependent shift parameters); and Gay–Berne–Vorobjev (the same as the preceding one, but with one more set of variable parameters). These functional forms were used to parameterize, within a short-distance range, the potentials of mean force for side-chain pair interactions that are related by the Boltzmann principle to the pair correlation functions determined from protein-crystal data. Parameter determination was formulated as a generalized nonlinear least-squares problem with the target function being the weighted sum of squares of the differences between calculated and ''experimental'' (i.e., estimated from protein-crystal data) angular, radial-angular, and radial pair correlation functions, as well as contact free energies. A set of 195 high-resolution nonhomologous structures from the Protein Data Bank was used to calculate the ''experimental'' values. The contact free energies were scaled by the slope of the correlation line between side-chain hydrophobicities, calculated from the contact free energies, and those determined by Fauchere and Pliška from the partition coefficients of amino acids between water and *n*-octanol. The methylene group served to define the reference contact free energy corresponding to that between the glycine methylene groups of backbone residues. Statistical analysis of the goodness of fit revealed that the Gay–Berne–Vorobjev anisotropic potential fits best to the experimental radial and angular correlation functions and contact free energies and therefore represents the free-energy surface of side-chain–side-chain interactions most accurately. Thus, its choice for simulations of protein structure is probably the most appropriate. However, the use of simpler functional forms is recommended, if the speed of computations is an issue.
 *J Comput Chem* **18**: 849–873, 1997

**Keywords:** protein structure prediction; united-residue representation of a polypeptide chain; potential of mean force; radial and angular distribution functions

## Introduction

The force fields that use a representation of amino-acid residues as one or two interaction sites, hereafter referred to as united-residue potentials, have long been of interest in theoretical simulations of protein structure.[1–39] The primary reason for this is that they involve much less computational effort than all-atom or united-atom representations of the polypeptide chain. This is especially important in protein-structure prediction, where extensive search of the conformational space of the polypeptide chain is required to locate its global minimum energy. After the global minimum energy has been found for the simplified chain, it can be converted to the all-atom chain, and limited exploration of the conformational space of the all-atom chain can then be carried out to locate the global minimum in the all-atom representation. Such a protocol has recently been developed and implemented with considerable success by Skolnick et al. in predicting the three-dimensional structures of model monomeric helical proteins,[15,16,18] crambin (which also contains a $\beta$-sheet section),[18] and the dimeric GCN4 leucine zipper.[19] These investigators used an on-lattice representation of the polypeptide chains to obtain united-residue structures which were then converted to full-atom chains by using a set of statistical rules determined from the Protein Data Bank. In our recent reports, we have described a similar procedure, based, however, on an off-lattice model of

polypeptide chains[38,39] and a dipole-path method (based on an optimal hydrogen-bond network) to convert the α-carbon trace to an all-atom backbone.[38] This procedure succeeded in predicting the three-dimensional structure of the avian pancreatic polypeptide.

As mentioned previously, there are two ways to explore the conformational space of polypeptide chains with the use of a united-residue potential: the on-lattice and the off-lattice approach. In the first case, the polypeptide chain is superposed on a discrete lattice, and the number of possible conformations is, therefore, finite. In the simplest approach, the interaction potential is reduced to a set of residue–residue contact free energies.[8–10] The rationale for such an approach was based on the assumption that side-chain packing is the principal driving force in protein folding; more recent studies, however, have shown that this assumption is probably not true.[13] The recent approach developed in Skolnick's group incorporates many different interactions that can be responsible for protein folding: side-chain packing; local interactions; hydrogen bonding; surface energy; and cooperativity in side-chain packing and hydrogen bonding.[12–19] The contact and hydrogen-bonding energies depend on the distance and orientation of the interacting sites. The resulting force field was able to locate the near-native structures of a number of test proteins as the lowest energy ones.[14–16,18,19] The parameters of the potentials for on-lattice simulations were determined from a statistical analysis of the distributions of interacting sites obtained from the crystal data of known proteins. Because the aforementioned force field expresses most of the energy components as analytical functions of geometry, it can also be used for off-lattice simulations.

For the sake of completeness, we mention here simple lattice models of proteins in which contact free energies and other interaction parameters are assigned arbitrary values (usually three types of contacts are chosen: hydrophobic–hydrophobic; hydrophobic–hydrophilic; and hydrophilic–hydrophilic); however, such models were used to study general statistical–mechanical characteristics of polypeptide chains and the folding process, but have not yet been used for predicting the three-dimensional structures of real proteins.[40–42]

The united-residue potentials for off-lattice simulations have an even longer history than the on-lattice ones.[1–7,24–37,39] They have also been used with considerable success to predict the three-dimensional structure of known proteins.[28–31,35,37,38] In contrast to the on-lattice potentials, they are functions of continuous variables. Therefore, the off-lattice approach to protein folding enables local energy minimization to be carried out for generated structures. Thus, many powerful techniques for global-search minimization, such as Monte Carlo with Minimization (MCM),[43,44] the diffusion equation method (DEM),[45] or the self-consistent mean torsional field (SCMTF)[46] method can be applied. This was the rationale for choosing the off-lattice potential in the present work.

The present work was aimed at determining the long-range potential for side-chain interactions. We parameterized several functional forms for the interaction potential that also include angular dependence. This was motivated by the results of a preliminary analysis of the average dimensions of the side chains as calculated from the Protein Data Bank, which show that ''geometrical'' anisotropy (which can be defined as the ratio of the long axis of a side chain to the geometric average of the two shorter axes) is pronounced in almost all cases. Also, Koliński et al. noticed that the pair distributions of side chains exhibit some anisotropy, and included this effect in their on-lattice statistical potential.[14] The short-range part of the potential is presented in the accompanying work.[47]

## Methods

### REPRESENTATION OF POLYPEPTIDE CHAINS AND INTERACTION SCHEME

The united-residue model of polypeptide chains adopted in this work is a natural extension of the model developed in our earlier studies.[38,39] The chain is represented by a sequence of α-carbons ($C^{\alpha}$) linked by virtual bonds with attached united side chains (SC) and united peptide groups (p) located in the middle between the consecutive α-carbons. Only the united peptide groups and united side chains serve as interaction sites, the α-carbons assisting in the definition of the geometry (Fig. 1). As in our previous model,[38,39] all the virtual bond lengths (i.e., $C^{\alpha}$—$C^{\alpha}$ and $C^{\alpha}$—SC) are fixed; the $C^{\alpha}$—$C^{\alpha}$ distance is taken as 3.8 Å, which corresponds to trans peptide bonds. We allow, however, for variation of the side-chain positions with respect to the backbone ($\alpha_{SC}$ and $\beta_{SC}$), and for the variation of the virtual-bond angles, $\theta$, which were assumed fixed in our earlier approach.[38,39]

The energy of the virtual-bond chain is expressed by:

$$U = \sum_{i<j} U_{SC_iSC_j} + \sum_{i\neq j} U_{SC_ip_j} + w_{el} \sum_{i<j-1} U_{p_ip_j} + w_{tor} \sum_i U_{tor}(\gamma_i) + w_{loc} \sum_i \left[ U_b(\theta_i) + U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) \right] + w_{corr} U_{corr} \quad (1)$$

where $U_{SC_iSC_j}$, $U_{SC_ip_j}$, and $U_{p_ip_j}$ denote the energies of the interactions between side chains, between side chains and peptide groups, and between peptide groups, respectively, $U_{tor}(\gamma_i)$ denotes the energy of variation of the virtual-bond dihedral angle $\gamma_i$, $U_b(\theta_i)$ denotes the ''bending energy of the virtual-bond angle $\theta_i$, $U_{rot}(\alpha_{SC_i}, \beta_{SC_i})$ is the local energy of side chain $i$, $U_{corr}$ includes cooperative terms (e.g., the four-body interactions considered by Skolnick et al.,[15] as will be shown in part III of the present work) and the $w$ values denote relative weights of the respective energy terms.

## GENERAL PROCEDURE OF PARAMETERIZATION

The following procedures are commonly used to parameterize united-residue potentials:

1. Direct averaging of the all-atom potentials over those degrees of freedom that are lost when passing from the all-atom to the united-residue representation of the polypeptide chain.[2–5] This method directly implements the assumption that united-residue potentials are formally all-atom potentials averaged over some ''less important'' degrees of freedom (such as the dihedral angles, $\chi$, of the side chains). In this approach, functional forms are assumed for all energy terms, with parameters determined by fitting energies computed at chosen values of the variables describing the geometry of interacting sites to those obtained by direct averaging of the all-atom potentials. However, even in the earliest attempts, it was recognized that some extra terms have to be added to account for hydrophobic interactions between the side chains; in the early work of Levitt and Warshell[1] and Levitt,[2] those terms were assigned according to side-chain hydrophobicities.

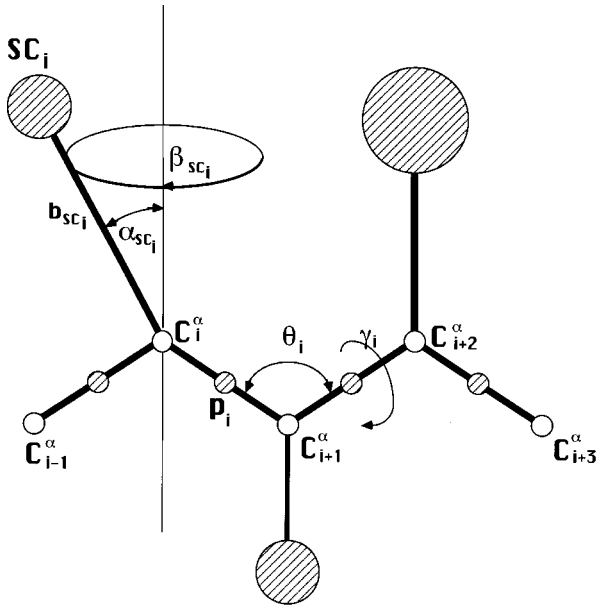


**FIGURE 1.** United-residue representation of a polypeptide chain. The interaction sites are side-chain centroids of different sizes (SC) and peptide-bond centers (p) indicated by dashed circles, whereas the $\alpha$-carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual $C^\alpha$—$C^\alpha$ bonds have a fixed length of 3.8 Å, corresponding to a trans peptide group; the virtual bond ($\theta$) and dihedral ($\gamma$) angles are variable. Each side chain is attached to the corresponding $\alpha$-carbon with a fixed ''bond length,'' $\mathbf{b}_{SC_i}$, variable ''bond angle,'' $\alpha_{SC_i}$, formed by $SC_i$ and the bisector of the angle defined by $C^\alpha_{i-1}$, $C^\alpha_i$, and $C^\alpha_{i+1}$, and with a variable ''dihedral angle'' $\beta_{SC_i}$ of counterclockwise rotation about the bisector, starting from the right side of the $C^\alpha_{i-1}$, $C^\alpha_i$, $C^\alpha_{i+1}$ frame.

2. Determination of the united-residue potentials so as to reproduce the single-body, pair, and possibly triplet distribution or correlation functions, as well as contact free energies determined from protein crystal data.[24–26,32–35] In this case, the potential is expressed either in a functional form,[24–26] or as a set of values (at given points) obtained by taking the negative logarithms of appropriately scaled correlation functions. The on-lattice potentials and the harmonic potentials of the distance-constraint approach of Goel and Yčas[21] and Wako and Scheraga[22,23] can be considered to belong to this class, although the latter[21–23] used the one- and two-body distribution functions directly.
3. A combination of the two preceding approaches in which some part of the potential is determined by direct averaging of the all-

atom potential, for example, the local and hydrogen-bonding interactions, and some estimated from protein crystal data. Such a division is motivated by the fact that, if direct averaging is computationally feasible, as in the case of the local and hydrogen-bonding interactions, the resulting potential will always be more accurate than that calculated from experimental distribution functions, whose accuracy is severely limited by the sparse number of protein crystal data. Conversely, obtaining the hydrophobic potential by direct averaging is in most cases not feasible, owing to the large number of degrees of freedom over which averaging must be carried out (i.e., the dihedral angles, $\chi$, for each side chain) and possibly to the necessity of including explicit water molecules in the averaging. Such a combination was implemented in our earlier work.[38] The local-interaction and backbone hydrogen-bonding terms were determined by direct averaging of the all-atom ECEPP/2[48,49] potential. This was motivated by the fact that local and hydrogen-bonding interactions are well represented in the ECEPP force field.[50] The hydrophobic potential was assumed to have a modified Lennard–Jones form, the parameters being assigned by the use of protein crystal data, namely the side-chain van der Waals radii[2] and the interresidue contact free energies.[9]

4. Determination of the parameters of the potential so as to locate the native structures as global minima for a set of training proteins and, simultaneously, introducing a large energy gap between the near-native and nonnative structures. For on-lattice simulations, such an approach based on spin-glass theory was developed by Wolynes and coworkers.[20] A similar method was developed for off-lattice simulations by Crippen and coworkers.[26–31] In both cases, the resulting potentials appeared successful in predicting the native structures of proteins that were not included in the training sets.

In this work, we have implemented procedure 3 to determine the parameters of individual energy terms (the $U$ values). The side-chain interaction and local terms are parameterized based on correlation functions collected from the Protein Data Bank (PDB). For the peptide-group interaction potential, $U_{pp}$, we use the energy function developed, and then parameterized through averaging of the all-atom ECEPP/2 potential,[48,49] in our earlier studies.[38,39] The derivation of local-interaction terms $U_b(\theta)$ and $U_{rot}(\alpha_{SC}, \beta_{SC})$ from protein-crystal data will be described in an accompanying article. In the accompanying work,[47] we also describe the procedure for the determination of the relative weights so as to locate the native structures of a set of training proteins as the lowest-energy ones. Therefore, our approach is a combination of all the procedures to determine the aforementioned potential. Use of distribution functions or averaging of all-atom potentials to obtain individual energy terms allows us to collect data from the PDB or from all-atom potential functions with meaningful statistics. The use of flexible weights, which constitute a small subset of adjustable parameters, enables us to scale the individual terms so as to obtain a folding potential. The procedure for weight determination is described in the accompanying article.[47]

## MODELING SIDE-CHAIN INTERACTIONS

The general form of the side-chain interaction ($U_{SC_iSC_j}$) parameterized in this work is given by:

$$U_{ij} = 4\left[|\epsilon_{ij}| x_{ij}^{12} - \epsilon_{ij} x_{ij}^{6}\right] \tag{2}$$

where $\epsilon_{ij}$ is the pair-specific van der Waals well depth, which depends on side-chain orientation for the potentials with angular dependence; as in our earlier work, $\epsilon > 0$ corresponds to hydrophobic–hydrophobic-type and $\epsilon < 0$ to hydrophobic-hydrophilic and hydrophilic-hydrophilic-type interactions. The quantity, $x_{ij}$, is the reciprocal of the reduced distance between side chains; for angular-dependent potentials, it also depends on the orientation of the side chains.

We first consider radial-only potentials. We assumed the following two functional forms for $x_{ij}$:

$$x_{ij} = \frac{\sigma_{ij}^0}{r_{ij}} \tag{3}$$

$$x_{ij} = \frac{r_{ij}^0}{r_{ij} + r_{ij}^0 - \sigma_{ij}^0} \tag{4}$$

Equation (3) corresponds to the Lennard–Jones (LJ)-type potential. The constant $\sigma_{ij}^0$, in this case, can be identified with the equilibrium van der Waals distance between side chains $i$ and $j$. Equation (4) corresponds to the shifted Lennard–Jones

potential of the form proposed by Kihara[51] (hereafter referred to as LJK). In this case, the quantities $\sigma_{ij}^0 - r_{ij}^0$ and $r_{ij}^0$ can be identified with the dimensions of the ''hard core'' and the ''soft core'' of the interacting bodies, respectively; we express the hard-core diameter as a combination of two terms, $\sigma_{ij}^0$ and $r_{ij}^0$, to maintain consistency of the notation with that corresponding to the angular potentials.

