Prediction of Residue−Residue Pair Frequencies in Proteins

M. Vijayakumar and Huan-Xiang Zhou*

Department of Physics, Drexel University, Philadelphia, Pennsylvania 19104

Received: May 11, 2000; In Final Form: July 27, 2000

Knowledge of how frequently different types of residues are found near each other in protein structures has been widely used in threading and in simulating protein folding. In this paper we show that the residue−residue pair frequencies can be reproduced by a simple, physical model. The central component is the nonpolar in−charge out character. This character was captured by obtaining for each type of residue the relative density at a given distance from the protein's center of geometry. These densities were conveniently fitted to exponential or linear functions of the radial distance and used to generate atomic positions. To account for chain connectivity, distances between residue pairs were constrained by independent Gaussian functions, which have increasing means and deviations for increasing sequence separations. Interactions between nonpolar residues and between charged residues were found to extend up to a distance of ∼7.5 Å and the interaction potentials extracted appear to be an intrinsic property. This radial-distance based model, constructed and tested on a set of 243 nonhomologous proteins, has a clear physical basis and may hold important clues for structure prediction.

Introduction

Knowledge of how frequently different types of amino acids are found near each other in protein structures has been widely used in threading sequences through structures[1−8] and in simulating protein folding.[9−11] Because of the inhomogeneous distributions of residues within each protein and the finite sizes of proteins, it can be argued that the pair frequencies (PFs) are not a direct measure of the intrinsic affinities between the residues.[12] This paper aims to elucidate the physical basis of PFs by reproducing them from a simple model. We show that the PF for any pair of residues beyond a distance $s = \sim 7.5$ Å can be completely explained by the inhomogeneous distributions of the residues, the finite size of the protein, and chain connectivity. When $s < 7.5$ Å, nonpolar residues and oppositely charged residues exhibit strong interactions. The interaction potentials extracted from two sets of proteins with chain lengths less and greater than 220 residues are identical, suggesting that they are intrinsic to the particular types of residues.

The inhomogeneous distributions of residues are most prominently manifested as nonpolar in and charge out, dictated by hydrophobicity and desolvation cost. The nonpolar in−charge out character motivated us to model the inhomogeneous distributions by the radial density $\rho_i(r)$, i.e., the relative density of type i residue at a distance r from a protein's center of geometry. Specifically, $\rho_i(r)$ is the number of type i residues in the spherical shell between $r - \Delta r/2$ and $r + \Delta r/2$ divided by the number of all atoms in that shell (not by the volume of that shell). This definition is used to account for the fact that a protein is never a sphere with a uniform atom-number density. The number $D(r)\Delta r$ of all atoms between $r - \Delta r/2$ and $r + \Delta r/2$ initially increases (since the shell volume increases as r^2) but eventually decreases to zero because of the finite size of the protein. Let the spherical coordinates of a residue be (r, θ, ϕ). Given $\rho_i(r)$ and $\rho_j(r')$ for the radial densities of types i and j residues and $D(r)$ for the atom-number density profile of the protein, the probability density for finding the two types of residues at a distance s is

$$p_{ij}^{\,0}(s) = \int \mathrm{d}r\, \mathrm{d}\Omega \int \mathrm{d}r'\, \mathrm{d}\Omega'\, \rho_i(r)D(r)\rho_j(r')D(r')\delta(|\mathbf{r} - \mathbf{r}'| - s)/(4\pi)^2 n_i n_j \quad (1)$$

where $\mathrm{d}\Omega$ represents the angular factor $\sin\theta\, \mathrm{d}\theta\, \mathrm{d}\phi$ and $n_i = \int \mathrm{d}r\, \rho_i(r)D(r)$ is total number of type i residue. The normalization condition for $p_{ij}^{\,0}(s)$ is $\int_0^\infty \mathrm{d}s\, p_{ij}^{\,0}(s) = 1$. The pair frequency $\mathrm{PF}^0(s)$ is the integration of $n_i n_j p_{ij}^{\,0}(s)$ from $s - \Delta s/2$ to $s + \Delta s/2$ and is given by the total number of ij pairs with distances within that range. Specifically

$$\mathrm{PF}^0(s) = \int_{s-\Delta s/2}^{s+\Delta s/2} \mathrm{d}s\, n_i n_j p_{ij}^{\,0}(s) = N_{ij}^{\,0}(s \in [s - \Delta s/2,\, s + \Delta s/2]) \quad (2)$$

Equation 1 does not account for the fact that the residues of a protein are connected to form a chain. Chain connectivity affects the PF since two residues closer along the sequence tend to have a smaller distance. In other words, statistically, the probability of forming close contacts by two residues decreases as they are farther and farther separated along the chain. Let $\mathbf{X}$ represent the coordinates $(r_1, \Omega_1, ..., r_{ni}, \Omega_{ni}, r'_1, \Omega'_1, ..., r'_{nj}, \Omega'_{nj})$ of the $n_i + n_j$ residues and $\mathbf{K}$ represent their identification numbers $(k_1, ..., k_{ni}, k'_1, ..., k'_{nj})$ along the sequence. Chain connectivity modifies the probability density of $\mathbf{X}$ from $Q(\mathbf{X}) = \rho_i(r_1)D(r_1)...\rho_i(r_{ni})D(r_{ni})\, \rho_j(r'_1)D(r'_1)...\rho_j(r'_{nj})D(r'_{nj})$ to $Q(\mathbf{X})$-$C(\mathbf{X}, \mathbf{K})$. The probability density for finding the two types of residues at a distance s is then

$$p_{ij}^{\,0}(s) = \frac{1}{N}\int \mathrm{d}\mathbf{X}\, Q(\mathbf{X})C(\mathbf{X}, \mathbf{K}) \frac{1}{n_i n_j} \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} \delta(|\mathbf{r}_l - \mathbf{r}'_m| - s) \quad (3)$$

where N is a normalization factor. If $C(\mathbf{X}, \mathbf{K})$ is set to 1, eq 3

* Corresponding author: phone (215) 895-2716; fax (215) 895-5934; e-mail hxzhou@einstein.drexel.edu.

10.1021/jp001757f CCC: $19.00 © 2000 American Chemical Society
Published on Web 09/22/2000

reduces to eq 1. We modeled $C(\mathbf{X}, \mathbf{K})$ loosely after the Gaussian chain. Each pair of nonidentical residues (of the *ij*, *ii*, or *jj* type) contributes a Gaussian factor. This has the form

$$c(\mathbf{r}_1,\mathbf{r}_2,k_1,k_2) = \exp[-(|\mathbf{r}_1 - \mathbf{r}_2| - b)^2/2\sigma^2] \qquad (4)$$

where the mean b and standard deviation σ depend on the number of residues separating the particular pair. This is $|k_1 - k_2| - 1 \equiv \Delta$ and is called sequence separation.

If two residues interact at short distances, the pair frequencies there will be affected. Suppose that the interaction energy of an *ij* pair located at $\mathbf{r}$ and $\mathbf{r}'$ is $w_{ij}(|\mathbf{r} - \mathbf{r}'|)$, then this pair will contribute a factor $\exp[-w_{ij}(|\mathbf{r} - \mathbf{r}'|)/k_BT]$ to the probability density of $\mathbf{X}$. The probability density for finding the two types of residues at a distance s now becomes

$$p_{ij}(s) = \frac{1}{N}\int d\mathbf{X}\, Q(\mathbf{X})C(\mathbf{X}, \mathbf{K})\frac{\exp[-w_{ij}(s)/k_BT]}{n_in_j} \times \sum_{l=1}^{n_i}\sum_{m=1}^{n_j}\delta(|\mathbf{r}_l - \mathbf{r}'_m| - s) \qquad (5a)$$

$$= \exp[-w_{ij}(s)/k_BT]p_{ij}^{\,0}(s) \qquad (5b)$$

In writing eq 5b we have neglected the effect of the factor $\exp[-w_{ij}(s)/k_BT]$ when evaluating any integral over the full range of the distance s. This is generally justified for potentials of mean force. With this neglect $p_{ij}(s)$ [like $p_{ij}^{\,0}(s)$] is normalized. Multiplying both sides of eq 5b by n_in_j and integrating from $s - \Delta s/2$ to $s + \Delta s/2$, we get

$$\mathrm{PF}(s) = \exp[-w_{ij}(s)/k_BT]\mathrm{PF}^0(s) \qquad (6)$$

where we have neglected the distance dependence of the interaction potential within the integration range.

