AMINO ACID SIDE-CHAIN PARTITION ENERGIES AND DISTRIBUTION OF RESIDUES IN SOLUBLE PROTEINS

H. ROBERT GUY
Laboratory of Biophysics, National Institute of Neurological, Communicative Diseases, and Stroke, National Institutes of Health, Bethesda, Maryland 20205

ABSTRACT Energies required to transfer amino acid side chains from water to less polar environments were calculated from results of several studies and compared with several statistical analyses of residue distributions in soluble proteins. An analysis that divides proteins into layers parallel with their surfaces is more informative than those that simply classify residues as exposed or buried. Most residues appear to be distributed as a function of the distance from the protein-water interface in a manner consistent with partition energies calculated from partitioning of amino acids between water and octanol phases and from solubilities of amino acids in water, ethanol, and methanol. Lys, Arg, Tyr, and Trp residues tend to concentrate near the water-protein interface where their apolar side-chain components are more buried than their polar side-chain components. Residue distributions calculated in this manner do not correlate well with side-chain solvation energies calculated from vapor pressures of side-chain analogs over a water phase. Results of statistical studies that classify residues as exposed to solvent or buried inside the protein interior appear to depend on the method used to classify residues. Data from some of these studies correlate better with solvation energies, but other data correlate better with partition energies. Most other statistical methods that have been used to evaluate effects of water on residue distributions yield results that correlate better with partition energies than with solvation energies.

INTRODUCTION

It is well known that interactions with water cause polar side chains of soluble proteins to be nearer the protein surface than are apolar side chains. Several efforts have been made recently to quantitate these solvent effects and use polarity scales to analyze packing in globular proteins (Rose and Roy, 1980; Kyte and Doolittle, 1982), to predict secondary structure (Cid et al., 1982; Eisenberg et al., 1984; Finer-Moore and Stroud, 1984), to predict transmembrane segments (Argos et al., 1982; Engleman and Steitz, 1981; Von Heijne, 1981*a*, *b*; Kyte and Doolittle, 1982; Guy, 1984), and to evaluate amphiphilicity of α-helices (Eisenberg et al., 1982; Finer-Moore and Stroud, 1984; Guy, 1984*a*). Polarity scales used in these studies often do not agree with each other and are sometimes based on arbitrary assumptions or averaging of different scales. The purposes of this paper are to suggest causes for some of the apparent discrepancies among scales, to emphasize the importance of recognizing that most residues are neither completely buried within the protein nor completely exposed to water and that some large residues concentrate near the water-protein interface where their apolar components are more buried than their polar components, and to develop an approach that evaluates these effects and allows solvent energy terms to be combined with other energy terms in more complex conformational energy analyses.

Most polarity scales for amino acid side chains are based on partitioning of amino acids or their analogs between water and a less pola solvent, and on statistical analyses of amino acid residue distributions within soluble proteins of known structures. Chothia (1976), Janin (1979), Wertz and Scheraga (1978), and Robson and Osguthorpe (1979) have performed statistical analyses in which residues in globular proteins of known structure are classified as buried or exposed to water. Data from the Chothia and Janin studies correlate better with solvation energies calculated from vapor pressures of side-chain analogs (Wolfenden et al., 1981) than with side-chain partition energies calculated from solubilities of amino acids in ethanol (Tanford, 1962; Nozaki and Tanford, 1971). This observation has led to claims that solvation energies better simulate the interior of proteins (Wolfenden et al., 1981; Wolfenden, 1983), and led Kyte and Doolittle (1982) to rely primarily on the Wolfenden et al., Janin, and Chothia data in their analysis of hydrophobicity of protein sequences.

The analysis presented here demonstrates that a preference for these data is not justified. The data obtained by Wertz and Scheraga and by Robson and Osguthorpe correlate better with calculated partition energies than with solvation energies. Differences among the data of the

Dr. Guy's present address is Laboratory of Mathematical Biology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205 four statistical studies are probably related to methods used to classify residues as buried or exposed, and to other problems that are inherent in a binary classification method. Prabhakaran and Ponnuswamy (1980) have analyzed amino acid residue distributions by dividing proteins into layers. This method avoids some of the problems of a binary classification and yields more information. Their data are analyzed here to determine the distribution of residues as a function of distance from the protein surface and relate these distributions to apparent partition energies. Results of this analysis correlate much better with partition energies calculated from partitioning of amino acids between organic solvents and water than with solvation energies of side-chain analogs. This method has the additional advantage that absolute values of the energies required to transfer residues from the surface of a protein to its interior can be estimated and the tendency of amphiphilic side chains to concentrate near the water-protein interface can be evaluated. Absolute energies are required to include solvent effects in conformational energy calculations.

GLOSSARY

B_i energy required to move residue i from water to the reference layer divided by the energy required to completely bury it (Eq. 9)

C constant that determines the steepness of the transition in Sig(x) (Eq. 10)

ΔF_{aa} mean partition energy scale for side chains determined from studies using amino acids and organic solvents. Values were normalized to those for octanol

ΔF_i energy calculated to transfer the ith amino acid side chain from water to an organic solvent (Eq. 1)

$\Delta F_i'$ apparent energy required to transfer the ith residue from the surface of a protein to its interior based on distribution of residues in soluble proteins of known structure (Eq. 6)

$\Delta f'(x)$ apparent energy required to transfer the ith residue from water to a distance x from the protein's surface (Eq. 5)

$\Delta f_i^*(x)$ apparent energy required to move the ith residue from the reference layer to a distance x from the protein's surface (Eq. 8)

M_i the mole fraction of the ith residue in the entire population of proteins used in the data base

$M_{i,b}$ the mole fraction of the ith residue in the population of residues that are completely buried inside the proteins

$M_{i,e}$ the mole fraction of the ith residue in the population of residues that are completely exposed to water on the surfaces of the proteins

$M_i(x)$ the mole fraction of the ith residue in a layer that is a distance x from the surface of the proteins

$R(\Delta F)$ the correlation coefficient between ΔF_{aa} values and polarity scales calculated from distributions of residues in soluble proteins

$R(\Delta SE)$ the correlation coefficient between ΔSE values and polarity scales calculated from distributions of residues in soluble proteins

Reference layer the layer at which the mole fraction of the ith residue equals M_i

ΔSE_i solvation energy required to transfer the ith amino acid side-chain analog from water to the vapor phase

x relative distance from the protein surface. It is zero at the surface and one at the center of the protein

x_o value of x at the inflection point of Sig(x) (Eq. 10). The residue can be considered half buried at x_o

x_r value of x at the reference layer.