The constants $r_{ij}^0$ and $\sigma_{ij}^0$, can be assumed to be pair-specific or calculated from the constants that pertain to single residues:

$$r_{ij}^0 = r_i^0 + r_j^0; \qquad \sigma_{ij}^0 = \sqrt{\sigma_i^{02} + \sigma_j^{02}} \qquad (5)$$

To include angular dependence, we considered three forms of anisotropic potentials derived on the basis of the Gaussian-overlap model[52]: modified Berne–Pechukas[52]; Gay–Berne,[53,54] which is used in liquid-crystal simulations[54,55]; and, finally, a potential developed by Vorobjev[56,57] for nucleic-acid simulations. The latter can be considered as a generalized form of the Gay–Berne potential. These three forms will hereafter be referred to as BP, GB, and GBV, respectively.

All these potentials assume that the interacting sites are ellipsoids of revolution. We placed the centers of the ellipsoids at the centers of mass of the side chains, the long axes being assumed to be collinear with the $C^\alpha$–SC axes. To describe the relative orientation of the ellipsoids, it is sufficient to define three angles that describe the relative orientation of their long axes: $\theta_{ij}^{(1)}$, $\theta_{ij}^{(2)}$, and $\phi_{ij}$ (Fig. 2). Although such a model of angular dependence is apparently a very simplified one, it keeps the number of orientational parameters at a reasonable minimum, which enables us to collect data with meaningful statistics from the PDB.

The expressions for all three potentials are obtained by introducing the angular dependence into $\epsilon_{ij}$ and $x_{ij}$ of the general expression given by eq. (2):

$$\epsilon_{ij} \equiv \epsilon\left(\omega_{ij}^{(1)}, \omega_{ij}^{(2)}, \omega_{ij}^{(12)}\right) = \epsilon_{ij}^0 \epsilon_{ij}^{(1)} \epsilon_{ij}^{(2)} \epsilon_{ij}^{(3)}$$

$$\epsilon_{ij}^{(1)} = \left[1 - \chi_{ij}^{(1)} \chi_{ij}^{(2)} \omega_{ij}^{(12)2}\right]^{-1/2}$$

$$\epsilon_{ij}^{(2)} = \left[1 - \frac{\chi_{ij}'^{(1)} \omega_{ij}^{(1)2} + \chi_{ij}'^{(2)} \omega_{ij}^{(2)2} - 2\chi_{ij}'^{(1)} \chi_{ij}'^{(2)} \omega_{ij}^{(1)} \omega_{ij}^{(2)} \omega_{ij}^{(12)}}{1 - \chi_{ij}'^{(1)} \chi_{ij}'^{(2)} \omega_{ij}^{(12)2}}\right]^2$$

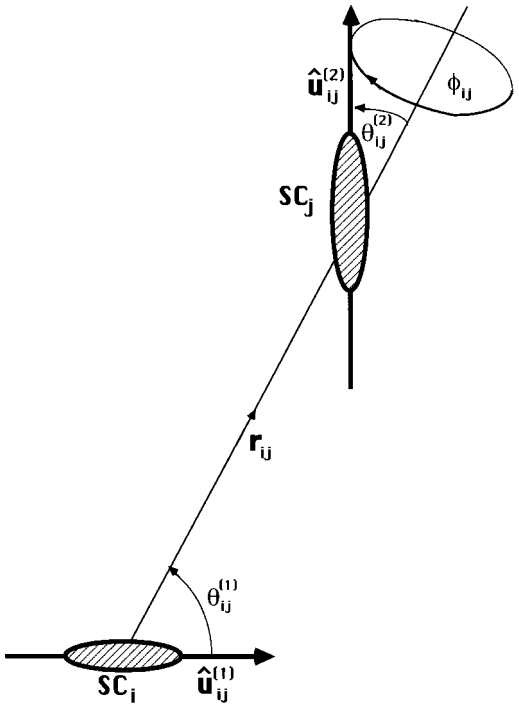


**FIGURE 2.** Definition of the orientation of two anisotropic side chains, $SC_i$ and $SC_j$, represented by ellipsoids of revolution. The relative position of the centers of the side chain is given by the vector $\mathbf{r}_{ij}$ (of length $r_{ij}$). The principal axes of the ellipsoids are assumed to be colinear with the $C^\alpha$—SC lines; their directions are given by the unit vectors $\hat{\mathbf{u}}^{(1)}$ and $\hat{\mathbf{u}}^{(2)}$. The variables defining the relative orientations of the ellipsoids are the angles $\theta_{ij}^{(1)}$ (the planar angle between $\hat{\mathbf{u}}_{ij}^{(1)}$ and $\mathbf{r}_{ij}$), $\theta_{ij}^{(2)}$ (the planar angle between $\hat{\mathbf{u}}_{ij}^{(2)}$ and $\mathbf{r}_{ij}$), and $\phi_{ij}$ (the angle of counterclockwise rotation of the vector $\hat{\mathbf{u}}_{ij}^{(2)}$ about the vector $\mathbf{r}_{ij}$ from the plane defined by $\hat{\mathbf{u}}_{ij}^{(1)}$ and $\mathbf{r}_{ij}$) when looking from the center of $SC_j$ toward the center of $SC_i$.

$$\epsilon_{ij}^{(3)} = \left[1 - \alpha_{ij}^{(1)} \omega_{ij}^{(1)} + \alpha_{ij}^{(2)} \omega_{ij}^{(2)} - 0.5\left(\alpha_{ij}^{(1)} + \alpha_{ij}^{(2)}\right) \omega_{ij}^{(12)}\right]^2 \qquad (6)$$

$$x_{ij} \equiv x\left(r_{ij}, \omega_{ij}^{(1)}, \omega_{ij}^{(2)}, \omega_{ij}^{(12)}\right) = \begin{cases} \dfrac{\sigma_{ij}}{r_{ij}} & \text{for the BP potential} \\ \dfrac{\sigma_{ij}^0}{r_{ij} - \sigma_{ij} + \sigma_{ij}^0} & \text{for the GB potential} \\ \dfrac{r_{ij}^0}{r_{ij} - \sigma_{ij} + r_{ij}^0} & \text{for the GBV potential} \end{cases} \qquad (7)$$

with:

$$\sigma_{ij} = \sigma_{ij}^0 \left[ 1 - \frac{\chi_{ij}^{(1)}\omega_{ij}^{(1)2} + \chi_{ij}^{(2)}\omega_{ij}^{(2)2} - 2\chi_{ij}^{(1)}\chi_{ij}^{(2)}\omega_{ij}^{(1)}\omega_{ij}^{(2)}\omega_{ij}^{(12)}}{1 - \chi_{ij}^{(1)}\chi_{ij}^{(2)}\omega_{ij}^{(12)2}} \right]^{-1/2} \quad (8)$$

$$\omega_{ij}^{(1)} = \hat{\mathbf{u}}_{ij}^{(1)} \cdot \hat{\mathbf{r}}_{ij} = \cos\theta_{ij}^{(1)}$$

$$\omega_{ij}^{(2)} = \hat{\mathbf{u}}_{ij}^{(2)} \cdot \hat{\mathbf{r}}_{ij} = \cos\theta_{ij}^{(2)}$$

$$\omega_{ij}^{(12)} = \hat{\mathbf{u}}_{ij}^{(1)} \cdot \hat{\mathbf{u}}_{ij}^{(2)} = \cos\theta_{ij}^{(1)} \cos\theta_{ij}^{(2)} + \sin\theta_{ij}^{(1)} \sin\theta_{ij}^{(2)} \cos\phi_{ij}$$

where $\hat{\mathbf{u}}_{ij}^{(1)}$ and $\hat{\mathbf{u}}_{ij}^{(2)}$ are unit vectors along the principal axes of the interacting sites (in this work identified with the $C^\alpha$–SC axes), $\mathbf{r}_{ij}$ is the vector linking the centers of the sites, $\hat{\mathbf{r}}_{ij}$ is the corresponding unit vector, $r_{ij}$ is the distance between the side-chain centers (Fig. 2), the constants $\chi_{ij}^{(1)}$ and $\chi_{ij}^{(2)}$ are the anisotropies of the van der Waals radius, and the constants $\chi_{ij}^{\prime(1)}$ and $\chi_{ij}^{\prime(2)}$ are the anisotropies of the van der Waals well depth.

The angular dependence of $\epsilon_{ij}^{(1)}$ and $x_{ij}$ arises from the extension of the Gaussian overlap potential to the LJ-type function.[53] Additional dependence of the van der Waals well depth on orientation in the form of $\epsilon_{ij}^{(2)}$ has been introduced by GB.[53] For the original BP potential, $\epsilon_{ij}^{(2)} = 1$, but we keep its orientational dependence to preserve the same form of the potential. The formulas are generalized in this work to the case of ellipsoids of revolution with different axes (the BP and GB potentials were originally derived for the interaction of identical ellipsoidal bodies). Finally, the function $\epsilon_{ij}^{(3)}$ with the constants $\alpha_{ij}^{(1)}$ and $\alpha_{ij}^{(2)}$ has been introduced in this work to account for the lower symmetry of the angular distribution functions observed in protein crystals than that implied by the three potentials outlined previously. Squaring in the expressions for $\epsilon_{ij}^{(2)}$ and $\epsilon_{ij}^{(3)}$ is done to keep them non-negative.

As in the case of the radial potentials, the constants $\sigma_{ij}^0$, $r_{ij}^0$, $\chi_{ij}^{(1)}$, $\chi_{ij}^{(2)}$, $\chi_{ij}^{\prime(1)}$, $\chi_{ij}^{\prime(2)}$, $\alpha_{ij}^{(1)}$, and $\alpha_{ij}^{(2)}$ can be assumed to be pair-dependent or constrained to be calculated from single-residue constants. In this work, we tried both procedures. For the constants $r_{ij}^0$ and $\sigma_{ij}^0$, the formulas are given by eq. (5). For the case of different ellipsoids, the anisotropies of the van der Waals distances can be expressed by eq. (9):

$$\chi_{ij}^{(1)} = \frac{\sigma_i^{\parallel 2} - \sigma_i^{\perp 2}}{\sigma_i^{\parallel 2} + \sigma_j^{\perp 2}}; \quad \chi_{ij}^{(2)} = \frac{\sigma_j^{\parallel 2} - \sigma_j^{\perp 2}}{\sigma_j^{\parallel 2} + \sigma_i^{\perp 2}} \quad (9)$$

Further, for the Gaussian overlap model, it follows[52] that $\sigma_i^\perp = \sigma_i^0$ and $\sigma_j^\perp = \sigma_j^0$. The constants $\sigma^\perp$ and $\sigma^\parallel$ can be identified with the lengths of the short and long axes of the ellipsoids, respectively. Our variable parameters were $\sigma^0$ and the ratio $(\sigma^\parallel/\sigma^\perp)^2$; the first one can be considered as a measure of the size, and the second of the anisotropy of a side chain.

The same type of dependence can be assumed for the constants $\chi'$, but then there would be too many parameters to be determined. Therefore, we assumed that the constants $\chi'$ and $\alpha$ depend on single-residue types, namely:

$$\chi_{ij}^{\prime(1)} = \chi_i', \quad \chi_{ij}^{\prime(2)} = \chi_j' \qquad \alpha_{ij}^{(1)} = \alpha_i, \quad \alpha_{ij}^{(2)} = \alpha_j \quad (10)$$

It should be noted that, in the case of isotropic interactions, the GB and BP potentials become the LJ potential, whereas the GBV potential becomes the LJK potential.

## PARAMETERIZING SIDE-CHAIN INTERACTION POTENTIALS

Similar to earlier work,[24–26] we determine the parameters of the potentials introduced in the preceding section by fitting them to correlation functions and contact free energies calculated from protein-crystal data. In doing this, we make the following two assumptions:

1. The correlation functions obtained by using a sufficiently large number of protein crystal data (each of which corresponds to a system at a free-energy minimum) are sufficiently good approximations to the correlation functions of a hypothetical ''stochastic'' mixture of nonconnected side chains. This approximation is justified by the observation that, although a crystal structure is at equilibrium as the whole structure, its individual parts can be forced to assume geometries far from locally equilibrated, locally lower energy conformations having, however, higher probability of occurrence in the whole structure.[58] For example, the distributions of $X$—H bond lengths obtained from large data bases of

crystal structures are qualitatively similar to those calculated from potential-energy surfaces of proton transfer.[58]

2. Interactions between the side chains can be described with sufficient accuracy by using the potential of mean force, $W_{ij}(r_{ij}, \theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \phi_{ij})$, related directly to the corresponding side-chain pair correlation functions, $g_{ij}(r_{ij}, \theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \phi_{ij})$:

$$g_{ij}\left(r_{ij}, \theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \phi_{ij}\right) = \exp\left[-W_{ij}\left(r_{ij}, \theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \phi_{ij}\right)/RT\right] \quad (11)$$

where $R$ and $T$ are the gas constant and the absolute temperature, respectively. According to point 1, the reference state in eq. (11) corresponds to a hypothetical polypeptide chain with noninteracting side chains (the unfolded state according to the classification of Godzik et al.[59]).

Because we want to exclude the effects of local interactions [since the local interactions are included in the terms $U_b(\theta)$, $U_{tor}(\gamma)$, and $U_{rot}(\alpha_{SC}, \beta_{SC})$ of eq. (1)], we consider only the side chains that are separated by at least ten peptide groups. This also makes it legitimate to disregard the direction of the chain separating the residues; therefore, we assume that $W_{i_k, j_l} = W_{i_l, j_k}$, where $i_k$ denotes a residue of type $i$ occupying the $k$th position in the chain.

To avoid the influence of many-body and boundary effects on the distributions at large distances, we confine our treatment to a short-range distance limit $r \le r_{ij}^{max}$, and we express the potential of mean force by one of the analytical forms given by eqs. (2)–(8). The upper distance limits $r_{ij}^{max}$ are defined so that only the regions of the first peak of the correlation functions are considered, in which the potential of mean force is unimodal in $r$:

$$r_{ij}^{max} = \min\left\{0.5\left(r_i^{0;L} + r_j^{0;L}\right) + 2\,\text{Å}, 8\,\text{Å}\right\} \quad (12)$$

where $r^{0;L}$ are the mean side-chain van der Waals radii for each of the 20 naturally occurring amino acids calculated by Levitt.[2]

For the LJ and LJK potentials [eqs. (3) and (4)], which depend only on the distance between the side chains, $g_{ij}(r_{ij}, \theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \phi_{ij})$ of eq. (11) is replaced by the radial-only correlation function, $g_{ij}(r_{ij})$ [which is equivalent to $g_{ij}(r_{ij}, \theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \phi_{ij})$ averaged over the angles $\theta_{ij}^{(1)}$, $\theta_{ij}^{(2)}$, and $\phi_{ij}$]. Thus, the potentials of mean force and, in turn, the correlation functions depend parametrically on the constants appearing in eqs. (3)–(8), which can be optimized so that the theoretical correlation functions given by eq. (11) fit best (in the least-squares sense) to the correlation functions determined from protein crystal data.