We can interpret eq 6 as the definition for the interaction potential in terms of the ratio of the PF functions in two states: $\mathrm{PF}(s)$ in the physical state in which the residues adopt the actual structures of the proteins and $\mathrm{PF}^0(s)$ in a reference state in which the interaction potential is eliminated but everything else stays exactly the same. This is precisely how the potential of mean force is defined in liquids. There, in the reference state the atoms are uniformly distributed when the influence of interactions is eliminated by design; hence $\mathrm{PF}^0(s)$ is proportional to $s^2\Delta s$. In proteins we cannot take the reference state to be one in which the residues are uniformly distributed. We only want to eliminate the influence of residue−residue interactions, not the direct residue−solvent interactions that give rise to the inhomogeneous distributions of residues within proteins. Neither do we want to eliminate the chain constraints. The desired reference state should include the effects of the amino acid composition, the inhomogeneous distribution of the different residue types, and chain constraints. Such a state is generated according to eq 3. In practice, for the purpose of determining the interaction potential, the inhomogeneous distributions of residues can be accounted for by keeping the radial distances at their values in the actual structures of the proteins but randomly selecting their polar and azimuthal angles. The chain constraints are then enforced. From these "reference structures" one calculates $\mathrm{PF}^0(s)$ (see eq 2) and from the actual structures one calculates $\mathrm{PF}(s)$. The interaction potential is finally obtained according to eq 6. Since the interaction potential is not expected to persist beyond a certain distance, an important indication for the validity of our approach is that $\mathrm{PF}^0(s)$ reproduces $\mathrm{PF}(s)$ at long distances.

Though the interaction potentials are extracted from actual protein structures, once they are determined, one can use them and eqs 3 and 6 to obtain PF functions with just sequence information. Generating mock structures according to eq 3 for the purpose of finding $\mathrm{PF}^0(s)$ requires three kinds of input: the Gaussian constraints, the radial densities of residues, and the atom-number densities of proteins. The Gaussian constraints are already determined in extracting the interaction potentials. The radial density, specific for each type of residue, can be predetermined from actual protein structures. The atom-number density of a protein is determined by the total number of residues. In this way of implementing eq 3 and applying eq 6, one "predicts" the PF function, since that quantity is obtained using just sequence information.

Experimental Section

List of Proteins. The 243 proteins used in this study are a subset of a set of 322 nonhomologous proteins used in a previous study for investigating the tendency of polar side chains to form hydrogen bonds with the peptide backbone.[13] The selected proteins are monomeric. The PDB names along with the residue numbers of the 243 proteins are as follows: 1aaj (105), 1aak (149)*, 1aba (86), 1abk (211)#, 1abt (74), 1ace (501)#, 1acx (106)*, 1add (349)#, 1ads (314)#, 1aep (153)*, 1afn (326)#, 1ak3 (226)#, 1ald (359)#, 1alk (445)#, 1aoz (551), 1apa (261)#, 1aps (98), 1arb (259)*, 1atr (380)#, 1avh (318)#, 1ayh (211)*, 1baa (241)#, 1bbh (131)*, 1bet (107)#, 1bge (153)*, 1bgh (85)*, 1bib (277)#, 1bnh (456), 1bov (68), 1brd (168)*, 1btc (489)#, 1c2r (115)*, 1c5a (65), 1cde (208)*, 1cew (108)*, 1chm (398), 1chr (365)#, 1cid (177)#, 1cmb (104)*, 1cob (149)*, 1col (197)*, 1crl (530), 1csc (424)#, 1ctf (68), 1ctm (246), 1cy3 (117)*, 1dhr (232)#, 1dsb (187)*, 1dxt (146)*, 1eaf (235)#, 1ede (309)#, 1end (137)*, 1f3g (145)*, 1fbp (316)#, 1fc1 (206), 1fha (151)#, 1fia (77)*, 1fkf (107)*, 1fnr (296)#, 1fxi (96), 1gal (576), 1gd1 (334)#, 1gdh (308), 1gf1 (70), 1gky (186)*, 1glt (296)#, 1gmf (121)*, 1gof (628), 1gox (349)#, 1gst (217)#, 1hbg (146)*, 1hc6 (632), 1hfh (120)#, 1hip (81), 1hmy (323)#, 1hoe (74), 1hpl (447), 1hrh (112)*, 1hsl (236)#, 1hst (73), 1huw (166)*, 1hyp (75), 1ifc (131)*, 1isu (60), 1lap (481), 1lba (145)*, 1le2 (144)#, 1lfb (77)*, 1lfi (684), 1lga (338)#, 1lis (129)*, 1lz1 (130)*, 1mat (236)*, 1mbc (153)*, 1mdc (129)*, 1mol (94)*, 1ms2 (127)#, 1msb (114)*, 1mup (150)*, 1nar (287)#, 1ndk (147)*, 1nip (248)#, 1npx (446), 1nsb (390)#, 1omf (336)#, 1omp (363)#, 1onc (103)*, 1osa (148)#, 1ova (363)#, 1paz (120)*, 1pda (292)#, 1pdg (83)#, 1pgd (467), 1pgx (67)*, 1phh (393)#, 1pi2 (59), 1pii (449), 1pkp (143)*, 1poc (134)*, 1poh (85), 1pox (582), 1psp (105)*, 1pts (118)*, 1r69 (63), 1rbp (174)*, 1rcb (129)*, 1rec (181)*, 1rfb (118)#, 1rib (340)#, 1rpa (340)#, 1rpr (63)*, 1rve (243)#, 1sac (203)*, 1sbp (304)#, 1sha (102)*, 1sim (378)#, 1snc (135)*, 1sry (419), 1sto (207)#, 1tah (315)#, 1tbp (180)#, 1tca (315)#, 1ten (89)*, 1tie (143)*, 1tlk (103)*, 1tml (286)#, 1tnf (150)*, 1tpk (87), 1tpl (422), 1trb (314)#, 1trk (675), 1ubq (75), 1ula (287)#, 1utg (70), 1vmo (161)*, 1vsg (360), 1ycc (106), 256b (106)*, 2apr (321)#, 2aza (128)*, 2bb2 (180), 2ca2 (255)#, 2cas (545), 2ccy (126)*, 2cmd (312)#, 2cpl (162)*, 2cpp (402)#, 2cyp (265)#, 2er7 (180), 2fb4 (226), 2fxb (81), 2gbp (309)#, 2gcr (173)*, 2hhm (269)#, 2hip (71), 2hpd (455), 2il8 (71)*, 2liv (342)#, 2mnr (355)#, 2pab (112)*, 2pcy (99), 2pia (319)#, 2pmg (561), 2pna (104)*, 2pol (364), 2por (297)#, 2rhe (113)*, 2sar (96), 2sas (184)*, 2scp (173)*, 2sga (165)*, 2sn3 (65), 2snv (148)*, 2spc (106), 2stv (184)#, 2tgi (112)#, 2tmd (723), 2trx (108), 2ts1 (291)#, 2tsc (263)#, 2wrp (101)*, 3adk (193)#, 3b5c (83), 3blm (256)#, 3cd4 (176)#, 3chy (127)*, 3cox (468)#, 3ebx (62), 3eca (324)#, 3gap

(208)#, 3grs (457), 3hhr (187)#, 3lzm (164)*, 3sdp (184)*, 3tgl (264)*, 4bp2 (114)*, 4cla (195)#, 4cpv (108), 4dfr (151)*, 4enl (433)#, 4fgf (123)*, 4fxn (138)*, 4mt2 (60), 4pfk (310), 4ptp (172)*, 4sbv (198)#, 5cpa (305)#, 5fd1 (106), 5hvp (92)*, 5p21 (165)*, 6ldh (329)#, 6tmn (313)#, 6xia (384), 7aat (401), 7rsa (124)*, 7tim (244)#, 8acn (746), 8adh (371)#, 8cat (496), 8i1b (146)*, 9pap (209)*, 9rnt (102), and 9wga (160)#. In all our work, residues with missing atoms were discarded and they are not counted in calculating the residue number. Hence the number shown is an underestimate. Of the 243 proteins, 142 have less than 220 residues and 101 have longer chains. In the above list, an asterisk means a protein has a radius of gyration between 13 and 17 Å (a total of 85 proteins) and a pound symbol means a proteins has a radius of gyration between 17 and 22 Å (a total of 87 proteins).