RESULTS

Partition and Solvation Energies

Energies required to move amino acid side chains from water to ethanol (Tanford, 1962; Nozaki and Tanford, 1971) and methanol (Gekko, 1981) have been calculated using the equation

$$\Delta F_i = RT \ln(\eta_{w,i}/\eta_{s,i}) - RT \ln(\eta_{w,Gly}/\eta_{s,Gly}), \qquad (1)$$

where ΔF_i is the partition energy of the ith side chain, R is the gas constant, T is temperature, $\eta_{w,i}$ and $\eta_{s,i}$ are solubilities of amino acid i in water and organic solvent respectively, and $\eta_{w,Gly}$ and $\eta_{s,Gly}$ are solubilities of Gly in water and organic solvent. Also, energies required to transfer amino acids from water to octanol can be calculated from results of two-phase partition experiments (Yunger and Cramer, 1981; Fauchere et al. 1980; Klein et al., 1971). Partition energies from these studies are compared in Table I. Data in Table I are for Glu, Lys, and Arg in octanol when the pH of the water phase is 7.0.

Data for these amino acids in ethanol probably reflect the noncharged side chains. One might expect the magnitude of amino acid side-chain partition energies to increase as solvents become less polar; however, this does not appear to be true. Energies calculated from solubility experiments using ethanol and methanol average 1.35 and 1.40 times those calculated for octanol. To calculate a single polarity energy scale, ΔF_{aa}, data for ethanol and methanol were normalized to those of octanol by dividing by the factors above. Without normalization, data for different amino acids would be weighted differently because they are not all represented in each study. Units of this scale are arbitrary because the data could have been normalized to values for ethanol or methanol rather than octanol (see Discussion). Solvation energies, ΔSE, calculated by Wolfenden et al. (1981) from side-chain analogs, are also listed in Table I. These values differ substantially from amino acid side-chain partition energies both in absolute magnitude and relative values.

Binary Classification

Several attempts have been made to relate solvent accessiblity of amino acid residues in proteins of known structure to transfer energies of their side chains from water to a less polar environment. These analyses are often an

TABLE I
SIDE CHAIN PARTITION ENERGIES CALCULATED BY DIFFERENT METHODS

Amino acid	Solubilities in ethanol	Solubilities in methanol‖	Partitioning to octanol	Normalized mean (ΔF_{aa})	Solvation of side-chain analogs§§
Ile	−2.97§	—	−1.93‡‡	−2.04 ± 0.11	−2.15
Leu	−2.42‡	−2.49	−1.43**, −2.16‡‡	−1.76 ± 0.26	−2.28
Val	−1.68‡	−1.69	−1.16¶	−1.18 ± 0.02	−1.99
Ala	−0.73‡	−0.77	−0.50¶ , −0.53‡‡	−0.52 ± 0.02	−1.94
Pro	—	—	−0.78¶	−0.78	—
Phe	−2.65‡	−2.77	−2.16**, −2.39‡‡	−2.09 ± 0.21	0.79
Met	−1.30§	—	−1.69‡‡	−1.32 ± 0.37	1.48
Trp	−3.00‡	−3.39	−2.72¶ , −2.79**	−2.51 ± 0.26	5.88
Tyr	−2.53‡	—	−1.44¶	−1.63 ± 0.19	6.11
His	−0.67‡	—	−1.58¶ , 0.80**	−0.95 ± 0.46	10.27
Thr	−0.44§	—	−0.23¶	−0.27 ± 0.04	4.88
Ser	−0.04§	—	−0.05¶	−0.04 ± 0.02	5.06
Asn	0.01§	—	—	0.01	9.68
Gln	0.10§	—	—	0.07	9.38
Asp	−0.54§	—	—	−0.38	6.68
Glu	−0.55§	—	—	−0.40	6.45
Lys	−1.50§	—	—	−1.08	4.37
Arg	−0.73§	—	—	−0.53	10.92
Asp*	—	—	—	—	10.95
Glu*	—	—	0.79¶	0.79	10.20
Lys*	—	—	−0.08¶	−0.08	9.52
Arg*	—	—	1.32¶	1.32	19.92

All energies are in kilocalories per mole. Values for methanol and ethanol were scaled to be approximately the same as those for octanol in calculating the mean.

*Data at pH = 7.0.

‡Nozaki and Tanford (1971).

§Tanford (1962).

‖Gekko (1981).

¶Yunger and Cramer (1981).

**Fauchere et al. (1980).

‡‡Klein et al., (1971).

§§Wolfenden et al. (1981).

attempt to determine whether a given type of amino acid side chain is, on the average, distributed between the surface and interior of the protein in the same way it would be distributed between water and a solvent with a polarity similar to that of the protein's interior. The simplest approach is to classify all residues as being either buried in the protein or exposed to water. These data can then be analyzed using the equation

$$\Delta F'_i = RT \ln(M_{i,e}/M_{i,b}), \qquad (2)$$

where $\Delta F'_i$ is an apparent transfer energy for the ith residue and $M_{i,e}$ and $M_{i,b}$ are its mole fractions that are exposed to water or buried in the protein, respectively (Janin, 1979). Table II shows results of four studies using this analysis. Comparison of these data to the partition energy and solvation energy scales yields ambiguous findings. Data from studies by Wertz and Scheraga (1978) and Robson and Osguthorpe (1979) correlate better with the partition energies (ΔF_{aa}) of Table I, whereas data obtained by Janin (1979) and Chothia (1976) correlate better with solvation energies(ΔSE)(see correlation coefficients in Table II).