The limited number of data that we are able to collect from protein crystals prohibits the direct use of the full radial-and-angular correlation function $g_{ij}(r_{ij}, \theta_{ij}^{(1)}, \theta_{ij}^{(2)}, \phi_{ij})$. Therefore, our target function for parameter estimation includes correlation functions that are averaged over some of the variables, and side-chain contact free energies. The side-chain contact free energies are the logarithms of the correlation functions averaged over the coordination sphere of the interacting side chains. To determine the parameters of the potentials, we minimized the weighted sum of the squares $\Phi(\mathbf{X})$ of the differences between histograms of the radial ($H_{ij;k}^{r}$), radial-angular ($H_{ij;klm}^{r\theta}$), and angular ($H_{ij;klm}^{\theta\phi}$) correlation functions, as well as contact free energies ($F_{ij}$), calculated as functions of the parameters, and determined from protein-crystal data, respectively:

$$\begin{aligned}\Phi(\mathbf{X}) = \sum_{i=1}^{20}\sum_{j=1}^{20} w_{ij}\Bigg\{ & w^{r}\sum_{k=1}^{nr_{ij}}\left[H_{ij;k}^{r}(\mathbf{X}) - \hat{H}_{ij;k}^{r}\right]^2 \\ & + w^{\theta\phi}\sum_{k=1}^{n_\theta}\sum_{l=1}^{n_\theta}\sum_{m=1}^{n_\phi}\left[H_{ij;klm}^{\theta\phi}(\mathbf{X}) - \hat{H}_{ij;klm}^{\theta\phi}\right]^2 \\ & + w^{r\theta}\sum_{k=1}^{nr_{ij}}\sum_{l=1}^{n_\theta}\sum_{m=1}^{n_\theta}\left[H_{ij;klm}^{r\theta}(\mathbf{X}) - \hat{H}_{ij;klm}^{r\theta}\right]^2 \\ & + w^{F}\left[F_{ij}(\mathbf{X}) - \hat{F}_{ij}\right]^2\Bigg\} \\ = {} & w^{r}\Phi^{r}(\mathbf{X}) + w^{\theta\phi}\Phi^{\theta\phi}(\mathbf{X}) \\ & + w^{r\theta}\Phi^{r\theta}(\mathbf{X}) + w^{F}\Phi^{F}(\mathbf{X})\end{aligned} \quad (13)$$

where the indices $i$, $j$ run over all 20 naturally occurring amino acids. Also:

$$w_{ij} = \sum_{p=1}^{np} w_p n_{ij} \Big/ \sum_{i\le j}\sum_{p=1}^{np} w_p n_{ij;p} \quad (14)$$

is the statistical weight of the pair of side chains of types $i$ and $j$ ($w_p$ and $np$ being the statistical weight of the $p$th protein and the total number of protein structures in the sample, respectively, and $n_{ij}$ and $n_{ij;p}$ being the total number of pairs of residues of types $i$ and $j$ and the number of such

pairs in protein $p$, respectively); $nr_{ij}$ is the number of distance values considered for a pair of side chains of types $i$ and $j$; $n_\theta$ and $n_\phi$ are the numbers of values of the angles $\theta^{(1)}$ or $\theta^{(2)}$; and $\phi$, $w^{\theta\phi}$, $w^r$, $w^{r\theta}$, and $w^F$ are weights of the histograms and of the free energy, respectively; $\mathbf{X}$ is a shorthand for the parameters of the target potential, and a ''hat'' (circumflex) over a quantity designates the value determined from the crystal data. For the radial-only potentials, LJ and LJK, the angular and radial-angular components $\Phi^{\theta\phi}(\mathbf{X})$ and $\Phi^{r\theta}(\mathbf{X})$ do not appear in the expression for $\Phi(\mathbf{X})$.

The histograms of the correlation functions $H^r_{ij;k}$, $H^{r\theta}_{ij;klm}$, and $H^{\theta\phi}_{ij;klm}$ are defined by:

$$H^r_{ij;k} = \frac{\bar{g}^r_{ij}(r_k)}{\sum_{k=1}^{nr_{ij}} \bar{g}^r_{ij}(r_k)}$$

$$H^{\theta\phi}_{ij;klm} = \frac{\bar{g}^{\theta\phi}_{ij}(\theta^{(1)}_k, \theta^{(2)}_l, \phi_m)\Delta\cos\theta^{(1)}_k \Delta\cos\theta^{(2)}_l}{\sum_{k=1}^{n_\theta}\sum_{l=1}^{n_\theta}\sum_{m=1}^{n_\phi} \bar{g}^{\theta\phi}_{ij}(\theta^{(1)}_k, \theta^{(2)}_l, \phi_m)\Delta\cos\theta^{(1)}_k \Delta\cos\theta^{(2)}_l}$$

$$H^{r\theta}_{ij;klm} = \frac{\bar{g}^{r\theta}_{ij}(r_k, \theta^{(1)}_l, \theta^{(2)}_m)\Delta\cos\theta^{(1)}_l \Delta\cos\theta^{(2)}_m}{\sum_{k=1}^{nr_{ij}}\sum_{l=1}^{n_\theta}\sum_{m=1}^{n_\theta} \bar{g}^{r\theta}_{ij}(r_k, \theta^{(1)}_l, \theta^{(2)}_m)\Delta\cos\theta^{(1)}_l \Delta\cos\theta^{(2)}_m} \quad (15)$$

where $\bar{g}^r_{ij}(r_k)$, $\bar{g}^{\theta\phi}_{ij}(\theta^{(1)}_k, \theta^{(2)}_l, \phi_m)$, and $\bar{g}^{r\theta}_{ij}(r_k, \theta^{(1)}_l, \theta^{(2)}_m)$ are the average values of the radial, angular and radial-angular correlation functions within bins defined as $[r_k - \Delta r/2, r_k + \Delta r/2]$, $[\theta^{(1)}_k - \Delta\theta/2, \theta^{(1)}_k + \Delta\theta/2] \times [\theta^{(2)}_l - \Delta\theta/2, \theta^{(2)}_l + \Delta\theta/2] \times [\phi_m - \Delta\phi/2, \phi_m + \Delta\phi]$, and $[r_k - \Delta r/2, r_k + \Delta r/2] \times [\theta^{(1)}_l - \Delta\theta/2, \theta^{(1)}_l + \Delta\theta/2] \times [\theta^{(2)}_m - \Delta\theta/2, \theta^{(2)}_m + \Delta\theta/2]$, respectively, with $\Delta r$, $\Delta\theta$, and $\Delta\phi$ being the dimensions of the bins. The definitions and method of calculation of the correlation functions from protein crystal data are given in the Appendix.

The purpose of the introduction of the factors $\Delta\cos\theta^{(1)}_k \Delta\cos\theta^{(2)}_l$ into the angular and radial-angular terms is to avoid overweighting the regions around $\theta^{(1)}$ or $\theta^{(2)} = 0°$ or 180° in which the number of counts is very small, which results in poor accuracy there of the angular-distribution functions.

The free energies of contact interaction were calculated from protein crystal data using the quasichemical approximation procedure developed by Miyazawa and Jernigan.[9] We chose this approach because it takes into account the fact that competitive interactions between residues of different types occur in real proteins. We chose a radius of 8 Å for the coordination sphere, which is greater than the 6.5-Å radius used by Miyazawa and Jernigan. The reason for this was that not many hydrophilic contacts are present within the 6.5 Å coordination sphere which results in poorer statistics. Then, we scaled these free energies to be compatible with free energies of transfer of amino-acid side chains from $n$-octanol to water determined by Fauchere and Pliška[60]; this will be described in the subsection ''Contact Free Energies'' of the ''Results'' section. The corresponding ''theoretical'' values of the free energies are calculated from:

$$F_{ij}(\mathbf{X}) = -RT \times \ln\left[(1/V^c_{ij})\int_{\Omega^c_{ij}} g_{ij}(\varrho, \vartheta^{(1)}, \vartheta^{(2)}, \varphi)\, dV\right] \quad (16)$$

where $\Omega^c_{ij} = S(0, R_c) - S(0, r^c_i + r^c_j)$ is the allowed coordination sphere corresponding to pair $ij$ [$S(\mathbf{a}, r)$ denoting the region of space bounded by a sphere of radius $r$ centered at point $\mathbf{a}$], $V^c_{ij}$ is the volume of $\Omega^c_{ij}$, $R_c$ is the maximum radius of the coordination sphere (assumed to be 8 Å in this work), $r^c_i$ and $r^c_j$ are the contact radii of the side chains calculated from their volumes (see Table III in ref. 9).

Because the relative weights of the angular, radial-angular, radial, and contact-free-energy terms in eq. (13) were not known *a priori*, estimation of the parameters by minimization of expression (13) is a generalized least-square problem.[61] In such a case, the weights are usually estimated as the inverses of the squares of the residuals, and the resulting sum of squares is minimized successively with iteratively updated weights, until the calculated and the assumed weights are consistent. Thus, the estimates $\tilde{w}^r$, $\tilde{w}^{r\theta}$, $\tilde{w}^{\theta\phi}$, and $\tilde{w}^F$ of the weights of the terms in eq. (13) can be expressed

by:

$$\tilde{w}^r = 1/\bar{\sigma}_r^2 = \left\{(1/\Sigma w)\sum_{i=1}^{20}\sum_{j=i}^{20} w_{ij}(1/nr_{ij}) \times \sum_{k=1}^{nr_{ij}}\left[H_{ij;k}^r(\mathbf{X}) - \hat{H}_{ij;k}^r\right]^2\right\}^{-1}$$

$$\tilde{w}^{\theta\phi} = 1/\bar{\sigma}_{\theta\phi}^2\left\{\left(1/\Sigma w n_\theta^2 n_\phi\right)\sum_{i=1}^{20}\sum_{j=i}^{20} w_{ij} \times \sum_{k=1}^{n_\theta}\sum_{l=1}^{n_\theta}\sum_{m=1}^{n_\phi}\left[H_{ij;klm}^{\theta\phi}(\mathbf{X}) - \hat{H}_{ij;klm}^{\theta\phi}\right]^2\right\}^{-1}$$

$$\tilde{w}^{r\theta} = 1/\bar{\sigma}_{r\theta}^2 = \left\{(1/\Sigma w n_\theta^2)\sum_{i=1}^{20}\sum_{j=i}^{20} w_{ij}(1/nr_{ij}) \times \sum_{k=1}^{nr_{ij}}\sum_{l=1}^{n_\theta}\sum_{m=1}^{n_\theta}\left[H_{ij;klm}^{r\theta}(\mathbf{X}) - \hat{H}_{ij;klm}^{r\theta}\right]^2\right\}^{-1}$$

$$\tilde{w}^F = 1/\bar{\sigma}_F^2 = \left\{(1/\Sigma w)\sum_{i=1}^{20}\sum_{j=i}^{20} w_{ij}\left[F_{ij}(\mathbf{X}) - \hat{F}_{ij}\right]^2\right\}^{-1} \quad (17)$$

where $\bar{\sigma}^2$ is a variance of the corresponding quantity, $\Sigma w = \sum_{i=1}^{20}\sum_{j=i}^{20} w_{ij}$, and the other symbols are defined in eq. (13).

The standard deviations of the parameters were estimated according to the Gauss–Markov formula[62]:

$$[\sigma(x_i)]^2 = \frac{\Phi(\mathbf{X}^*)}{N-p}[\mathbf{J}^T(\mathbf{X}^*)\mathbf{W}(\mathbf{X}^*)\mathbf{J}(\mathbf{X}^*)]_{ii}^{-1} \quad (18)$$

where $N = \{210n_\theta^2 n_\phi + (n_\theta^2 n_\phi + 1)\ \sum_{i=1}^{20}\sum_{j=i}^{20} nr_{ij} + 210\}$ is the total number of terms in eq. (13), $p$ is the total number of parameters, $J_{ij} = \partial\delta_i/\partial x_j$ is the element of the first derivatives of the residuals $\delta_1, \delta_2, \ldots, \delta_N$ that occur in the sum of the squares; $\mathbf{W}$ is the corresponding matrix of weights, and $\mathbf{X}^*$ denotes the vector of the parameters at the minimum.

### SELECTION OF PROTEIN STRUCTURES

The protein crystal structures were taken from the Brookhaven Protein Data Bank. First, the list of available structures obtained from the PDB server pdb.pdb.bnl.gov (as of June 25, 1994) was scanned and, those with a resolution not exceeding 2 Å and having a chain length of at least 100 amino-acid residues, were selected. Then, the percentage of sequence homology was calculated for all pairs of sequences using the FASTA program[63,64] available on anonymous ftp from uvaarpa.virginia.edu. Then, cluster analysis was carried out with the minimal-tree algorithm,[65] taking the values of (100% − percentage homology) as distances between pair of structures. This grouped the selected proteins into 154 families of homologous structures. From each family, those structures were taken that had the highest resolution or, if the resolution was the same, the longest chain(s). In several cases, however, both criteria were satisfied by more than one structure. In such a case, we took all the structures satisfying the criteria, diminishing their statistical weights when calculating the histograms of pair-correlation functions and contact free energies. The final list contained 195 structures, whose identities are summarized in Table I of the Supplementary Material.

## Results and Discussion

### DISTRIBUTION AND CORRELATION FUNCTIONS

In all calculations, we assumed that the centers of interactions are in the geometric centers of the side chains, calculated from the coordinates of the nonhydrogen atoms, including $C^\alpha$, as expressed by:

$$\mathbf{R}_i = \frac{1}{NH_i}\sum_{j=1}^{NH_i}\mathbf{r}_{ji} \quad (19)$$

where $\mathbf{R}_i$ represents the coordinates of the geometric center of the $i$th side chain, $\mathbf{r}_{ji}$ represents the coordinates of the $i$th nonhydrogen atom of the $i$th side chain, and $NH_i$ is the number of nonhydrogen atoms in side-chain $i$. The index $i$ denotes an *individual* side chain in the data base and not the side-chain type. For glycine the position of the side-chain atom coincides with the position of $C^\alpha$.

When calculating the pair distributions and contact free energies, we excluded disulfide-bonded cystine pairs; however, the nonbonded cysteine pairs were included. The weights of the structures were calculated from the following formula:

$$w_p = \frac{1}{n_{chain}n_{hom}res^2} \quad (20)$$

where $n_{chain}$ is the number of equivalent chains in a protein, $n_{hom}$ is the number of homologous structures, if more than one was taken from a family, and *res* is the crystallographic resolution. The weights, $w_p$, appear in equations in the Appendix.

The single-body densities of the amino-acid side chains were calculated from eq. (A-12) of the Appendix, and were subsequently used in the evaluation of the factor $T_{ij}$ used for the calculation of reference (i.e., in the absence of any side-chain interactions) radial and radial-angular probability distributions [eqs. (A-7) and (A-9) of the Appendix]. A sample collection of radial pair densities together with the reference and total pair distribution functions is shown for the Leu–Leu pair in Figure 3. The radial distribution (curve C of Fig. 3) qualitatively resembles that calculated theoretically by Gan and Eu[66] in their study of model van der Waals polymer chains. As shown, the distribution calculated from single-body density and the Markovian factor (curve B of Fig. 3) approximates quite well the Leu–Leu pair distribution function (curve C of Fig. 3) for distances longer than 10 Å. The correlation function (curve A of Fig. 3) at distances longer than 10 Å is almost constant. Greater deviations occur only at very long distances; this can be explained by the fact that the single-body density is determined with poor accuracy at longer distances.

To calculate the reference angular distribution functions, we averaged the computed angular distribution functions over all pairs of side chains, using the method of Hao et al.[67] (see eq. (A-13) and the following text in the Appendix for details).

However, the angular pair correlation functions, calculated from eq. (A-5), still exhibited the behavior of the ''background'' correlation function (solid curve of Fig. 4) for many pairs of residues. By least-squares fitting, we found that the ''background'' angular pair correlation function can be described by $\epsilon^{(3)}$ of eq. (6). Therefore, we included $\epsilon^{(3)}$ in the angular potentials.