Representation of a Residue by a Single Atom. It is impractical and perhaps not useful to explicitly include every atom of a residue in studying residue−residue PFs. In other studies, residues have been represented by CB[1,8] or side-chain centroids.[3,8,14] We chose to represent a residue by one of the outmost side-chain atoms that characterize that type of residue, because these "tip" atoms are the ones that are most frequently involved in residue−residue contacts. They are as follows: Leu CD1, CD2; Val CG1, CG2; Ile CD1; Phe CZ; Cys SG; Met CE; Ala CB; Gly CA; Pro CD; Asp OD1, OD2; Glu OE1, OE2; Arg NH1, NH2; Lys NZ; Ser OG; Thr OG1; Asn ND2; Glu NE2; His NE2; Tyr OH; and Trp NE1. For a residue with a symmetrically branched tip (e.g., Leu with CD1 and CD2), the statistics of the PF function were improved by averaging those obtained for the individual tip atoms.

For the purpose of obtaining the radial density, the representation of a residue was somewhat different. Some additional atoms were put into the lists of tip atoms. These include Ile CG2; Phe CE1, CE2; Pro CB, CG; Arg NE; Asn OD1; Gln OE1; and His ND1. A residue now is represented by the tip atom that has the shortest radial distance.

Calculation of PF Functions. The calculation of PF functions followed a simple procedure. For each protein, the coordinates of the n_i type-i residues and n_j type-j residues were either obtained from the PDB file or reconstructed. The distances of the $n_i n_j$ pairs were then calculated. Out of these some were discarded, either because the sequence separation between a pair was ≤ 50 or because one partner was in the periphery of the protein. The remaining distances from all the proteins in a particular set were pooled and the total numbers in individual distance bins were obtained. For the PF functions displayed in Figures 1 and 6, the bins have a width of $\Delta s = 0.5$ Å and are centered at $s = 0.25, 0.75, 1.25, ...,$ Å.

Parametrization of Chain Constraints. For the purpose of parametrizing the chain constraints, the PF function of the Ser-Asn pair from the 243 proteins was used. Now the bin width was increased to $\Delta s = 2$ Å (for better statistics) and the bins were centered at $s = 1, 3, 5, ...,$ Å. All the Ser-Asn pairs from the 243 proteins were found and segregated into eight groups according to chain separations. The chain separations of the eight groups were 0−5, 6−10, 11−20, 21−50, 51−100, 101−150, 151−200, and >200. For each group the PF function was obtained.

The corresponding PF^0 function was then obtained from reference structures generated according to eq 3 with the radial distances kept at their values in actual structures of the proteins. The chain constraints were enforced by restraining the residue−residue distance according to a Gaussian probability function (see eq 4). We used the same Gaussian constraint in each of the above intervals of the chain separation Δ. The mean b and standard deviation σ for each interval were assigned initial values and were used to generate residue positions and reconstruct PF functions. Then the b and σ values for $\Delta \in [0, 5]$ were adjusted to best reproduce the PF function of the $\Delta \in [0, 5]$ group. This was followed by adjustment for $\Delta \in [6, 10]$, by adjustment for $\Delta \in [11, 20]$, and so on. The adjustments in later intervals hardly affected the reconstructed PF functions in earlier intervals. A set of b and σ values were easily obtained that yielded accurate reproduction of the PF functions in all the eight intervals of the chain separation.

Generation of Reference Structures. The residues positions of the reference structures were generated according to eq 3, with radial distances taken from the actual structure of a protein while enforcing the chain constraints. To that end, residue positions were generated sequentially. For generating the position of the kth residue, constraints between that residue and residues 1 through $k - 1$ were invoked. Initially 100 trial positions were generated. These had the predetermined radial distance but random polar and azimuthal angles θ and ϕ (with $\cos\theta$ taken from a uniform distribution between -1 and 1 and ϕ taken from a uniform distribution between 0 and 2π). The weight of each trial position was then calculated as the product of the Gaussian constraints between that position and the positions of the first $k - 1$ residues. The position of the kth residue was selected from the 100 trial positions according to their weights. Specifically, we randomly sampled the 100 trial positions and compared the normalized weight of each sampled position (by the maximum of the 100 weights) to a random number that is uniformly distributed between 0 and 1. A trial position was selected when the random number was smaller. To improve statistics, 10 copies of reference structures were generated for each protein and the final PF function was divided by that number.

It should be noted that, while the position of each residue was generated, distances to all other residues were constrained. Once the positions for all the residues were generated, we could then choose to calculate the PF function for pairs within a certain range of chain separations. Indeed, other than the Ser-Asn pair when used for the purpose of parametrizing the chain constraints, we discarded pairs with chain separation ≤ 50 in calculating PF functions.

Generation of Mock Structures with Residues Distributed According to Radial Densities. For the purpose of implementing eq 3 using just sequence information, we had to select the radial distance of a residue from the distribution $\rho_i(r)D(r)$. This was implemented as follows. First a trial value of r was generated according to the Gaussian function $D(r)$ with peak position $r_M = 13.4(N/100)^{1/3}$ Å and width $d_1 = 4.94$ or 6.42 Å [depending on $R_G \in (13, 17)$ or $(17, 22)$]. A lower bound of 0 and upper bound of 35 or 40 Å were placed. Then $\rho_i(r)$ normalized by its maximum within the bounds was compared to a random number that is uniformly distributed between 0 and 1. The trial value was accepted when the random number was smaller. The rest of the procedure described under Generation of Reference Structures was then followed to enforce chain constraints and generate mock structures.

Results

Reproduction of Pair Frequencies at Long Distances. The key to the elucidation of the physical basis of PFs came from the recognition that the radial distance is important for representing the inhomogeneous distributions of residues. The role of the radial distance was established when we tried to reproduce

PFs from eq 3 by keeping the actual radial distance of each residue. The θ and ϕ angles were randomly generated. By selecting a set of parameters for the constraints due to chain connectivity (see below), PFs were indeed accurately reconstructed. Figure 1 illustrates the agreement between 15 representative PF functions obtained from a set of 243 nonhomologous proteins and those reconstructed from fixed radial distances. The deviations at $s < 7.5$ Å in some cases reflect interactions between residues and will be further examined later.

Chain Constraints. We used the actual PF function of two particular residues from the 243 proteins to parametrize the chain constraints. Ser and Asn were chosen because they are more or less "neutral" with regard to any preference for the inside or the exterior of a protein and thus interference from inhomogeneous residue distributions is least likely. We decided to partition the full range of the sequence separation Δ into a number of intervals and use the same Gaussian constraint in each interval. The intervals were 0−5, 6−10, 11−20, 21−50, 51−100, 101−150, 151−200, and > 200. The values of the mean b of the Gaussian constraints in these intervals were found to be 5, 17, 24, 25, 32, 42, 45, and 60 Å, respectively. The values of the standard deviation σ are 5, 7, 13, 22, 23, 25, 25, and 30 Å, respectively. This set of constraints allowed for the reconstruction of the Ser-Asn PF function not only in the global sense, i.e., when the PF function was obtained for all the Ser-Asn pairs, but also in the local sense, i.e., when the PF function was obtained for only those Ser-Asn pairs with sequence separations within each of the above intervals.

The reconstruction of all the PF functions shown in Figure 1 used the same set of Gaussian constraints. However, only the pairs with sequence separations > 50 are included. We found that the Gaussian constraints parametrized on the Ser-Asn pair do not work for all pairs when the sequence separation is less than 50. In other words, there does not appear to exist a set of universal constraints for pairs close along the sequence. This is very reasonable since distances between such pairs are likely dictated by secondary structures and different types of residues have different preferences for secondary structures. All the PF functions presented in this paper exclude pairs with sequence separation ≤ 50. We emphasize that, except for the Ser-Asn pair, the reconstruction of all the other PF functions shown in Figure 1 did not involve adjusting any parameters.

Interaction Potentials. According to eq 6, the interaction potential for a pair of residues can be determined by taking the ratio of the actual and reconstructed PF functions. To most accurately extract the pair interaction potential, we excluded another small group of pairs in calculating the PF functions shown in Figure 1. These were the ones with at least one partner in the periphery of a protein. Peripheral residues were defined as those with radial distances that exceed the expected radius, $16.9(N/100)^{1/3}$ Å, of a protein with N residues and they constituted about 10% of all residues. These residues are most displaced by angular scrambling and thus the reconstructed pair distances involving them are least reliable.