Other Statistical Studies

Three methods that do not classify residues as buried or exposed have been used to analyze effects of side-chain polarities on their distributions in proteins of known structure. Correlation coefficients between any scale and ΔF_{aa} and ΔSE values of Table I will be called $R(\Delta F)$ and $R(\Delta SE)$. Meirovitch et al. (1980) determined the radius of gyration, $\langle r \rangle$, of side chains and α-carbons relative to the radius of gyration of the proteins. The $\langle r \rangle$ values for side chains in small proteins correlate better with the partition energy scale, ΔF_{aa}, $[R(\Delta F) = 0.80]$ than with the solvation energy scale $[R(\Delta SE) = 0.60]$; however, for side-chains in large proteins, $\langle r \rangle$ correlates slightly better with solvation energies $[R(\Delta SE) = 0.73, R(\Delta F) = 0.70]$. It thus appears that protein size may affect correlation of these data.

Ponnuswamy et al. (1980) developed a hydrophobicity index, $\langle H^f \rangle$, by summing the hydrophobic indices (as

TABLE II
APPARENT PARTITION ENERGIES CALCULATED BY CLASSIFYING RESIDUES AS EXPOSED OR BURIED*

Amino acid	Wertz and Scheraga (1978)	Robson and Osguthorpe (1979)	Janin (1979)	Chothia (1976)
Ile	−0.69	−2.15	−0.66	−0.80
Leu	−0.62	−1.08	−0.53	−0.44
Val	−0.46	−0.75	−0.62	−0.65
Ala	0.05	0.54	−0.31	−0.27
Gly	0.31	—	−0.33	−0.22
Pro	0.46	−0.22	0.34	0.36
Phe	−1.03	−1.51	−0.45	−0.55
Met	−0.59	−0.97	−0.38	−0.31
Trp	−0.98	−1.61	−0.27	0.05
Tyr	−0.25	−1.13	0.40	0.48
His	−0.41	−0.59	0.13	0.37
Thr	0.38	0.27	0.21	0.18
Ser	0.12	0.65	0.10	0.17
Asn	0.29	0.38	0.49	0.61
Gln	0.46	0.05	0.70	1.00
Asp	0.41	0.65	0.58	0.50
Glu	0.38	0.38	0.68	0.33
Lys	0.57	0.48	1.79	1.17
Arg	0.12	−0.16	1.30	2.00
Cys	−0.84	−1.13	−0.89	−0.23
$R(\Delta SE)$	0.62	0.53	0.82	0.89
$R(\Delta F)$	0.95	0.92	0.60	0.56

*All apparent energies are in kilocalories per mole and were calculated from Eq. 2. $R(\Delta SE)$ and $R(\Delta F)$ are correlation coefficients with solvation energies of Wolfenden et al. (1981) and with the mean ΔF_{aa} values of Table I.

given by Tanford, 1962, and Jones, 1975) of residues that have an α-carbon within 8 Å of the α-carbon of each residue in a population of globular proteins. These data correlate slightly better with ΔF_{aa} than with ΔSE [$R(\Delta F) = 0.77$, $R(\Delta SE) = 0.74$]. This may not be relevant since the data may be biased by using some of the Tanford data in the method.

Miyazawa and Jernigan (1985) have analyzed residue-residue and water-residue contacts in 41 proteins. The data base for this study was larger than for any of the other studies described here. All proteins contained >100 residues and none of them was closely homologous. Contacts among residues were defined as those in which the centers of the side chains are within 6.5 Å. Effective water molecules were given the volume of an average residue, and residue-water contacts were defined in a manner similar to that of residue-residue contacts. These data were used to approximate the energy required to replace a water contact on the ith residue with a residue contact. Average values of this energy correlate with ΔF_{aa} values much better than with solvation energies [$R(\Delta F) = 0.86$, $R(\Delta SE) = 0.61$].

Layer Analysis

The binary classification method has the following shortcomings: (*a*) Most side chains are neither entirely buried nor entirely exposed to water. Data obtained in this manner reflect the energy required to move residues from an environment in which the ratio of water to protein is high to an environment in which the ratio is low. These data are difficult to compare with experimental results in which amino acid side-chains are maximally exposed to each solvent. (*b*) Results obtained with this approach appear to depend greatly upon the method used to classify residues as buried or exposed. (*c*) Amphiphilic side chains may concentrate near the water-protein interface so that their apolar moiety can be buried while their polar moiety is exposed. This situation is impossible to represent with a binary classification method.

These difficulties are reduced or eliminated with the layer analysis presented here. The purpose of this analysis is to determine whether residues in globular proteins are statistically distributed as a function of their distance from the protein's surface in the manner expected if the protein's interior acted as an apolar solvent and residues were free to move between the interior and exterior of the protein. To analyze residue distributions, proteins can be divided into layers that are parallel to the protein's surface. The center of each layer is a fixed distance, x, from the protein's surface. It is assumed here that residues exist in a dynamic equilibrium and that after the initial folding process the volume of each layer does not change. This assumption requires that the movement of the ith residue from layer x_1 to a more interior layer x_2 be accompanied by movement of another residue, say j, from x_2 to x_1. This exchange can be expressed by the reaction scheme

$$i(x_1) + j(x_2) \rightarrow i(x_2) + j(x_1). \tag{3}$$

The energy change caused by this exchange is given by

$$\Delta f_i'(x_1 \rightarrow x_2) - \Delta f_j'(x_1 \rightarrow x_2) = RT\ln[M_i(x_1)/M_i(x_2)] - RT\ln[M_j(x_1)/M_j(x_2)], \tag{4}$$

where $\Delta f_i'(x_1 \rightarrow x_2)$ is the energy required to move i from x_1 to x_2 and $M_i(x)$ is the mole fraction of i in the x layer. This analysis is equivalent to treating each layer as a solvent of different polarity, and M_i as proportional to the concentration of i in each layer. Note that in this model the number of residue backbone components remains constant in each layer and thus the apparent partition energy, $\Delta f_i'(x_1 \rightarrow x_2)$, depends only on the relative polarity of the residue side-chain components.