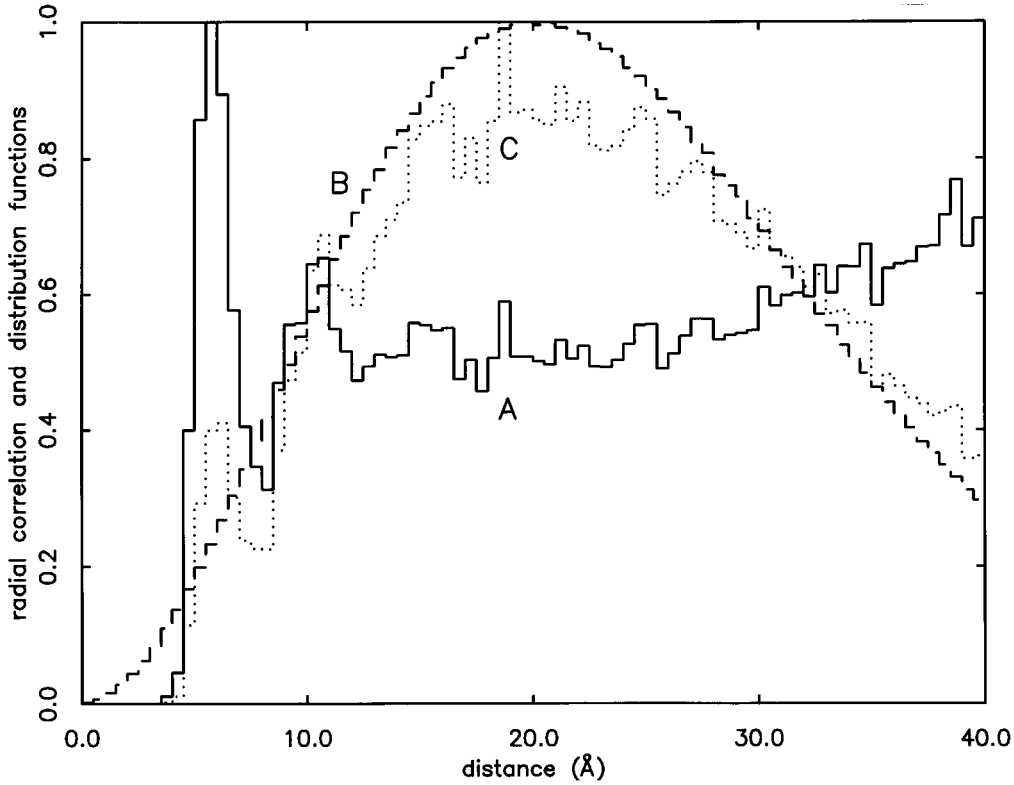


**FIGURE 3.** Sample pair-distribution and pair-correlation functions for the Leu–Leu pair averaged over consecutive 0.5-Å shells. (A) Radial pair correlation functions $g_{ij}^{r}$; (B) the reference pair distribution function $\nu^{(2,0,r)}$ [denominator in eq. (A-4)]; and (C) the total pair distribution function $\nu^{(2,r)}$ [numerator in eq. (A-4)]. All graphs were normalized to the maximum values of 1.0.

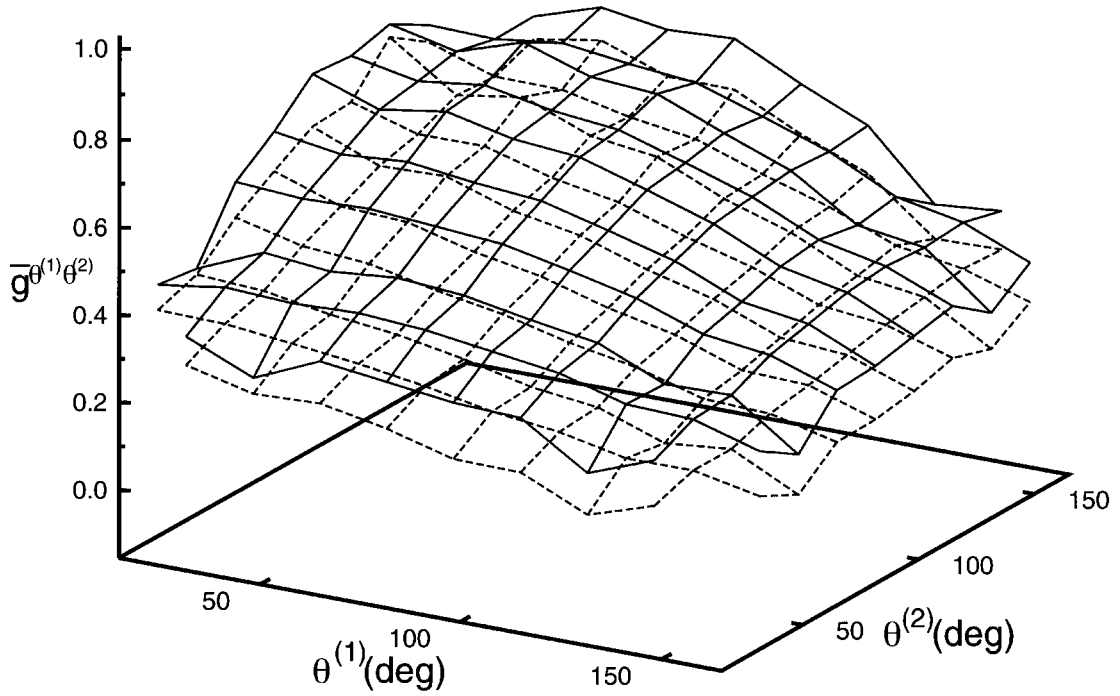


**FIGURE 4.** The angular correlation functions $\bar{g}^{\theta\phi}(\theta^{(1)}, \theta^{(2)}, \phi)$ averaged over all pairs of side-chain types and over the angle $\phi$ for displaying purpose. The surfaces were normalized so that 1 is the maximum value for both. Solid surface: the function obtained from the PDB averaged over all the pairs of side chains. Dashed surface: the function obtained in simulation studies. The latter were carried out by generating a total of 1000 50-residue energy-minimized chains with random sequence confined to the ellipsoid characteristic of proteins of this size, according to the method of Hao et al.,[66] with the united-residue representation of polypeptide chains and the energy function developed in our earlier work[38,39]; that is, the side-chain interaction potential did not have explicit angular dependence. The simulation study shows that the background angular pair correlation function is not constant, even for a radial side-chain interaction potential.

## CONTACT FREE ENERGIES

The calculated contact free energies, together with the numbers of contacts, are summarized in Table I.

Even though the set of 195 structures (see Table 1 of the Supplementary Material) has almost no structure in common with that used by Miyazawa and Jernigan (MJ),[9] the computed contact free energies correlate well with those determined by MJ, the correlation equation being:

$$e_{ij}^{MJ} = 1.580(0.028)e_{ij}\left(R_c = 8\ \text{Å}\right) + 2.135(0.093); \quad R = 0.9689 \quad (21)$$

where $R$ denotes the correlation coefficient, and the numbers in parentheses the standard deviation of the slope and intercept, respectively. When the radius of the coordination sphere was taken as 6.5 Å (as used by MJ), and the contact free energies re-evaluated, the correlation equation became:

$$e_{ij}^{MJ} = 0.844(0.026)e_{ij}\left(R_c = 6.5\ \text{Å}\right) + 0.45(0.11); \quad R = 0.9159 \quad (22)$$

Thus, the slope from eq. (22) is closer to 1.0, as expected.

The correlation coefficient of our contact free energies with the contact free energies derived by Tanaka and Scheraga[8] with a smaller data base, with the definition of contact similar to that used in our work except that their distances were measured between the α-carbons, is 0.8670. By contrast, the correlation of our contact free energies with those determined by Koliński et al.[14] or Gregoret and Cohen[10] is quite weak, the correlation coefficients being 0.6019 and 0.2261, respectively. This is understandable because the reference state in the latter two potentials corresponds to side chains arranged in a hydrophobic core and a hydrophilic exterior (the $U_{phil;phob}$ state according to

**TABLE I.**

**Contact Free Energies (*RT* Units; Diagonal and Upper Triangle) and Total Number of Counts Within the 8-Å Coordination Sphere (Lower Triangle, and the Last Line for Diagonal Elements) for Pairs of Amino-Acid Residues.**

| | Cys | Met | Phe | Ile | Leu | Val | Trp | Tyr | Ala | Gly | Thr | Ser | Gln | Asn | Glu | Asp | His | Arg | Lys | Pro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cys | −4.54 | −4.72 | −4.94 | −4.90 | −4.72 | −4.53 | −4.60 | −4.08 | −3.93 | −3.68 | −3.73 | −3.56 | −3.31 | −3.14 | −2.97 | −3.12 | −3.87 | −3.02 | −2.71 | −3.50 |
| Met | 198 | −4.80 | −5.03 | −4.95 | −4.91 | −4.59 | −4.87 | −4.30 | −3.93 | −3.40 | −3.62 | −3.35 | −3.24 | −3.04 | −2.90 | −2.80 | −3.80 | −3.15 | −2.58 | −3.43 |
| Phe | 476 | 735 | −5.22 | −5.25 | −5.18 | −4.89 | −5.07 | −4.43 | −4.04 | −3.52 | −3.70 | −3.46 | −3.25 | −3.17 | −2.89 | −2.95 | −3.95 | −3.26 | −2.70 | −3.54 |
| Ile | 541 | 817 | 1920 | −5.29 | −5.21 | −5.01 | −4.98 | −4.51 | −4.18 | −3.61 | −3.87 | −3.58 | −3.30 | −3.13 | −3.11 | −3.09 | −3.74 | −3.42 | −2.95 | −3.62 |
| Leu | 739 | 1365 | 3123 | 3816 | −5.03 | −4.85 | −4.87 | −4.32 | −4.06 | −3.56 | −3.61 | −3.47 | −3.17 | −3.09 | −2.85 | −2.84 | −3.69 | −3.24 | −2.74 | −3.55 |
| Val | 598 | 937 | 2365 | 3175 | 4731 | −4.64 | −4.63 | −4.04 | −3.93 | −3.42 | −3.58 | −3.35 | −3.05 | −3.02 | −2.84 | −2.82 | −3.44 | −3.00 | −2.69 | −3.39 |
| Trp | 131 | 240 | 549 | 601 | 924 | 741 | −4.74 | −4.20 | −3.81 | −3.47 | −3.44 | −3.30 | −3.25 | −3.23 | −3.04 | −3.12 | −3.94 | −3.42 | −2.84 | −3.60 |
| Tyr | 283 | 464 | 962 | 1227 | 1753 | 1327 | 366 | −3.69 | −3.41 | −3.12 | −3.16 | −2.97 | −2.90 | −2.82 | −2.71 | −2.84 | −3.42 | −3.01 | −2.58 | −3.23 |
| Ala | 679 | 1105 | 2071 | 2889 | 4762 | 4279 | 639 | 1445 | −3.28 | −2.91 | −2.97 | −2.76 | −2.67 | −2.59 | −2.45 | −2.52 | −2.95 | −2.56 | −2.28 | −2.81 |
| Gly | 788 | 750 | 1391 | 1878 | 3126 | 2944 | 561 | 1273 | 3750 | −2.62 | −2.78 | −2.61 | −2.39 | −2.41 | −2.12 | −2.34 | −2.77 | −2.43 | −2.08 | −2.60 |
| Thr | 446 | 612 | 1220 | 1644 | 2140 | 2225 | 348 | 877 | 2543 | 2438 | −2.81 | −2.74 | −2.46 | −2.43 | −2.33 | −2.43 | −2.97 | −2.55 | −2.13 | −2.68 |
| Ser | 550 | 579 | 1068 | 1543 | 2335 | 2131 | 381 | 927 | 2379 | 2344 | 1737 | −2.47 | −2.34 | −2.36 | −2.22 | −2.31 | −2.76 | −2.46 | −2.01 | −2.57 |
| Gln | 251 | 329 | 472 | 644 | 1056 | 899 | 219 | 475 | 1319 | 1146 | 801 | 840 | −1.94 | −2.26 | −1.93 | −2.09 | −2.48 | −2.25 | −1.84 | −2.35 |
| Asn | 250 | 290 | 659 | 766 | 1215 | 1238 | 295 | 631 | 1598 | 1566 | 1093 | 1172 | 598 | −2.24 | −2.14 | −2.21 | −2.60 | −2.19 | −1.95 | −2.28 |
| Glu | 281 | 392 | 673 | 1071 | 1356 | 1354 | 281 | 679 | 1957 | 1447 | 1241 | 1407 | 604 | 919 | −1.60 | −1.77 | −2.49 | −2.61 | −2.20 | −2.09 |
| Asp | 308 | 343 | 718 | 1004 | 1238 | 1241 | 336 | 813 | 2096 | 1910 | 1410 | 1389 | 675 | 1103 | 867 | −1.85 | −2.70 | −2.67 | −2.22 | −2.16 |
| His | 218 | 287 | 597 | 556 | 939 | 708 | 238 | 469 | 868 | 893 | 689 | 618 | 289 | 443 | 555 | 721 | −3.27 | −2.60 | −2.02 | −2.69 |
| Arg | 203 | 312 | 561 | 831 | 1252 | 932 | 269 | 603 | 1312 | 1343 | 953 | 1043 | 538 | 637 | 1308 | 1394 | 375 | −2.03 | −1.47 | −2.29 |
| Lys | 262 | 394 | 705 | 1122 | 1481 | 1595 | 355 | 782 | 2077 | 1758 | 1324 | 1385 | 609 | 995 | 1721 | 1778 | 483 | 542 | −1.14 | −1.95 |
| Pro | 305 | 391 | 771 | 1000 | 1628 | 1371 | 354 | 743 | 1604 | 1507 | 1100 | 1143 | 561 | 763 | 821 | 803 | 434 | 600 | 865 | −2.53 |
| | 260 | 216 | 944 | 1237 | 3213 | 2185 | 97 | 360 | 2534 | 1793 | 842 | 971 | 223 | 452 | 421 | 504 | 261 | 270 | 443 | 396 |

the terminology of Godzik et al.[59]), rather than to a completely unarranged polypeptide chain (the *U* state[59]), which is the reference state in the Tanaka–Scheraga,[8] MJ,[9] and our approach. In the Koliński–Skolnick[14] and Gregoret–Cohen[10] potentials, the nonspecific grouping of side chains into the hydrophobic core and hydrophilic exterior is accounted for by one-body centrosymmetric potentials, whereas in our approach it is encoded in the side-chain pair potentials.

According to Miyazawa and Jernigan,[9] the quantities $0.5q_ie_i$, where $q_i$ is the coordination number of residue of type $i$ and $e_i = \sum_{i=1}^{20} N_{ij}e_{ij}/\sum_{i=1}^{20} N_{ij}$ is the average contact free energy of residue of type $i$, can be regarded as hydrophobicities of the corresponding types of residues. Therefore, we correlated these quantities with side-chain hydrophobicities determined by Fauchere and Pliška who measured the partition coefficients of amino acids between *n*-octanol and water,[60] and obtained the following correlation equation:

$$-RT \times (0.5q_ie_i) = 1.60(0.15) \times [RT\ln(10)\pi_i] - 10.50(0.23); \quad R = 0.9278 \quad (23)$$

where $T = 298$ K and $\pi_i$ is the contribution of the side chain of type $i$ to the logarithm of the partition coefficient between *n*-octanol and water, as determined by Fauchere and Pliška.[60] The correlation graph is shown in Figure 5.

Similar correlation also holds with the diagonal contact free energies, $e_{ii}$:

$$-RTe_{ii} = 0.528(0.046) \times [RT\ln(10)\pi_i] - 1.197(0.070); \quad R = 0.9372 \quad (24)$$

The correlation with other hydrophobicity scales derived on the basis of thermodynamic data, for example, those of Nozaki and Tanford,[68] is worse (with $R = 0.8019$). The correlation is also worse ($R = 0.8518$) when the contact free energies obtained with $R_c = 6.5$ Å are used. The latter fact is understandable in view of the fewer number of contacts and therefore poorer statistics, especially for hydrophilic pairs.