A major concern is whether the extracted interaction potential is an intrinsic property of the residues or depends on the particular set of proteins used. We thus divided the 243 proteins into two smaller subsets: one has 142 proteins with chain lengths less than 220 residues, and the other has 101 proteins with chain lengths greater than 220 residues. The interaction potentials extracted from the two subsets of proteins for the Leu-Ile, Arg-Asp/Glu, and Lys-Asp/Glu pairs are shown in Figure 2a. Very good agreement is seen. The interaction potentials extracted thus appear to be an intrinsic property. The interaction potentials extracted from the full set of proteins for the Asp-Glu, Arg-Lys, Leu-Ser, Val-Cys/Met, and Tyr-Asp/Glu

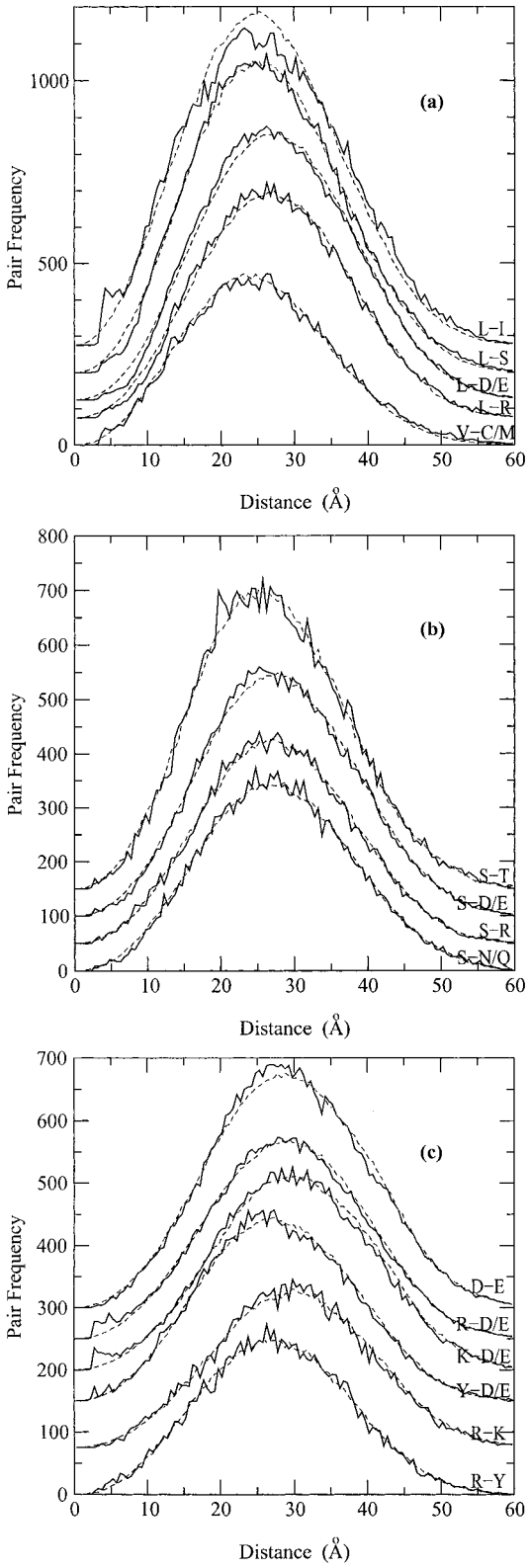

Figure 1. PF(s) obtained from 243 proteins (solid curves) and PF^0(s) reconstructed using radial distances of residues from protein structures (dashed curves). (a) Pairs involving Leu or Val. (b) Pairs involving Ser. (c) Pairs involving charged residues. Each curve actually has a zero value at $s = 0$. In each panel all curves but one are shifted upward by different amounts for clarity. Val-Cys and Val-Met pairs have similar PF functions and are combined, as are X-Asp and X-Glu pairs and X-Asn and X-Gln pairs.

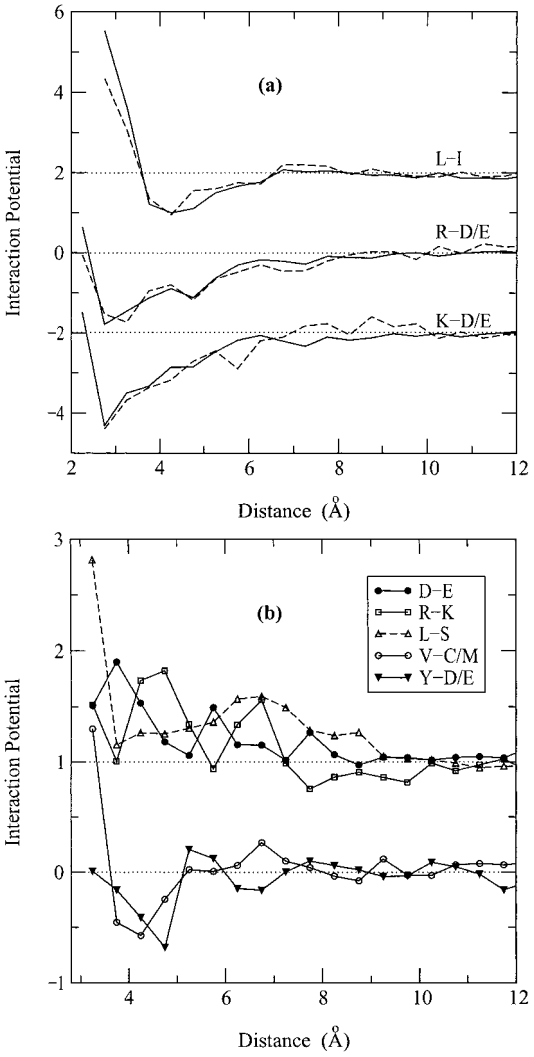

Figure 2. Interaction potentials of residue pairs (in units of kBT). (a) Comparison of results for Leu-Ile, Arg-Asp/Glu, and Lys-Asp/Glu pairs extracted from the 101 larger proteins (solid curves) and the 142 smaller proteins (dashed curves). (b) Results of five other pairs extracted from all the 243 proteins. Some curves are shifted upward or downward for clarity, with the amounts indicated by the dotted lines.

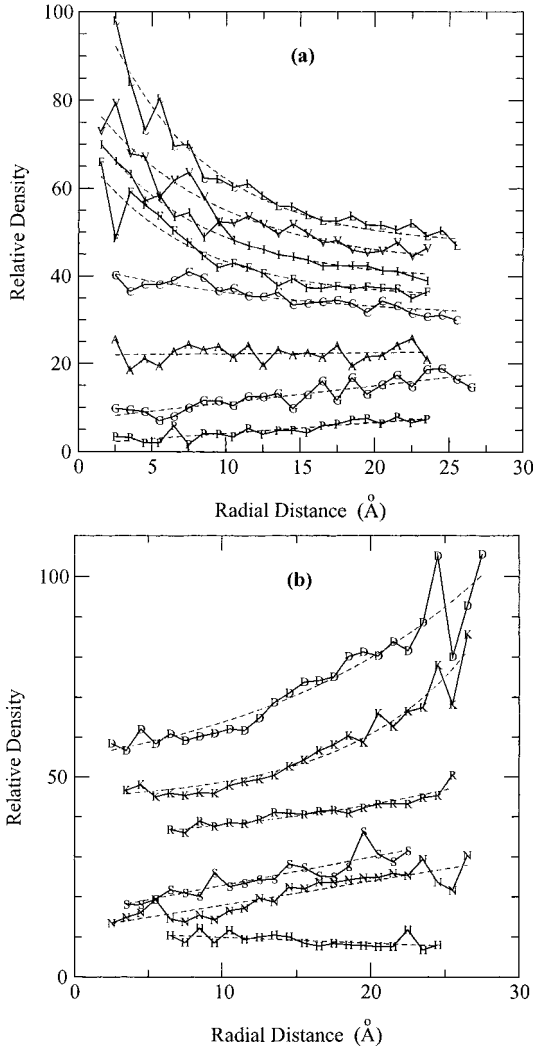

Figure 3. Radial densities of the 20 types of residues in the RG13−17 subset of proteins. (a) Nonpolar residues and Ala, Gly, Pro. (b) Polar and charged residues and His/Tyr/Trp. For each residue the one-letter code is drawn at each data point. Some curves are due to the pooling of two or three residues. Hence C actually means Cys/Met, D means Asp/Glu, S means Ser/Thr, N means Asn/Gln, and H means His/Tyr/Trp. The fitting of the radial densities is shown as dashed curves. Curves are shifted upward different amounts for clarity.

pairs are shown in Figure 2b. They are much smaller in magnitude than those of the three residue pairs shown in Figure 2a.

Radial Densities of Residues. We expected that the size of a protein would affect the distributions of residues. In particular, in a small protein the relative density of a nonpolar residue would exhibit a sharp decay as the radial distance increases. A large protein has a core region that is isolated from the solvent. The relative density of a nonpolar residue there would exhibit at most a moderate decay. Therefore we decided to obtain the radial densities on proteins with similar sizes. Specifically, two subsets of the 243 proteins were selected. One has 85 proteins with the radius of gyration R_G between 13 and 17 Å, and the other has 87 proteins with R_G between 17 and 22 Å.

