Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation

Sanzo Miyazawa* and Robert L. Jernigan

Building 10, Room 4B-56, Laboratory of Mathematical Biology, DCBD, National Institutes of Health, Bethesda, Maryland 20205. *Received December 5, 1983*

ABSTRACT: Effective interresidue contact energies for proteins in solution are estimated from the numbers of residue–residue contacts observed in crystal structures of globular proteins by means of the quasi-chemical approximation with an approximate treatment of the effects of chain connectivity. Employing a lattice model, each residue of a protein is assumed to occupy a site in a lattice and vacant sites are regarded to be occupied by an effective solvent molecule whose size is equal to the average size of a residue. A basic assumption is that the average characteristics of residue–residue contacts formed in a large number of protein crystal structures reflect actual differences of interactions among residues, as if there were no significant contribution from the specific amino acid sequence in each protein as well as intraresidue and short-range interactions. Then, taking account of the effects of the chain connectivity only as imposing a limit to the size of the system, i.e., the number of lattice sites or the number of effective solvent molecules in the system, the system is regarded to be the mixture of unconnected residues and effective solvent molecules. The quasi-chemical approximation, that contact pair formation resembles a chemical reaction, is applied to this system to obtain formulas that relate the statistical averages of the numbers of contacts to the contact energies. The number of effective solvent molecules for each protein is chosen to yield the total number of residue–residue contacts equal to its expected value for the hypothetical case of hard sphere interactions among residues and effective solvent molecules; the expected number of residue–residue contacts at this condition has been crudely estimated by means of a freely jointed chain distribution and an expansion originating in hard sphere interactions. Each residue is represented by the center of its side chain atom positions, and contacts among residues and effective solvent molecules are defined to be those pairs within 6.5 Å, a distance that has been chosen on the basis of the observed radial distribution of residues; nearest-neighbor pairs along a chain are explicitly excluded in counting contacts. Coordination numbers, for each type of residue as well as for solvent molecules, are estimated from the mean volume of each type of residue and used to evaluate the numbers of residue–solvent and solvent–solvent contacts from the numbers of residue–residue contacts. The estimated values of contact energies have reasonable residue-type dependences, reflecting residue distributions in protein crystals; nonpolar-residue-in and polar-residue-out are seen as well as the segregation of those residue groups. In addition, there is a linear relationship between the average contact energies for nonpolar residues and their hydrophobicities reported by Nozaki and Tanford; however, the magnitudes on average are about twice as large. The relevance of results to protein folding and other applications are discussed.

Introduction

A complete treatment of protein conformations in solution requires inclusion of solvent effects. Solvent molecules interact with atoms in proteins not only in short-range interactions such as hydrogen-bond formation and van der Waals interactions but modify electrostatic interactions between protein atoms. Also the entropy of water molecules around protein molecules differs from that of bulk water by forming more ordered cagelike structures or binding to specific sites. As originally pointed out by Kauzmann,[1] hydrophobic interactions, which would occur explicitly because of the nonspecific solvent effects, might be a principal force in leading to a collapsed protein molecule. Hydrophobic energies have been evaluated, among other ways, as the free energy changes of transfer of amino acids from ethanol or dioxane to water[2] and of liquid hydrocarbons into water.[3-6] Chothia[7-11] evaluated the contributions of hydrophobic energy to the formation of secondary, tertiary, and quaternary structures by employing the estimates in the reference[2] quoted above for values of the hydrophobic energy of interfacial areas exposed to water. His and others[12] estimates indicate that the hydrophobic energies, or the solvent effects, are a major contributor to the energetics of protein folding, essentially because large surface areas of protein molecules become buried in the interior upon folding. However, there is the fundamental question of whether liquid hydrocarbons and the organic solvents can completely represent a protein interior.[13] Lee[14,15] has pointed out on the basis of a scaled particle theory that thermodynamic properties such as the partial molecular volume of the solute in dilute binary solutions[14] and the change in the Ben-Naim local standard chemical potential of a solute molecule upon transferring it from the gas phase to a liquid phase[15] depend significantly on both the packing density of pure solvent and the ratio of the size of the solvent molecule to that of the solute molecule. Then, he has claimed that an obvious major difference of the high packing density and solidlike rigidity of protein interiors from small nonpolar solvents and even simple polymers makes it difficult to justify using the transfer data generally in quantitative studies of protein folding. Thus, estimates of hydrophobic interactions which are specific to protein molecules would be desirable.

Protein folding processes include a wide range of protein conformations from denatured to native states. The conformational freedom of a protein is vast. This makes it difficult to simulate the whole process of protein folding, if all atoms of a protein and solvent molecules are to be included in a detailed energy calculation. The geometry of molecules and interaction potentials require some simplification. The principal purpose of the present work is to include solvent effects into effective interresidue contact energies, which can then provide a crude estimate of the long-range component of conformational energies. Tanaka and Scheraga[16] estimated contact energies by a method which may appear to be similar but ignores solvent and is different in essence from the present one; incidentally, their method yielded extremely large magnitudes for contact energies.

Here the effective contact energies between residues in proteins will be estimated directly from the numbers of residue–residue contacts observed in protein crystal structures by regarding them as statistical averages in the quasi-chemical approximation[17-21] with an approximate treatment of the effects of chain connectivity. Estimated contact energies will be compared with experimental values

of hydrophobic energies. Also, the relevance of results to protein folding and other applications will be discussed.

Lattice Model

Let us consider a single protein molecule in solution. In order to take account of hard-core repulsions among residues and solvent molecules, residues of a protein are assumed to occupy lattice sites or cells in a linear chain fashion. Each of the vacant cells is regarded to be occupied by an effective solvent molecule; an effective solvent molecule would correspond to a group of actual solvent molecules whose total size is equal to the average size of a residue. This is an idealization from using a lattice model. As a result, volume change due to the conformational change of a protein is completely neglected in this model. Interactions are assumed to occur only between nearest-neighbor pairs of residues and effective solvent molecules.

Protein conformations may not be well represented by simple regular lattices. However, the details of lattice structure are unimportant here, because an approximation that is employed to estimate effective interresidue contact energies does not depend on the details of lattice structure but only on the coordination number of the lattice, i.e., the number of nearest neighbors around a lattice site. In protein structures, the number of nearest neighbors around a residue or the number of contacts including residue–solvent contacts will depend on the type of the given residue and its surrounding residues because of differences in residue sizes; a residue's position is taken herein as the center of its side chain atom positions, and contacting residues and effective solvent molecules are simply defined to be close pairs whose centers fall within the distance R_c. However, it is simply assumed here that the average number of nearest neighbors or contacts per residue depends only on the central residue type. Neighboring residues along an amino acid sequence tend to be in contact with each other. Contacts between nearest-neighbor residues along the sequence are explicitly excluded in counting contacts, and for convenience, the coordination numbers for residues are defined to omit those contacts. Thus the coordination numbers for terminal residues in a chain should be larger than for middle residues; however, this minor end effect is neglected here. In the result, coordination numbers are regarded to depend only upon the type of residue and effective solvent molecule.

If q_i is the coordination number for residues of type i, then the following relationship between the number of residues of the ith type, n_i, and the numbers of contacts will be satisfied.

$$q_i n_i/2 = \sum_{j=0}^{20} n_{ij} \tag{1}$$

where

$$n_{ij} = n_{ji}$$

n_{ii} and $2n_{ij}$ for $i \neq j$ are defined as the numbers of contacts between two residues of the ith type and between the ith and jth types of residues, respectively; the subscript 0 represents effective solvent molecules, whereas the other indices from 1 to 20 represent the types of amino acids. For convenience, let

$$n_{ir} = n_{ri} \equiv \sum_{j=1}^{20} n_{ij} \qquad n_{rr} \equiv \sum_{i=1}^{20} n_{ir} \qquad n_{r0} = n_{0r} \equiv \sum_{i=1}^{20} n_{i0}$$
$$n_r \equiv \sum_{i=1}^{20} n_i \tag{2}$$

n_{rr} and $2n_{r0}$ are the total numbers of residue–residue and residue–solvent contacts, and n_r is the total number of residues.

With an appropriate value of R_c, the numbers of all residue–residue contacts, n_{ij}, in a protein conformation are counted. For this specific value of R_c, the average coordination numbers, q_i, for residues and an effective solvent molecule are estimated for protein molecules. Then the number of residue–solvent contacts, $2n_{i0}$, is calculated straightforwardly with eq 1. The number of solvent–solvent contacts, n_{00}, is calculated from eq 1 with the number of effective solvent molecules, n_0, and the coordination number, q_0; the total number of effective solvent molecules and residues, $(n_0 + n_r)$, is equal to the volume of the system divided by the mean residue volume. The procedures to determine an appropriate value of R_c and to estimate coordination numbers will be deferred until the Results section.

In the present model, interactions are assumed to occur only among residues and effective solvent molecules that are in contact with each other, ignoring longer range interactions. Hence, the total contact energy of the system is taken to be

$$E_C = \sum_{i=0}^{20} \sum_{j=0}^{20} E_{ij} n_{ij} \tag{3}$$

where

$$E_{ij} = E_{ji}$$

E_{ij} is the contact energy between the ith and jth types of residues. By using eq 1, eq 3 is transformed to

$$E_C = \sum_{i=0}^{20} (2E_{i0} - E_{00}) q_i n_i/2 + \sum_{i=1}^{20} \sum_{j=1}^{20} e_{ij} n_{ij} \tag{4a}$$

$$= \sum_{i=0}^{20} E_{ii} q_i n_i/2 + \sum_{\substack{i=0 \\ (i \neq j)}}^{20} \sum_{j=0}^{20} e_{ij}' n_{ij} \tag{4b}$$

where

$$e_{ij} \equiv E_{ij} + E_{00} - E_{i0} - E_{j0} \tag{5a}$$

$$e_{ij}' \equiv E_{ij} - (E_{ii} + E_{jj})/2 \tag{5b}$$

Therefore

$$e_{ij} = e_{ij}' + e_{00}' - e_{i0}' - e_{j0}' \tag{6a}$$

$$e_{ij}' = e_{ij} - (e_{ii} + e_{jj})/2 \tag{6b}$$

Here, it is clear that only the last terms in eq 4a and 4b depend on the protein conformation. Thus, in order to discuss the dependence of energy on protein conformations, a knowledge of the absolute contact energies E_{ij} is not necessary but only the relative energies e_{ij} or e_{ij}', termed here both effective contact energies, must be known. The expression of eq 4b is more common in lattice theories than eq 4a, but eq 4a is more appropriate for calculating the total contact energy of protein conformations, because the numbers of residue–residue contacts can be calculated more directly than residue–solvent contacts. The principal purpose of the present work is to estimate e_{ij} and e_{ij}' from known crystal structures of proteins. In all following discussions, energies are represented in dimensionless RT units, unless otherwise specified, where R is the gas constant and T is absolute temperature.

Approximation of Ignoring Chain Connectivity

In the present work, it is intended to estimate contact energies from the numbers of contacts observed in protein crystal structures by regarding them as statistical averages. The numbers of contacts, n_{ij}, formed in each protein

structure might depend significantly on the order of residues in the amino acid sequence, because its particular amino acid sequence must lead to the unique native structure. However, for a large sample of proteins the effects of specific sequences should be averaged out, and then the numbers of specific residue–residue contacts would represent only the intrinsic differences of interactions among residues in proteins. This is the implicit assumption in the present work. Of course, nearest-neighbor contacts along chains probably reflect the amino acid sequences of proteins rather than intrinsic differences of interresidue interactions. Therefore, in the present analysis, the contacts between nearest-neighbor residues along chains are explicitly excluded in the counting of contacts. We do not intend to deny the contributions of intraresidue, short-range, and specific long-range interactions in the formation of each native structure, but here it is assumed that on the average intrinsic contact interactions are consistent with the stability of native structures; Go[22] pointed out that various types of interactions appear to be almost consistent with each other and therefore with native conformations and proposed to call this fact the consistency principle in protein folding. In other words, it is assumed that the average characteristics of residue–residue contacts formed in a large number of protein native structures reflect actual differences of interactions among residues and solvent molecules, as if there were no significant contribution from the specific amino acid sequence in each protein as well as intraresidue and short-range interactions. This assumption insists that the chain connectivity may be neglected to determine the relative values of effective contact energies e_{ij}. In order to determine those absolute values, however, the elastic energy originating in the chain entropy must be taken into account.

The dependence of the size of a flexible chain molecule on intramolecular interactions is hard to obtain. Several theories have been proposed to estimate the average molecular expansion of a chain molecule in good solvents; see ref 23–25 for reviews. Unlike a chain molecule in good solvents, in which intramolecular interactions are effectively repulsive, a protein molecule is usually under effectively attractive interresidue interactions. In other words, the circumstance for a protein molecule appears to correspond to poor solvents. The conformational characteristics of single-chain molecules in poor solvents, are barely studied, probably because, in practice, in such poor solvents a polymer coil will join with other coils to form a separate, more concentrated phase in preference to appreciable contraction below the unperturbed size.[23] The most important conformational characteristic which distinguishes proteins from simple polypeptides is, however, that under proper conditions a globular protein takes a highly compact form that is still soluble; the information for forming such a native conformation is, of course, coded into the amino acid sequence particular to each protein. Edwards[26] discussed briefly the case of simply changing the sign of the excluded-volume parameter in the same model as for molecular expansion. It can be shown that the optimum radius of the collapsed state considered by Edwards is zero.[27] This fact indicates that a more detailed description of intramolecular interactions, specifically taking account of hard-core repulsions as well as attractions, is required in order to reach a meaningful result.[27]

Here hard-core repulsions are explicitly taken into account in a lattice model and short-range attractive interactions are included as effective contact energies between residues. It would be difficult to evaluate the requisite combinatory factor, which is defined as the number of conformations with a certain number of contacts, even if simple lattices are employed to represent the conformations of chain molecules. Approximations, more or less ignoring the chain connectivity, have been used often in the evaluations of combinatory factors in lattice theories[23] of polymer solutions. Likewise, the chain connectivity in the amino acid sequence of a protein is neglected here; a system consisting of a single protein in solution is to be regarded as the mixture of unconnected residues and effective solvent molecules. It should be noted here that because the chain connectivity is ignored in single-chain systems, the dimension of the system is the size of a protein molecule; if it were to be taken as the actual dimension like many-chain systems, the residue solution would become unrealistically dilute. Now, in the mean field approximation, which has been used often in polymer solution theories, contact formation is approximated to be random. In order to relate the statistical average of the numbers of contacts to the contact energies, a next order approximation must be used. Thus the Bethe approximation or quasi-chemical approximation[17–21] which is well-known as a next order approximation is employed here. The number of effective solvent molecules, that is, the system size, for each protein is adjusted to yield the number of residue–residue contacts equal to its expected value for the hypothetical case of no interactions except hard sphere volume exclusions of residues and effective solvent molecules; the details will be deferred.