The slopes of eqs. (23) and (24) were used to estimate the ''true'' free energies of contacts implemented in the sum of squares given by eq. (13). As in our earlier work,[38,39] we considered the residue contact free energies to be composed of the parts due to side-chain–side-chain and backbone–backbone interaction. To obtain the peptide-group–peptide-group interaction free energy for any amino acid, we subtract from $e_{GlyGly}$ (obtained from the PDB) the contribution of the $CH_2$ group of Gly. We take the latter as $-0.528RT\ln(10)\pi_{CH_2}$ from eq. (24), with $\pi_{CH_2} = 0.41$ (Ref. 60), which can be considered as an estimate of the glycine ''side-chain–side-chain'' contact free energy. Then, we rescale our contact free energies by introduction of the factor 1.60 of eq. (23) for nonproline residues. Further, because Pro has no backbone NH donor group, we have to reduce the corresponding estimate of the peptide-group–peptide-group interaction free energy by a factor[39] $f_{Pro}$ or $f_{ProPro}$. Thus, the estimates of experimental contact free energies can be expressed by eq. (25):

$$\hat{F}_{ij} = \frac{RT}{1.60} \times \begin{cases} e_{ij} - \left(e_{GlyGly} + 0.528\ln(10)\pi_{CH_2}\right), & \text{if both } i \text{ and } j \neq \text{Pro} \\ e_{ij} - f_{Pro}\left(e_{GlyGly} + 0.528\ln(10)\pi_{CH_2}\right), & \text{if only one of } i \text{ or } j = \text{Pro} \\ e_{ij} - f_{ProPro}\left(e_{GlyGly} + 0.528\ln(10)\pi_{CH_2}\right), & \text{if both } i \text{ and } j = \text{Pro} \end{cases} \quad (25)$$

Finally, it should be noted that the computed contact free energies are additive to a good approximation, which is reflected in the following correlation equation:

$$e_{ij} = 1.050(0.020)(e_{ii} + e_{jj})/2 + 0.072(0.068); \quad R = 0.9669 \quad (26)$$

in which the slope and intercept do not differ significantly from 1 and 0, respectively. The quantity $(e_{ii} + e_{jj})/2$ is called the *ideal* pair-interaction free energy, while the quantity $e_{ij} - (e_{ii} + e_{jj})/2$ is called the *excess* pair-interaction free energy.[59] Equation (26), together with eq. (24) can serve to estimate the interresidue contact free energies of non-natural amino acids which do not occur in the data base of the structures of known proteins, but for which the water–*n*-octanol transfer free energies can be measured directly or estimated from QSAR equations. It must be noted, however, that, for quite a number of side-chain pairs, there are several outliers that depart from eq. (26) by more than the standard deviation; this is illustrated in

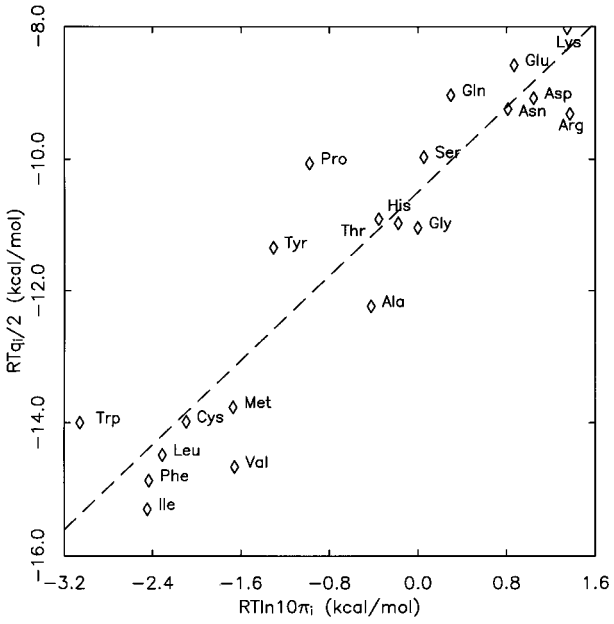


**FIGURE 5.** Correlation between side-chain hydrophobicities calculated from contact energies (abscissae) and those determined by Fauchere and Pliška,[59] as given by eq. (22).

| | C | M | F | I | L | V | W | Y | A | G | T | S | Q | N | E | D | H | R | K | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | | | | | | | | | | | | | | 1 | | | | 1 | | |
| M | | | | | | | | | | 1 | | 1 | | 2 | 1 | 2 | | 1 | 1 | |
| F | | | | | | | | | | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 |
| I | | | | | | | | | | 1 | | 1 | 1 | 2 | 1 | 2 | 2 | | 1 | 1 |
| L | | | | | | | | | | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | |
| V | | | | | | | | | | | | | | 1 | 1 | 1 | 2 | 1 | | |
| W | | | | | | | | | | | 1 | 1 | | 1 | | | | | | |
| Y | | | | | | | | | | | | | | | | | | | | |
| A | | | | | | | | | | | | | | | | | 1 | | | |
| G | | 1 | 1 | 1 | 1 | | | | | | | | | | | | | | | |
| T | | | 1 | | 1 | | 1 | | | | | | | | | | | | | |
| S | | 1 | 1 | 1 | 1 | | 1 | | | | | | | | | | | | | |
| Q | | | 1 | 1 | 1 | | | | | | | | | | | | | -1 | -1 | |
| N | 1 | 2 | 2 | 2 | 2 | 1 | 1 | | | | | | | | | | | | -1 | |
| E | | 1 | 2 | 1 | 1 | 1 | | | | | | | | | | | | -3 | -3 | |
| D | | 2 | 2 | 2 | 2 | 1 | | | | | | | | | | | | -3 | -3 | |
| H | | | 1 | 2 | 1 | 2 | | | 1 | | | | | | | | | | | |
| R | 1 | 1 | 1 | | 1 | 1 | | | | | | | -1 | | -3 | -3 | | | | |
| K | | 1 | 2 | 1 | 1 | | | | | | | | -1 | -1 | -3 | -3 | | | | |
| P | | | 1 | 1 | | | | | | | | | | | | | | | | |

**FIGURE 6.** A diagram showing the pairs of side chains with large excess free energies computed from the data in Table I. The one-letter code of amino-acid residues is used. The numbers in each box are the integers of the ratio of the excess contact free energy to the mean-square excess contact free energy, $\sqrt{\langle e^2_{excess} \rangle} = 0.23$ *RT* [the standard deviation from the ideal contact free energy in eq. (26)].

Figure 6. As shown, the excess free energy is less than the standard deviation from the ideal free energy only when both side chains in a pair are hydrophobic, both are hydrophilic with zero net charge, or both have charges of the same sign. Therefore, eqs. (24) and (26) should, in principle, be applicable only to such pairs of side chains. For pairs composed of one hydrophobic and one hydrophilic side chain the excess free energy is positive, which means that making such a contact is less favorable than predicted by the ideal contact free energy. This is understandable because, for example, for hydrophilic side chains bearing hydrogen-bonding groups, making a contact with a nonpolar side chain, as opposed to making a contact with another hydrogen-bonding group, results in breaking of hydrogen bonds. The excess free energies of the interaction of charged side chains of opposite signs are strongly negative, about $-0.7$ *RT*, which can be explained in terms of the formation of salt bridges. Finally, some small negative excess contact free energies occur for pairs composed of side chains with carboxyamide groups and positively charged side chains.

It should also be noted that the Arg−Arg contact free energy is considerably more negative than the contact free energies of the other pairs of residues with equal charges: Lys−Lys, Asp−Asp, and Glu−Glu. This is consistent with the results of the work of Magalhaes et al.,[69] in which it was demonstrated that water bridges constitute an important factor stabilizing the spatially close configurations of the charged guanidino groups of the arginine side chains. Because the other charged residues do not possess so many groups capable of forming hydrogen bonds with water, the exceptionally low value of the Arg−Arg contact free energy is understandable.

## DETERMINATION OF PARAMETERS AND STATISTICAL EVALUATION OF THE FIT OF POTENTIALS TO EXPERIMENTAL DATA

The sum of the squares defined by eq. (13) was minimized for all five potentials given by eqs. (2)−(8). The angular and radial-angular terms were not considered in determining the parameters of radial-only potentials, namely LJ and LJK [eqs. (3) and (4)]. We used the Marquardt algorithm,[70] which is especially designed for minimizing the sums of squares. This method requires only the first derivatives of the components of the sums of squares, from which both the gradient and the positive-definite approximate to the Hessian matrix are constructed.

Numerical integration to calculate the average correlation functions and free energies was carried out with step sizes of $d\varrho = 0.25$ Å, $d\vartheta = \pi/24$, and $d\varphi = \pi/6$, by taking the value of the function in the center of the respective bin as the average

value of the function in the bin. These step sizes were chosen as a compromise between computational efficiency and the error caused by too coarse a grid in the integration. Use of a finer grid resulted in differences in free energies and histogram values less than 1%. To increase computational efficiency, minimization was first carried out with a coarse grid (i.e., $d\varrho = \Delta r = 0.5$ Å, $d\vartheta = d\varphi = \pi/6$) and then completed (starting from the computed parameters) using a finer grid ($d\varrho = 0.125$ Å, $d\vartheta = \pi/24$, and $d\varphi = \pi/6$).

The starting values of $\epsilon^\circ$ were the free energies of contacts calculated from eq. (25). The values of $\sigma^\circ$ in the LJK, GB, and GBV potentials and the values of $r^\circ$ in the LJ potential were initially assigned half the side-chain van der Waals distances calculated by Levitt.[2] The initial values of $r_{ij}^\circ$ in the GBV and LJK potentials were 1.3 Å, this being the approximate van der Waals radius of the ''outer'' atoms of the side chains. For the anisotropic parameters $(\sigma^\parallel/\sigma^\perp)^2$ and $\chi'$, one choice was based on the ratio of the long and the geometric mean of the shorter principal axes of the moments of inertia of the side chains calculated by averaging their geometry, and another start was from isotropic potentials. Both gave the same final results.

Based on eqs. (17), the final estimated ratios of the weights of eq. (13) were $w^r:w^{\theta\phi}:w^{r\theta}:w^F = 1:20:20:20$ for all models.

Equation (13) contains pair-specific and single-residue-specific terms. We initially made trial runs by assuming that all the parameters are pair-specific; that is, we avoided the relations in eqs. (5), (9), and (10). However, for most of the side-chain pairs, the results were unreasonable, with the standard deviations exceeding the parameter values. Therefore, we decided to use eqs. (5), (9), and (10) to express all the constants except $\epsilon_{ij}$ in terms of single-body constants.

The fit of the functional forms of the potentials considered in this work to experimental data is compared in terms of the $F$-test[71] in Table II for the radial and anisotropic potentials, respectively. In the case of the radial potentials, the LJK model (shifted Lennard–Jones) appears clearly superior to the simple LJ model, the level of significance of introducing the ''shift'' parameters $r^\circ$ being close to 100%. A similar situation occurs for the anisotropic potential where the GBV functional

**TABLE II.**
**Comparison of Fit of Various Radial and Anisotropic Potentials to Experimental Data.**

| Potential | $\Phi$[a] | $\Phi^{\theta\phi}$ | $\Phi^{r\theta}$ | $\Phi^r$ | $\Phi^F \times 1000$ | $p$[b] | $\Delta\Phi$[c] | $F$[d] |
|---|---|---|---|---|---|---|---|---|
| LJK | 0.39 | — | — | 0.383 | 0.233 | 250 | — | — |
| LJ | 1.35 | — | — | 1.340 | 0.433 | 230 | 0.960 | 468 |
| GBV | 9.26 | 0.223 | 0.220 | 0.380 | 1.053 | 310 | — | — |
| LJK[e] | 10.23 | 0.230 | 0.261 | 0.384 | 0.952 | 250 | 0.975 | 871 |
| GB | 9.87 | 0.222 | 0.247 | 0.465 | 1.031 | 290 | 0.615 | 549 |
| BP | 9.75 | 0.222 | 0.241 | 0.462 | 1.625 | 290 | 0.493 | — |

[a] $\Phi$, $\Phi^{\theta\phi}$, $\Phi^{r\theta}$, $\Phi^r$, $\Phi^F$ are defined in eq. (13). For the LJ and LJK (radial) potentials $\Phi = w^r\Phi^r + w^F\Phi^F$, for the remaining (anisotropic) potentials $\Phi = w^{\theta\phi}\Phi^{\theta\phi} + w^{r\theta}\Phi^{r\theta} + w^r\Phi^r + w^F\Phi^F$, where $w^r = 1$ and $w^{\theta\phi} = w^{r\theta} = w^F = 20$ have been assigned based on the ratios of the respective residual variances (see text).

[b] The number of adjustable parameters.

[c] The difference between the fit corresponding to the potential giving an inferior fit to experimental data with that of the potential giving the best fit (i.e., LJK for the radial and GBV for the anisotropic potential).

[d] The $F$-test value to compare the goodness of fit of the inferior potential with that of the best one is:

$$F_i = \frac{N - p_i}{p^* - p_i} \frac{\Phi(\mathbf{X}_i) - \Phi(\mathbf{X}^*)}{\Phi(\mathbf{X}^*)}$$

where $N$ is the number of terms in the expression for $\Phi$; $N = 3451$ for the radial and 165,487 for the anisotropic potentials, $\mathbf{X}^* = (x_1^*, x_2^*, \ldots, x_{p^*}^*)^T$ is the vector of the parameters of the best model, and $\mathbf{X}_i = (x_{i;1}, x_{i;2}, \ldots, x_{i,p_i})^T$ is the vector of the parameters of $i$th inferior model $p_i < p^*$ (see ref. 70). With the large values of $N$ taken in this study, the best-fitting potentials are effectively different from the inferior ones at the 100% significance level. Because the model with the BP potential is not nested in the model with the GBV potential, the $F$-test value is not given in this case.

[e] The whole sum of squares was minimized, but with the radial LJK potential, which can be considered as the GBV potential devoid of the angular terms.

form, which allows for free ''shifting'' of $r$ in eq. (7), appears superior to both the BP and GB forms (however, BP is a model not nested in GBV and, therefore, we did not carry out its statistical comparison in terms of the $F$-test). Again, the significance level of introducing the new class of parameters $r°$, when passing from GB to GBV, is effectively 100%. Of the two potentials with fewer parameters, the BP form gives a better fit to the experimental data. From Table II, it also follows that the LJK potential gives a significantly poorer fit to the data that contain both radial and angular terms, that is, the angular terms are statistically significant.

The lesser adequacy of the GB form, compared with the BP and GBV ones, also follows from the plot of the residuals in the contact free energies shown in Figure 7. For the GB potential, all residuals show an apparent trend: the negative ones occur for negative and positive ones for positive contact free energies. This trend is partially eliminated for the BP potential and remains only for strongly hydrophilic pairs in the case of the GBV potential. The last observation may indicate that none of the models is fully adequate for hydrophilic pairs. On the other hand, such pairs interact weakly and therefore this should not cause great concern.

Sample theoretical and experimental histograms of the radial and angular correlation functions (the latter being averaged for visualizing purposes over the dihedral angle $\phi$) corresponding to the Leu–Leu pair are shown in Figure 8A and 8B.