The radial densities of the 20 types of residues in the R_G-13−17 subset are displayed in Figure 3. The spherical shells in which the relative densities were calculated are centered at $r =$ 0.5, 1.5, 2.5, ..., Å and have a thickness of 1 Å. The value displayed is the ratio between the number of a particular type of residue and the number of all atoms in each shell, magnified by 1000. The statistics of the results near the protein centers and surfaces were poor and are not displayed for clarity. Residues similar in nature show similar radial density profiles and their radial densities were pooled for better statistics. Four pairs of residues were pooled: Cys and Met, Asp and Glu, Asn and Gln, and Ser and Thr. The three aromatic residues, His, Tyr, and Trp, were also pooled.

The radial densities of the nonpolar residues (Leu, Val, Ile, Phe, and Cys/Met) can be fitted to a decaying exponential function $a_0 \exp(-r/a_1)$ quite well. The fitted values of (a_0, a_1) are (65.14, 6.84), (40.64, 8.78), (37.79, 7.42), (34.31, 7.06), and (12.40, 14.71), respectively. For charged residues (Lys, Arg, and Asp/Glu), a growing exponential function $a_0[\exp(r/a_1) - 1]$ gives reasonable fitting. The fitted values of (a_0, a_1) are (1.58, 8.34), (4.27, 18.91), and (10.36, 16.34), respectively. The radial densities of all other residues were fitted to a linear function $a_0 + a_1 r$. The fitted values of (a_0, a_1) are as follows: Ala (9.90, 0.03); Gly (4.21, 0.39); Pro (1.73, 0.25); Ser/Thr (5.76, 0.71); Asn/Gln(1.87, 0.60); and His/Tyr/Trp (11.43, −0.15).

The radial densities of the 20 types of residues in the R_G-17−22 subset are displayed in Figure 4. For the nonpolar residues, an exponential decay after $r = 10.5$ Å is still seen. That part of the radial density was thus fitted to the function a_0

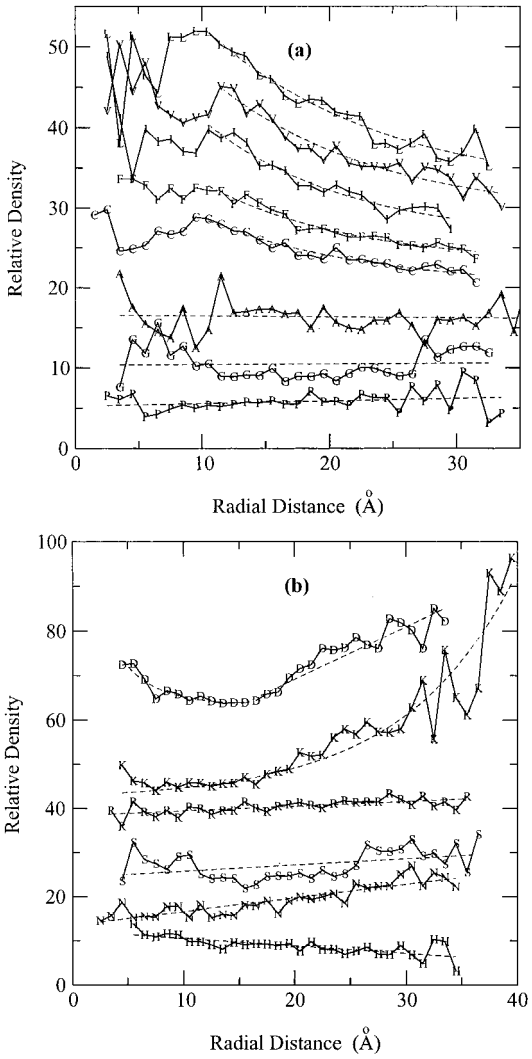

Figure 4. Radial densities of the 20 types of residues in the R_G17-22 subset of proteins. See caption of Figure 3 for details.

$\exp[-(r - 10.5)/a_1]$. The fitted values of (a_0, a_1) for Leu, Val, Ile, Phe, and Cys/Met are (19.87, 13.69), (15.93, 12.76), (14.38, 11.33), (9.80, 11.27), and (8.37, 12.82), respectively (the starting point for Val was actually 11.5 Å). Inside $r = 10.5$ Å the radial densities display fluctuations around the fitted values of a_0. We represented that portion of the radial density simply by a_0. The near-constant value of the radial density of a nonpolar residue in the core of the subset of larger proteins confirms our earlier expectation.

The radial densities of the charged residues also display more complicated behavior than in the R_G13-17 subset. For Lys the growing exponential function $a_0[\exp(r/a_1) - 1]$ with $a_0 = 1.01$ and $a_1 = 10.18$ still gave good fitting. For Arg we had to use a linear function $a_0 + a_1 r$ instead. The fitted values are $a_0 = 3.41$ and $a_1 = 0.11$. Asp/Glu actually show an initial decay. The radial density was fitted to $a_0 \exp(-r/a_1) + a_2$ with $a_0 = 42.0$, $a_1 = 3.11$, and $a_2 = 8.52$ for $r < 15.5$ Å. For $r > 15.5$ Å we used a linear function $a_4 + a_5(r - 15.5)$ with $a_4 = a_0 \exp(-15.5/a_1) + a_2 = 8.81$ to ensure continuity and $a_5 = 1.18$. The radial densities of all other residues were fitted to a linear function $a_0 + a_1 r$. The fitted values of (a_0, a_1) are as follows: Ala (11.61, −0.01), Gly (10.33, 0.01), Pro (5.25, 003), Ser/Thr (14.35, 0.14), Asn/Gln(5.50, 0.31), and His/Tyr/Trp (12.22, −0.17).

Atom-Number Density of Proteins. The atom-number densities of the R_G13-17 and R_G17-22 subsets are shown in Figure 5. A Gaussian function $d_0 \exp[-(r - r_M)^2/2d_1^2]$ appears to fit both density profiles quite well. The fitted values of (d_0, r_M, d_1) are (7615.3, 14.38, 4.94) for the R_G13-17 subset and (11995.1, 18.78, 6.42) for the R_G17-22 subset. Up to $r = 10$ Å, the two density profiles agree with each other to within 5% and can be fitted to $a_0 r^2$ even better, indicating that the core regions of the proteins in both subsets are well packed.

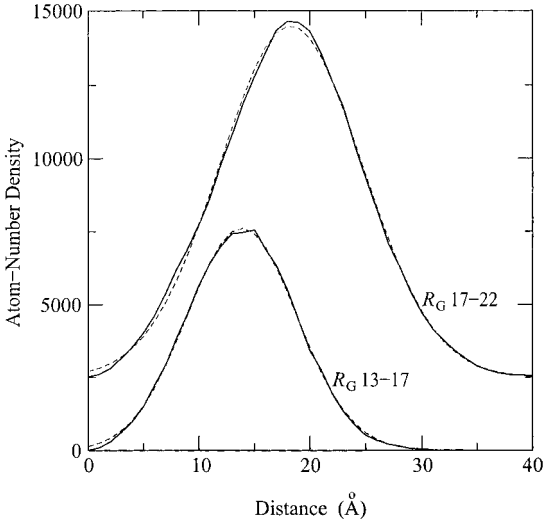

Figure 5. Atom-number densities of the R_G13-17 and R_G17-22 subsets of proteins (solid curves) and their fitting to a Gaussian function (dashed curves). The R_G17-22 curves are shifted upward by 2500 units for clarity.

The results shown in Figure 5 are those when all the proteins in each subset are pooled. The range of R_G in each subset is relatively wide (to ensure reasonable statistics for the radial densities). The atom-number density profiles of individual proteins will vary somewhat and will be different from those displayed in Figure 5. To account for this variation, we individualized the peak position r_M of the Gaussian function for each protein. For a protein with atom-number density given by $D(r) = d_0 \exp[-(r - r_M)^2/2d_1^2]$, the radius of gyration is $R_G = (r_M^2 + d_1^2)^{1/2} \approx r_M$. We thus modeled r_M by the radius of gyration expected for a protein with N residues: $f(N/100)^{1/3}$. Linear regression analysis indicated that the R_G values of the 243 proteins and $(N/100)^{1/3}$ are highly correlated and the coefficient f is bounded by 13 and 14 Å. Hence there is a small degree of freedom in selecting an f value to set the peak positions of the atom-number densities for individual proteins. No variation was allowed for the width d_1 of the atom-number densities for the proteins within each subset ($d_1 = 4.94$ for the R_G13-17 subset and 6.42 for the R_G17-22 subset).