The energy required to move i from an aqueous environment to x is given by

$$\Delta f_i'(x) = RT\ln[M_i(e)/M_i(x)], \tag{5}$$

where $M_i(e)$ is the mole fraction of i in a layer in which residues are maximally exposed to solvent. The energy required to move i from an aqueous environment to an environment in which it is completely buried in the protein is given by

$$\Delta F_i' = RT\ln[M_i(e)/M_i(b)], \tag{6}$$

where $M_i(b)$ is the mole fraction of i in completely buried layers. $\Delta F'_i$ is the apparent energy term that is best related to energies calculated from partitioning of amino acids or their side-chain analogs between water and less polar phases. $\Delta F'_i$ is difficult to determine directly because very few residues are maximally exposed to water or completely buried inside proteins. It can be estimated by making a few simple assumptions. Consider the case in which a residue, which is treated as a sphere of finite size and uniform polarity, is moved from water through each layer to the center of the protein. When the residue is far from the surface, the energy required to move it will be relatively independent of x. The environment experienced by the residue will be less polar in each successive layer until it reaches layers in which residues are never exposed to water where, if the protein is sufficiently large, the energy to move it will be independent of x again. The transition will be gradual because of the finite size of the residue, the gradual change in average polarity of side chains in the proteins, and statistical variation introduced by the method used to approximate protein surfaces. With this rationale, Eq. 5 can be rewritten

$$\Delta f'_i(x) = \Delta F'_i \mathrm{Sig}(x), \tag{7}$$

where Sig(x) is a sigmoidal curve that equals the energy required to move i from water to x, divided by the energy required to completely bury it. Sig(x) may be considered as the fraction of i that is buried at x. A residue can be considered completely exposed when its accessible surface area equals its accessible area in a single extended strand, and completely buried when its accessible surface area is zero. To analyze statistical data from proteins with this equation, apparent energies must be calculated relative to some point in the proteins. The completely exposed or completely buried regions are not good reference points because the surface is difficult to define and very few residues are either completely buried inside the protein or completely exposed to water. The proteins should contain a layer in which the mole fraction of i equals its mole fraction in the entire data base, M_i [see Fig. 1 where $f^*_i(x) = 0$]. This layer, which will be called the reference layer, is not subject to the statistical problems of completely exposed or completely buried regions. The apparent energy, $\Delta f^*_i(x)$, required to move i from the reference layer to x can be easily calculated from the equation

$$\Delta f^*_i(x) = RT\ln[M_i/M_i(x)]. \tag{8a}$$

The sum of $\Delta f^*_i(x)$ for all residues in a given protein should approximate the energy difference between the actual structure and the mean of all structures with the same surface-to-volume ratio but in which residues are randomly distributed. $\Delta f^*_i(x)$ equals the energy required to move residue i from water to x minus the energy required to move i from water to the reference layer. Thus, it can be related to $\Delta F'_i$ by the equations,

$$\Delta f^*_i(x) = RT\ln[M_i(e)/M_i(x)] - RT\ln[M_i(e)/M_i], \tag{8b}$$

which, from Eqs. 5–7, can be rewritten as

$$\Delta f^*_i = \Delta F'_i[\mathrm{Sig}(x) - B_i], \tag{9}$$

where $B_i = \{\ln[M_i(e)/M_i]\}/\{\ln[M_i(e)/M_i(b)]\}$. B is a constant equal to the energy required to move i from water to the reference layer divided by the energy required to completely bury it.

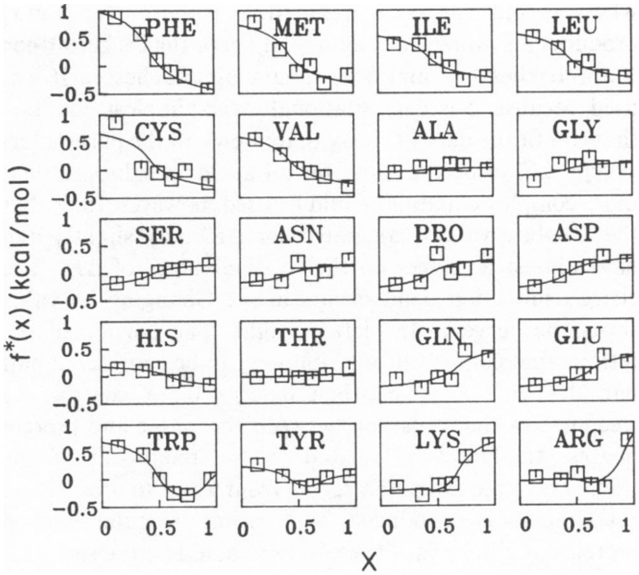

FIGURE 1 Apparent energies required to move the ith residue from the reference layer to x vs. the relative distance, x, of the residue from the protein surface. $\Delta f^*_i(x)$ values were calculated by multiplying information parameters of Prabhakaran and Ponnuswamy (1980) by $RT/\ln 2$. Curves were determined by fitting these data with Eqs. 10 or 11 using the parameter values listed in Table III.

Prabhakaran and Ponnuswamy (1980) performed an analysis that can be evaluated with this approach. They determined the three axes of ellipsoids that best represent the shapes of 19 proteins. These proteins were then divided into six concentric ellipsoidal layers of equal volume that had the same axial ratios as the outer ellipsoid. The mole fraction of α-carbons of each residue was then determined for each layer. These data were used to calculate information parameters for each residue. They did not attempt to fit these data with theoretical curves or relate them to partition energies or other statistical studies of soluble proteins. Their information parameters are proportional to $\Delta f^*_i(x)$ values calculated by Eq. 8. Fig. 1 shows $\Delta f^*_i(x)$ plotted as a function of relative distance, x, of each layer from the proteins' outer surfaces to their centers and curves used to fit these data.