Quasi-Chemical Approximation

First, let us consider the mixture of unconnected residues and effective solvent molecules, each of which is present in the amount of n_i molecules. Each residue occupies a site on the lattice, and neighboring pairs of residues of the ith and jth types are assumed to interact with energy E_{ij}. In the Bethe approximation, only the occurrence probabilities of specific site pairs and no larger clusters are taken into account; therefore, in this approximation, lattice structures are represented only by the coordination number of the lattices. Then, the partition function of this system is approximated by[17]

$$Z = \text{const}\sum_{\{n_{ij}\}} \frac{(\sum_{i=0}\sum_{j=0} n_{ij})!}{n_{00}!n_{r0}!n_{0r}!n_{rr}!} \frac{n_{r0}!n_{0r}!}{\prod_{i=1} n_{i0}! \prod_{j=1} n_{0j}!} \frac{n_{rr}!}{\prod_{j=1}\prod_{i=1} n_{ij}!} \times \exp(-\sum_{i=0}\sum_{j=0} E_{ij}n_{ij}) \quad (7)$$

The combinatory factor in eq 7 is simply the number of combinations of distributing the total number of contacts into the certain numbers, $\{n_{ij}\}$, of contact pairs i–j. The statistical averages, $\bar{n}_{ij}$, of n_{ij} are derived by maximizing the partition function with respect to n_{ij}. This approximation is equivalent to the assumption that the neighboring site pairs, i–j and k–l, are in quasi-chemical equilibrium with one another as follows.[19-20]

$$i\text{–}j + k\text{–}l \leftrightarrow i\text{–}k + j\text{–}l \quad (8)$$

In other words, it is assumed that the following relations are satisfied.

$$\frac{\bar{n}_{ij}\bar{n}_{00}}{\bar{n}_{i0}\bar{n}_{j0}} = \exp(-e_{ij}) \quad (9a)$$

$$\frac{\bar{n}_{ij}^2}{\bar{n}_{ii}\bar{n}_{jj}} = \exp(-2e_{ij}') \quad (9b)$$

where $\bar{n}_{ij}$ is the statistical average of n_{ij}, and energies e_{ij} and e_{ij}' are defined by eq 5a and 5b and as usual repre-

sented in RT units. Here it is noteworthy that eq 9b with $i \neq 0$ and $j \neq 0$ can be obtained by maximizing the third and the last factors in eq 7 even for fixed n_{i0}. Similarly, an equation related to $(e_{i0}' - e_{k0}')$ which is derived from eq 9b can be obtained by maximizing eq 7 even under the restriction of fixing the first factor, that is, fixing the total number of residue–solvent contacts, $2n_{r0}$.

A basic assumption introduced here is that the partition function for a protein can be approximated by eq 7 with a crude estimate of the number of effective solvent molecules, n_0; of course, const in eq 7 cannot be the same for both systems of protein solutions and simple monomer solutions, because residues in the protein system must be connected as a linear chain. The number of effective solvent molecules for each protein is chosen to yield the total number of residue–residue contacts equal to its expected value for the hypothetical case of hard sphere interactions among residues and effective solvent molecules; the expected number of residue–residue contacts at this condition will be crudely estimated by means of a freely jointed chain distribution and an expansion originating in hard sphere interactions. In the result, in this approximation, the effects of chain connectivity are taken into account only in the definitions of the coordination numbers, q_i ($i \neq 0$), for residues which exclude contacts between nearest neighbors along a chain and of the size of the system, that is, the number of effective solvent molecules, n_0, in the system.

Estimating e_{i0}' requires the estimation of n_0 or n_{00} which represents the effects of chain connectivity. Estimates of the values of e_{i0}' may be inexact, because of the crude account of chain connectivity and also the intrinsic limitations of the quasi-chemical approximation. The quasi-chemical approximation is appropriate for systems of molecules interacting weakly with each other, i.e., high-temperature limit, but inappropriate for strongly interacting molecular systems, i.e., low-temperature limit, because only the occurrence probabilities of pairs are taken into account. Interactions, e_{i0}', must be strong enough to make protein native structures compact; hydrophobic interactions are likely to be responsible for such strongly attractive interactions. Interactions of higher order than binary clusters might play significant roles in such systems. For these reasons, the estimate of the absolute values of e_{i0}' may be crude. On the other hand, e_{ij}' and the relative values of e_{i0}' can be estimated without any knowledge of n_0; this, of course, results from the fact that the effects of specific amino acid sequences on the formation of contacts are completely neglected.

Estimation of Effective Contact Energies

One simple way to estimate contact energies according to eq 9 from the observed numbers of contacts among residues and effective solvent molecules in protein crystal structures is to use the actual sum of contact pairs i–j in the sample of proteins for $\bar{n}_{ij}$ in eq 9. However, this yields a biased estimate of e_{ij}'; for example, in the case of $e_{ij} = 0$ in which residues and effective solvent molecules are randomly mixed, the sum of n_{ij} expected for proteins would be different from that calculated from the average composition over all proteins, because of differences in amino acid composition among proteins. To remove such biases, contact energies e_{ij}' between residues are estimated in the following manner.

$$\exp(-2e_{ij}') = \frac{N_{ij}^2}{N_{ii}N_{jj}} \frac{C_{ii}C_{jj}}{C_{ij}^2} \qquad \text{for } i, j \neq 0 \qquad (10)$$

where

$$N_{ij} \equiv \sum_p n_{ij;p} \qquad (11)$$

$$C_{ij} \equiv \sum_p \frac{n_{ir;p}n_{jr;p}}{n_{rr;p}} \qquad (12)$$

The supbscript p is used to indicate each protein. The second factor in eq 10 is a correction factor to remove the biases so that the right-hand side of eq 10 will give the correct value of one for the case of $e_{ij}' = 0$. C_{ij} is the sum over all proteins of contact pairs i–j expected for that case; eq 12 is derived from eq 9b. The contact energies e_{i0}' between residues and effective solvent molecules are estimated by

$$\exp(-2e_{i0}') = \frac{N_{i0}^2}{N_{ii}N_{00}} \frac{C_{ii}'C_{00}'}{C_{i0}'^2} \qquad (13)$$

where

$$C_{ii}' \equiv \sum_p n_{rr;p}\left(\frac{q_i n_{i;p}}{\sum_{i=1} q_i n_{i;p}}\right)^2 \qquad (i \neq 0) \qquad (14)$$

$$C_{i0}' \equiv \sum_p n_{r0;p}\frac{q_i n_{i;p}}{\sum_{i=1} q_i n_{i;p}}$$

$$C_{00}' \equiv \sum_p \frac{n_{r0;p}^2}{n_{rr;p}} \qquad (15)$$

The correction factors C_{ii}' and C_{i0}' are the expected numbers of contact pairs i–i and i–0 for the case of $e_{ij}' = 0$ and $e_{i0}' = e'$. C_{00}' is the expected number of solvent–solvent contacts for the case of $e_{ij}' = e_{i0}' = 0$. Equations 14 and 15 are derived from eq 9b. Thus, for the case of $e_{ij}' = 0$ and $e_{i0}' = e'$, eq 13 assumes the reasonable form, $\exp(-2e') = C_{00}'/N_{00}$. A useful alternative representation of contact energies, e_{ij}, is given by eq 6a in terms of values of e_{ij}'.

Estimation of the Number of Effective Solvent Molecules, n_0

Let us consider a hypothetical system of a single protein molecule in solution, in which there are no explicit interactions except hard sphere repulsions among residues and effective solvent molecules, although intraresidue interactions are implicitly present; all e_{ij} are zero. For this system, the quasi-chemical approximation, eq 9a or 9b, becomes equivalent to the mean field or random mixture approximation[17,18,23] $\bar{n}_{ij} = q_i n_i q_j n_j/(\sum_{i=0} q_i n_i)$. Thus n_0 in the present approximation may be chosen so as to yield n_{ij} or $\bar{n}_{rr}$, equal to their expected values for the hypothetical molecules with $e_{ij} = 0$, representing the effects of chain connectivity; here it should be noted that hard sphere interactions are implicitly taken into account by the use of coordination numbers. In other words, n_0 in eq 9a and 9b corresponds not to the overall concentration of solvent molecules in solution but reflects the local concentration of effective solvent molecules in the vicinity of residues in the hypothetical protein. When interactions between residues, e_{ij}, are introduced to this hypothetical molecule, the molecule collapses by excluding solvent molecules from the interior volume of molecules until a sufficient number of contacts among residues are formed, so that eq 9 is satisfied. Thus, n_0 is estimated from the expected value $\bar{n}_{rr}$ for this hypothetical case.

Equation 9a with $e_{ij} = 0$ can be represented as

$$\bar{n}_{rr}\bar{n}_{00} = \bar{n}_{r0}^2 \qquad (16)$$

From eq 1, 2, and 16, the following equation is derived from

which n_0 can be evaluated when n_i and $\bar{n}_{rr}$ are known.

$$n_0 = \frac{(q_r n_r)^2}{2q_0\bar{n}_{rr}} - \frac{q_r}{q_0}n_r \tag{17}$$

where

$$q_r \equiv (\sum_{i=1} q_i n_i)/n_r \tag{18}$$

In the following, $\bar{n}_{rr}$ for $e_{ij} = 0$ will be estimated with two different approximations, a smoothed density approximation and a random flight approximation for chain molecules; the former serves only to clarify the meaning of n_0, and the latter is actually employed to evaluate n_0, because the effects of chain connectivity are more realistically taken into account. Both approximations were employed[24] to evaluate intramolecular interactions in attempts to obtain closed expressions for the expansion factor of a polymer in good solvents.

(1) Smoothed Density Approximation. $\bar{n}_{rr}$ is represented in the smoothed density approximation as follows.

$$\bar{n}_{rr} = \frac{1}{2}\int \frac{(q_r n_r \rho(s))^2}{(q_0(1/v_r - n_r\rho(s)) + q_r n_r \rho(s))}\,ds \tag{19}$$

where

$$v_r \equiv \sum_{i=1} v_i n_i / n_r \tag{20}$$

v_i is the average volume occupied by the ith type of residue and v_r is the average of v_i in a protein. $\rho(s)$ is the normalized density of residues at distance s from the center of mass. $(1/v_r - n_r\rho(s))$ is the number density of effective solvent molecules with volume v_r at distance s. Let us approximate the density $\rho(s)$ for the hypothetical protein having $e_{ij} = 0$ by the Gaussian distribution with the second moment $\langle s^2 \rangle$ equal to the mean square radius of gyration. In addition, the gyration radius is assumed to be uniformly expanded by a factor α_s due to hard sphere volume exclusions. With these approximations, eq 19 becomes in a series expansion in powers of $(1 - q_r/q_0)n_r v_r/V$

$$\bar{n}_{rr} = \frac{(q_r n_r)^2}{2q_0}\left(\frac{3}{4\pi}\right)^{1/2}\frac{v_r}{V}\left[1 + \frac{4}{3\pi^{1/2}}\left(1 - \frac{q_r}{q_0}\right)\frac{n_r v_r}{V} + \ldots\right] \tag{21}$$

where V is the volume of a sphere whose radius is equal to the root mean square radius of gyration.

$$V \equiv \frac{4\pi}{3}(\alpha_s b)^3\left(\frac{n_r - 1}{6}\right)^{3/2} \tag{22}$$

b is the equivalent virtual bond length between residues for this unperturbed protein molecule; $(n_r - 1)b^2$ is equal to the mean square end-to-end distance of the unperturbed chain. Equation 21 indicates that $\bar{n}_{rr}$ depends on chain length as $n_r^{1/2}\alpha_s^{-3}$. From eq 17 and 21, an expression for n_0 is derived.

$$n_0 = \left(\frac{4\pi}{3}\right)^{1/2}\frac{V}{v_r}\left[1 - \left(\frac{4}{3\pi^{1/2}}\left(1 - \frac{q_r}{q_0}\right) + \left(\frac{3}{4\pi}\right)^{1/2}\frac{q_r}{q_0}\right)\frac{n_r v_r}{V} - \ldots\right] \tag{23}$$

The contribution from the second terms in eq 21 and 23 is negligible for this hypothetical protein. Equation 23 indicates that the effective solvent molecules to be taken into account are those within a sphere whose radius is equal to the root mean square radius of gyration of the hypothetical protein with $e_{ij} = 0$. This interpretation of n_0 is intuitively reasonable because conformations of a single molecule are considered. However, in this approximation, the effects of chain connectivity are taken into account only in a one-particle distribution function, that is, the density distribution of residues. Therefore $\bar{n}_{rr}$ is underestimated in this approximation and n_0 is overestimated, because the local concentration of residues in the vicinity of a given residue is significantly higher than given by this approximation.

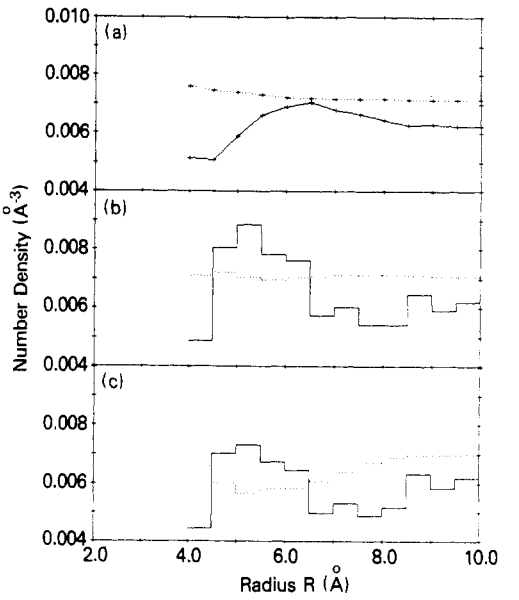

Figure 1. Residue packing around interior residues in protein crystal structures. The solid lines represent (a) the average number density of residues in a sphere of radius R centered at interior residues, (b) that in a shell between spheres of radius R and $R + 0.5$, and (c) the average number density of residues excluding nearest neighbors along a chain in each shell; interior residues are defined to be residues within 7 Å of the center of a protein subunit. Only protein subunits consisting of more than 100 residues are used in this calculation; the total number of interior residues is 393. The dotted lines represent the corresponding quantities calculated by assuming that residues are distributed with smoothed densities; refer to eq 32 for the definition.

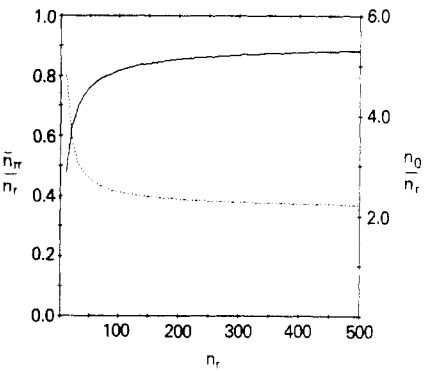

Figure 2. Expected number, $\bar{n}_{rr}$, of residue–residue contacts calculated with eq 26 for the case of hard sphere interactions, i.e., $e_{ij} = 0$, and the number n_0 of effective solvent molecules calculated with eq 17 are shown as functions of the chain length, n_r, by the solid and dotted lines, respectively. The amino acid compositions of the chains are the same as the occurrence frequencies of amino acids in the entire group of proteins used.