Predicted Pair Frequencies. The reconstruction of the PF functions displayed in Figure 1 is a special implementation of eq 3. It used structural information of the proteins, namely, the radial distances of the residues. The reconstruction by such a means cannot be called a prediction. However, once this reconstruction yielded results for the chain constraints and the interaction potentials, we can use them together with independently obtained radial densities of residues and atom-number densities of proteins to predict PF functions. Of course we have yet to select a precise value of the parameter f between 13 and 14 Å for fixing the peak positions of the atom-number densities for individual proteins.

The prediction works quite well with $f = 13.4$ Å. In Figure 6 we show the PF functions for the Leu-Ile, Ser-Asn/Gln, and Arg-Asp/Glu pairs obtained from the R_G13-17 and R_G17-22 subsets of proteins and those predicted by eqs 3 and 6. Very

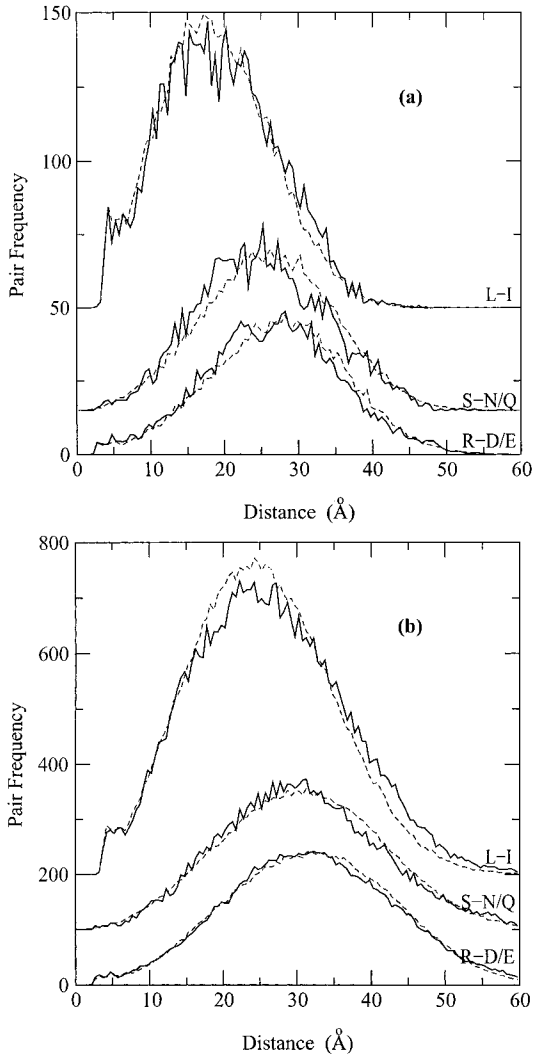

Figure 6. PF functions of Leu-Ile, Ser-Asn/Gln, and Arg-Asp/Glu pairs from two subsets of proteins (solid curves) and those predicted by eqs 3 and 6 (dashed curves). (a) RG13−17 subset. (b) RG17−22 subset. Some curves are shifted upward for clarity. The interaction potentials of used were extracted from the 243 proteins (similar to those shown in Figure 2a). For the Ser-Asn/Gln pair, the interaction potential extracted was not statistically significant and no interaction potential was included in the predicted PF function.

good agreement is seen. As before, pairs between residues with sequence separation ≤ 50 were excluded, but otherwise all pairs were included (i.e., even pairs involving peripheral residues were kept).

Universality of Extracted Interaction Potentials and Radial Densities. To what extent are the extracted interaction potentials and radial densities universal (i.e., invariant among different sets of proteins)? To address this question, we applied our extraction procedures to the PDB Select list of representative proteins with 25% sequence identity.[15] The 874 chains with ≥90 residues were retained. In extracting the interaction potentials, the chain constraints listed earlier were used. Hence we did not adjust a single parameter. In Table 1 we compare the interaction potentials extracted from the earlier 243 proteins and those extracted from the new set of 874 protein chains. The four interaction potentials with the largest magnitudes are shown. Results from the two protein sets are very consistent. The root-mean-square deviation (RMSD) is only $0.16k_BT$.

We also found the radial densities of residues to be invariant in different protein sets. This is demonstrated in Figure 7 by the results for seven residues in proteins with the radius of

TABLE 1: Interaction Potentials (in Units of k_BT) Extracted from Two Sets of Proteins

distance s (Å)	Leu-Ile	Arg-Asp/Glu	Lys-Asp/Glu	Asp-Glu
2.75	3.10 (−0.34)[a]	−1.93 (0.03)	−2.14 (−0.17)	1.02 (0.20)
3.25	1.44 (0.35)	−1.46 (0.05)	−1.60 (−0.19)	0.51 (−0.18)
3.75	−0.72 (−0.11)	−1.09 (−0.09)	−1.26 (−0.21)	0.90 (0.01)
4.25	−1.05 (−0.06)	−0.89 (−0.18)	−0.93 (−0.15)	0.53 (−0.25)
4.75	−0.71 (−0.02)	−1.13 (−0.16)	−0.75 (0.06)	0.18 (−0.13)
5.25	−0.48 (−0.08)	−0.61 (0.16)	−0.41 (0.18)	0.06 (−0.06)
5.75	−0.33 (−0.02)	−0.36 (−0.18)	−0.40 (−0.14)	0.49 (0.33)
6.25	−0.24 (−0.06)	−0.17 (−0.07)	−0.11 (0.11)	0.16 (−0.08)
6.75	0.07 (0.09)	−0.28 (−0.06)	−0.16 (−0.16)	0.15 (0.11)
7.25	0.04 (−0.05)	−0.27 (−0.18)	−0.23 (−0.02)	0.02 (−0.23)

[a] The number before the parentheses is the potential extracted from the 243 proteins. The difference between that number and the potential extracted from the 874 protein chains is given in the parentheses.

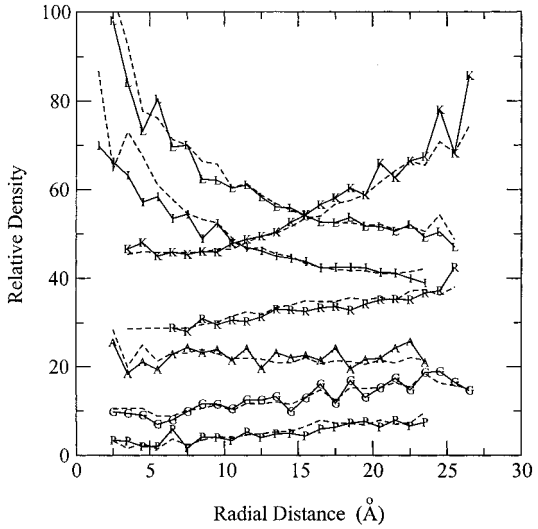

Figure 7. Radial densities of Leu, Ile, Lys, Arg, Ala, Gly, and Pro in proteins with radius of gyration between 13 and 17 Å. The solid curves are the same as those displayed in Figure 3, except that the Arg curve is shifted downward slightly for clarity. These represent the results from the R_G13-17 subset of the 243-protein set. The dashed curves are the results from the R_G13-17 subset of the 874 PDB Select protein chains.

gyration between 13 and 17 Å. The solid curves are the same as those displayed in Figure 3 and represent the radial densities obtained from the 85-protein R_G13-17 subset of the 243-protein set. The dashed curves represent the radial densities obtained from the R_G13-17 subset of the 874 PDB Select protein chains. The latter subset has 333 protein chains. No statistically significant differences between the two sets of results can be detected.

Structural Alignment of Side-Chain Tip Atoms. In structural alignments of proteins, the corresponding backbone atoms are generally much closer to each other than the atoms of corresponding side chains. This is especially true for proteins with low sequence identities. The tip atoms, being farthest from the back, may appear to have a disadvantage in structural alignment and, by implication, in sequence-structure threading. We therefore investigated the use of side-chain tip atoms in structural alignment. Specifically, we wanted to compare the use of tip atoms to the use of CB atoms. For this purpose we chose eight protein pairs that were studied in a recent threading work by Mirny and Shakhnovich[7] (to be referred to as MS).