To test how the parameters of the side-chain interaction potentials depend on the choice of the data base, we evaluated the parameters of the

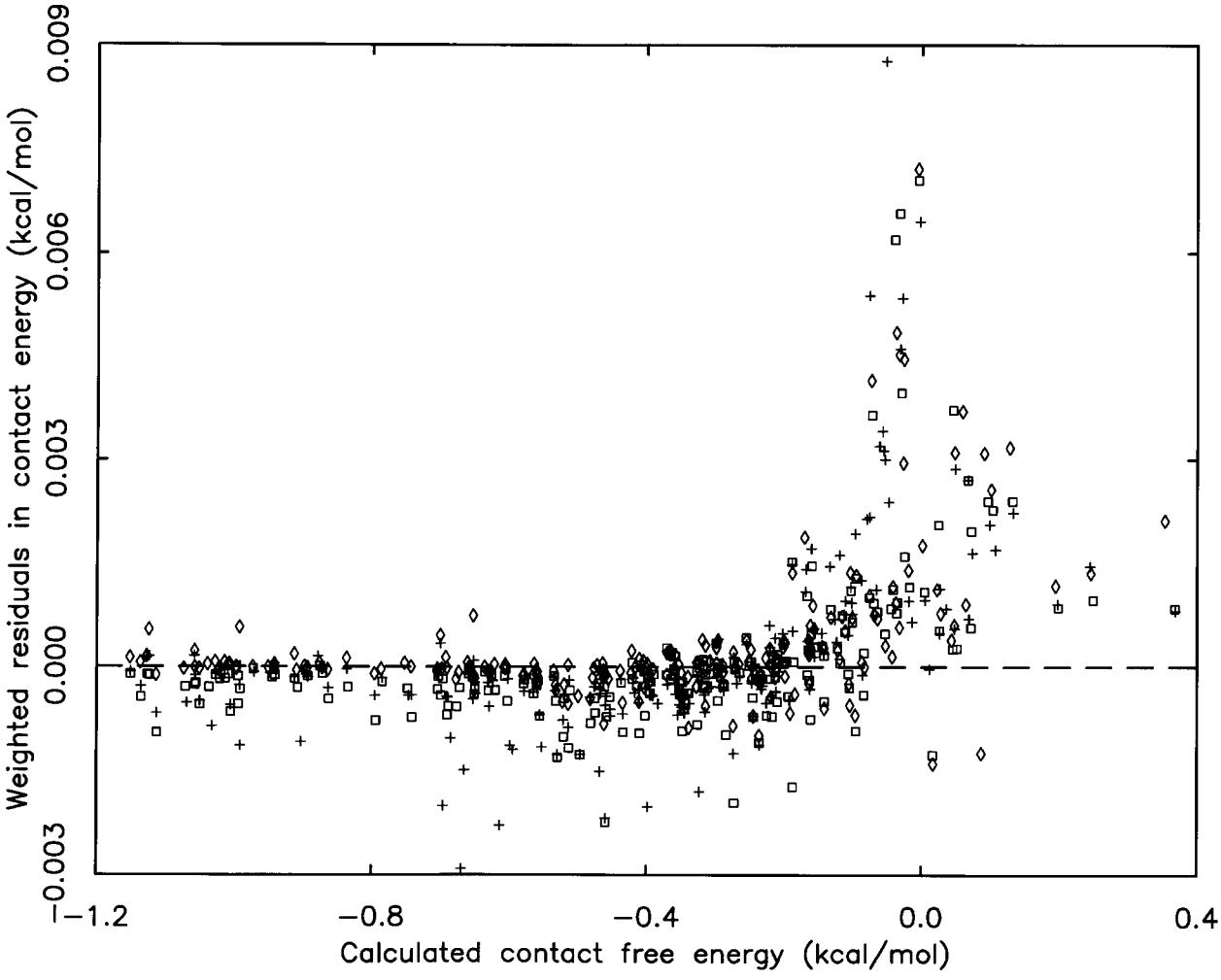


**FIGURE 7.** Plots of weighted residuals of the contact energies corresponding to the BP (crosses), GB (squares), and GBV (diamonds) potentials.

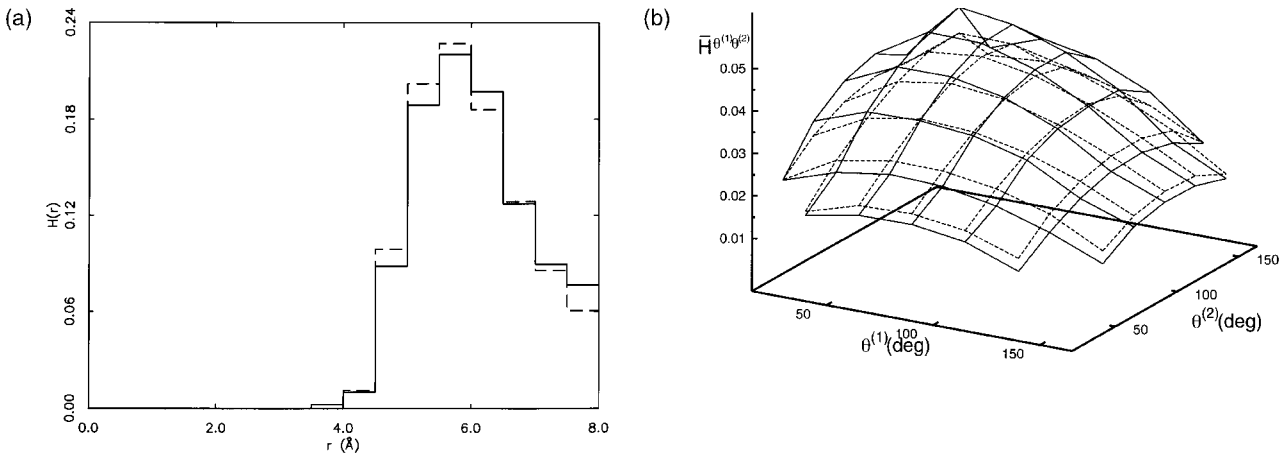


**FIGURE 8.** Sample calculated (dashed line and dashed surface) and experimental (solid line and solid surfaces) histograms of radial (A) and angular (B) correlation function for the Leu–Leu pair. For visualizing, the histograms of the angular correlation functions are averaged over the dihedral angle $\phi$.

GBV potential, using the set of 42 protein structures of Miyazawa and Jernigan[9]; the GBV potential contains the greatest number of adjustable parameters and should, therefore, be the most sensitive to data-base selection. (In the section ''Contact Free Energies,'' we have already shown that the contact free energies determined from our data base of 195 protein structures are in very good agreement with those determined by Miyazawa and Jernigan.) For the values of $\epsilon_\circ$ of eq. (6), which range from −12 to +1.6, the correlation coefficient was 0.8569 and the mean-square difference was 0.3 kcal/mol. For other parameters for which the range is not so extensive, correlation coefficients of approximately 0.8 were obtained. In view of the fact that the two data bases have almost no structure in common and the MJ data base is much smaller than ours, the parameters of the GBV potential determined from the two data bases are reasonably consistent.

## DISCUSSION OF COMPUTED PARAMETERS

The computed values of $\epsilon^\circ_{ij}$ and the single-body parameters of eq. (6) and their standard deviations for the GBV side-chain interaction potential considered in this work are given in Tables IIIa and IIIb. The parameters for the other four simpler potential functions are included in Tables 2a, 2b to 5a, 5b of the Supplementary Material, which also contains the parameters for all five potentials in machine-readable form.

Except for $\epsilon^\circ_{LysLys}$ of the GBV model, and the constants $\chi'$ and $\alpha$, the parameters are well determinable and significantly greater than their standard deviations. For hydrophobic residues, the values of the well-depth anisotropy $\chi'$ are small, although the anisotropy measures of the van der Waals radii $\sigma^\parallel/\sigma^\perp$ are significantly different from 1.0. Well-depth anisotropies appear significant for neutral and hydrophilic residues.

It is interesting to compare the computed parameters with contact free energies and estimates of the van der Waals radii and anisotropies that can be determined from the geometrical characteristics of the side chains. Such comparison is presented for the GBV model in Figure 9A–D.As shown, the contact free energies of hydrophobic pairs (for which $\epsilon_{ij} > 0$) correlate quite well with their van der Waals well depths determined by minimizing $\Phi$ of eq. (13). The values of $r^{\circ;L}$, used as the van der Waals distances in our earlier work, correlate with the values of $\sigma^\circ$, with the definite exception of aromatic residues and arginine (Fig. 9B). A similar situation occurs when the values corresponding to the LJK potential are taken into account. The correlation is even better (with the exception of Lys) when the values of $\sigma^\parallel$ calculated from $\sigma^\circ$ and the ratio $\sigma^\parallel/\sigma^\perp$ are considered (Fig. 9C). On the other hand, there is no correlation between the values of $r^\circ$ from our earlier work and the constants $r^\circ$ of Eq. (7).

There is no correlation between the parameters and the ratio of the long to the short axes of the side chains determined by diagonalizing the average matrices of the moments of inertia determined from the PDB. Thus, estimating these parameters based on the average dimensions of a side chain is incorrect. On the other hand, it is interesting to note that anisotropy parameters correlate with the

**TABLE IIIa.**

**Calculated Values of $\epsilon_{ij}^{\circ}$ (Kilocalories per Mole) for the GBV Potential (Diagonal and Lower Triangle) and Their Standard Deviations (Upper Triangle; Last Line Contains the Standard Deviations of the Diagonal Constants).**

| | Cys | Met | Phe | Ile | Leu | Val | Trp | Tyr | Ala | Gly | Thr | Ser | Gln | Asn | Glu | Asp | His | Arg | Lys | Pro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cys | 1.05 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| Met | 1.26 | 1.45 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 |
| Phe | 1.19 | 1.34 | 1.27 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 |
| Ile | 1.30 | 1.47 | 1.41 | 1.58 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 |
| Leu | 1.25 | 1.51 | 1.40 | 1.59 | 1.55 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 |
| Val | 1.17 | 1.38 | 1.31 | 1.52 | 1.50 | 1.40 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 |
| Trp | 0.99 | 1.17 | 1.15 | 1.21 | 1.18 | 1.10 | 0.97 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| Tyr | 0.92 | 1.15 | 1.05 | 1.22 | 1.18 | 1.04 | 0.87 | 0.81 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 |
| Ala | 0.98 | 1.20 | 0.99 | 1.24 | 1.26 | 1.19 | 0.77 | 0.81 | 1.02 | 0.01 | 0.01 | 0.01 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 |
| Gly | 0.98 | 1.03 | 0.84 | 1.06 | 1.13 | 1.01 | 0.71 | 0.72 | 0.82 | 0.56 | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.00 | 0.01 |
| Thr | 0.80 | 0.91 | 0.76 | 0.98 | 0.90 | 0.87 | 0.56 | 0.58 | 0.64 | 0.55 | 0.43 | 0.01 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.00 | 0.01 |
| Ser | 0.77 | 0.86 | 0.68 | 0.90 | 0.93 | 0.83 | 0.51 | 0.52 | 0.59 | 0.47 | 0.45 | 0.28 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.00 | 0.01 |
| Gln | 0.81 | 1.02 | 0.72 | 0.95 | 0.98 | 0.85 | 0.59 | 0.63 | 0.75 | 0.33 | 0.35 | 0.26 | −0.28 | 0.06 | 0.03 | 0.02 | 0.03 | 0.04 | 0.01 | 0.02 |
| Asn | 0.73 | 0.95 | 0.70 | 0.87 | 1.00 | 0.89 | 0.60 | 0.60 | 0.77 | 0.49 | 0.38 | 0.38 | 0.53 | 0.66 | 0.01 | 0.04 | 0.03 | 0.05 | 0.00 | 0.02 |
| Glu | 0.64 | 0.81 | 0.53 | 0.88 | 0.79 | 0.72 | 0.52 | 0.51 | 0.47 | −0.06 | 0.20 | 0.04 | −0.23 | −0.02 | −1.58 | 0.10 | 0.03 | 0.04 | 0.04 | 0.02 |
| Asp | 0.68 | 0.64 | 0.52 | 0.79 | 0.68 | 0.62 | 0.53 | 0.57 | 0.51 | 0.23 | 0.29 | 0.12 | −0.12 | 0.27 | −0.93 | −0.66 | 0.03 | 0.03 | 0.05 | 0.02 |
| His | 0.91 | 1.05 | 0.92 | 0.94 | 0.98 | 0.83 | 0.82 | 0.76 | 0.65 | 0.56 | 0.57 | 0.49 | 0.38 | 0.59 | 0.42 | 0.62 | 0.80 | 0.03 | 0.00 | 0.02 |
| Arg | 0.58 | 0.87 | 0.69 | 0.96 | 0.95 | 0.75 | 0.66 | 0.67 | 0.53 | 0.38 | 0.43 | 0.40 | 0.36 | 0.33 | 1.01 | 1.00 | 0.49 | −0.02 | 0.09 | 0.02 |
| Lys | 0.59 | 0.81 | 0.55 | 0.96 | 0.97 | 0.85 | 0.50 | 0.62 | 0.68 | −0.01 | 0.00 | −0.01 | −0.02 | 0.00 | 1.30 | 1.09 | −0.01 | −0.48 | −11.96 | 0.02 |
| Pro | 0.82 | 0.95 | 0.81 | 0.98 | 1.00 | 0.92 | 0.77 | 0.79 | 0.74 | 0.69 | 0.57 | 0.58 | 0.62 | 0.62 | 0.42 | 0.42 | 0.61 | 0.53 | 0.56 | 0.82 |
| | 0.03 | 0.03 | 0.02 | 0.01 | 0.01 | 0.01 | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.07 | 0.10 | 0.25 | 0.09 | 0.04 | 0.01 | 18. | 0.02 |

**TABLE IIIb.**
**Calculated Values of Single-Body Parameters of the GBV Potential. (Standard Deviations in Parentheses.) Except for $\sigma^0$ and $r^0$ All Quantities Are Dimensionless.**

| Residue | $\sigma^\circ$ (Å) | $r^\circ$ (Å) | $(\sigma^\parallel/\sigma^\perp)^2$ | $\chi'$ | $\alpha$ |
|---|---|---|---|---|---|
| Cys | 2.3204 (0.0382) | 5.7866 (0.2123) | 2.6006 (0.2042) | −0.0025 (0.0155) | 0.0299 (0.0070) |
| Met | 2.4984 (0.0237) | 3.5449 (0.1299) | 4.4303 (0.2018) | 0.0968 (0.0108) | 0.0878 (0.0058) |
| Phe | 2.2823 (0.0245) | 6.3367 (0.1459) | 3.9640 (0.1815) | 0.0491 (0.0077) | 0.0801 (0.0039) |
| Ile | 2.5919 (0.0150) | 4.4859 (0.0860) | 3.2406 (0.0926) | 0.0897 (0.0061) | 0.0664 (0.0031) |
| Leu | 2.8905 (0.0098) | 3.3121 (0.0548) | 2.3636 (0.0406) | 0.0749 (0.0047) | 0.1108 (0.0027) |
| Val | 2.7251 (0.0125) | 3.7770 (0.0629) | 2.0347 (0.0514) | 0.0770 (0.0062) | 0.0679 (0.0032) |
| Trp | 1.6947 (0.0644) | 9.2904 (0.3832) | 7.5089 (1.0060) | 0.0731 (0.0170) | 0.0549 (0.0076) |
| Tyr | 2.1346 (0.0271) | 4.8607 (0.1529) | 5.9976 (0.3578) | 0.1177 (0.0134) | 0.0438 (0.0064) |
| Ala | 2.4366 (0.0100) | 2.1574 (0.0423) | 1.8090 (0.0396) | 0.0333 (0.0077) | 0.1052 (0.0041) |
| Gly | 2.3359 (0.0169) | 2.5197 (0.0532) | 1.0429 (0.0498) | 0.2238 (0.0127) | −0.1277 (0.0073) |
| Thr | 2.6047 (0.0188) | 3.0723 (0.0996) | 2.2451 (0.0899) | 0.0236 (0.0162) | −0.0264 (0.0075) |
| Ser | 2.4471 (0.0203) | 2.2432 (0.0770) | 1.6795 (0.0749) | −0.0029 (0.0184) | −0.0348 (0.0083) |
| Gln | 2.6269 (0.0229) | 1.1813 (0.1189) | 2.6172 (0.1455) | 0.2960 (0.0305) | 0.0505 (0.0165) |
| Asn | 2.6954 (0.0165) | 0.7634 (0.0826) | 2.0433 (0.0850) | 0.2732 (0.0286) | −0.0064 (0.0152) |
| Glu | 2.5933 (0.0191) | 1.2819 (0.0874) | 2.5707 (0.1327) | 0.4904 (0.0332) | −0.0266 (0.0175) |
| Asp | 2.5098 (0.0192) | 1.4061 (0.0804) | 1.9262 (0.0925) | 0.3090 (0.0299) | 0.0250 (0.0164) |
| His | 2.3409 (0.0323) | 3.3570 (0.1817) | 3.6263 (0.2703) | 0.1351 (0.0245) | 0.0589 (0.0124) |
| Arg | 2.3694 (0.0214) | 1.8119 (0.1201) | 6.6061 (0.3758) | 0.2624 (0.0270) | 0.0062 (0.0130) |
| Lys | 2.7249 (0.0161) | 0.2712 (0.0913) | 8.0078 (0.2948) | 0.5790 (0.0364) | 0.0115 (0.0161) |
| Pro | 2.7230 (0.0228) | 3.3320 (0.1059) | 1.7905 (0.0759) | −0.1105 (0.0160) | −0.0190 (0.0065) |

dimensions of the side chains; larger side chains are more likely to exhibit more pronounced anisotropy (Fig. 9D).