(2) Random Flight Approximation. $\bar{n}_{rr}$ is represented as

$$\bar{n}_{rr} = \sum_{j=2}^{n_r-1}\sum_{i=1}^{n_r-j}\int_{|R_{i,i+j}|\le R_c} P(R_{i,i+j})\,dR_{i,i+j} \tag{24}$$

in a two-particle distribution approximation, where $P(R_{i,i+j})$ is the probability density of the ith and $(i + j)$th residues at separation $R_{i,i+j}$; it should be noted that nearest-neighbor contacts are not included in eq 24. Here we have chosen to approximate $P(R_{i,i+j})$ with the end-to-end distance distribution for a freely jointed chain of length j with the equivalent virtual bond length $b(j)$ and the expansion factor $\alpha_r(j)$. The equivalent virtual bond length b for unperturbed chains must include the effects of intraresidue interactions; $jb(j)^2$ is equal to the mean square end-to-end distance of the unperturbed chain of length j. The expansion factor α_r is defined as the ratio of the root mean square end-to-end distance perturbed by hard sphere volume exclusion to that of an unperturbed chain molecule.

$P(R_{i,i+j})$ is approximated here as

$$P(R_{i,i+j}) \simeq \left(\frac{3}{2\pi j(\alpha_r(j)b(j))^2}\right)^{3/2} \exp\left(-\frac{3X^2}{2}\right) \times \left[1 - j^{-1}\left(\frac{3}{4} - \frac{3}{2}X^2 + \frac{9}{20}X^4\right) + j^{-2}\left(\frac{29}{160} - \frac{69}{40}X^2 + \frac{981}{400}X^4 - \frac{1341}{1400}X^6 + \frac{81}{800}X^8\right) - j^{-3}\left(-\frac{351}{3200} - \frac{387}{1600}X^2 + \frac{9369}{3200}X^4 - \frac{4293}{1120}X^6 + \frac{37449}{22400}X^8 - \frac{15633}{56000}X^{10} + \frac{243}{16000}X^{12}\right) + \mathcal{O}(j^{-4})\right] \quad (25)$$

for $j \gg 1$ and $X^2/j \ll 1$, where

$$X^2(j) \equiv \frac{(R_{i,i+j})^2}{j(\alpha_r(j)b(j))^2}$$

The higher order terms in eq 25 have been obtained from eq 5.30 of ref 24. With this density function eq 24 becomes in a series expansion

$$\bar{n}_{rr} = \left(\frac{6}{\pi}\right)^{1/2} n_r \sum_{j=2}^{n_r-1}\left(1 - \frac{j}{n_r}\right) X_c^3 \left[\left\{1 - \frac{9}{10}X_c^2 + \frac{27}{56}X_c^4 - \frac{3}{16}X_c^6 + \frac{81}{1408}X_c^8 - \frac{243}{16640}X_c^{10} + \frac{81}{25600}X_c^{12} - \mathcal{O}(X_c^{14})\right\} - j^{-1}\left\{\frac{3}{4} - \frac{63}{40}X_c^2 + \frac{243}{160}X_c^4 - \frac{297}{320}X_c^6 + \frac{1053}{2560}X_c^8 - \frac{729}{5120}X_c^{10} + \frac{4131}{102400}X_c^{12} - \mathcal{O}(X_c^{14})\right\} + j^{-2}\left\{\frac{29}{160} - \frac{1917}{1600}X_c^2 + \frac{100683}{44800}X_c^4 - \frac{39897}{17920}X_c^6 + \frac{207441}{143360}X_c^8 - \frac{987309}{1433600}X_c^{10} + \frac{7351803}{28672000}X_c^{12} - \mathcal{O}(X_c^{14})\right\} - j^{-3}\left\{-\frac{351}{3200} - \frac{297}{6400}X_c^2 + \frac{34749}{25600}X_c^4 - \frac{1007721}{358400}X_c^6 + \frac{74250999}{31539200}X_c^8 - \frac{753158817}{372736000}X_c^{10} + \frac{16718157}{16384000}X_c^{12} - \mathcal{O}(X_c^{14})\right\} + \mathcal{O}(j^{-4})\right] \quad (26)$$

where

$$X_c^2(j) \equiv \frac{R_c^2}{j(\alpha_r(j)b(j))^2}$$

The terms explicitly listed above will suffice for the convergence of the series at small values of R_c (≤7.0 Å). In eq 26, the total number of contacts is proportional to the chain length, n_r, in the long-chain limit, indicating that the solution surrounding residues in a chain is much denser than given by the smoothed density approximation. This chain length dependence of $\bar{n}_{rr}$ is reasonable because most contacts are short range in the hypothetical protein expanded with $e_{ij} = 0$. Equation 26 is employed together with eq 17 to estimate n_0; the approximation of random flight chains for protein molecules is not entirely satisfactory for the estimation of short-range contacts, which are the main contributors in eq 24, because the intraresidue interactions in a protein molecule would significantly affect the formation of short-range contacts.

The required expansion factor $\alpha_r(j)$ and the equivalent virtual bond length $b(j)$ between residues are calculated as follows. The expansion factor α_r is calculated with Flory's equation[23,28,29] modified to give the exact first-order perturbation theory.[30,31]

$$\alpha_r(j)^5 - \alpha_r(j)^3 = {}^4/_3 z \quad (27)$$

$$z \equiv \left(\frac{3}{2\pi b(j)^2}\right)^{3/2} \beta j^{1/2} \quad (28)$$

The excluded-volume parameter β is the negative of a binary cluster integral with the pair potential $u(r)$ of mean force between residues; the mean force potential with $u(\infty) \equiv 0$ is obtained by integrating a Boltzmann factor over the phase space of all solvent molecules.[18]

$$\beta \equiv \int \left(1 - \exp\left(-\frac{u(r)}{kT}\right)\right) dr \quad (29a)$$

In Flory's theory,[23] β is represented in terms of an entropy parameter ψ_1 and an "ideal" temperature Θ; ψ_1 and Θ include the entropic and enthalpic contributions of solvent effects, respectively.

$$\beta \equiv 2\frac{v_r^2}{v_0}\psi_1\left(1 - \frac{\Theta}{T}\right) \quad (29b)$$

v_r and v_0 are volumes of a residue and a solvent molecule. ψ_1 is equal to ${}^1/_2$ in the formalism of Flory's polymer lattice model based on the mean field approximation; however, the values of ψ_1 and Θ are difficult to evaluate for actual systems. For the present hypothetical protein with $e_{ij} = 0$, Θ/T should be zero. Since the present model is just a lattice model, ψ_1 should be taken as ${}^1/_2$, and the volume of an equivalent solvent molecule, v_0, should be taken to be equal to the average volume of a residue, v_r. Thus, the excluded volume β is taken here to be the average volume of a residue, v_r.

$$\beta = v_r \quad \text{for the present case} \quad (29c)$$

The equivalent virtual bond length, $b(j)$, for unperturbed protein molecules has been approximated as follows to account for calculated chain length dependence of dimensions and experimental values of chain dimensions; $b(j)$ is defined so that $jb(j)^2$ is equal to the mean square end-to-end distance of the unperturbed chain of length j.

$$\frac{1}{b(j)} \simeq \frac{1}{b(\infty)}\left(1 + \frac{C_\infty^{1/2} - 1}{2.04}(3.17j^{-1} - 1.13j^{-2})\right) \quad (30)$$

where

$$b(\infty) \equiv C_\infty^{1/2} l = 3.8C_\infty^{1/2} \quad (\text{Å})$$

l is the customary virtual bond length, 3.8 Å, between C^α atoms. Equation 30 with $C_\infty = 9.27$ has been obtained by curve-fitting from the chain length dependence of the root mean square end-to-end distance for polyalanine unper-

turbed chains (Figure 15 on p 279 of ref 32) where only intraresidue interactions are included. The characteristic ratio, C_∞ is almost constant for homopolypeptide chains with various side chains except for polyglycine and polyproline.[32] Because the presence of glycine significantly reduces chain dimensions, C_∞ is given as a function of the fraction P_{Gly} of glycine residues.

$$C_\infty^{1/2} \simeq 2.98 - 3.80P_{Gly} + 2.34P_{Gly}^2 \quad (31)$$

This equation has been obtained by curve-fitting the characteristic ratios of random copolymers of glycine and alanine (Figure 16 on p 283 of ref 32).

Results

A. Protein Coordinates. Proteins used here are listed in Table I and their coordinates are taken from the Brookhaven Protein Data Bank;[33] 42 globular proteins are used, including 30 monomeric proteins. These proteins have been chosen with the criteria that their chain lengths are longer than 100 residues and that few atomic positions are missing; small proteins have been excluded because they are often inhibitors or act in their functional state by binding to other proteins, and hence interprotein contacts may be important in stabilizing them. Also, in cases where coordinates are available for several closely homologous proteins, only one representative has been used; the minimum difference between amino acid sequences for homologous proteins included here is 50%. For proteins that are polymeric or bind to an inhibitor or a substrate in their functional state, the numbers of contacts are calculated for the complete assembly. There are several proteins composed of identical chains; however, the three-dimensional structures of their subunits may differ. In an attempt to avoid sampling biases, weights W_p have been used in the sums over proteins such as in eq 11, 12, 14, and 15; for example, the weight is taken as $^1/_2$ for proteins composed of two identical chains. Weights W_p are given in Table I.

B. Definition of Contacts. Each residue is represented by the center of its side chain atom positions; the positions of C^α atoms are used for glycines. Residues whose centers are closer than R_c are defined to be in contact. This kind of simple method to evaluate the number of residue–residue contacts in proteins has often been used.[34-41] One difference between the present method and others is that most others have employed the positions of C^α atoms to represent residues. The choice of residue positions at the centers of their side chain atom positions is more appropriate for evaluating side chain–side chain contacts than either C^α atom positions or the centers of all residue atoms including backbone atoms. More long-range contacts are obtained by using the centers of side chain atoms than with these other definitions.

In order to determine an appropriate value of R_c, residue packing in the interior of protein molecules has been examined in terms of the number of residues within a sphere of radius R centered at interior residues. Interior residues are defined as residues within a distance R_I of the center of each protein subunit. R_I must be chosen to be small enough so that there are no voids for effective solvent molecules but only residues within a distance $(R_I + R)$ from the center of a protein molecule. We have used 7.0 Å for R_I; this value would be sufficiently small unless R were large, because the radius of a sphere whose volume is almost equal to that of a protein consisting of 100 residues is about 14.9 Å. In this paper only protein subunits consisting of more than 100 residues have been used. The total number of interior residues for the present sample is 393. Table II shows the average numbers of residues including or excluding a central residue and nearest neighbors along a chain, within spheres of radius R = 4–10 Å centered at the interior residues. For comparison, their expected values under the assumption of smoothed densities for residues are also shown in Table II; the expected number of residues within a sphere of radius R centered at the ith type of residue under the smoothed density assumption is equal to

$$\left(\frac{4\pi}{3}R^3 - v_i\right) \Big/ v_{ir}(R) + 1 \quad (32)$$

where v_i is the mean volume of the ith type of residue and v_{ir} is the average volume of residues surrounding the ith type of residues. The mean volumes occupied by buried residues in the interiors of nine proteins (Table 2 of ref 8) are used as v_i except for arginine whose volume is taken to be the mean volume of arginine residues located on subunit–subunit interfaces (Table 4 of ref 9). These mean residue volumes, v_i, are given in Table III. The average volume of surrounding residues, $v_{ir}(R)$, is calculated as the average volume of residues observed within the distance R of the ith type of residue in protein crystal structures; nearest neighbors along chains are included in the calculation of v_{ir}. For the cases in which the nearest-neighbor residues are excluded (see the right side of Table II), the average number of nearest neighbors along chains within the distance R is subtracted from eq 32. Those numbers for R = 6.5 Å are shown in Table III. Values listed in Table II are averages weighted with the numbers of each type of interior residues.

From these data, the average number densities of residues within a sphere of radius R and within the shell between spheres of radius R and R + 0.5 are calculated and shown in Figure 1a,b, respectively. Figure 1c represents the average number density within each shell in which nearest neighbors along a chain are excluded. The solid and dotted lines in Figure 1 represent the observed values and expected values with the smoothed density assumption, respectively; radial distributions correspond to the solid lines divided by the dotted lines in Figure 1b,c. The first peak in the radial distribution occurs at the shell between 5.0 and 5.5 Å, indicating that the average distance between nearest-neighbor residues falls in this region; the peaks are certainly broadened by the heterogeneity of residue size. In Figure 1a, the density within a sphere attains its maximum value near 6.5 Å, and at this point also achieves its closest approach to the smoothed density curve. The number densities at large values of radius are significantly smaller than values calculated with the assumption of smoothed density. This is attributable to including space outside the surfaces of proteins. The densities and the radial distribution in the case of excluding a central residue and nearest neighbors along a chain show the same behavior; the dotted line in Figure 1c becomes concave because the nearest-neighbor residues along a chain are mostly located in the range of distance from about 4.5 to 7.5 Å. Thus, it appears that the radius 6.5 Å is an appropriate value for R_c to define contacts. This value of 6.5 Å for R_c is also appropriate with respect to residue volumes. The average volume of a residue is 139.6 Å^3. If the packing density of residues, the ratio of the actual volume of an object to the volume of space occupied, is taken as 0.74, then the mean size of a residue will be 139.6·0.74 Å^3 corresponding to a sphere of radius, 2.91 Å. The mean packing density of interior protein atoms is essentially identical with that reported for crystals of small organic molecules, and the latter is close to the theoretical value, 0.74, for close-packed spheres.[8,42] The average