The structural alignment between proteins A and B was obtained by optimizing a distance-based score:

$$F = \sum_{i_A, i'_A} E(|\mathbf{r}_{i_A} - \mathbf{r}_{i'_A}|, |\mathbf{r}_{j_B} - \mathbf{r}_{j'_B}|) \qquad (7)$$

TABLE 2: RMSDs and Lengths of Structure−Structure and Sequence−Structure Alignments[a]

	structure−structure alignment[c]			sequence−structure alignment[d]		
protein pair[b]	Dali−CA	contact−tip	contact−CB	eq 10		MJ96
1ten-1ten				0.0 (89)	0.0 (84)	3.27 (84)
1ten-1fnf	0.97 (89)	0.97 (89)	0.97 (89)	1.38 (88)	0.88 (82)	2.64 (82)
1ten-3hhrB	2.00 (87)	2.04 (86)	2.11 (87)	7.51 (88)	5.70 (80)	4.26 (80)
1ten-1hnf	2.77 (74)	3.44 (72)	2.80 (74)	14.78 (77)	13.75 (73)	16.89 (73)
1ten-1cidA	3.47 (83)	3.74 (83)	3.57 (84)	15.37 (89)	13.15 (69)	13.96 (69)
1ten-1tit	2.99 (75)	4.89 (82)	4.56 (85)	6.04 (86)	4.93 (75)	5.39 (75)
1ash-1ash				0.0 (146)	0.0 (141)	0.77 (141)
1ash-2fal	1.85 (140)	1.95 (140)	2.01 (140)	3.47 (142)	2.44 (123)	3.44 (123)
1ash-2hbg	2.68(137)	2.76 (137)	3.34 (137)	6.53 (140)	5.87 (115)	5.87 (115)
1ash-1colA	3.23(112)	3.85 (113)	6.39 (123)	10.15 (123)	7.04 (102)	15.30 (102)

[a] RMSDs are in angstroms and alignment lengths are the number of aligned residues. The result for each alignment is reported in the form of RMSD (aligned length). [b] Each protein is specified by the PDB entry name (with chain number). [c] The structure−structure alignment is done by using the Dali scoring function (eq 8) with distances measured between CA atoms or the contact scoring function (eq 9) with distances measured either between side-chain tip atoms or between CB atoms. [d] The sequence−structure alignment is either done in the present study with the energy function of eq 10 or by Mirny and Shakhnovich[7] with the potential function of Miyazawa and Jernigan.[14] MS chose to report their alignment results by calculating the RMSD over a portion of the aligned residues. To make it easy to compare with their results, we report both the RMSD calculated over all the aligned residues and the RMSD over the same number of residues for which MS calculated their RMSD. These are shown side by side under the heading of eq 10.

where i_A and i'_A refer to residues of protein A and j_B and j'_B are the corresponding residues in protein B in a particular alignment. The Dali scoring function introduced by Holm and Sander[16] is

$$E(s_A,s_B) = (0.2 - |s_A - s_B|/d) \exp[-(d/20)^2] \qquad (8)$$

where the distance between two residues is measured between the CA atoms and $d = (s_A + s_B)/2$. Following MS, we used a contact scoring function

$$E(s_A, s_B) = 1 \quad \text{if both } s_A \text{ and } s_B < s_{cut} \qquad (9)$$
$$= 0 \quad \text{otherwise}$$

where $s_{cut} = 8$ Å. The weaker distance dependence of eq 9 makes it more suitable for aligning side-chain atoms, which are not superimposed as well as the corresponding CA atoms. MS applied eq 9 to the distance between the CB atoms of two residues. We now apply it to the distance between the side-chain tip atoms. The same Monte Carlo procedure with simulated annealing is used to optimize the score in eq 7.

In Table 2 we compare the results of structure−structure alignment on the eight protein pairs by using the contact scoring function (eq 9) with distances measured either between side-chain tip atoms or between CB atoms. In general the results by the two kinds of distances are comparable and, as expected, are somewhat worse than those of the Dali scoring function with distances measured between CA atoms. A notable difference occurs in the alignment of 1ash and 1colA, where the RMSD by using the CB−CB distance is much larger than that by using the tip−tip distance. We conclude that, for the purpose of structure comparison, the tip−tip distance is not inferior to the commonly used CB−CB distance.

Use of Residue−Residue Interaction Potentials in Threading. Although the main aim of this paper is to elucidate the physical basis of residue−residue pair frequencies and interaction potentials, because of the wide use of interaction potentials in sequence−structure threading, it is obviously interesting to see how our interaction potentials perform in threading experiments. Our preliminary work focused on two types of tests: template recognition in gapless threading and alignment accuracy of gapped threading.

Equation 5a suggests that the probability density for finding the residues in positions **X** can be cast into a Boltzmann distribution $\exp(-E/k_BT)$, with the energy function given by

$$\epsilon/k_BT = -\sum_{k=1}^{N} \ln[\rho_{i_k}(r_k)] + \sum_{k,k'=1}^{N} w_{i_k i_{k'}}(|\mathbf{r}_k - \mathbf{r}_{k'}|)/k_BT \qquad (10)$$

where i_k is the residue type (1−20) of the kth residue. This energy function is analogous to the Hamiltonian of electrons moving around a nucleus. Here the one-body potentials (contained in the first term) are not due to attraction by the nucleus but model the interactions of the residues with the solvent and account for the nonpolar in−charge out character. We used the energy function of eq 10 for threading. Specifically, we equated $\rho_i(r)$ to the fitted functions of the radial densities of the R_G13−17 proteins (shown in Figure 3). Only five distinct types of residue−residue interaction potentials were used. The potential for a pair of any two nonpolar residues (i.e., Leu, Val, Ile, Phe, Cys, and Met) was taken to be that for the Leu-Ile pair shown in Figure 2a. Four types of interaction potentials were used for charged groups: the ones between Arg and Asp or Glu and between Lys and Asp or Glu are shown in Figure 2a, and the ones between two identically charged residues were taken to be those for the Asp-Glu and the Arg-Lys pairs shown in Figure 2b. Four of these potentials are also listed in Table 1. The interaction potential for any other pair was set to zero. In threading, one matches the sequence of a query sequence with a target structure. It is not reasonable to emphasize the subtle difference between the interaction potential for the Leu-Ile pair and, say, that for the Val-Ile pair, since a Leu residue may be replaced by a Val residue in the actual alignment between the query and the target.

We carried out gapless threading on those PDB_Select protein chains with 90−220 residues (a total of 466 chains). A standard set of structures compiled by Maiorov and Crippen[17] was used as decoys. Of the 466 chains, 414 (or 89%) were correctly identified (i.e., had the lowest energy when threaded to its own structure). A similar test was carried out by Thoma and Dill[18] on 121 protein chains by using an energy function optimized iteratively. Their success rate was around 88%. Though the success rates of the two tests are similar, the procedures for creating and the resulting energy functions are quite different. We would like to think of our procedure as based on a physical understanding of protein structures rather than merely as a training process.

A major problem in current threading studies is the accuracy of the query-target alignment obtained by threading. MS tested the alignment accuracy of threading by the function of Miyazawa and Jernigan[14] and suggested that more accurate alignment may require energy functions with higher quality. Their results for threading 1ten to six targets and 1ash to four targets are shown in Table 2. We carried out threading of the two query proteins to the same targets by the Monte Carlo procedure with simulated annealing. Gaps were introduced by the stipulation that those residues in gaps (either on the query or on the target) do not contribute to the energy function (analogous to the way that gaps are treated in the structure−structure alignment). The alignment accuracy of threading by the energy function of eq 10 is reported in Table 2 in the form of RMSD. Of the 10 alignments, our energy function outperforms that of Miyazawa and Jernigan[14] in eight, yields an identical RMSD in one, and performs slightly worse in one. The better performance of our energy function is especially noticeable in cases where query and target structures deviate by less than 2 Å (e.g., 0 versus 3.3 Å RMSD for 1ten self-threading), but improvement is also observed when where query and target structures deviate much more (e.g, 7 versus 15.3 Å RMSD for threading 1ash to 1colA).

Discussion

Physical Basis of Pair Frequencies. We have accurately reconstructed and even predicted the PF functions (see Figures 1 and 6). This indicates that PFs can really be explained by inhomogeneous distributions of residues, finite sizes of proteins, chain connectivity, and residue−residue interactions at short distances. More importantly, it appears that the radial distance provides a simple and powerful means for modeling the inhomogeneous distributions of residues and the finite sizes of proteins.