Finally, it should be noted that, in the case of the LJK and GBV potentials, for many of the side chains, the constants, $r^\circ$, exceed $\sigma^\circ$ (see Table 3b and Table 3b of the Supplementary Material). Particularly large $r^\circ$ values occur for the aromatic residues which exhibit broad radial distributions. This means that the potential will rarely tend to infinity as side-chain separation approaches zero. This does not seem to be the result of inadequacy of the fitting procedure, because we included the regions in which the radial-correlation function is zero. We have also carried out additional trial runs by assuming lower exponents in eq. (2) than 6 and 12, which results in broadening the potential wells. However, we still obtained $r^\circ > \sigma^\circ$ for most of the side chains. Thus, to use the LJK and GBV potentials in simulations, in these two cases we changed the general form of the potential given by eq. (2) to:

$$U_{ij} = 4\left[|\epsilon_{ij}|x_{ij}^{12} - \epsilon_{ij}x_{ij}^{6}\right] + 4\epsilon_{ij}^{\circ}h(r_{ij}^{\circ} - \sigma_{ij}^{\circ})\left(\frac{\frac{1}{2}\sigma_{ij}^{\circ}}{r_{ij}}\right)^{12} \quad (27)$$

where $x_{ij}$ is defined by eqs. (4) and (7) for the LJK and GBV potentials, respectively, and $h(y)$ is the step function of $y$; $h(y) = 0$ for $y \leq 0$ and $h(y) = 1$ otherwise.

Thus, the modified expression includes a short-range ''repulsive core'' potential which prevents the collapse of the side chains. Introduction of this repulsive core does not impair the fit of the potentials to the experimental data.

## Conclusions

We have parameterized several functional forms for the potential of mean force of side-chain–side-chain interactions that are based on reasonable site–site interaction potentials used in molecular simulations. The parameters of the potentials have been determined consistently by fitting the energy expressions to the correlation functions and contact free energies obtained from high-resolution protein crystal data. Compared to related work on deriving the mean-field potentials from protein-crystal data,[24–31,34–36] our approach has two new features: inclusion of anisotropy of the free-energy surface, and explicit use of thermodynamic data to rescale the free energies of contacts [eq. (25)]. The

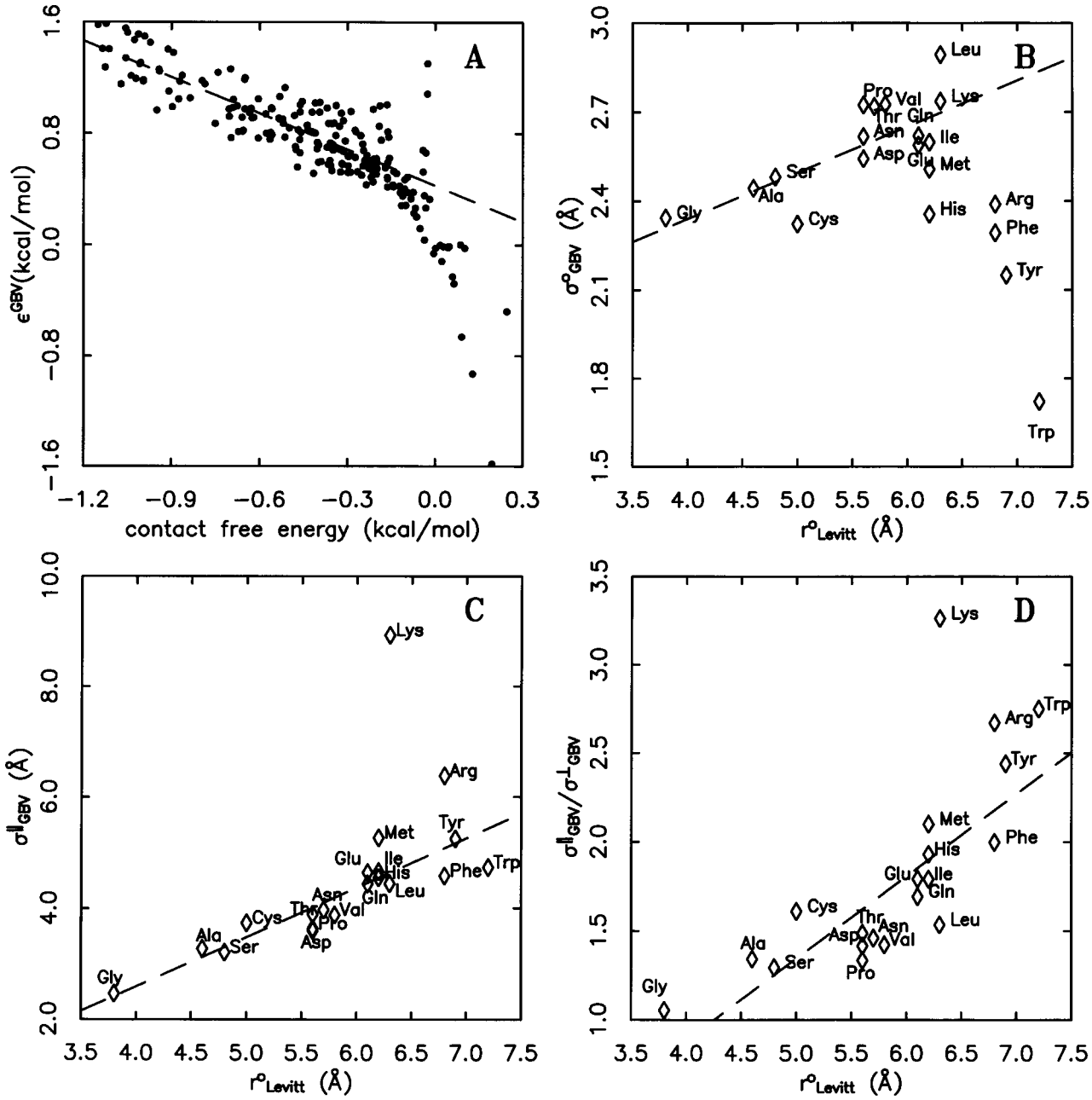


**FIGURE 9.** (A) Correlation between the hydrophobicity-scaled contact free energies, $F_{ij}$ (abscissae), and the corresponding van der Waals well depths, $\epsilon^{\circ}_{ij}$, corresponding to the GBV potential (ordinates). The straight line is the least-squares line calculated for the ''definitely hydrophobic'' pairs with $\epsilon_{ij} \geq 0$ kcal / mol; its equation is $\epsilon = -0.865(0.040)F + 0.425(0.022)$; $R = -0.8417$. (B) Correlation between the mean radii of side-chain contacts derived by Levitt[2] (abscissae) with the computed values of $\sigma^{\circ}$ of the GBV potential. After eliminating five apparent outliers: Phe, Tyr, Trp, His (aromatic side chains), and Arg, the equation is $\sigma^{\circ} = 0.1552(0.041)r^{0;L} + 1.72(0.23)$; $R = 0.7254$. (C) Correlation between Levitt's mean side-chain contact distances and the values of $\sigma^{\parallel}$ calculated from the parameters of the GBV potential. After eliminating lysine, the equation is $\sigma^{\parallel} = 0.89(0.13)r^{0;L} - 0.94(0.75)$; $R = 0.8619$. (D) Correlation between Levitt's mean side-chain contact distances and the values of $\sigma^{\parallel}/\sigma^{\perp}$ corresponding to the GBV potential. After eliminating lysine, the equation is $\sigma^{\parallel}/\sigma^{\perp} = 0.463(0.072)r^{0;L} - 0.97(0.43)$; $R = 0.8417$.

second feature enables us to consider the computed free energies of side-chain interactions as absolute values that can be compared directly with experimental data and/or the results of calculations with all-atom potentials (including hydration).

The choice as to which potential to use in simulations should be based on the balance between the accuracy of the representation of free-energy surface and computational effort. Regarding these two issues, the potentials can be ordered as follows: GBV, BP and GB, LJK, LJ. The GBV potential (that includes angular dependence) most accurately represents the free-energy surface, but involves the greatest computational effort, whereas the LJ potential (radial-only) is the simplest, but least accurate representation of the free-energy surface, and should be used when the computation time is a significant issue.

## Acknowledgments

This work was supported by Grant PB 190/T09/96/10 from the Polish State Committee for Scientific Research (KBN) (to A.L. and S.O.), by Grant AG 00322 from the National Institute on Aging (to S.R.), by Grant GM-14312 from the National Institute of General Medical Sciences, by Grant MCB95-13167 from the National Science Foundation (to H.A.S.), and by Grant CA 42500 from the National Cancer Institute (to M.R.P.).

Computations were carried out with one processor of the IBM-SP2 computer at the Cornell National Supercomputer Facility, a resource of the Center for Theory and Simulation in Science and Engineering at Cornell University, which is funded by the National Science Foundation, New York State, the IBM Corporation, and members of its Corporate Research Institute, with additional funds from the National Institutes of Health.

## Appendix: Definition and Calculation of Side-Chain Pair Correlation Functions from Protein-Crystal Data

Assume that we have a data base of $np$ protein structures. Let $\nu^{(2)}_{ij;p}(r, \theta^{(1)}, \theta^{(2)}, \phi)$ denote the number density of pairs of side chains of types $i$ and $j$ at a distance $r$ and orientation defined by the angles $\theta^{(1)}$, $\theta^{(2)}$, and $\phi$ for protein $p$, all assumed to be at the same temperature (for brevity of notation, we omit the side-chain-pair subscripts $ij$ in the symbols of the variables throughout the Appendix). Because we cannot determine the actual density at a point from experimental data, instead we will consider $\bar{\nu}^{(2)}_{ij;p}(r, \theta^{(1)}, \theta^{(2)}, \phi; T)$ defined as the average number density in the bins $b_{klmn} = b(r_k, \theta^{(1)}_l, \theta^{(2)}_m, \phi_n) = [r_k - \Delta r/2, r_k + \Delta/2] \times [\theta^{(1)}_l - \Delta\theta/2, \theta^{(1)}_l + \Delta\theta/2] \times [\theta^{(2)}_m - \Delta\theta/2, \theta^{(2)}_m + \Delta\theta/2] \times [\phi_n - \Delta\phi/2, \phi_n + \Delta\phi/2]$, $k = 1, 2, \ldots, nr$, $l = 1, 2, \ldots n_\theta$, $m = 1, 2, \ldots n_\theta$, $n = 1, 2, \ldots n_\phi$:

$$\bar{\nu}^{(2)}_{ij;p}(r_k, \theta^{(1)}_l, \theta^{(2)}_m, \phi_n) = \frac{\text{number of pairs of side chains of types } i \text{ and } j \text{ within } b_{klmn}}{\text{volume of } b_{klmn}} \qquad \text{(A-1)}$$

where $r_k = (k - 1/2)\Delta r$, $\theta^{(1)}_l = (l - 1/2)\Delta\theta$, $\theta^{(2)}_m = (m - 1/2)\Delta\theta$, $\phi_n = (n - 1/2)\Delta\phi$, $\Delta r = 0.5$ Å, $\Delta\theta = 30^\circ$, $\Delta\phi = 30^\circ$, $nr_{ij} = \text{int}(r^{max}_{ij}/\Delta r)$, $n_\theta = \pi/6$, $n_\phi = 2\pi/6$, where int is the integer part of a number; the values of $r^{max}$ are defined by eq. (12).

The pair number density $\bar{\nu}^{(2)}$ can be decomposed into the pair correlation function for residues of types $i$ and $j$, $\bar{g}_{ij}(r, \theta^{(1)}, \theta^{(2)}, \phi)$ (assumed to depend only on the types of the side chains and not on the protein in which they reside) and the reference-state pair number density $\bar{\nu}^{2,0}_{ij;p}(r, \theta^{(1)}, \theta^{(2)}, \phi)$, corresponding to a hypothetical chain with noninteracting side chains. Thus, given the actual and reference pair number densities that can be evaluated from the data base of protein structures, the pair correlation function can be calculated as follows:

$$\bar{g}_{ij}(r, \theta^{(1)}, \theta^{(2)}, \phi) = \frac{\sum_{p=1}^{np} w_p \bar{\nu}^{(2)}_{ij;p}(r, \theta^{(1)}, \theta^{(2)}, \phi)}{\sum_{p=1}^{np} w_p \bar{\nu}^{(2,0)}_{ij;p}(r, \theta^{(1)}, \theta^{(2)}, \phi)} \qquad \text{(A-2)}$$

where $w_p$ is the statistical weight of the $p$th protein in the sample; the choice of weights is discussed in the Results section [eq. (20)].

Clearly, $\bar{g}$ is the *average* pair correlation function corresponding to bin $b_{klmn}$:

$$\bar{g}_{ij}(r, \theta^{(1)}, \theta^{(2)}, \phi) = (1/\Delta V) \times \int_{r_-}^{r^+} \int_{\theta_-^{(1)}}^{\theta_+^{(1)}} \int_{\theta_-^{(2)}}^{\theta_+^{(2)}} \int_{\phi_-}^{\phi_+} g_{ij}(\varrho, \vartheta^{(1)}, \vartheta^{(1)}, \varphi)\, \mathrm{d}V \tag{A-3}$$

where:

$$r_- = r - \Delta r/2, \quad r_+ = r + \Delta r/2,$$
$$\theta_-^{(1)} = \theta^{(1)} - \Delta\theta/2, \quad \theta_+^{(1)} = \theta^{(1)} + \Delta\theta/2,$$
$$\theta_-^{(2)} = \theta^{(2)} - \Delta\theta/2, \quad \theta_+^{(2)} = \theta^{(2)} + \Delta\theta/2,$$
$$\phi_- = \phi - \Delta\phi/2, \quad \phi_+ = \phi + \Delta\phi/2,$$
$$\mathrm{d}V = \varrho^2 \sin\vartheta^{(1)} \sin\vartheta^{(2)} \mathrm{d}\varrho \mathrm{d}\vartheta^{(1)} \mathrm{d}\vartheta^{(2)} \mathrm{d}\varphi$$

and:

$$\Delta V = (1/3)(r_+^3 - r_-^3)(\cos\theta_-^{(1)} - \cos\theta_+^{(1)}) \times (\cos\theta_-^{(2)} - \cos\theta_+^{(2)})\Delta\phi = (1/3)\Delta r^3 \Delta\cos\theta^{(1)} \Delta\cos\theta^{(2)} \Delta\phi$$