Table I
Proteins Used in the Present Analyses

code[a]	protein	Quaternary structure, n_r[b]	weight W_p	n_0/n_r	n_{rr}	$2n_{r0}$	e_r[c]	e_v[c]	e_s[c]	$2\tilde{n}_{r0}$[d]
1. α proteins										
3CPV	Calcium-binding parvalbumin B (Carp)	108(109)	1	2.52	181	322	-3.30	-2.81	-2.27	273
4CYT	Reduced cytochrome C (Tuna)	103(104)	1	2.39	172	308	-2.92	-2.57	-2.18	266
1CCY	Cytochrome C' (R. Molischianum)	128(A)+128(B)	1/2	2.44	462	689	-2.97	-2.63	-2.17	659
1C2C	Ferricytochrome C_2 (R. Rubrum)	112	1	2.54	191	326	-2.62	-2.45	-2.26	295
1ECD	Erythrocruorin (Deoxy) (C. Thummi)	135(136)	1	2.42	249	349	-3.58	-3.11	-2.45	312
2MHB	Hemoglobin A (Aquo met) (Horse)	141(A)*2+146(B)*2	1/2	2.40	1194	1213	-3.32	-2.97	-2.28	1322
1MBD	Myoglobin (Deoxy) (Sperm whale)	153	1	2.46	271	418	-3.46	-3.00	-2.41	367
1LHB	Hemoglobin (Met, Cyanide V) (Lamprey)	148	1	2.55	239	452	-3.50	-3.01	-2.49	357
1HBL	Leghemoglobin (Accetate, Met) (Yellow Lupin)	153	1	2.51	285	389	-3.41	-3.02	-2.46	361
1BP2	Phospholipase A_2 (E.C.3.1.1.4) (Bovine)	123	1	2.63	232	318	-3.13	-2.77	-2.23	312
2. β proteins										
2GCH	γ Chymotrypsin A (E.C.3.4.21.1) (Bovine)	236(245)	1	2.29	540	411	-3.17	-2.90	-2.17	543
1EST	Tosyl-elastase (E.C.3.4.21.11) (Porcine)	240	1	2.26	533	448	-3.17	-2.91	-2.30	533
1PTC	Beta-trypsin (E.C.3.4.21.4) and Inhibitor (Bovine)	223(E)+56(59)(I)	1	2.32	617	532	-3.12	-2.84	-2.19	655
2SOD	Cu,Zn Superoxide dismutase (E.C.1.15.1.1) (Bovine)	151(152)(O)+151(152)(Y)	1/2	2.11	652	604	-3.09	-2.73	-1.97	709
1REI	Bence-Jones immunoglobulin REI (variable part) (Human)	107(A)+107(B)	1/2	2.54	407	539	-3.17	-2.76	-2.15	532
1FC1	Immunoglobulin Fc fragment (Ig-G1 class) (Human)	206(224)(A)+206(224)(B)	1/2	2.54	745	1099	-3.05	-2.70	-2.23	1035
1APP	Penicillopepsin (E.C.3.4.23.7) (P. Janthinellum)	323	1	2.14	704	634	-3.16	-2.80	-1.98	747
2SGB	Proteinase B (Streptomyces Griseus)	185	1	2.02	416	340	-2.81	-2.65	-2.28	415
1ALP	α Lytic protease (E.C.3.4.21.12) (Myxobacter 495)	198	1	2.07	449	355	-3.05	-2.81	-2.22	440
3. α + β proteins										
2ACT	Actinidin (Sulfhydryl Proteinase) (Kiwifruit)	217(220)	1	2.18	471	426	-3.15	-2.90	-2.34	485
2FD1	Ferredoxin (Azotobacter Vinelandii)	106	1	2.68	118	429	-3.15	-2.85	-2.69	254
1LZM	Lysozyme (E.C.3.2.1.17) (T4 phage)	164	1	2.49	299	436	-3.47	-3.00	-2.36	394
2LYZ	Lysozyme (E.C.3.2.1.17) (Hen egg white)	129	1	2.45	263	292	-3.23	-2.84	-2.15	310
8PAP	Papain (E.C.3.4.22.2) (Papaya)	212	1	2.19	454	425	-2.99	-2.78	-2.34	481
1RN3	Ribonuclease A (E.C.3.1.4.22) (Bovine)	124	1	2.70	236	317	-2.95	-2.62	-2.12	330
2SNS	Staphylococcal nuclease (E.C.3.1.4.7) (S. Aureus)	141(149)	1	2.53	246	397	-2.88	-2.60	-2.26	371
3TLN	Thermolysin (E.C.3.4.24.4) (B. Thermoproteolyticus)	316	1	2.20	703	588	-2.90	-2.71	-2.26	735
4. α/β proteins										
2ADK	Adenylate kinase (E.C.2.7.4.3) (Porcine)	194(195)	1	2.35	292	639	-3.14	-2.79	-2.48	472
1ABP	L-Arabinose-binding protein (E. Coli)	306	1	2.26	629	666	-3.23	-2.88	-2.20	725
4CPA	Carboxypeptidase α (E.C.3.4.17.1) (Bovine) and Inhibitor (Potato)	307+37(38)(I)+1(G)	1	2.42	769	633	-3.18	-2.92	-2.31	805
4FXN	Flavodoxin (Semiquinone form) (Clostridium MP)	138	1	2.37	275	317	-3.65	-3.08	-2.10	319
4ADH	Apo-Liver alcohol dehydrogenase (E.C.1.1.99.8) (Horse)	374*2	1/2	2.20	1758	1186	-3.36	-3.03	-2.05	1666
1GPD	D-Gyceraldehyde-3-phosphate dehydrogenase (E.C.1.2.1.12) (Lobster)	333(334)(G)*2+333(334)(R)*2	1/4	2.26	2797	2789	-3.16	-2.91	-2.41	3059
4LDH	Lactate dehydrogenase (Apo, M4) (E.C.1.1.1.27) (Dogfish)	329(330)*4	1/4	2.34	3028	2224	-3.38	-3.06	-2.18	2980
3PGK	Phosphoglycerate kinase (E.C.2.7.2.3) (Bakers Yeast)	415(416)	1	2.24	779	1051	-3.23	-2.89	-2.39	972
3PGM	Phosphoglycerate mutase (E.C.2.7.5.3) (Bakers Yeast)	230(241)*4	1/4	2.43	1738	2299	-2.96	-2.74	-2.42	2261
1RHD	Rhodanese (E.C.2.8.2.1) (Bovine)	293	1	2.29	574	686	-3.27	-2.96	-2.44	668
1TIM	Triose phosphate isomerase (E.C.5.3.1.1) (Chicken)	247(A)+247(B)	1/2	2.23	1031	1045	-3.23	-2.90	-2.23	1140
2TAA	Taka-amylase A (E.C.3.2.1.1) (A. Oryzae)	478(A)	1	2.25	919	1176	-3.13	-2.85	-2.41	1098
1CAC	Carbonic anhydrase form C (E.C.4.2.1.1) (Human)	256(260)	1	2.32	465	677	-3.09	-2.78	-2.35	617
3DFR	Dihydrofolate reductase (E.C.1.5.1.3) (Lactobacillus Casei)	161(162)	1	2.48	301	408	-3.41	-3.02	-2.44	376
4DFR	Dihydrofolate reductase (E.C.1.5.1.3) (E. coli B)	157(159)(B)	1	2.46	305	372	-3.53	-3.12	-2.44	359

[a] Protein codes which are used in Brookhaven Protein Data Bank.[33] [b] The number of residues used for each protein or subunit is shown in this column; if it is different from the actual chain length because of missing atom positions or the presence of a N-terminal acetyl base, the latter is shown in parenthesis. The characters in parentheses are chain identification codes used in Brookhaven Protein Data Bank.[33] [c] See eq 38 and 39 for the definitions of e_r, e_v, and e_s; these energies are in RT units. [d] The expected number of residue-solvent contacts calculated by solving the nonlinear simultaneous equations (1) and (9a) with the estimated values of e_{ij}.

distance between residues in contact is estimated to be 5.82 Å, which is only slightly larger than the position of the first peak in the radial distribution; the average distance should be somewhat shorter than 5.82 Å because the center of side chain atom positions is used as a residue position. The bulkiest residue is tryptophan whose volume is 237.6 Å^3, and its side chain volume is 171.2 Å^3 obtained by subtracting the volume of glycine. Thus 6.5 Å for R_c is sufficiently large even to detect contacts between such bulky side chains; the side chain–side chain distance for tryp-

Table II
Residue Packing around Interior Residues

radius R (A)	#residues within a sphere of radius R centered at an interior residue[a] excluding a central residue: observed mean	observed s.d.	by smoothed density assump.[b]	and nearest neighbors: observed mean	observed s.d.	by smoothed density assump.[b]
4.0	1.38	0.64	2.04	0.29	0.52	1.00
4.5	1.94	1.00	2.85	0.79	0.85	1.75
5.0	3.08	1.37	3.87	1.79	1.13	2.60
5.5	4.62	1.69	5.10	3.06	1.43	3.58
6.0	6.25	1.75	6.55	4.46	1.59	4.79
6.5	8.13	1.87	8.28	6.04	1.72	6.22
7.0	9.77	2.00	10.31	7.47	1.86	7.96
7.5	11.77	2.30	12.66	9.23	2.13	10.08
8.0	13.82	2.54	15.36	11.07	2.41	12.62
8.5	16.14	2.80	18.41	13.29	2.72	15.55
9.0	19.26	3.15	21.84	16.33	3.11	18.91
9.5	22.44	3.37	25.64	19.47	3.36	22.68
10.0	26.16	3.79	29.88	23.17	3.79	26.90

[a] The number of interior residues, which are defined to be residues within 7.0 Å of the center of a protin subunit, is 393.0. [b] See eq 32 and the text.

Table III
Coordination Numbers, q_i, for R_c = 6.5 Å

	#residues	#surrounding residues[a]	v_i[b]	v_{ir}[c]	$q_{n;i}$[d]	q_i[e]	from interior residues[f] mean	s.d.	#residues
GLY	823.0	4534.5	66.4	133.13	1.754	6.388	6.14	1.60	50.5
ALA	779.0	4294.0	91.5	136.54	1.460	6.295	6.06	1.72	35.0
SER	672.0	3249.5	99.1	132.88	1.333	6.579	6.68	1.88	28.0
CYS	183.0	1224.0	111.65[b]	133.87	1.148	6.612	4.96	1.35	14.0
THR	570.0	2821.0	122.1	133.63	1.191	6.504	6.74	2.10	19.0
ASP	531.0	2169.0	124.5	134.70	1.001	6.615	6.06	1.90	9.0
PRO	367.0	1639.0	129.3	140.63	1.448	5.812	6.21	1.52	7.0
ASN	404.0	1670.0	135.2	137.14	0.901	6.502	5.88	1.11	8.0
VAL	720.0	4513.0	141.7	140.67	1.068	6.102	5.70	1.73	48.0
GLU	454.0	1678.5	155.1	142.41	0.721	6.267	5.14	1.64	7.0
GLN	326.0	1287.5	161.1	136.22	0.739	6.523	6.60	1.85	5.0
HIS	199.0	954.5	167.3	142.11	0.734	6.184	6.11	1.78	14.0
LEU	675.0	4171.0	167.9	144.10	0.743	6.075	6.27	1.59	45.5
ILE	452.0	2919.5	168.8	141.20	0.920	6.031	6.08	1.34	37.0
MET	139.0	858.5	170.8	146.88	0.594	6.076	5.25	1.92	6.0
LYS	632.0	1821.5	171.3	138.06	0.539	6.553	2.67	2.36	3.0
ARG	292.0	1116.5	202.1[b]	140.75	0.346	6.391	6.00	1.26	5.0
PHE	342.0	2061.0	203.4	147.02	0.561	5.880	6.07	1.73	34.0
TYR	331.0	1715.0	203.6	143.31	0.542	6.064	6.13	0.99	8.0
TRP	149.0	843.0	237.6	144.68	0.450	5.859	6.05	0.67	10.0
g	9040.0	45540.5	139.6	138.52	1.013	6.298	6.05	1.72	393.0
SLV[h]			139.6	139.60	0.0	7.240			

[a] The total number of residues within the distance R_c = 6.5 Å from residues, including nearest neighbors along chains. [b] The volumes v_i except for ARG have been taken from Table 2 of ref 8, and v_i for ARG from Table 4 of ref 9. v_i for CYS is the mean volume of cysteine and half cystine. [c] The average volume of surrounding residues. [d] The average number of nearest neighbors along a chain within the distance R_c = 6.5 Å from a residue. [e] See eq 33 for the definition. [f] The mean and standard deviation of the observed number of surrounding residues, excluding nearest neighbors along the chain, within a sphere R_c centered at interior residues, and the number of interior residues that are defined as those within 7 Å of the center of a protein subunit. [g] Total or weighted average. [h] SLV stands for an effective solvent molecule.

tophan–tryptophan contacts is estimated to be about 6.23 Å. For all of these circumstantial reasons, we have chosen 6.5 Å for R_c.

C. Coordination Numbers, q_i. Residue packing around interior residues cannot be used here to determine the coordination number, q_i, for each type of residue, because of the small numbers of interior residues. Instead, q_i has been estimated from the average residue volume in a manner similar to eq 32, because the number density within the sphere of R_c = 6.5 Å around interior residues is near the mean density; see Figure 1a.

$$q_i = \left(\frac{4\pi}{3}R_c^3 - v_i\right) \Big/ v_{ir}(R_c) - q_{n;i}(R_c)$$
$$q_0 = \left(\frac{4\pi}{3}R_c^3 - v_0\right) \Big/ v_{0r} \qquad (33)$$

where

$$R_c = 6.5 \quad (\text{Å})$$

$$v_0 \equiv v_r \qquad v_{0r} \equiv v_0$$

$q_{n;i}(R_c)$ is the average number of nearest neighbors along a chain within a sphere of R_c centered at the ith type of residue. The volume, v_0, of an effective solvent molecule is defined to be equal to the mean volume v_r of a residue and v_{0r} is assumed to be equal to v_0. The values v_i, v_{ir}, $q_{n;i}$ and q_i for R_c = 6.5 Å are shown in Table III. Although the first term in eq 33 tends to take larger values for smaller residues, the variation of q_i among residue types is not large, because whenever the side chain of a central residue is small, then more nearest neighbors along the chain tend to be located within the sphere of R_c; here it should be noted that nearest neighbors along a chain are excluded in the counting of contacts, and the coordination numbers, q_i ($i \neq 0$), for residues are reduced by the presence of these nearest-neighbor residues. The deviations of the coordination numbers from their means are not small as shown for interior residues in Tables II and III. These relatively large ranges will cause some errors in the following estimates.

D. Evaluation of the Number of Effective Solvent Molecules, n_0. The expected number $\bar{n}_{rr}$ of total residue–residue contacts for a hypothetical protein with e_{ij} = 0 has been evaluated with eq 26–31 based on the random flight approximation for peptide chains. The values of residue volumes v_i shown in Table III are employed to calculate the excluded volume β with eq 29c and 20. For polymeric proteins, $\bar{n}_{rr}$ has been taken to be equal to the sum of $\bar{n}_{rr}$ for each subunit. Then n_0 has been calculated from eq 17 with the values of q_i defined in the preceding section; n_0/n_r for each protein is listed in Table I. The chain length dependences of $\bar{n}_{rr}$ and n_0 are shown in Figure 2 for a chain whose amino acid composition is that of the average composition of the present sample of proteins. As expected, $\bar{n}_{rr}$ and n_0 are almost proportional to the chain length, n_r, in the long-chain limit. The ratios $\bar{n}_{rr}/n_r$ and n_0/n_r are 0.855 and 2.33 for a chain of 200 residues, indicating that the local density of residues in the vicinity of each residue is high even in the case of no attractive interactions between residues.

E. Contact Energies, e_{ij}' and e_{ij}. In the lower triangular part of Table IV are shown the sums, N_{ij}, of the numbers of contact pairs i–j over all proteins and in the upper triangular part their expected numbers, C_{ij}, C_{ii}', C_{i0}', and C_{00}', for the case of random mixing; see eq 11, 12, 14, and 15 for the definitions of these quantities. In order to remove biases arising from the short-range order of amino acid sequences, nearest neighbors along a chain have been excluded in counting contacts. The contact energies, e_{ij}' and e_{ij}, calculated with eq 10, 13, and 6a are shown in Table V; note that $e_{i0}' = -0.5e_{ii}$ according to eq 6. The small numbers of contacts sampled may limit the precision of the estimated contact energies.

Table IV indicates that the correction factors in eq 10 and 13 are not negligible; the correction factors for all i and j have values larger than one, mainly due to the difference between auto- and cross-correlations of amino acid composition. The largest corrections are found for cysteine, because of its S–S bond capability. The values of 0.5 ln $(C_{ii}C_{jj}/C_{ij}^2)$ as required in eq 10 and 0.5 ln $(C_{ii}'C_{00}'/C_{i0}'^2)$ for eq 13 range from 0.26 to 0.70 for Cys-X pairs, from 0.20 to 0.48 for His-X, from 0.19 to 0.48 for Met-X, from 0.18 to 0.31 for Trp-X, from 0.17 to 0.39 for Gln-X, from 0.13 to 0.37 for Tyr-X, and from 0.06 to 0.28 for all others.