Reference State for Determining Interaction Potentials. In the present study the interaction potentials are defined in terms of the ratio of the PF functions in the physical state and in a reference state. This is precisely the way potentials of mean force are defined in liquids. The construction of the reference state is based on a physical understanding of protein structures. This is that inhomogeneous distributions of residues, the finite sizes of proteins, chain constraints, and residue−residue interactions are major determinants. We took great care to ensure that in the reference state the influence of residue−residue interactions is eliminated but nothing more. This is verified by the fact that at long distances $PF^0(s)$ in the reference state reproduces $PF(s)$ in the physical state. The chain constraints required in constructing the reference state were parametrized by using the PF function of just one pair of residues: Ser-Asn. Without adjusting any parameters, the PF functions at long distances ($s > 7.5$ Å) for all the other 209 distinct pairs of residues can be reproduced by those of the reference state. This is a remarkable success in passing an important internal check.

Interaction Potentials. The internal check on the reference state and the finding that the resulting interaction potentials are invariant among different sets of proteins give us confidence to argue that these interaction potentials reflect intrinsic properties of the residue pairs. They extend up to a distance of ∼7.5 Å and are particularly large for pairs of nonpolar residues and pairs of oppositely charged residues. The behavior of the interaction potentials is consistent with our expectations of the different types of residues. Roughly speaking, $w_{ij}(s)$ represents the difference in free energy between two states. The final state is one in which a type *i* residue and a type *j* residue are separated by *s*. In the initial state, the two residues are well separated and each is located in an "average" environment in the protein. For two nonpolar residues (such as Leu and Ile), the contact between them in the final state provides additional isolation from the solvent and a negative interaction potential is in accord with the hydrophobic nature of the residues. In the case of two oppositely (or identically) charged residues, the electrostatic interaction between them in the final state gives rise to the negative (or positive) potential. For a polar residue located in an "average" environment in the protein, the nature of the environment perhaps is not much perturbed when another polar residue is brought in. Hence the interactions between polar residues are apparently weak. The interaction potential extracted for the Ser-Asn/Gln pair was not statistically significant (see Figures 1b and 6). Similar features of interaction potentials were reported in a recent work by Bahar and Jernigan.[19]

The behavior of the radial densities also agrees with our expectations of the 20 types of residues. The nonpolar residues (Leu, Val, Ile, Phe, Cys, and Met) show a decaying radial density because of hydrophobicity, whereas charged residues show a growing radial density because of desolvation cost. That the radial densities from two different sets of proteins are virtually identical suggests that these results are universal among proteins with similar sizes. These results are also in good accord with the spatial preferences of residues obtained by Prabhakaran and Ponnuswamy[20] using a cruder approach. These authors modeled each protein as an ellipsoid and used a scaled distance from the ellipsoid surface as the measure of the burial of a residue. Possibly our results may become even closer to reality if we consider not only the radial distance but also the distance to the actual surface of the protein in that particular radial direction.

Comments on Previous Work. That the interaction potentials extracted here extend to a distance of ∼7.5 Å is in contrast to earlier results indicating that residue−residue interactions extend to 10 Å and beyond.[21,22] Bryant and Lawrence[21] derived the interaction potential of a charged pair essentially by taking the ratio of the PF functions of that pair and a "reference" pair with Asp/Glu replaced by Asn/Gln. Figures 3 and 4 show that the radial densities of Asp/Glu and Asn/Gln are quite different. This difference will contribute to the difference in the two PF functions and thus result in systematic errors in the interaction potential obtained by Bryant and Lawrence. Sippl[22] obtained the PF function of backbone N and O atoms up to a separation of 12 Å. The PF function increased gradually from 4 to 12 Å. This mainly is due to the fact that $PF(s)$ includes all the pairs that fall into a spherical shell whose volume increases as s^2. However, Sippl chose to attach great physical meaning to the apparent "correlation" between backbone N and O atoms at large distances. It is difficult to envision any physical reason a backbone N atom will specifically react to a backbone O atom that is 12 Å away. The 7.5 Å cutoff found here for residue−residue interactions lends some support to the practice of including only "contact" pairs in building energy functions.[6,14,23]

The volume for finding residue pairs that contribute to $PF(s)$ is $4\pi s^2\Delta s$. At short distances (e.g., $s < 4$ Å) it is not unreasonable to use $4\pi s^2\Delta s$ as an approximation for $PF^0(s)$. In their recent work Bahar and Jernigan[19] essentially set $PF_{ij}0(s)$ to $n_{ij}s^2\Delta s$ for all distances (where n_{ij} is a normalization factor). This amounts to assuming that in the reference state residues are uniformly distributed throughout the solvent. In reality $PF_{ij}(s)$ [and thus $PF_{ij}{}^0(s)$] at long distances must decrease to zero because of the finite sizes of proteins and cannot increase as s^2. The interaction potential of Bahar and Jernigan is $E_{ij}(s) = -k_BT \ln [PF_{ij}(s)/n_{ij}s^2\Delta s]$. This function will not be a constant

even at long distances. To avoid this unphysical result, Bahar and Jernigan (and others in earlier studies, e.g., Sippl[24]) introduced an interaction potential for a pair of generic residues: $E_{XX}(s) = -k_BT \ln [\sum_{i \geq j = 1}^{20} PF_{ij}(s)/210 n_{ij} s^2 \Delta s]$. The interaction potential for an *ij* pair was then taken to be $\Delta E_{ij}(s) = E_{ij}(s) - E_{XX}(s)$. It is interesting that the $\Delta E_{ij}(s)$ results for different pairs of residues found by Bahar and Jernigan actually share similarities to the interaction potentials displayed in Figures 2a and 2b. While the "peculiar" shapes of proteins present obstacles for theoretical treatments (the nonspherical shape for our treatment and the finite size for the treatment of Bahar and Jernigan), the similarities between the interaction potentials from the two different approaches suggest that physically meaningful results can nonetheless be obtained.

DeBolt and Skolnick[5] have used a one-body potential similar to the radial density introduced here. Two differences are worth commenting. The distribution of a particular residue will depend on the size of the protein. We specifically took this into consideration by working with subsets of proteins with similar sizes. In the work of DeBolt and Skolnick, the one-body and pair potentials were treated as individual screening tools. Here we show that they model complementary aspects of residue properties and are integral parts of a single energy function (see eqs 5a and 10).

Implication for Structure Prediction. We have suggested that the interaction potentials reported in the present study are an intrinsic property of residues and have demonstrated that they indeed lead to improved alignment accuracy in sequence−structure threading. The utility of these interaction potentials in threading is under further investigation.

Acknowledgment. This work is supported in part by NIH Grant GM58187.

References and Notes

(1) Hendlich, M.;, Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J. *J. Mol. Biol.* **1990**, *216*, 167.
(2) Jones, D.T..; Taylor, W. R.; Thornton, J. M. *Nature* **1992**, *358*, 86.
(3) Bryant, S. H.; Lawrence, C. E. *Proteins: Struct., Funct., Genet.* **1993**, *16*, 92.
(4) Kocher, J. P.; Rooman, M. J.; Wodak, S. J. *J. Mol. Biol.* **1994**, *235*, 1598.
(5) DeBolt, S. E.; Skolnick, J. *Protein Eng.* **1996**, *9*, 637.
(6) Mirny, L. A.; Shakhnovich, E. I. *J. Mol. Biol.* **1996**, *264*, 1164.
(7) Mirny, L. A.; Shakhnovich, E. I. *J. Mol. Biol.* **1998**, *283*, 507.
(8) Rooman, M.; Gilis, D. *Eur. J. Biochem.* **1998**, *254*, 135.
(9) Skolnick, J.; Kolinski, A. *Science* **1990**, *250*, 1121.
(10) Sun, S. *Protein Sci.* **1993**, *2*, 762.
(11) Kolinski A, Skolnick J. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 338.
(12) Thomas, P. D.; Dill, K. A. *J. Mol. Biol.* **1996**, *257*, 457.
(13) Vijayakumar, M.; Qian, H.; Zhou, H.-X. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 497.
(14) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623.
(15) Hobohm, U.; Sander, C. *Protein Sci.* **1994**, *3*, 522.
(16) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *23*, 123.
(17) Maiorov, V. N.; Crippen, G. M. *J. Mol. Biol.* **1992**, *227*, 876.
(18) Thoma, P. D.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 11628.
(19) Bahar, I.; Jernigan, R. L. *J. Mol. Biol.* **1997**, *266*, 195.
(20) Prabhakaran, M.; Ponnuswamy, P. K. *J. Theor. Biol.* **1980**, *87*, 623.
(21) Bryant, S. H.; Lawrence, C. E. *Proteins: Struct., Funct., Genet.* **1991**, *9*, 108.
(22) Sippl, M. J. *J. Mol. Biol.* **1996**, *260*, 644.
(23) Huang, E. S.; Subbiah, S.; Tsai, J.; Levitt, M. *J. Mol. Biol.* **1996**, *257*, 716.
(24) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859.