Data for most residues were fit with the equation

$$\Delta f^*_i = \Delta F'_i[\mathrm{Sig}(x) - B]$$
$$\mathrm{Sig}(x) = 0.5C(x - x_o)/(1 + C|x - x_o|) + 0.5, \tag{10}$$

where C and x_o are constants. This equation for Sig(x) produces a symmetrical sigmoidal curve that is zero when x approaches $-\infty$ and one when x approaches ∞. It was used because it is computationally the simplest equation that can fit the data in a way that allows all the parameters except $\Delta F'_i$ to be the same for all apolar residues. Other, more complex equations could be used; however, they alter the absolute values calculated for $\Delta F'$ only slightly and have almost no effect on the relative values of $\Delta F'$. The curve's inflection point occurs at x_o. Changing its value shifts the curve to the left or right. The portion of the residue that controls its distribution can be considered half buried when $x = x_o$; i.e., when Sig(x) = 0.5. C controls the steepness of the transition between the water and protein phases. It should be related to the reciprocal of the thickness of the transition region from water to protein and randomization introduced by treating irregular shaped proteins as ellipsoids. These factors should be the same for all residues. A value of 4 was determined for C by first fitting all the data with no constraints. The location of the reference layer, x_r, can be determined by solving Eq. 10 for x when $f_i^*(x)$ is zero.

Prabhakaran and Ponnuswamy's data were analyzed in four categories according to their side chains. Those with side chains that have no distinct polar atoms are shown in the first two rows of Fig. 1. When data for these residues were fit with Eq. 10 using least-square analysis and no constraints, values of B, C, and x_o were about the same for all residues. This suggests that these parameters are protein dependent and should be the same for all apolar residues. If so, improved values of $\Delta F'_i$, which is the parameter of most interest, should be obtained for each residue by constraining the other parameters to be the same for all residues. Values of $B = 0.79$ and $x_o = 0.43$ were calculated by weighting the $\Delta f'_i(x)$ values for apolar residues to reflect their frequency of occurrence in proteins, summing all of these values, and then fitting these data with Eq. 10 with $C = 4$. $\Delta F'_i$ values of Table III were then determined by constraining these parameters. If this analysis is correct and Sig(x) approximates the fraction of the residue that is buried, then apolar residues in the outer layer ($x = 0.14$) are 23% buried and those in the inner layer ($x = 0.95$) are 84% buried. These values appear reasonable.

Residues in the second category (Ser, Asn, and Asp) have small apolar side-chain components and polar components near the α-carbon. Their distributions are shown in the third row of Fig. 1. These data were fit with the same x_o and C values as above; however, a slightly more positive value for B yielded better fits and the reference layer appears to be slightly nearer the surface than that for apolar side chains (Table III). Pro and Gly, which distribute as if they were polar, were included in this category.

Residues in the third category (Glu, Gln, Thr, and His) have side chains with both polar and apolar components. Alkyl side-chain partition energies are directly proportional to their accessible surface areas (Chothia, 1974). Surface areas of α- plus β-carbons and associated hydrogens of Glu, Gln, Thr, and His are near that of an Ala side chain (Lee and Richards, 1971). Ala distributes evenly throughout soluble proteins. Thus, only the polar side-chain components of these third-category residues should affect their distribution. Curves fitting data for these residues are shown in the fourth row of Fig. 1. These curves were calculated from Eq. 10 by allowing x_o and B to be adjustable parameters. The x_o values of these residues are all greater than that of apolar and short chain residues. This result is expected if the polar component of the side chain is nearer the surface of the protein than is the α-carbon. Other studies (Meirovitch et al., 1980) indicate this to be the case. The difference between these values, i.e., $x_o - 0.43$, should be proportional to the average displacement near the interface of the polar side-chain component from the α-carbon. With no constraints, the value of x_o calculated for Thr was unrealistically large. Thus, values of x_o and B for Thr were constrained so that the distance, x_r, of its reference layer from the surface was the same as that for the other residues in this category. The reference layer for these residues appears to be more buried than for other residues.

TABLE III
PARAMETERS FOR EQS. 10 AND 11 AND DISTANCE x_r OF REFERENCE LAYER FROM PROTEIN SURFACE

Amino acid	$\Delta F'_i$	x_o	B	x_r
	kcal/mol			
Phe	−2.12	0.43	0.79	0.49
Met	−1.59	0.43	0.79	0.49
Ile	−1.13	0.43	0.79	0.49
Leu	−1.18	0.43	0.79	0.49
Cys	−1.42	0.43	0.79	0.49
Val	−1.27	0.43	0.79	0.49
Ala	0.10	0.43	0.79	0.49
Gly	0.33	0.43	0.61	0.46
Ser	0.52	0.43	0.61	0.46
Asn	0.48	0.43	0.61	0.46
Pro	0.73	0.43	0.61	0.46
Asp	0.78	0.43	0.61	0.46
His	−0.50	0.59	0.48	0.58
Thr	0.07	0.60*	0.43	0.58
Gln	0.95	0.68	0.21	0.58
Glu	0.83	0.63	0.33	0.58
Trp(apl)	−2.19	0.48	0.79	—
Trp(pol)	1.68	0.92	0.14	—
Tyr(apl)	−1.08	0.43	0.79	—
Tyr(pol)	0.87	0.69	0.40	—
Lys(apl)	−0.73	0.43	0.79	—
Lys(pol)	2.13	0.68	0.31	—
Arg(apl)	−0.87	0.57	0.79	—
Arg(pol)	2.78	0.92	0.07	—

*Value constrained so that $x_r = 0.58$.

$C = 4$ for all residues. (apl) and (pol) indicate the parameters are for apolar and polar side-chain components.

A model that treats residues as single spheres of uniform polarity is not adequate for some residues. Tyr, Trp, Lys, and Arg side chains have polar components and apolar components with surface areas substantially larger than that of an Ala side chain. These residues would be expected to concentrate near the interface so that their apolar components can be buried while their polar components are exposed. Their distribution reflects this expectation (Fig. 1). These residues can be represented by a two component, or two sphere, model in which the polar component is nearer the surface and thus more exposed than the apolar component. Thus, data for these residues were fit by the following equation which is the sum of two sigmodial curves with the curve for the apolar component shifted along the x-axis relative to that for the apolar component

$$\Delta f_i^*(x) = \{\Delta F_i'[\mathrm{Sig}(x) - B]\}_a + \{\Delta F_i'[\mathrm{Sig}(x) - B]\}_p, \quad (11)$$

where a and p subscripts indicate that the parameters are for apolar and polar side-chain components, respectively. In fitting the data, all parameters for the apolar component were constrained: $x_{o,a}$ was constrained to be ≥ 0.43 and B_a was held at the value used for apolar residues. Values of $\Delta F'_{i,a}$ were approximated by relating surface areas to $\Delta F_i'$ values for alkyl and phenyl side chains. Parameters of the polar components were given complete freedom, except for C, which was held at 4 (Table III). Note that $x_{o,p}$ is always greater than $x_{o,a}$, indicating that the polar component is nearer than the apolar component to the surface. $\Delta F_i'$ values of Table III, which represent differences between extrapolated values of the curves in Fig. 1, correlate substantially better with the partition energy scale, ΔF_{aa}, [$R(\Delta F) = 0.84$] than with solvation energies [$R(\Delta SE) = 0.65$].