The estimated values of contact energies, e_{ij}', display many of the expected characteristics; here the definition of e_{ij}', eq 5b, should be recalled, that is, e_{ij}' is the energy difference accompanying the formation of a contact pair i–j from contact pairs i–i and j–j. (1) The formation of Cys-X contacts from Cys-Cys and X-X contacts represents a relatively large energy loss, because Cys-Cys contacts often form disulfide bonds. (2) The contact formations between negatively charged (Glu, Asp) and positively charged residues (Arg, Lys) are preferable to contacts between residues of the same type because of favorable electrostatic interactions. The magnitudes of the interactions of glutamic acid and aspartic acid with histidine are smaller than for lysine and arginine because of its smaller average charge. (3) Tyrosine and to a smaller extent tryptophan prefer contacts with polar residues probably because of the presence of a polar atom in their side chains, although they have hydrophobic characteristics as indicated by large negative values of e_{ij}. (4) The segregation of hydrophobic and hydrophilic residues can be seen directly from the values of e_{ij}'. e_{ij}' among hydrophobic residues (Met, Phe, Ile, Leu, and Val) takes small positive or negative values, indicating that these residues do not have strong specific preferences but are almost randomly mixed in protein structures. Hydrophilic residues (Thr, Ser, Asn, Gln, His, Arg, Lys, and Pro) for the most part prefer contacts with each other to those between the same type of residues; in the case of charged residues, the subtracted unfavorable electrostatic interactions would in part be responsible for this. The large positive values of e_{ij}' among pairs composed of a hydrophobic and a hydrophilic residue are a manifestation of the segregation between them, that is, nonpolar-residue-in and polar-residue-out, although this originates principally in the differences of e_{i0}' among residues. (5) The values of $e_{i0}' = -0.5e_{ii}$ coincide with the general characteristics of hydrophobicity and hydrophilicity of each residue; however, it should be noted that the e_{i0}' does not properly represent mean characteristics such as hydrophobicity but is directly related to the energy change on transfer of the ith type of residue from its pure state to water, and therefore e_{i0}' for charged residues would include removing unfavorable electrostatic energies specific to the same residue–residue pair.

F. Partition Energies of Residues to Protein Interior. In the following, a simple quantity which is related to the propensity of residues to be exposed to water in protein structures is presented. Equation 9a is transformed with eq 1 and 2 as follows.

$$\frac{\bar{n}_{i0}}{\bar{n}_{r0}} = \frac{\sum_{j=1}\bar{n}_{ij}\exp(e_{ij})}{\sum_{i=1}\sum_{j=1}\bar{n}_{ij}\exp(e_{ij})} = \frac{(q_i/2)n_i}{\sum_{i=1}(q_i/2)n_i}\left[1 + \left(1 - \frac{\bar{n}_{r0}}{\sum_{i=1}(q_i/2)n_i}\right)(\exp(e_{rr} - e_{ir}) - 1)\right]^{-1} \quad (34)$$

where

$$\exp(-e_{ir}) \equiv \left[\frac{\sum_{j=1}\bar{n}_{ij}\exp(e_{ij})}{\bar{n}_{ir}}\right]^{-1} = \frac{\sum_{j=1}\bar{n}_{j0}\exp(-e_{ij})}{\bar{n}_{r0}} = \frac{\bar{n}_{ir}\bar{n}_{00}}{\bar{n}_{i0}\bar{n}_{r0}} \quad (35a)$$

$$\exp(-e_{rr}) \equiv \left[\frac{\sum_{i=1}\bar{n}_{ir}\exp(e_{ir})}{\bar{n}_{rr}}\right]^{-1} = \frac{\sum_{i=1}\bar{n}_{i0}\exp(-e_{ir})}{\bar{n}_{r0}} = \frac{\bar{n}_{rr}\bar{n}_{00}}{\bar{n}_{r0}\bar{n}_{r0}} \quad (35b)$$

Table IV
Numbers of Contacts: Upper Triangle for Random Mixing and Lower Triangle for Actual Counts in Protein Sample

		SLV[a]	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO		
		64671	166	58	260	480	1024	1184	58	315	1488	1788	896	1294	322	479	543	760	128	255	1134	283		$C_{ii}'\times10$
			4271	3098	7374	9705	15027	15604	3105	7154	17889	18296	13120	15371	7594	9279	10909	13024	4471	6721	15670	7839	SLV[a]	$C_{i0}'\times20$
SLV[a]	210696	659988	303	203	447	728	861	1001	217	451	791	845	571	700	331	446	337	446	200	260	375	325	CYS	
CYS	1830	1959	790	124	393	578	782	795	158	299	676	612	450	484	198	262	313	344	160	229	345	230	MET	
MET	1390	685	160	120	576	1290	1916	1910	370	751	1664	1530	1090	1185	513	630	729	870	411	498	784	560	PHE	
PHE	3420	1418	535	600	890	993	2424	2517	536	1050	2137	2133	1424	1593	723	882	952	1132	520	681	1014	735	ILE	
ILE	4520	2226	590	765	1560	1250	2057	3822	799	1431	3258	3008	2051	2307	1025	1252	1421	1628	863	1030	1590	1113	LEU	
LEU	6750	4312	885	960	2580	3650	2700	2103	783	1438	3200	3228	2208	2464	1035	1291	1370	1631	871	1032	1488	1150	VAL	
VAL	7200	6496	895	920	2375	3250	5115	2800	111	335	671	631	459	498	237	284	281	331	168	228	311	244	TRP	
TRP	1490	970	190	310	435	530	875	845	125	440	1265	1297	945	1009	505	613	519	733	326	437	589	484	TYR	$C_{ii}\times10$
TYR	3310	4717	410	340	750	940	1120	1405	350	465	1492	2648	1844	2000	888	1111	1199	1417	693	882	1322	913	ALA	
ALA	7790	17471	755	530	1615	2680	3290	3555	735	1160	1703	1499	1909	2051	877	1105	1056	1365	681	892	1135	910	GLY	or
GLY	8230	21662	860	500	1035	1655	2375	2715	525	1120	2815	2160	716	1404	646	770	738	960	454	607	809	648	THR	
THR	5700	15651	450	370	795	1245	1670	1720	370	825	2020	2460	695	829	675	869	819	1039	509	648	911	719	SER	$C_{ij}\times20$
SER	6720	20669	680	420	885	1295	1805	2065	355	925	2120	2385	1520	1010	206	384	348	468	206	267	394	334	GLN	
GLN	3260	10799	300	125	315	505	795	815	220	615	825	925	740	775	155	277	453	606	286	355	515	391	ASN	
ASN	4040	13206	370	240	520	515	920	780	265	740	835	1250	850	990	490	465	321	596	299	375	640	422	GLU	
GLU	4540	14943	160	200	550	670	920	970	210	570	920	810	730	1055	410	615	325	423	366	444	655	499	ASP	
ASP	5310	18751	410	180	585	630	830	910	260	730	1260	1680	1280	1420	555	815	690	405	134	228	343	232	HIS	
HIS	1990	4221	230	60	480	320	735	720	190	460	495	555	550	535	210	360	465	540	180	178	402	303	ARG	
ARG	2920	8507	195	170	330	450	725	770	270	565	575	920	860	550	450	405	785	920	260	160	363	462	LYS	
LYS	6320	26603	175	345	505	710	1310	960	250	725	1145	1000	880	1030	575	705	1665	1390	275	190	280	200	PRO	
PRO	3670	10256	310	325	460	575	735	1055	325	675	830	1005	695	710	510	465	465	480	285	445	415	155		
		SLV[a]	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO		

$N_i\times10$ $N_{ii}\times10$ or $N_{ij}\times20$

Total are: N_t= 9040, N_{rr}= 18192.25, $2\times N_{r0}$= 20552.1

[a] SLV stands for effective solvent molecules.

Table V
Contact Energies in RT Units: e_{ij} for Upper Half and Diagonal and e_{ij}' for Lower Half

e_{ij}' \ e_{ij}	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO
CYS	-5.44	-5.05	-5.63	-5.03	-5.03	-4.46	-4.76	-3.89	-3.38	-3.16	-2.88	-2.86	-2.73	-2.59	-2.08	-2.66	-3.63	-2.70	-1.54	-2.92
MET	0.70	-6.06	-6.68	-6.33	-6.01	-5.52	-6.37	-4.92	-3.99	-3.75	-3.73	-3.55	-3.17	-3.50	-3.19	-2.90	-3.31	-3.49	-3.11	-4.11
PHE	0.52	-0.22	-6.85	-6.39	-6.26	-5.75	-6.02	-4.95	-4.36	-3.72	-3.76	-3.56	-3.30	-3.55	-3.51	-3.31	-4.61	-3.54	-2.83	-3.73
ILE	0.80	-0.18	0.14	-6.22	-6.17	-5.58	-5.64	-4.63	-4.41	-3.65	-3.74	-3.43	-3.22	-2.99	-3.23	-2.91	-3.76	-3.33	-2.70	-3.47
LEU	0.59	-0.09	0.06	-0.16	-5.79	-5.38	-5.50	-4.26	-3.96	-3.43	-3.43	-3.16	-3.09	-2.99	-2.91	-2.59	-3.84	-3.15	-2.63	-3.06
VAL	0.73	-0.02	0.14	0.00	-0.01	-4.94	-5.05	-4.05	-3.62	-3.06	-2.95	-2.79	-2.67	-2.36	-2.56	-2.25	-3.38	-2.78	-1.95	-2.96
TRP	0.67	-0.63	0.12	0.19	0.11	0.13	-5.42	-4.44	-3.93	-3.37	-3.31	-2.95	-3.16	-3.11	-2.94	-2.91	-4.02	-3.56	-2.49	-3.66
TYR	0.60	-0.12	0.25	0.25	0.41	0.19	0.04	-3.55	-2.85	-2.50	-2.48	-2.30	-2.53	-2.47	-2.42	-2.25	-3.33	-2.75	-2.01	-2.80
ALA	0.59	0.29	0.31	-0.05	0.19	0.10	0.03	0.18	-2.51	-2.15	-2.15	-1.89	-1.70	-1.44	-1.51	-1.57	-2.09	-1.50	-1.10	-1.81
GLY	0.64	0.37	0.79	0.55	0.56	0.50	0.43	0.36	0.19	-2.17	-2.03	-1.70	-1.54	-1.56	-1.22	-1.62	-1.94	-1.68	-0.84	-1.72
THR	0.70	0.16	0.52	0.23	0.33	0.38	0.26	0.15	-0.04	-0.09	-1.72	-1.59	-1.59	-1.51	-1.45	-1.66	-2.31	-1.97	-1.02	-1.66
SER	0.61	0.22	0.61	0.42	0.48	0.42	0.50	0.21	0.11	0.13	0.00	-1.48	-1.37	-1.31	-1.48	-1.46	-1.94	-1.22	-0.83	-1.35
GLN	0.43	0.30	0.56	0.33	0.25	0.24	-0.01	-0.31	0.00	-0.01	-0.29	-0.18	-0.89	-1.36	-1.33	-1.26	-1.85	-1.85	-1.02	-1.73
ASN	0.93	0.33	0.67	0.91	0.70	0.91	0.39	0.10	0.61	0.32	0.14	0.23	-0.13	-1.59	-1.43	-1.33	-2.01	-1.41	-0.91	-1.43
GLU	1.23	0.43	0.50	0.47	0.58	0.49	0.36	-0.06	0.34	0.45	0.00	-0.15	-0.30	-0.04	-1.18	-1.23	-2.27	-2.07	-1.60	-1.40
ASP	0.54	0.61	0.59	0.68	0.79	0.71	0.28	0.01	0.16	-0.05	-0.32	-0.23	-0.33	-0.06	-0.16	-0.96	-2.14	-1.98	-1.32	-1.19
HIS	0.48	1.11	0.21	0.75	0.44	0.48	0.08	-0.17	0.55	0.53	-0.06	0.19	-0.02	0.18	-0.29	-0.26	-2.78	-2.12	-1.09	-2.17
ARG	0.71	0.23	0.58	0.48	0.43	0.38	-0.16	-0.28	0.44	0.10	-0.42	0.21	-0.72	0.07	-0.79	-0.80	-0.04	-1.39	-0.06	-1.85
LYS	1.11	-0.15	0.53	0.34	0.20	0.45	0.15	-0.31	0.08	0.18	-0.23	-0.15	-0.65	-0.19	-1.08	-0.90	0.24	0.57	0.13	-0.67
PRO	0.40	-0.49	0.29	0.23	0.42	0.10	-0.36	-0.43	0.03	-0.04	-0.21	-0.02	-0.69	-0.04	-0.22	-0.11	-0.19	-0.56	-0.15	-1.18
e_{rr} -2.34 e_{ir}	-3.36	-4.22	-4.37	-4.17	-3.93	-3.49	-3.82	-2.91	-2.36	-2.06	-2.04	-1.81	-1.75	-1.70	-1.74	-1.67	-2.41	-1.92	-1.23	-1.89
e_r -3.18 e_i	-4.00	-4.91	-5.12	-4.88	-4.65	-4.17	-4.36	-3.24	-2.82	-2.34	-2.30	-2.07	-1.98	-1.90	-1.94	-1.81	-2.75	-2.18	-1.50	-2.22

Table VI
Values of $e_{ij} + e_{rr} - e_{ir} - e_{jr}$

CYS	-1.06																			
MET	0.19	0.04																		
PHE	-0.23	-0.42	-0.44																	
ILE	0.16	-0.28	-0.19	-0.22																
LEU	-0.08	-0.20	-0.30	-0.41	-0.27															
VAL	0.06	-0.14	-0.22	-0.25	-0.29	-0.29														
TRP	0.08	-0.67	-0.16	0.02	-0.09	-0.07	-0.12													
TYR	0.04	-0.13	0.00	0.11	0.24	0.02	-0.04	-0.06												
ALA	0.00	0.25	0.03	-0.22	-0.01	-0.10	-0.09	0.09	-0.13											
GLY	-0.08	0.19	0.38	0.25	0.23	0.16	0.18	0.14	-0.07	-0.38										
THR	0.19	0.19	0.31	0.14	0.20	0.25	0.22	0.13	-0.09	-0.26	0.03									
SER	-0.02	0.14	0.29	0.21	0.25	0.18	0.34	0.09	-0.06	-0.16	-0.08	-0.20								
GLN	0.05	0.46	0.49	0.36	0.26	0.24	0.08	-0.20	0.08	-0.06	-0.14	-0.14	0.29							
ASN	0.13	0.08	0.18	0.53	0.30	0.50	0.06	-0.20	0.28	-0.14	-0.11	-0.14	-0.25	-0.53						
GLU	0.69	0.44	0.27	0.35	0.43	0.34	0.29	-0.10	0.26	0.25	0.00	-0.26	-0.17	-0.32	-0.03					
ASP	0.03	0.65	0.39	0.59	0.67	0.58	0.24	0.00	0.12	-0.22	-0.29	-0.31	-0.17	-0.30	-0.15	0.04				
HIS	-0.19	0.99	-0.16	0.49	0.16	0.19	-0.12	-0.34	0.34	0.20	-0.19	-0.05	-0.02	-0.24	-0.45	-0.39	-0.29			
ARG	0.24	0.31	0.41	0.42	0.35	0.30	-0.16	-0.25	0.43	-0.04	-0.35	0.17	-0.52	-0.14	-0.74	-0.72	-0.12	0.11		
LYS	0.71	0.00	0.44	0.36	0.19	0.44	0.22	-0.21	0.14	0.11	-0.09	-0.13	-0.38	-0.33	-0.97	-0.76	0.22	0.75	0.25	
PRO	0.00	-0.34	0.20	0.25	0.42	0.09	-0.28	-0.33	0.10	-0.11	-0.07	0.01	-0.42	-0.18	-0.10	0.04	-0.21	-0.38	0.11	0.26
	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO

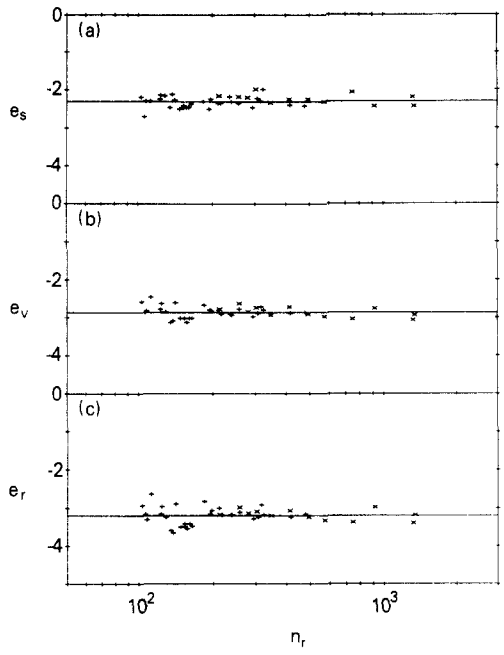

Figure 3. Values of e_s, e_v, and e_r are plotted for each protein against chain length, n_r. e_r is the average energy per contact, and e_s and e_v correspond to the following averages of the average contact energies e_i: over residues located on protein surface, and over the entire amino acid composition, respectively; refer to eq 38 and 39 for exact definitions. The solid lines in (a), (b), and (c) represent the weighted average of e_s, e_v, and e_r over all proteins, respectively; $e_s = -2.300 \pm 0.024$, $e_v = -2.865 \pm 0.023$, and $e_r = -3.184 \pm 0.032$. The marks + and × correspond to monomeric and polymeric proteins, respectively.

e_{ir} and e_{rr} are average energy changes accompanying the contact formations, i–0 + r–0 ↔ i–r + 0–0 and r–0 + r–0 ↔ r–r + 0–0, respectively; r represents the average residue. Positive or negative values of $(e_{ir} - e_{rr})$ indicate whether the ith type of residue tends to be exposed to solvent or buried in the interior of proteins. In this sense, e_{ir} might be termed an effective partition energy for residues in protein structures. Here e_{ir} and e_{rr} have been calculated from the expected values $\bar{n}_{ij}$ of residue–residue contacts calculated by solving the nonlinear simultaneous equations (1) and (9a) or (9b) with the estimated values of e_{ij} or e_{ij}' for a hypothetical protein that consists of 200 residues with the same amino acid composition as the average composition over all proteins listed in Table IV. These values of e_{ir} and e_{rr} are shown at the bottom of Table V.

Residues Phe, Met, Ile, Leu, Trp, Val, and Cys in order tend to be buried in protein interiors with large negative values of $(e_{ir} - e_{rr})$. Tyrosine is less buried than those residues, probably because of its polar hydroxyl side chain. The values of $(e_{ir} - e_{rr})$. for histidine and alanine are almost zero, indicating that it is equally probable whether they are exposed or buried. Glycine and threonine have a weak tendency to be exposed to solvent. Other residues, Lys, Asp, Asn, Glu, Gln, Ser, Pro, and Arg in order from most to least exposed, tend to be more exposed; lysine especially has a strong propensity for exposure. The preference of proline for exposure to water can be attributed to the fact that proline is often observed at bends or turns, which are usually located on protein surface.

The quantities defined in eq 35 are also useful because they are related to an average energy change accompanying the contact formation, i–r + j–r ↔ i–j + r–r, as

$$\frac{\bar{n}_{ij}\bar{n}_{rr}}{\bar{n}_{ir}\bar{n}_{jr}} = \exp(-(e_{ij} + e_{rr} - e_{ir} - e_{jr})) \qquad (36)$$

In other words, $(e_{ij} + e_{rr} - e_{ir} - e_{jr})$ represents the preference for the specific contact pair i–j over the average contacts of the ith and the jth types of residues. Values are shown in Table VI. This quantity takes negative values for most contacts between hydrophobic pairs of residues and between hydrophilic pairs but positive values for most contacts between hydrophobic and hydrophilic pairs, showing that contacts within each of these groups are more favorable than between the two groups. Also, contacts between positively charged residues (Arg-Arg, Arg-Lys, and Lys-Lys) are unfavorable; those for Asp-Asp and Glu-Glu, although not strong, are not favored. This quantity takes a large negative value for Cys-Cys because of its frequent disulfide bond formation.

G. Total Contact Energies of Protein Native Structures. The total contact energy of each protein native structure has been calculated by eq 4a with the values of e_{ij} shown in Table V. Here it is defined as the energy difference, ΔE_C, between crystal structures and completely extended forms with no residue–residue contacts; contacts between nearest-neighbor residues along a chain are assumed to exist in equal amount in both the native and extended conformations. Hence, the total contact energies are represented by

$$\Delta E_C = E_C(\text{for native structure}) - E_C(\text{for extended conformation}) = \sum_{i=1}^{20}\sum_{j=1}^{20} e_{ij}n_{ij} =$$

$$\sum_{i=1} e_i n_{ir} = \sum_{i=1} e_i\left(\frac{q_i}{2}n_i - n_{i0}\right) = e_v\left(\sum_{i=1}\frac{q_i n_i}{2}\right) - e_s n_{r0} \qquad (37a)$$

or

$$= e_r n_{rr} = e_r\left(\sum_{i=1}\frac{q_i n_i}{2} - n_{r0}\right) \qquad (37b)$$

The equations above serve to define e_i, e_v, e_s, and e_r.

$$e_i \equiv \frac{\sum_{j=1} e_{ij}n_{ij}}{n_{ir}} \qquad e_r \equiv \frac{\sum_{i=1} e_i n_{ir}}{n_{rr}} \qquad (38)$$

$$e_v \equiv \frac{\sum_{i=1} e_i(q_i/2)n_i}{\sum_{i=1}(q_i/2)n_i} \qquad e_s \equiv \frac{\sum_{i=1} e_i n_{i0}}{n_{r0}} \qquad (39)$$

e_i corresponds to the average contact energy for the ith type of residue, and e_r is the average energy per contact. e_v and e_s correspond roughly to the average of e_i over amino acid composition and over only residues located on a protein's surface. The weighted averages of e_i and e_r over all proteins are listed in Table V; N_{ij} is used instead of n_{ij} in eq 38; here it should be noted that e_i and e_r take more negative values than e_{ir} and e_{rr} because of differences in averaging. The values of e_v, e_s, and e_r for each protein are listed together with n_{rr} and $2n_{r0}$ in Table I, and plotted against its chain length, in Figure 3. The values of e_v, e_s, and e_r are almost constant for the proteins examined here. The weighted averages of eq 37 over all proteins are

$$\Delta E_C(\text{for native structures}) \simeq (-2.865 \pm 0.023)\sum_{i=1}\frac{q_i n_i}{2} - (-2.300 \pm 0.024)n_{r0} \qquad (40a)$$

or

$$\simeq (-3.184 \pm 0.032)\left(\sum_{i=1}\frac{q_i n_i}{2} - n_{r0}\right) \qquad (40b)$$

The standard deviations of e_v, e_s, and e_r are 0.136, 0.144, and 0.187, respectively; the root mean square errors of eq

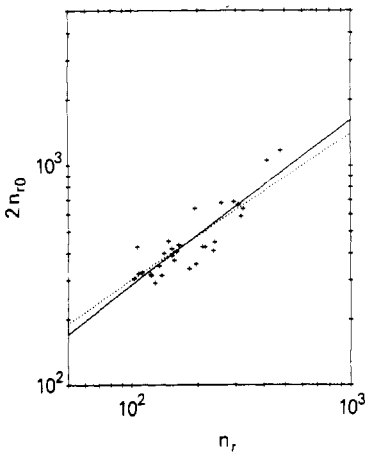

Figure 4. Dependence of the observed number of residue–solvent contacts, $2n_{r0}$, in each monomeric protein on its chain length, n_r. The solid line shows eq 41a in which the power dependence of n_{r0} on n_r corresponds to the slope 0.751 ± 0.078 of the regression line in the log ($2n_{r0}$) vs. log (n_r) plot; the correlation coefficient is 0.876. The dotted line shows eq 40b in which the $^2/_3$ power dependence of n_{r0} on n_r is assumed.

40a and 40b are 74.4 and 76.1 for monomeric proteins, and 89.3 and 76.2 for all proteins, respectively. e_s is about 20% less negative and e_r about 11% more negative than e_v. These differences result from the fact that hydrophilic residues tend to be exposed to solvent and hydrophobic residues tend to be buried in the interior of proteins.

The relationship between the total number of residue–solvent contacts, $2n_{r0}$, and the chain length, n_r, is shown in Figure 4 for 30 monomeric proteins. Least-squares analysis of log ($2n_{r0}$) as a function of log (n_r) yields 0.751 ± 0.078 as the slope; the correlation coefficient is 0.876. This value of the slope is slightly larger than $^2/_3$ for figures of identical shape. The dependence of the surface area on the volume of a protein is not clear-cut; the surface area has been alleged to be proportional to $^2/_3$ power[43] of volume or 0.77 ± 0.02[44] which indicates that larger proteins tend to be more aspherical. The present result is more consistent with the latter; however, it is not certain because of relatively large deviations. Thus, for monomeric proteins the arithmetic mean of the ratio $2n_{r0}/n_r^{0.751}$ gives

$$2n_{r0} \simeq (1.431 \pm 0.046)q_r n_r^{0.751\pm0.078} \quad (41a)$$

or if $^2/_3$ power is assumed, $2n_{r0}$ will be approximated by

$$2n_{r0} \simeq (2.229 \pm 0.074)q_r n_r^{2/3} \quad (41b)$$

where $q_r = 6.298$. The standard deviations of the coefficients in eq 41a and 41b are 0.251 and 0.405, respectively. Then the average number of residue–residue contacts in protein native structures is estimated from eq 41a to be in the range of about 1.71–2.06 per residue for monomeric proteins with $100 < n_r < 300$. From eq 40a and 41, the total contact energy of the native structure for monomeric proteins can be approximated by

$$\Delta E_C(\text{for native structures of monomeric proteins}) \simeq (-2.865 \pm 0.023)\sum_{i=1}\frac{q_i n_i}{2} - (-1.646 \pm 0.069)q_r n_r^{0.751\pm0.078} \quad (42a)$$

or

$$\simeq (-2.865 \pm 0.023)\sum_{i=1}\frac{q_i n_i}{2} - (-2.563 \pm 0.108)q_r n_r^{2/3} \quad (42b)$$

where $q_r = 6.298$. The root mean square errors for eq 42a and 42b are 144.4 and 154.4, respectively.

To discuss the energy gain accompanying protein folding, the contact energy of the denatured state must be estimated; here the denatured state is defined to be the conformational state of unfolded proteins at the midpoint of folding–unfolding transition. Let us think about the folding–unfolding process of proteins induced by increasing temperature. As temperature increases, a protein becomes unfolded and will continue to expand. The size of an unfolded protein at its transition midpoint is characteristic of the protein; however, its size must be smaller than at $e_{ij} = 0$, that is, in the case of no attractive interactions between residues. An upper bound to the contact energy of the denatured state can then be estimated roughly from the total number of residue–residue contacts formed at $e_{ij} = 0$. From the average number of residue–residue contacts at $e_{ij} = 0$ shown in Figure 2

$$\Delta E_C(\text{at denatured state}) < -2.59\bar{n}_{rr} = -2.59 \cdot (0.817 \text{ to } 0.870)n_r \quad (43)$$

for $100 < n_r < 300$. The contact energy per contact in eq 43 has been estimated by assuming random mixing; in this case, e_v, e_s, and e_r are all equal. Then from eq 42a and 43 the upper limit of the energy gain accompanying protein folding is estimated to be in the range 3.60–4.25 per residue for monomeric proteins with $100 < n_r < 300$. The conformational energy gain and entropy loss in protein folding are balanced against each other at the transition midpoint. Therefore the above estimate of an upper bound of conformational energy gain could also correspond to an upper bound of conformational entropy loss accompanying protein folding. With attractive interresidue interactions e_{ij} at the transition midpoint, a denatured protein is expected to be more compact than at $e_{ij} = 0$, and therefore the energy gain and the conformational entropy loss for protein folding would be significantly less than the values above.

H. Intersubunit Contact. For polymeric proteins and protein–inhibitor complexes, intersubunit contact energies have been estimated from the numbers of residue–residue contacts between subunits (Table VII). The average energy per intersubunit contact for each protein is compared with the values of e_r, e_v, and e_s for that protein; $e_r < e_v < e_s$. The average contact energies on major interfaces among subunits are more negative than the value of e_s, indicating that the subunits are associated by more favorable contacts than residue–residue contacts observed on the protein surface. Intersubunit contacts at about half of the interfaces examined here are as favorable as residue–residue contacts observed in the interiors of proteins, because the values of the average energies per contact at these interfaces are more negative than e_r, the average energy over all contacts in the protein complex. These results indicate that interresidue interactions can make favorable contributions to the proper association of subunits or proteins; however, complementarity of the molecular surfaces is also required to yield close-packing[9] and may be essential to obtain sufficient contact energy for association.