The limited number of available protein structures still makes it impossible to determine the pair correlation functions with reasonable accuracy. This is easily realized because, even the choice of a coarse grid of $\Delta r = 0.5$ Å, $\Delta\theta = \Delta\phi = 30°$ with implementation of the symmetry of the hypersurface in $\phi$ (only the interval $[0°, 180°]$ needs to be considered) yields $16 \times 6 \times 6 \times 6 = 3456$ bins [according to eq. (12) we take a maximum 8-Å coordination sphere for a residue] for which the average correlation functions would have to be determined. Within this coordination sphere, we have at best about 5000 points per residue pair, which would mean an average of about 1.4 counts per bin. Therefore, in the fitting procedure we use the correlation functions averaged over some radial and angular variables, respectively:

$$\bar{g}_{ij}^{r}(r) = \frac{1}{8\pi\Delta r^3} \int_{r_-}^{r^+} \int_0^{\pi} \int_0^{\pi} \int_0^{2\pi} g_{ij}(\varrho; \vartheta^{(1)}, \vartheta^{(2)}, \varphi)\, \mathrm{d}V \approx \frac{\sum_{p=1}^{np} w_p \bar{\nu}_{ij;p}^{(2,r)}(r)}{\sum_{p=1}^{np} w_p \bar{\nu}_{ij;p}^{(0,2,r)}(r)} \tag{A-4}$$

$$\bar{g}_{ij}^{\theta\phi}(\theta^1, \theta^2, \phi) = \frac{3}{r_{max}^3 \Delta\cos\theta^{(1)} \Delta\cos\theta^{(2)} \Delta\phi} \times \int_0^{r_{max}} \int_{\theta_-^{(1)}}^{\theta_+^{(1)}} \int_{\theta_-^{(2)}}^{\theta_+^{(2)}} \int_{\phi_-}^{\phi^+} g_{ij}(\varrho; \vartheta^{(1)}, \vartheta^{(2)}, \varphi)\, \mathrm{d}V \approx \frac{\sum_{p=1}^{np} w_p \bar{\nu}_{ij;p}^{(2,\theta\phi)}(\theta^{(1)}, \theta^{(2)}, \phi)}{\sum_{p=1}^{np} w_p \bar{\nu}^{(0,2,\theta\phi)}(\theta^{(1)}, \theta^{(2)}, \phi)} \tag{A-5}$$

$$\bar{g}_{ij}^{r\theta}(r, \theta^{(1)}, \theta^{(2)}) = \frac{1}{2\pi\Delta r^3 \Delta\cos\theta^{(1)} \Delta\cos\theta^{(2)}} \times \int_{r_-}^{r_+} \int_{\theta_-^{(1)}}^{\theta_+^{(1)}} \int_{\theta_-^{(2)}}^{\theta_+^{(2)}} \int_0^{2\pi} g_{ij}(r; \vartheta^{(1)}, \vartheta^{(2)}, \varphi)\, \mathrm{d}V \approx \frac{\sum_{p=1}^{np} w_p \bar{\nu}_{ij;p}^{(2,r\theta)}(r, \theta^{(1)}, \theta^{(2)})}{\sum_{p=1}^{np} \bar{\nu}^{(0,2,r\theta)}(\theta^{(1)}, \theta^{(2)}, \phi)} \tag{A-6}$$

where $\bar{g}^{r}$, $\bar{g}^{\theta\phi}$, and $\bar{g}^{r\theta}$ denote the correlation functions averaged over all angles ($\theta^{(1)}$, $\theta^{(2)}$ and $\phi$), $r$, and the rotation angle, $\phi$, respectively; we noticed that the dependence of the distribution function on $\phi$ is the weakest and, therefore, chose to average over $\phi$ to obtain a mixed radial and angular correlation function [eq. (A-6)]. Likewise, $\bar{\nu}^{(2,r)}(r)$, $\bar{\nu}^{(2,\theta\phi)}(\theta^{(1)}, \theta^{(2)}, \phi)$, and $\bar{\nu}^{(2,r\theta)}(r, \theta^{(1)}, \theta^{(2)})$ denote the average number densities within $[r_-, r_+]$, $[\theta_-^{(1)}, \theta_+^{(1)}] \times [\theta_-^{(2)}, \theta_+^{(2)}] \times [\phi_-, \phi_+]$, and $[r_-, r_+] \times [\theta_-^{(1)}, \theta_+^{(1)}] \times [\theta_-^{(2)}, \theta_+^{(2)}]$, respectively.

The reference pair distribution functions must still be defined. We assume that they can be decomposed into the radial and angular part, and that the radial part can be expressed as a product of the Markovian factor, $M_{ij;p}(r)$, arising from the fact that the side chains are on a polypeptide chain,[9] a factor accounting for the finite dimensions and nonuniform residue density, $T_{ij;p}(r)$,[72] of protein molecules, and the ''background'' angular distribution function, $\Omega_{ij}(\theta^{(1)}, \theta^{(2)}, \phi)$, as given by eqs. (A-7)–(A-9).

$$\bar{\nu}^{(0,2,r)}(r) = \sum_{p=1}^{np} w_p M_{ij;p}(r) T_{ij;p}(r) \tag{A-7}$$

$$\bar{\nu}^{(0,2,\theta\phi)}(\theta^{(1)}, \theta^{(2)}, \phi) = N_{ij}(r \le R_c)\Omega(\theta^{(1)}, \theta^{(2)}, \phi) \tag{A-8}$$

$$\bar{\nu}^{(0,2,r\theta)}(r, \theta^{(1)}, \theta^{(2)}) = (1/2\pi) M_{ij;p} T_{ij;p}(r) \int_0^{2\pi} \Omega(\theta^{(1)}, \theta^{(2)}, \varphi)\, \mathrm{d}\varphi \tag{A-9}$$

where $N_{ij}(r \le R_c) = \sum_{i=1}^{np} w_p n_{ij;p}(r \le R_c)$ is the weighted total number of side chains of types $i$ and $j$ in the protein-structure data base, which are separated by at least 10 peptide groups and whose distance is less than the assumed radius, $R_c$, of the coordination sphere of a residue (assumed to be 8 Å in this work).

We assumed the reference function to be independent of side chain and protein; that is, it includes all side chains in all the proteins used.

The components $M_{ij;p}(r)$ and $T_{ij;p}(r)$ of the radial reference functions are defined as follows[9,72]:

$$M_{ij;p}(r) = \sum_{k \ge 10} n_{ij;k;p} P(r;k) \qquad \text{(A-10)}$$

where $k$ is the number of peptide groups separating side chains of types $i$ and $j$, $n_{ij;k;p}$ is the total number of pairs of side chains of types $i$ and $j$ separated by $k$ peptide groups in the data base of protein crystal structures, and $P(r;k)$ is the Markovian probability density that two side chains separated by $k$ peptide groups are at the distance $r$; we assumed the form given by eq. (25) of ref. 9 for $P(r;k)$. From ref. 72:

$$T_{ij;p}(r) = \frac{1}{4\pi r^2 V_p} \int_{S_p} \int_{S_p \cap S(\mathbf{x}, r)} \rho_{i;p}(\mathbf{x}) \rho_{j;p}(\mathbf{y}) \mathrm{d}^2\mathbf{y} \mathrm{d}^3\mathbf{x} \qquad \text{(A-11)}$$

where $S_p$ and $V_p$ denote the region of space occupied by the $p$th protein and the volume of this region, respectively, $S(\mathbf{x}, r)$ is the sphere of radius $r$ centered at the point $\mathbf{x}$, and $\rho_i$ and $\rho_j$ are the average single-body densities of residues of types $i$ and $j$; we assume that they depend only on the ratio of the distance from the center of a protein to the end of its radius of gyration. The corresponding average density, used for $\rho_{i;p}$ and $\rho_{j;p}$, is calculated from eq. (A-12).

$$\bar{\rho}_i(\xi) = \frac{\sum_{p=1}^{np} w_p \bar{\nu}_{i;p}(\xi - \Delta\xi/2 \le r/r_p^{gy} \le \xi + \Delta\xi/2)}{\sum_{p=1}^{np} w_p n_{i;p}} \qquad \text{(A-12)}$$

where $\xi$ is the distance from the center of a protein scaled by the radius of gyration, $r^{gy}$, $\bar{\nu}_{i;p}(r/r_p^{gy})$ is the average number density of residues $i$ at $r/r_p^{gy}$, and $n_{i;p}$ is the total number of residues of type $i$ in the $p$th protein. We used a step size of $\Delta\xi = 0.1$.

The angular reference function $\Omega_{ij}(\theta^{(1)}, \theta^{(2)}, \phi)$, was calculated as the average of the angular correlation functions, $\bar{g}(\theta^{(1)}, \theta^{(2)}, \phi)$, averaged over side-chain types:

$$\Omega(\theta^{(1)}, \theta^{(2)}, \phi) = 1/210 \sum_{i=1}^{20} \sum_{j=i}^{20} \bar{g}_{ij}^{\theta\phi}(\theta^{(1)}, \theta^{(2)}, \phi) \qquad \text{(A-13)}$$

This ''background'' angular correlation function, $\nu^{(0,2,\theta,\phi)}$, is qualitatively similar to the function that we obtained in a test Monte Carlo simulation of 1000 random-sequence and random-conformation 50-residue polypeptide chains confined to an average volume[67] characteristic of a 50-residue protein, using the united-residue potential of our earlier work[38,39] and a procedure developed by Hao et al. for confined-space simulations.[67] The united-residue force field did not include any side-chain anisotropy.[38,39] Nevertheless, the average angular correlation function exhibits some anisotropy (Fig. 4). This results from the fact that one end of each side chain is tethered to the backbone and therefore the ''free'' ends of the side chains can easily approach each other, whereas the tethered ends cannot. The similarity of the ''background'' angular correlation function obtained from protein crystal data to the function obtained in simulations with a radial potential (shown in Fig. 4) justifies its use as the reference angular correlation function.

## References


1. M. Levitt and A. Warshel, *Nature*, **253**, 694 (1975).
2. M. Levitt, *J. Mol. Biol.*, **104**, 59 (1976).
3. M. R. Pincus and H. A. Scheraga, *J. Phys. Chem.*, **81**, 1579 (1977).
4. P. R. Gerber, *Biopolymers*, **32**, 1003 (1992).
5. A. Wallqvist and M. Ullner, *Proteins*, **18**, 267 (1994).
6. A. Rey and J. Skolnick, *Proteins*, **16**, 8 (1993).
7. A. Rey and J. Skolnick, *J. Chem. Phys.*, **100**, 2267 (1994).
8. S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 945 (1976).
9. S. Miyazawa and R. L. Jernigan, *Macromolecules*, **18**, 534 (1985).
10. L. M. Gregoret and F. E. Cohen, *J. Mol. Biol.*, **211**, 959 (1990).
11. D. G. Covell, *Proteins*, **14**, 409 (1992).
12. J. Skolnick and A. Koliński, *Science*, **250**, 1121 (1990).
13. A. Koliński and J. Skolnick, *J. Chem. Phys.*, **97**, 9412 (1992).
14. A. Koliński, A. Godzik, and J. Skolnick, *J. Chem. Phys.*, **98**, 7420 (1993).
15. A. Godzik, A. Koliński, and J. Skolnick, *J. Comput.-Aid. Mol. Des.*, **7**, 397 (1993).

16. J. Skolnick, A. Koliński, C. L. Brooks, III, A. Godzik, and A. Rey, *Cur. Biol.*, **3**, 414 (1993).
17. A. Koliński and J. Skolnick, *Proteins*, **18**, 338 (1994).
18. A. Koliński and J. Skolnick, *Proteins*, **18**, 353 (1994).
19. M. Vieth, A. Koliński, C. L. Brooks, III, and J. Skolnick, *J. Mol. Biol.*, **237**, 361 (1994).
20. R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, **89**, 9029 (1992).
21. N. S. Goel and M. Yčas, *J. Theor. Biol.*, **77**, 253 (1979).
22. H. Wako and H. A. Scheraga, *J. Prot. Chem.*, **1**, 5 (1982).
23. H. Wako and H. A. Scheraga, *J. Prot. Chem.*, **1**, 85 (1982).
24. G. M. Crippen and V. N. Viswanadhan, *Int. J. Peptide Prot. Res.*, **24**, 279 (1984).
25. G. M. Crippen and V. N. Viswanadhan, *Int. J. Peptide Prot. Res.*, **25**, 487 (1985).
26. G. M. Crippen and P. K. Ponnuswamy, *J. Comput. Chem.*, **8**, 972 (1987).
27. G. M. Crippen and M. E. Snow, *Biopolymers*, **29**, 1479 (1990).
28. P. Seetharamulu and G. M. Crippen, *J. Math. Chem.*, **6**, 91 (1991).
29. V. N. Maiorov and G. M. Crippen, *J. Mol. Biol.*, **227**, 876 (1992).
30. V. N. Maiorov and G. M. Crippen, *Proteins*, **20**, 167 (1994).
31. G. M. Crippen and V. N. Maiorov, In *Protein Structure Distance Analysis*, H. Bohr and S. Brunak, Eds., IOS Press, Amsterdam, 1994, p. 158.
32. C. Wilson and S. Doniach, *Proteins*, **6**, 193 (1989).
33. K. Nishikawa and Y. Matsuo, *Prot. Eng.*, **6**, 811 (1993).
34. M. J. Sippl, *J. Mol. Biol.*, **213**, 859 (1990).
35. G. Casari and M. J. Sippl, *J. Mol. Biol.*, **224**, 725 (1992).
36. M. J. Sippl, *J. Comput.-Aid. Mol. Des.*, **7**, 473 (1993).
37. S. Sun, *Prot. Sci.*, **2**, 762 (1993).
38. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *Prot. Sci.*, **2**, 1697 (1993).
39. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *Prot. Sci.*, **2**, 1715 (1993).
40. K. A. Dill, *Biochemistry*, **29**, 7133 (1990).
41. E. I. Shakhnovich and A. M. Gutin, *Proc. Natl. Acad. Sci. USA*, **90**, 7195 (1993).
42. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.*, **98**, 4940 (1994).
43. Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **84**, 6611 (1987).
44. Z. Li and H. A. Scheraga, *J. Mol. Struct.* (*Theochem*), **179**, 333 (1988).
45. J. Kostrowicki and H. A. Scheraga, *J. Phys. Chem.*, **96**, 7442 (1992).
46. K. A. Olszewski, L. Piela, and H. A. Scheraga, *J. Phys. Chem.* **97**, 267 (1993).
47. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Ołdziej, and H. A. Scheraga, *J. Comput. Chem.* (accompanying article).
48. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361 (1975).
49. G. Némethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87**, 1883 (1983).
50. I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, *J. Biomol. Struct. Dyn.*, **7**, 421 (1989).
51. Cited in: H. Margenau and N. R. Kestner, *Theory of Intermolecular Forces*, Pergamon Press, Oxford, p. 107, 1st ed. (1969).
52. B. J. Berne and P. Pechukas, *J. Chem. Phys.*, **56**, 4213 (1972).
53. J. G. Gay and B. J. Berne, *J. Chem. Phys.*, **74**, 3316 (1981).
54. G. R. Luckhurst, R. A. Stephens, and R. W. Phippen, *Liquid Cryst.*, **8**, 451 (1990).
55. A. P. J. Emerson, R. Hashim, and G. R. Luckhurst, *Mol. Phys.*, **76**, 241 (1992).
56. Y. N. Vorobjev, *Biopolymers*, **29**, 1503 (1990).
57. Y. N. Vorobjev, *Biopolymers*, **29**, 1519 (1990).
58. H. B. Bürgi and J. D. Dunitz, *Acc. Chem. Res.*, **16**, 153 (1983).
59. A. Godzik, A. Koliński, and J. Skolnick, *Prot. Sci.*, **4**, 2107 (1995).
60. J.-L. Fauchere and V. Pliška, *Eur. J. Med. Chem.*, **18**, 369 (1983).
61. R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, Chapman and Hall, New York, 1988, p. 13.
62. D. A. Ratkowsky, *Handbook of Nonlinear Regression Models*, Marcel Dekker, New York, 1990, p. 38.
63. D. J. Lipman and W. R. Pearson, *Science*, **227**, 1435 (1985).
64. W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. USA*, **85**, 2444 (1988).
65. H. Späth, *Cluster Analysis Algorithms*, Halsted Press, New York, 1980, p. 170.
66. H. H. Gan and B. C. Eu, *J. Chem. Phys.*, **100**, 5922 (1994).
67. M. H. Hao, S. Rackovsky, A. Liwo, M. R. Pincus, and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **89**, 6614 (1992).
68. Y. Nozaki and C. Tanford, *J. Biol. Chem.*, **246**, 2211 (1971).
69. A. Magalhaes, B. Maigret, J. Hoflack, J. N. F. Gomes, and H. A. Scheraga, *J. Prot. Chem.*, **13**, 195 (1994).
70. D. W. Marquardt, *J. Soc. Indust. Appl. Math.*, **11**, 431 (1963).
71. G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, Wiley, New York, 1989, p. 228.
72. J. Edelman, *Biopolymers*, **32**, 3 (1992).