TABLE IV
POLARITY SCALES CALCULATED FROM DISTRIBUTIONS OF RESIDUES IN PROTEINS

Data source	$\Delta F_i'$	A	B	C	Mean
Ile	−1.13	−1.38	−1.55	−1.16	−1.31 ± 0.17
Leu	−1.18	−0.99	−1.61	−1.04	−1.21 ± 0.24
Val	−1.27	−1.55	−0.78	−0.77	−1.09 ± 0.33
Ala	0.10	0.07	0.00	0.08	0.06 ± 0.04
Gly	0.33	0.23	0.54	0.52	0.41 ± 0.13
Pro	0.73	0.71	0.60	0.77	0.70 ± 0.06
Phe	−2.12	−0.60	−2.26	−1.73	−1.68 ± 0.65
Met	−1.59	−1.12	−1.37	−0.99	−1.27 ± 0.23
Trp	−0.51	−0.32	−1.02	−1.65	−0.88 ± 0.52
Tyr	−0.21	−0.51	−0.18	−0.42	−0.33 ± 0.14
His	−0.50	−0.25	−0.54	−0.67	−0.49 ± 0.15
Thr	0.07	0.44	−0.06	0.64	0.27 ± 0.28
Ser	0.52	0.67	0.60	0.20	0.50 ± 0.18
Asn	0.48	0.82	0.12	0.48	0.48 ± 0.25
Gln	0.95	0.66	0.54	0.77	0.73 ± 0.15
Asp	0.78	0.84	0.90	0.69	0.80 ± 0.08
Glu	0.83	0.71	0.90	0.64	0.77 ± 0.10
Lys	1.40	0.93	1.43	0.96	1.18 ± 0.24
Arg	1.91	0.53	0.72	0.20	0.84 ± 0.65
Cys	−1.42	−1.47	−1.13	−1.41	−1.36 ± 0.13
R(ΔSE)	0.65	0.74	0.63	0.62	0.66 —
R(ΔF)	0.84	0.77	0.86	0.95	0.90 —
R($\Delta F'$)*	—	0.96‡	0.97	0.92	— —

The sources of data used to calculate the polarity scales are (A) Ponnuswamy et al. (1980), (B) Meirovitch et al. (1980), and (C) Wertz and Scheraga (1978). Methods used to calculate the polarity scales from the original data are described in the text. $R(\Delta SE)$ and $R(\Delta F)$ are the same as in Table II and $R(\Delta F')$ is the correlation coefficient with the $\Delta F'$ values calculated by fitting the Prabhakaran and Ponnuswamy data.
*Data for Arg was not included.
‡Data for Phe not included.

Comparison of Statistical Data that Correlate with Partition Energies

Statistical fluctuation of data is a major difficulty with the kinds of analyses described above. To determine whether deviation of $\Delta F'$ values calculated from the data of Prabhakaran and Ponnuswamy from ΔF_{aa} values are indicative of proteins in general, data from the studies described above that correlate better with partition energies were converted to scales that are approximately equivalent to $\Delta F'$ values (Table IV). The scale from the Ponnuswamy et al. (1980) data equals $-0.77(\langle H^f \rangle - 12.3)$, the scale from the study of Meirovitch et al. (1980) for small proteins equals $6.66(1 - \langle r \rangle)$, and the scale from Wertz and Scheraga (1978) is 2.25 times the values in Table II. The conversions were made to allow easier comparison of the data and do not alter correlation values because the scaled values are proportional to the original values. Although the new scales should not be construed to be equal to transfer energies, a mean of these scales correlates slightly better with partition energies than the $\Delta F'$ scale does. The correlation coefficient of the Miyazawa and Jernigan (1985) data with the mean scale of Table IV is 0.97 and with $\Delta F'$ is 0.95. Their data suggest that His is more polar than indicated by the mean scale or $\Delta F'$. Apparent transfer energies for Arg calculated from these studies are substantially less than those calculated by fitting the Prabhakaran and Ponnuswamy data or from partitioning of Arg into octanol. This result is probably related to the long Arg side chain, most of which may be buried even though its most polar end portion is exposed.

$\Delta F_i'$ and the mean scale from Table IV are plotted in Fig. 2 as a function of ΔF_{aa} from Table I. For Ile, Leu, Ala, Gly, Phe, His, Thr, Ser, Asn, and Gln the correlation coefficient relating the mean values is 0.994, the slope of the best fit line is 1.0, and its y-intercept is 0.53 kcal/mol. The slope of one suggests that the energies required to move the side chains from water to the interiors of soluble proteins are about the same as those calculated from partitioning of amino acids between water and octanol. The y-intercept of 0.53 kcal/mol suggests that this much energy is required to move Gly into the interior of the protein. Four residues (Trp, Tyr, Pro, and Lys) fall substantially above the line in Fig. 2, and three (Val, Met, and Glu) fall below it. Deviation of Pro is expected since it disrupts α-helices and β-sheets and is often found in reverse turns on or near the protein surface. Deviation of Lys and Glu may be related to the difficulty in determining partition energies of titratable groups. Trp and Tyr residues appear to be exposed to water more than expected from their ΔF_{aa} values. Values of ΔF_{aa} for Cys and Asp at pH 7 are not available.