Whether or not the present estimates of the intersubunit contact energies approximate the free energies required for those subunit–subunit associations must be examined. The free energy change that originates in the loss of translational and rotational freedoms by protein association can be roughly estimated in the ideal gas approximation. From the experimental values of the dissociation constants, Chothia and Janin[9] estimated the free energies required for association to be about 45 kcal/mol for the trypsin–inhibitor association and greater than 38 kcal/mol for the hemoglobin α–β dimer. The present estimates of the total interprotein contact energies are 61 kcal/mol for the trypsin–inhibitor and 66 kcal/mol for the hemoglobin

Table VII
Intersubunit Contacts

code	protein[a]	#contacts	total contact energy[b]	average energy per contact[b]	c
1CCY	Cytochrome c' dimer A-B	42	-130.9	-3.12	< -2.97 (e_r)
2MHB	Met-hemoglobin α1-β1, α2-β2	41	-110.7	-2.70	< -2.28 (e_s)
	α1-β2, α2-β1	17	-38.5	-2.26	
	α1-α2	6	-13.6	-2.26	
	β1-β2	4	-10.5	-2.61	
1PTC	Trypsin-inhibitor(PTI)	39	-101.5	-2.60	< -2.19 (e_s)
2SOD	Superoxide dismutase dimer O-Y	34	-94.6	-2.78	< -2.73 (e_v)
1REI	Bence-Jones immunogloblin REI dimer A-B	29	-89.8	-3.10	< -2.76 (e_v)
1FC1	Immunogloblin Fc fragment dimer A-B	55	-121.8	-2.21	$\simeq -2.23$ (e_s)
4CPA	Carboxypeptidase α - inhibitor	27	-90.8	-3.36	< -3.18 (e_r)
4ADH	Apo-liver ADH dimer	68	-226.8	-3.34	< -3.03 (e_v)
1GPD	GPDH tetramer G1-R1, G2-R2	49	-117.7	-2.40	
	G1-R2, G2-R1	78	-225.8	-2.90	< -2.41 (e_s)
	G1-G2	2	-4.5	-2.25	
	R1-R2	19	-41.9	-2.21	
4LDH	LDH tetramer 1-3, 2-4	114	-389.9	-3.42	< -3.38 (e_r)
	1-2, 3-4	61	-205.9	-3.38	
	1-4, 2-3	61	-186.8	-3.06	
3PGM	Phosphoglycerate mutase 1-3, 2-4	29	-59.2	-2.04	
	1-4, 2-3	28	-81.3	-2.90	< -2.74 (e_v)
	1-2, 3-4	0	0.0		
1TIM	Triose phosphate isomerase dimer A-B	81	-275.0	-3.39	< -3.23 (e_r)

[a] The subunit interface is specified by the chain identification code and/or number. [b] Energies are in RT units. [c] See eq 38 and 39 for the definitions of e_r, e_v, and e_s; $e_r < e_v < e_s$. Those values are given for each protein in Table I.

dimer; the estimates of the hydrophobic energy gains by Chothia and Janin[9] are 35 and 43 kcal/mol for these molecules, respectively. A definite conclusion cannot be drawn because of the paucity of data and the crude estimates of translational and rotational entropy losses. It is also possible that the free subunits may assume different conformations than in the complex.

I. Comparisons of Estimated Contact Energies with Experimental Data. Many experimental and theoretical works have been performed to estimate hydrophobic energies. Nozaki and Tanford[2] estimated the free energies of transfer of amino acids from solubilities of amino acids in ethanol, dioxane, and water. Similarly, solubilities of liquid hydrocarbons in water were measured by Hermann.[3-5] From these data, hydrophobic energies were analyzed as a linear function of the surface areas of molecules,[3,5-7] although there are controversies[13,15] about whether or not the linear relationship between surface area and free energy change associated with hydrophobic effects is supported on theoretical bases. The proportionality constant between hydrophobic energies for nonpolar side chains (Phe, Leu, Val, and Ala) and their surface areas is 22 (cal/mol)/Å^2 in the analysis of Nozaki's data by Chothia[7] and 33[3] (cal/mol)/Å^2 or 31[5] (cal/mol)/Å^2 for hydrocarbons in Hermann's analysis; Reynolds et al.[6] obtained the different values, 20–25 (cal/mol)/Å^2, from the same data of Hermann.[3] On the other hand, in the study of liquid–crystal phase transitions in fatty acid bilayers, Parsegian[45] assumed the water–hydrocarbon interactions to be in the form of interfacial tension and obtained 18.5–19.5 dyn/cm (27–28 (cal/mol)/Å^2) as the tension energy. Lee[46] has reported that the experimental values of the compressibilities of proteins will be consistent with the values, 25–46 (cal/mol)/Å^2, for the proportionality constant, if the volume fluctuations of proteins are assumed to be subject to a potential that is proportional to the protein surface area.

In the following, estimated contact energies are compared with the experimental values of hydrophobic energies. However, there is no reason to expect an exact correlation, because the contact energies are effective interaction energies between residues including not only hydrophobic energies but other interaction energies specific to proteins such as hydrogen bonding and electrostatic energies as well. Ethanol, dioxane, and liquid hydrocarbons used in the experiments may not be good models for a protein's interior.[13,15] Here it would be important to point out that effective intramolecular interactions in simple polypeptides such as homo- and copolypeptides may also not be the same as those in protein molecules, because the amino acid sequences of globular proteins are highly heterogeneous and therefore the local environment surrounding a residue might be significantly different. In addition, most polypeptides cannot realize so high a

packing density as in proteins; as stated in the Introduction, it has been indicated[15] from liquid theories that the free energy change of transfer would depend significantly on the packing density. Estimated contact energies e_{ij} or e_{ij}' include the effects of the environment specific to protein molecules as mean effects, and the values of the coordination numbers q_i used here are also specific to protein interiors. Thus, the following is a comparison between two estimations of solvent effects, one theoretical and the other experimental.

First, let us compare the average contact energy with the proportionality constant for hydrophobic energy. To compare with each other, the average contact energy must be represented in terms of an interfacial tension, that is, as energy per contact area; here it must be noted that the individual values of e_{ij} are not expected to be proportional to contact areas but rather to depend significantly on the types of residues. The accessible surface area A_S of monomeric proteins and the total area A_T in their extended conformations can be well approximated[43] by

$$A_S = (11.116 \pm 0.161)M^{2/3} = 255n_r^{2/3} \quad (\text{Å}^2) \quad (44a)$$

$$A_T = (1.449 \pm 0.006)M = 159n_r \quad (\text{Å}^2) \quad (44b)$$

where M is the molecular weight of a protein; we have used 110 for the average molecular weight of a residue.[11] Here it should be noted that the total accessible surface area, A_T, in the extended conformations was evaluated as the sum of the surface areas of residues X in the extended conformation of Gly-X-Gly with the trans conformation of its side chain.[8] By assuming that the surface area of a protein is proportional to the number of residue–solvent contacts on the average, A_S follows directly from eq 41b for monomeric proteins. A_T may be approximated by assuming that there are no residue–residue contacts in the extended conformations except nearest-neighbor contacts.

$$A_S \simeq 2a_c n_{r0} = 2.229a_c q_r n_r^{2/3} \quad (\text{Å}^2) \quad (45a)$$

$$A_T \simeq a_c \sum_{i=1} q_i n_i = a_c q_r n_r \quad (\text{Å}^2) \quad (45b)$$

where $q_r = 6.298$, and a_c is the mean area per residue–residue or residue–solvent contact; $2n_{r0}$ is the total number of residue–solvent contacts. a_c can be crudely estimated by equating eq 44 to eq 45.

$$a_c \simeq 18.2 \text{ to } 25.3 \quad (\text{Å}^2) \quad (46)$$

Then eq 37 can be transformed to a form consistent with the definition of hydrophobic energy.

$$(-\Delta E_C) = \frac{-e_v}{2a_c}(a_c \sum_{i=1} q_i n_i) - \frac{-e_s}{2a_c}(2a_c n_{r0}) \quad (47a)$$

$$= \frac{-e_r}{2a_c}(2a_c n_{rr}) \quad (47b)$$

The terms in parentheses in eq 47a and 47b represent the total contact area in the extended conformation, the total residue–solvent contact area, and the total area buried by forming residue–residue contacts, respectively. With the values of e_v, e_s, and e_r shown in eq 40a and 40b, the proportionality constants in the first and second terms of eq 47a and in eq 47b are found to be 0.0566–0.0788, 0.0454–0.0633, and 0.0629–0.0876 Å^{-2}. If RT is taken as 0.6 kcal/mol, they are 34.0–47.3, 27.3–38.0, and 37.7–52.5 (cal/mol)/Å^2, larger than but less than twice the typical values, 25–30 (cal/mol)/Å^2, of hydrophobic energy, but notably closer to the range of 25–46 (cal/mol)/Å^2 derived by Lee.[46]

Next, let us compare contact energies for several amino acids with the experimental values of their hydrophobic

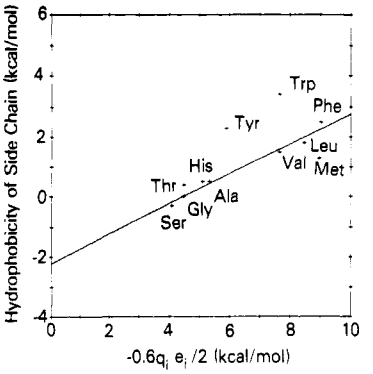

Figure 5. Hydrophobicities of amino acid side chains estimated by Nozaki and Tanford[2] and their corresponding values of $-0.6q_ie_i/2$ with the present model; $RT = 0.6$ kcal/mol for temperature has been employed to translate the contact energies into kcal/mol units. The hydrophobicity of glycine side chain is taken to be zero. $-0.6q_ie_i/2$ corresponds to the average contact energy gain of the ith type of a residue completely surrounded by other residues in protein crystal structures; e_i is defined by eq 38 with N_{ij} instead of n_{ij} and listed in Table V. The solid line shows a regression line for nonpolar residues, Phe, Leu, Val, and Ala, and passes close to the point for Gly. The slope and intercept of the regression line are 0.50 ± 0.08 and −2.20 ± 0.62, respectively. The correlation coefficient is 0.975. Although some polar amino acids are plotted in this figure, the comparison is meaningful only for nonpolar residues.

energies. The energy change of transfer of the ith type of residue from its pure state to water is represented in the present formalism as $q_ie_{i0}' = q_i(-e_{ii}/2)$. A solubility measurement of crystalline amino acids in water yields 13–15[42] (cal/mol)/Å^2 as the proportionality constant for hydrophobic energy. Richards[42] has considered this value to be smaller than other estimates in part because of the entropy difference between the amino acid in the crystal and in organic solvents. As pointed out by Richards[42], a crystal might not be a good model because there is more motion in a protein molecule than in most simple organic crystals. In Nozaki and Tanford's experiments,[2] the organic solvents were used to represent the protein interior. In Figure 5, their experimental free energies of transfer for amino acid side chains are plotted against $-q_ie_i/2$ that corresponds to the average energy gain of the ith type of a residue completely surrounded by other residues in protein crystal structures; e_i is defined by eq 38 with N_{ij} instead of n_{ij} and given in Table V. For comparison, $RT = 0.6$ kcal/mol for temperature has been employed in this figure to express the contact energies in kcal/mol. Although some polar amino acids are plotted in this figure, the comparison is meaningful only for nonpolar residues because the organic solvents cannot represent circumstances surrounding polar residues in the protein native structures; e_i for polar residues includes not only hydrophobic energies but also the average of other interaction energies with surrounding residues such as hydrogen bonds and electrostatic energies specific to this type of residue. For nonpolar residues (Phe, Leu, Val, and Ala), whose hydrophobicities were found by Chothia[7] to be proportional to the accessible surface areas of their side chains, there is a linear relationship between their hydrophobicities and the values of $-0.6q_ie_i/2$; the correlation coefficient is 0.975. The regression line with the slope 0.50 ± 0.08 and the intercept −2.20 ± 0.62 passes close to the point for glycine whose side chain hydrophobicity is plotted as zero. Here it should be noted that e_{ir} could be employed instead of e_i in this analysis; the slope and the intercept of a regression line in a similar plot with e_{ir} are 0.57 ± 0.08 and −2.10 ± 0.53, and the correlation coefficient is 0.980. This figure lends support to the

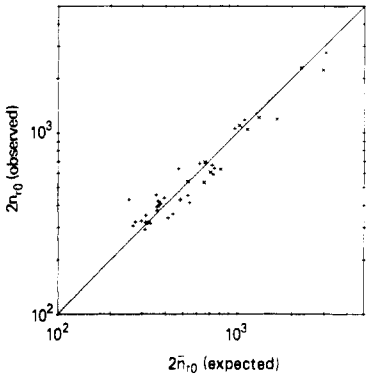

Figure 6. Comparison of the observed number of residue–solvent contacts, $2n_{r0}$, with its expected number, $2\bar{n}_{r0}$, is shown for each protein. The expected numbers of residue–solvent contacts have been calculated by solving the simultaneous equations (1) and (9a) with the estimated values of e_{ij}. The marks + and × are used to represent monomeric and polymeric proteins, respectively. The data points are expected to fall on the solid line with the slope of 1; however, the actual slope of the reression line is 0.82 ± 0.08 for monomeric proteins and 0.87 ± 0.04 for all proteins.

present estimates of relative contact energies.

J. Limitations of the Present Approximation. Effective interresidue contact energies have been estimated from the observed numbers of contacts in protein crystal structures. Conversely, the expected values, $\bar{n}_{ij}$, of contacts for each protein can be calculated by solving the simultaneous equations (1) and (9a) or (9b) with the estimated values of contact energies, e_{ij} or e_{ij}'. This process provides a direct test of the present approximation. Figure 6 shows the plot of the observed number, $2n_{r0}$, of residue–solvent contacts against the expected value, $2\bar{n}_{r0}$, for each protein; the values are listed in Table I. If the approximation were good, the data points would fall on the solid line of slope one; however, the actual slope of the regression line is 0.82 ± 0.08 for monomeric proteins and 0.87 ± 0.04 for all proteins. The deviation of the slope from one comes from the fact that the power dependence of the expected value $2\bar{n}_{r0}$ on chain length is 0.93 ± 0.02 for monomeric proteins and 0.97 ± 0.01 when polymeric proteins are included; whereas, the power dependence of the observed $2n_{r0}$ is 0.751 ± 0.078 for monomeric proteins. In the following, we will consider why the correct surface–volume ratio cannot be predicted. From eq 35b, the expected value $\bar{n}_{r0}$ can be represented as follows.

$$2\bar{n}_{r0} = \frac{q_0 n_0 q_r n_r}{q_0 n_0 + q_r n_r} \frac{2}{1+\gamma} \tag{48}$$

where

$$\gamma \equiv \left[1 + 4\left(\frac{q_0 n_0 q_r n_r}{(q_0 n_0 + q_r n_r)^2}\right)(\exp(-e_{rr}) - 1)\right]^{1/2} \tag{49}$$

The definitions of q_r and e_{rr} are from eq 18 and 35b. Equations 48 and 49 indicate that if e_{rr} is almost constant, the expected value $2\bar{n}_{r0}$ will be roughly proportional to chain length, because n_0 is known from Figure 2 to be approximately proportional to chain length except for chains shorter than 100 residues. If the composition of each type of amino acid and effective solvent molecules is constant, strictly $q_i n_i / \sum_{i=0} q_i n_i$ for all i is constant, then $\bar{n}_{ij} / \sum_{i=0} q_i n_i$ will not depend on chain length; see eq 1 and 9; therefore e_{ir} and e_{rr} are constant for this case. Of course, if the fraction of hydrophilic residues were proportional to the surface–volume ratio of proteins, then e_{rr} could become more negative for larger proteins, making the power dependence of $\bar{n}_{r0}$ on chain length smaller. However,

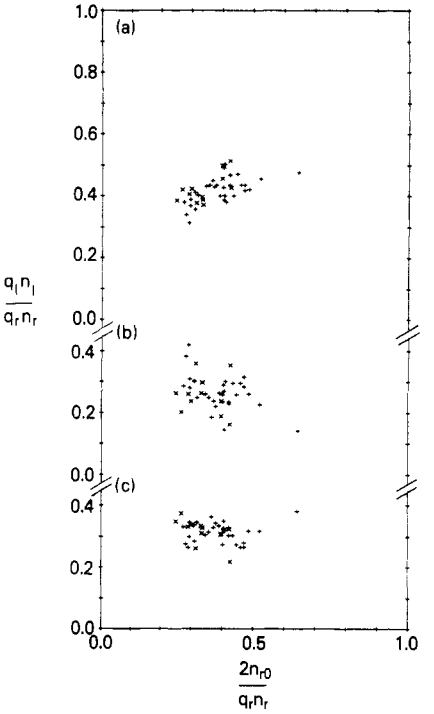

Figure 7. Fractions of (a) hydrophilic, (b) neutral, and (c) hydrophobic residues, strictly $q_I n_I/(q_r n_r)$ for each group I, are plotted against the fractions of residue–solvent contacts, $2n_{r0}/(q_r n_r)$, for each protein. Residues are classified into three groups according to the values of e_{ir} in Table V: hydrophobic residues: Phe, Met, Ile, Leu, Trp, Val, Cys, and Tyr; neutral residues: His, Ala, Gly, and Thr; hydrophilic residues: Lys, Asp, Asn, Glu, Gln, Ser, Pro, and Arg. The marks + and × are used to represent monomeric and polymeric proteins, respectively.

as we will see below, there appears to be no significant dependence of amino acid composition on surface–volume ratio or chain length.

To examine the dependence of amino acid composition on the surface–volume ratio, residues have been classified into three groups according to the values of e_{ir} from Table V: hydrophobic residues consisting of Phe, Met, Ile, Leu, Trp, Val, Cys, and Tyr; neutral residues consisting of His, Ala, Gly, and Thr; and hydrophilic residues consisting of Lys, Asp, Asn, Glu, Gln, Ser, Pro, and Arg. The sum, $q_I n_I$, of $q_i n_i$ over residues within each of these groups I has been calculated for each protein and the dependence of the fraction $q_I n_I/(q_r n_r)$ on the fraction of residue–solvent contacts, $2n_{r0}/(q_r n_r)$, is shown in Figure 7. There appears to be no significant correlation between amino acid composition and surface–volume ratio; the correlation coefficients between them are 0.57 for hydrophilic residues, −0.38 for neutral residues, and −0.10 for hydrophobic residues.

The present approximation yields $2\bar{n}_{r0}$ proportional to chain length and cannot reproduce the correct surface–volume ratio. This is most likely because the quasi-chemical approximation is inadequate for systems in which molecules interact strongly with one another: only the occurrence probabilities of contact pairs and no larger clusters are taken into account. Interactions of higher order than binary clusters are likely to play a role in a close-packed protein structure formed with strongly attractive interactions between residues.

Next, it is of interest to examine how well the present approximation can reproduce the observed preference for partitioning of each group of residues into the interior or onto the surface of proteins. The fraction $2n_{I0}/(q_I n_I)$ of residue–solvent contacts in each group of residues has been calculated for each protein, where n_{I0} and $q_I n_I$ are defined

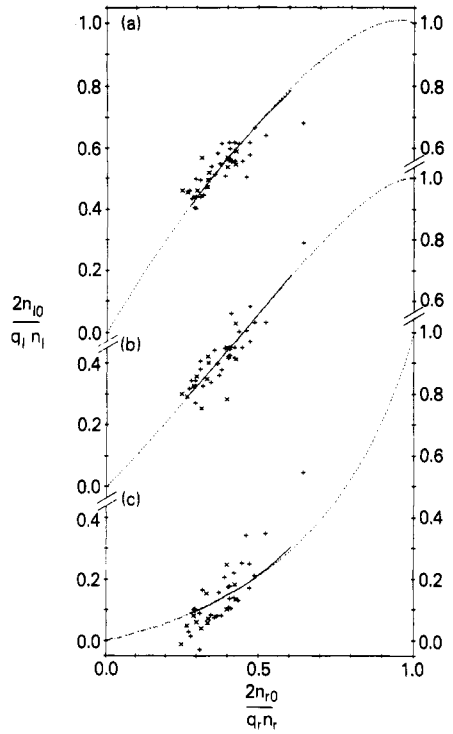

Figure 8. Fractions of exposure to water in (a) hydrophilic, (b) neutral, and (c) hydrophobic residues, $2n_{I0}/(q_I n_I)$ for each group I, are plotted against the fraction of residue–solvent contacts, $2n_{r0}/(q_r n_r)$, for each protein; see the legend of Figure 7 for the classification of the residue groups. The marks + and × are used to represent monomeric and polymeric proteins, respectively. The dotted lines show the expected values calculated from eq 34 with the values of e_{ir} and e_{rr} given in Table V, and the solid lines those obtained by solving the simultaneous equations (1) and (9a) with the estimated values of e_{ij}; $2\bar{n}_{r0}/(q_r n_r)$ is treated as a parameter and the amino acid composition is taken to be equal to the occurrence frequencies of amino acids over all proteins. The dotted line for hydrophilic residues takes unrealistic values larger than 1 near the right side. This is an artifact due to the approximation of treating $2\bar{n}_{r0}/(q_r n_r)$ as a parameter in eq 34. The crudeness of the present method to evaluate the numbers of residue–solvent contacts is responsible for the small negative values of n_{I0} for hydrophobic residues in a few proteins.

to be the sums of n_{i0} and $q_i n_i$ over residues within each group I above. Their dependences on the surface–volume ratios of proteins are shown in Figure 8 as plots of $2n_{I0}/(q_I n_I)$ against $2n_{r0}/(q_r n_r)$. The dotted lines in Figure 8 show the expected values calculated from the sum of eq 34 for group I with the values of e_{ir} and e_{rr} shown in Table V by treating $2\bar{n}_{r0}/(q_r n_r)$ as a parameter; eq 34 has been modified by including a normalization constant, and the amino acid composition has been taken to be equal to the occurrence frequencies of residues over all proteins. The solid lines that almost overlap the dotted lines show the expected values calculated by solving the simultaneous equations (1) and (9a) or (9b) with the estimated values of e_{ij} or e_{ij}' under a restriction to yield specified values for $2\bar{n}_{r0}/(q_r n_r)$; again the amino acid composition is taken as the average over all proteins. The observed values for neutral residues fall relatively near the dotted line, but the observed values for hydrophilic and hydrophobic residues consistently deviate with somewhat smaller and larger slopes than the dotted lines, respectively. These persistent deviations also indicate some limitation to the present approximation.

Discussion

The effective contact energies between residues in protein molecules have been estimated from the numbers of residue–residue contacts observed in protein crystal structures and estimates of the numbers of effective solvent molecules by means of the quasi-chemical approximation with a crude approximation for the effects of chain connectivity. The estimated values of contact energies have reasonable residue-type dependences, and also there is a linear relationship between the average contact energies for nonpolar residues and their hydrophobicities evaluated by Nozaki and Tanford[2]; however, the average contact energy is about twice as large as previous estimates of hydrophobic energies except those derived by Lee.[46] This difference in magnitude may be attributable to the crude approximation for the effects of chain connectivity and the limitations of the quasi-chemical approximation. However, the two quantities compared are substantively different so that in principal there is no reason to expect coincidence of the values of contact energies estimated in the present model with those hydrophobic energies, especially since the former includes not only hydrophobic energy but also average contributions of electrostatic, hydrogen bonds, and van der Waals energies in circumstances different than for the hydrophobic energies. In addition, there is the fundamental question of whether liquid hydrocarbons and organic solvents such as ethanol and dioxane can serve as adequate models for protein interiors.[13,15] It is noteworthy that Lee's estimates[46] are not hydrophobic energies obtained from bulk solvent transfer data but effective interfacial tension energies of globular proteins obtained from compressibility data. Therefore, it would be reasonable that the present estimates are closer to Lee's estimates than to others.

Chothia et al.[7-11] evaluated the hydrophobic energy gain accompanying protein folding on the basis of differences in surface areas between protein crystal structures and their completely extended forms. In those analyses, the proportionality constant between hydrophobic energy and surface area has been assumed to be the same for all types of residues with the actual value taken from nonpolar residues. This assumption appears to be supported[10] by the observations that polar groups, if buried in proteins, are almost always hydrogen bonded and polar atoms, if hydrogen bonded, resemble nonpolar atoms in their hydrophobicities. However, there is the report of Finney et al.[12] that the distortions from the ideal geometry of internal hydrogen bonds can yield large energy penalties almost comparable to the entropic gain from the release of water molecules bound to polar atoms in the process of protein folding. Even though the assumption of Chothia[10] is reasonable, the observed radial distribution of residues in protein crystal structures, that is, polar-residue-out and nonpolar-residue-in, cannot be predicted from his considerations alone. Contact energies estimated in the present model depend strongly on residue type and clearly reflect this distribution of residues in protein structures. Thus use of contact energies of the present type may be useful in achieving this ubiquitous feature of the tertiary structures of globular proteins.

In order to quantify the stabilities of protein native structures, one must know the conformational characteristics of a protein at the denatured state. Chothia et al.[8-11] assumed denatured proteins to be in the extended conformation, and then proceed to estimate the hydrophobic energy gain accompanying protein folding to be in the range of 2.5 (kcal/mol)/residue for 100 residue chains to 2.9 (kcal/mol)/residue for 300 residue chains (eq 4 in ref 11). However, protein conformations in the denatured state are highly unlikely to exist in the extended form and are certainly more compact on the average; the present analysis indicates that the total number of residue–residue contacts formed at $e_{ij} = 0$ amounts to about 48–42% of

the residue–residue contacts formed in the native structure for proteins consisting of 100–300 residues. Thus Chothia's estimate of hydrophobic energies accompanying protein folding must be an overestimate, even if the estimation of solvent effects were comprehensive. Here, on the other hand, although the average contact energy is about twice as large as estimates of hydrophobic energies, an upper bound of the contact energy gain in protein folding processes takes similar values, from about 2.2 to 2.6 (kcal/mol)/residue for 100–300 residue chains; see eq 42a and 43. This provides some indirect verification that the estimated values of contact energies are not actually too large. Further testing of these values of interresidue contact energies remains for the future.

Registry No. Cys, 52-90-4; Met, 63-68-3; Phe, 63-91-2; Ile, 73-32-5; Leu, 61-90-5; Val, 72-18-4; Trp, 73-22-3; Tyr, 60-18-4; Ala, 56-41-7; Gly, 56-40-6; Thr, 72-19-5; Ser, 56-45-1; Gln, 56-85-9; Asn, 70-47-3; Glu, 56-86-0; Asp, 56-84-8; His, 71-00-1; Arg, 74-79-3; Lys, 56-87-1; Pro, 147-85-3.

References and Notes

(1) Kauzmann, W. *Adv. Protein Chem.* **1959**, *14*, 1–63.
(2) Nozaki, Y.; Tanford, C. *J. Biol. Chem.* **1971**, *246*, 2211–2217.
(3) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754–2759.
(4) Hermann, R. B. *J. Phys. Chem.* **1975**, *79*, 163–169.
(5) Hermann, R. B. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 4144–4145.
(6) Reynolds, J. A.; Gilbert, D. B.; Tanford, C. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 2925–2927.
(7) Chothia, C. *Nature (London)* **1974**, *248*, 338–339.
(8) Chothia, C. *Nature (London)* **1975**, *254*, 304–308.
(9) Chothia, C.; Janin, J. *Nature (London)* **1975**, *256*, 705–708.
(10) Chothia, C. *J. Mol. Biol.* **1976**, *105*, 1–14.
(11) Janin, J.; Chothia C. "Protein: Structure, Function and Industrial Applications"; Aurich, H., Ed.; Pergamon Press: New York, 1979, pp 227–238.
(12) Finney, J. L.; Gellatly, B. J.; Golton, I. C.; Goodfellow, J. *Biophys. J.* **1980**, *32*, 17–33.
(13) Lesk, A. M.; Chothia, C. *Biophys. J.* **1980**, *32*, 35–47 (see discussion following this paper (pp 44–47)).
(14) Lee, B. *J. Phys. Chem.* **1983**, *87*, 112–118.
(15) Lee, B. "Mathematics and Computers in Biomedicine"; Eisenfeld, J. and DeLisi, C., Eds.; Elsevier North Holland: Amsterdam, 1985, in press.
(16) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945–950.
(17) Hill, T. L. "Statistical Mechanics"; McGraw-Hill: New York, 1956; "Introduction to Statistical Thermodynamics"; Addison-Wesley: Reading, MA, 1960.
(18) Saito, N. "Polymer Physics" (in Japanese); Shokabo: Tokyo, 1967.
(19) Fowler, R. H.; Guggenheim, E. A. "Statistical Thermodynamics"; Cambridge University Press: London, 1939.
(20) Guggenheim, E. A. *Proc. R. Soc. London, Ser. A* **1944**, *183*, 213–227.
(21) Miyazawa, S. *Biopolymers* **1983**, *22*, 2253–2271.
(22) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.
(23) Flory, P. J. "Principles of Polymer Chemistry"; Cornell University Press: Ithaca, NY, 1953.
(24) Yamakawa, H. "Modern Theory of Polymer Solutions"; Harper and Row: New York, 1971.
(25) de Gennes, P.-G. "Scaling Concepts in Polymer Physics"; Cornell University Press: Ithaca (NY) and London, 1979.
(26) Edwards, S. F. *NBS Spec. Publ. (U.S.)* **1966**, *No. 273*, 225 (appendix).
(27) de Gennes, P.-G. *Rep. Prog. Phys.* **1969**, *32*, 187–205.
(28) Flory, P. J. *J. Chem. Phys.* **1949**, *17*, 303–310.
(29) Flory, P. J.; Fox, T. G. *J. Am. Chem. Soc.* **1951**, *73*, 1904–1920.
(30) Stockmayer, W. H. *J. Polym. Sci.* **1955**, *15*, 595–598.
(31) Stockmayer, W. H. *Makromol. Chem.* **1960**, *35*, 54–74.
(32) Flory, P. J. "Statistical Mechanics of Chain Molecules"; Interscience: New York, 1969.
(33) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.
(34) Nishikawa, K.; Ooi, T. *Int. J. Pept. Protein Res.* **1980**, *16*, 19–32.
(35) Manavalan, P.; Ponnuswamy, P. K. *Arch. Biochem. Biophys.* **1977**, *184*, 476–487.
(36) Crippen, G. M. *Biopolymers* **1977**, *16*, 2189–2201.
(37) Go, M.; Miyazawa, S. *Int. J. Pept. Protein Res.* **1978**, *12*, 237–241.
(38) Miyazawa, S.; Jernigan, R. L. *Biopolymers* **1982**, *21*, 1333–1363.
(39) Miyazawa, S.; Jernigan, R. L. *Biochemistry* **1982**, *21*, 5203–5213.
(40) Jernigan, R. L.; Miyazawa, S. *Biopolymers* **1983**, *22*, 79–85.
(41) Miyazawa, S.; Jernigan, R. L. *J. Stat. Phys.* **1983**, *30*, 549–559.
(42) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.
(43) Teller, D. C. *Nature (London)* **1976**, *260*, 729–731.
(44) Gates, R. E. *J. Mol. Biol.* **1979**, *127*, 345–351.
(45) Parsegian, V. A. *Trans. Faraday Soc.* **1966**, *62*, 848–860.
(46) Lee, B. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 622–626.