DISCUSSION

It has been asserted that distributions of residues in soluble proteins do not correlate well with partitioning of amino acids between water and organic solvents (Janin, 1979) but do correlate with energies required to transfer side-chain analogs from water to vapor (Wolfenden et al., 1981; Wolfenden, 1983) These claims do not appear justified. More of the studies described here correlate better with partition energies than with solvation energies. Apparent transfer energies calculated from binary classification of residues as buried or exposed appear to depend on the method used to classify residues. Wertz and Scheraga (1978) used a seven-step algorithm based on the number of times a grid of lines parallel of each of the three axes intersected the solvent exposed van der Waals surface of each residue. Apparent transfer energies calculated from their data correlate with the partition energies, ΔF_{aa}, and apparent transfer energies, $\Delta F'$, calculated by fitting data of Prabhakaran and Ponnuswamy, better than any of the other binary classification data. Robson and Osguthorpe (1979) classified only those residues that are highly solvated as exposed and only those that are mostly buried as buried. This method appears to classify small residues as slightly more exposed and large ones as slightly more buried than predicted by fitting the Prabhakaran and Ponnuswamy data. Janin classified residues with <20 Å^2 accessible surface area as buried and the remainder as exposed. This method appears to classify small residues as more buried and large ones as more exposed than predicted from Prabhakaran and Ponnuswamy's data. This result might be anticipated since the accessible area of a relatively exposed Gly might be <20 Å^2 whereas that for a relatively buried Trp could easily be >20 Å^2. Data from Chothia's study was calculated by classifying as buried those residues with >95% of their maximum accessible surface area buried, and by classifying the remainder as exposed. This method yields results similar to the Janin data. Deviation of these data from apparent energies, $\Delta F'$, calculated from the Prabhakaran and Ponnuswamy data, is more difficult to rationalize.

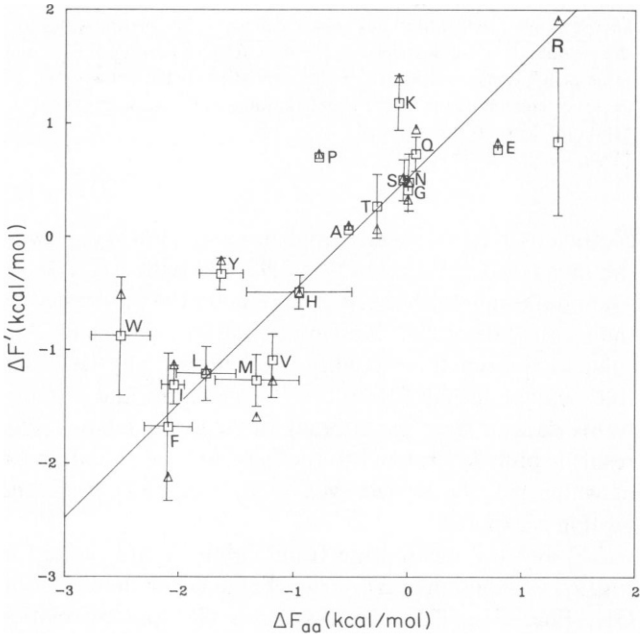

FIGURE 2 Apparent transfer energies calculated from residue distributions in soluble proteins ($\Delta F'$ and mean scale of Table IV) vs. mean side-chain partition energies (ΔF_{aa}) calculated from partitioning of amino acids in octanol and from solubilities of amino acids in ethanol and methanol (the latter two scaled to values for octanol). Values calculated from data of Prabhakaran and Ponnuswamy (1980) (column 1 of Table IV) are represented by Δ, and mean values (column 4 of Table IV) are represented by □. The ΔF_{aa} values are from column 4 of Table I. The line slope is 1 and the y-intercept is 0.53 kcal/mol. Letters represent the single letter code for amino acids.

The layer analysis in which residue distributions are analyzed as a function of distance from the protein surface has the following advantages: (*a*) ambiguities involved in classifying residues as buried or exposed are avoided, (*b*) absolute values of transfer energies can be approximated, and (*c*) concentration of residues that have substantial polar and apolar side-chain components near the interface can be detected. $\Delta F'_i$ values and/or mean scale of Table IV may serve as a reasonable approximation for the energy required to transfer amino acid residues from water to the interior of a protein. Any approximation that ignores specific atomic interactions in proteins is crude; however, when long range interactions are not known, this approximation should be better than using amino acid partition energies because it avoids zwitterionic effects of amino acids, avoids assumptions about the polarity of the protein interior, and yields data for all residues at a pH of 7. When using results presented here of the layer analysis, one should be aware that some effects are not represented in the data. Data for the layer analysis could be improved in the following ways: (*a*) The data base could be separated according to molecular weight or surface-to-volume ratio of the proteins. Protein size probably affects the data because a higher fraction of residues are buried in larger proteins (Janin, 1979) and residues inside large proteins are more polar than inside small proteins (Meirovitch et al., 1980). (*b*) Data could be separated according to secondary structure. Random coil and turn regions are more exposed than α-helices or β-sheets. Curves used to fit data for α-helices, in which residues may be exposed or buried to any degree will probably differ from those for

β-sheets, in which residues may tend to be more completely buried or exposed. (*c*) The data base could be increased. This is probably essential for the improvements suggested above to yield statistically significant data. (*d*) Distances could be analyzed in absolute instead of relative units. (*e*) A method that simulates the surface better than the ellipsoids or that relates layers to the fraction of each residue's surface that is exposed to water could be used.

The analysis presented here was undertaken to help develop a method to approximate solvent and long range protein interactions in models of protein structures in which precise long range interactions may not be known. A method using Eqs. 10 and 11, i.e., energy-distance relationships similar to those in Fig. 1, has been developed to predict the orientation and degree of solvent exposure of amphiphilic α-helices at a water-protein or a protein-lipid interface. In spite of the limitations of the data and analysis described above, this method correctly classifies almost all residues in hemoglobin as buried, partially buried, or exposed (unpublished observation). It has been used with other factors to predict which segments in acetylcholine receptor channel (Guy, 1984*a*) and colicin E1 (Guy, 1983) and colicin A channel (Guy, 1984*b*) sequences are likely to be transmembrane helices, and to predict which portions of these helices should be exposed to water inside the channel, to other helices, and to lipid. The polarity scale presented here was not used in these studies to evaluate protein-lipid interactions because the lipid environment is probably substantially less polar than the interior of soluble proteins.

The analysis presented here indicates that most residues distribute as a function of the relative distance from the surface of soluble proteins in a manner consistent with side-chain partition energies calculated from partitioning of amino acids between water and octanol phases. This does not indicate, however, that octanol is a good model for the interior of the protein. Yunger and Cramer (1981) noted that side-chain partition energies calculated from the studies using amino acids are ~0.6 times those calculated for transfer of side chains only to octanol using the method of Hansch and Leo (1979). Also, absolute values of side-chain partition energies to octanol calculated from studies using amino acids are less than those calculated from amino acid solubilities in ethanol and methanol even though ethanol and methanol are more polar solvents. These anomalous effects may be caused by the zwitterionic nature of amino acids. Thus, the absolute magnitude of energies required to transfer amino acid side chains to organic solvents is subject to uncertainty. In addition, absolute values of $\Delta F_i'$ are dependent upon the sigmoidal curve used in Eq. 10; e.g., if the same equation had been modified so that Sig(x) is 0 when $x < x_o - 0.5$ and 1 when $x > x_o + 0.5$, then $\Delta F_i'$ values would be two-thirds those reported but the relative values would be the same. Absolute values are not important for many applications. In fact, the Prabhakaran and Ponnuswamy data were originally obtained from information theory and no attempt was made to relate the terms to energies. The advantages of absolute energies are that they can be better compared with experimentally determined energies and can be combined with other energy terms. The finding that, in most studies, residue distributions do not correlate well with solvation energies does not indicate that solvation energies do not have important applications. Terms derived from solvation energies may be appropriate when attempting to add effects of water to other conformation energy terms that treat proteins as if they exist in a vacuum. Thus, selection of the appropriate energy scale depends upon what one is trying to calculate.

Received for publication 10 January 1984 and in final form 13 August 1984.

REFERENCES

Argos, J. K., J. K. M. Rao, and P. A. Hargrave. 1982. Structural predictions of membrane proteins. *Eur. J. Biochem.* 128:565–575.

Chothia, C. 1974. Hydrophobic bonding and accessible surface area in proteins. *Nature* (*Lond.*). 248:338–339.

Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105:1–14.

Cid, H., M. Bunster, E. Arriagada, and M. Campos. 1982. Prediction of secondary structure of proteins by means of hydrophobicity profiles. *FEBS* (*Fed. Eur. Biochem. Soc.*) *Lett.* 150:247–254.

Eisenberg, D., R. M. Weiss, and T. C. Terwilliger. 1982. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* (*Lond.*). 299:371–374.

Eisenberg, D., R. M. Weiss, and T. C. Terwilliger. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA.* 81:140–144.

Engelman, D. M., and T. A. Steitz. 1981. The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. *Cell.* 23:411–422.

Fauchere, J. L., K. Q. Do, P. Y. C. Jow, and C. Hansch. 1980. Unusually strong lipophilicity of "fat" or "super" amino-acids, including a new reference value for glycine. *Experientia* (*Basel*). 36:1203–1204.

Finer-Moore, J., and R. M. Stroud. 1984. Amphipathic analysis and possible formation of the ion channel in an acetylcholine receptor. *Proc. Natl. Acad. Sci. USA.* 81:155–159.

Gekko, K. 1981. Mechanism of polyol-induced protein stabilization: solubility of amino acids and diglycine in aqueous polyol solutions. *J. Biochem.* 90:1633–1641.

Guy, H. R. 1983. A model of colicin E1 membrane channel protein structure. *Biophys. J.* 41(2, Pt. 2):363*a*. (Abstr.)

Guy, H. R. 1984*a*. A structural model of the acetylcholine receptor channel based on partition energy and helix packing calculations. *Biophys. J.* 45:249–261.

Guy, H. R. 1984*b*. A model of colicin A membrane channel protein structure. *Biophys. J.* 45(2, Pt. 2):123*a*. (Abstr.)

Hansch, C., and A. Leo. 1979. Substituent Constants for Correlation Analysis in Chemistry and Biology. John Wiley & Sons, Inc., New York.

Janin, J. 1979. Surface and inside volumes in globular proteins. *Nature* (*Lond.*). 277:491–492.

Jones, D. 1975. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. *J. Theor. Biol.* 50:167–183.

Klein, R., M. Moore, and M. Smith. 1971. Selective diffusion of neutral amino acids across lipid bilayers. *Biochim. Biophys. Acta.* 233:420–433.

Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of proteins. *J. Mol. Biol.* 157:105–132.

Lee, B., and F. M. Richards. 1971. Interpretation of protein structure: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.

Meirovitch, H., S. Rackovsky, and H. A. Scheraga. 1980. Empirical studies of hydrophobicity. 3. Radial distribution of clusters of hydrophobic and hydrophilic amino acids. *Macromolecules.* 13:1398–1405.

Miyazawa, S., and R. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* In press.

Nozaki, Y., and C. Tanford. 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. *J. Biol. Chem.* 246:2211–2217.

Ponnuswamy, P. K., M. Prabhakaran, and P. Manavalan. 1980. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta.* 623:301–316.

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

Robson, B., and D. J. Osguthorpe. 1979. Refined models for computer simulation of protein folding. *J. Mol. Biol.* 132:19–51.

Rose, G. D., and S. Roy. 1980. Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci. USA.* 77:4643–4647.

Tanford, C. 1962. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* 84:4240–4247.

Von Heijne, G. 1981*a*. On the hydrophobic nature of signal sequences. *Eur. J. Biochem.* 116:419–422.

Von Heijne, G. 1981*b*. Membrane proteins: the amino acid composition of membrane-penetrating segments. *Eur. J. Biochem.* 120:275–278.

Wertz, D. H., and H. A. Scheraga. 1978. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules.* 11:9–15.

Wolfenden, R. 1983. Waterlogged molecules. *Science* (*Wash. DC*). 222:1087–1093.

Wolfenden, R., L. Andersson, P. M. Cullis., and C. C. B. Southgate. 1981. Affinities of amino acid side-chains for solvent water. *Biochemistry.* 20:849–855.

Yunger, L. M., and R. D. Cramer, III. 1981. Measurement and correlation of partition coefficients of polar amino acids. *Mol. Pharmacol.* 20:602–608.