Protein Science (1996), 5:1676–1686. Cambridge University Press. Printed in the USA.

Hydrophobic regions on protein surfaces. Derivation of the solvation energy from their area distribution in crystallographic protein structures

FRANK EISENHABER[1,2]

[1] Institut für Biochemie der Charité, Medizinische Fakultät, Humboldt-Universität zu Berlin, Hessische Straße 3-4, D-10098 Berlin-Mitte, Germany
[2] European Molecular Biology Laboratory, Meyerhofstraßel, Postfach 10.2209, D-69012 Heidelberg, Germany

(RECEIVED March 26, 1996; ACCEPTED May 16, 1996)

Abstract

For the first time, a direct approach for the derivation of an atomic solvation parameter from macromolecular structural data alone is presented. The specific free energy of solvation for hydrophobic surface regions of proteins is delineated from the area distribution of hydrophobic surface patches. The resulting value is 18 cal/(mol·Å^2), with a statistical uncertainty of ±2 cal/(mol·Å^2) at the 5% significance level. It compares favorably with the parameters for carbon obtained by other authors who use the the crystal geometry of succinic acid or energies of transfer from hydrophobic solvent to water for small organic compounds. Thus, the transferability of atomic solvation parameters for hydrophobic atoms to macromolecules has been directly demonstrated.

A careful statistical analysis demonstrates that surface energy parameters derived from thermodynamic data of protein mutation experiments are clearly less confident.

Keywords: atomic solvation parameter; conformational energy; hydrophobic effect, hydrophobic surface region, molecular recognition, solvent-accessible surface of proteins

The free energy of protein folding $\Delta G_{u\to f}$ in solution (u and f denote the unfolded and folded states, respectively) has contributions from many types of interactions, including those of intramolecular origin or resulting from solvent effects. It is generally accepted that protein–solvent interactions play a major role and may be even considered as the driving force in protein folding and in binding processes (Chothia, 1976; Dill, 1990; Honig & Yang, 1995). Whereas the intramolecular energy of a protein can be computed more easily in the approximation of atom–atom potential functions and monopole charges at atomic positions, the solvent contributions still remains a severe problem due to the large number of degrees of configurational freedom for water molecules (Harvey, 1989; Richards et al., 1989). Nevertheless, inclusion of hydration effects are essential in calculations involving docking of proteins (Korn & Burnett, 1991; Covell et al., 1994; Young et al., 1994; Jackson & Sternberg, 1995) and ligand binding (Leckband et al., 1994; Jones et al., 1995), as well as in protein fold recognition, protein modeling, engineering, and design (Eisenhaber et al., 1995b).

The characteristics of bulk water are only preserved at relatively large distances from the protein surface. Different physicochemical methods (spectroscopy, calorimetry, crystallography, etc.) have identified several hydration layers of structurally and dynamically altered water (Rupley & Careri, 1991; Otting & Liepinsh, 1995; Phillips & Pettitt, 1995). At the same time, experiments on hydrophobic model compounds (hydrocarbons with and without functional groups) have demonstrated that the energetic effect of their hydration is largely confined to a single hydration shell (Herrmann, 1972, 1977; Reynolds et al., 1974; Amidon et al., 1975; Valvani et al., 1976). In this case, the solvation free energy G_s of a molecule can be calculated simply as a linear function of the atomic solvent-accessible surface areas A_i in accordance with

$$G_s = \sum_i \sigma_i A_i \qquad (1)$$

(Chothia, 1974; Eisenberg & McLachlan, 1986). The coefficients σ_i are atom-type-specific solvation parameters and the summation is over all exposed atoms. The hydration free energy has contributions from the dispersion effect, cavity creation (excluded volume of the solute), and electrostatic polarization. Though the first two components can be well described by a function of type (1), the linear approximation fails in the case of highly polar or charged atomic groups (Still et al., 1990; Hasel et al., 1988). Thus, Equation 1 appears suitable mainly for com-

Reprint requests to: Frank Eisenhaber, European Molecular Biology Laboratory, Meyerhofstraße 1, Postfach 10.2209, D-69012 Heidelberg, Germany; e-mail: eisenhaber@embl-heidelberg.de.

puting the contribution of hydrophobic solvent-accessible surfaces (mostly formed by carbon) to the solvation free energy.

The atomic solvation parameters (ASP) for hydrophobic atoms have been derived almost exclusively from experimental data on small organic compounds by: (1) using transfer energies from a hydrophobic environment to water (Herrmann, 1972, 1977; Reynolds et al., 1974; Amidon et al., 1975; Valvani et al., 1976; Eisenberg & McLachlan, 1986; Ooi et al., 1987; Wesson & Eisenberg, 1992); (2) measuring surface tension at hydrocarbon–water interfaces (Tanford, 1979); or (3) relying on small molecule crystal geometry [succinic acid crystals (Rees & Wolfe, 1993)]. The transferability of the latter parameters to biological macromolecules was assumed but not proven. The derivation of an atomic solvation parameter for hydrophobic atoms was also attempted with site-directed protein mutation studies (Eriksson et al., 1992; Blaber et al., 1993; Pinker et al., 1993) and with a neural network trained to distinguish native and nonnative protein conformations (Wang et al., 1995), albeit with considerably less confidence than in the case of small organic molecules (see the Discussion).

In this work, I show that such an atomic solvation parameter can be derived directly from the distribution function of the areas of hydrophobic solvent-accessible surface regions (patches) in three-dimensional protein structures. This result demonstrates that atomic solvation parameters for hydrophobic atoms are transferable from low-molecular compounds to macromolecules.

Theory

It is paradoxical: The solvent-accessible surface as defined by Lee and Richards (1971), which is traditionally used for the analysis of solvation properties of proteins, is not informative for the determination of hydrophobic surface clusters. The hydrophobic surface of a protein is organized into regions (or patches) consisting of neighboring spherical atomic faces sharing common arcs. For a typical medium-sized globular protein, there is only a single large and topologically interconnected hydrophobic surface region, consisting of faces of exposed apolar atoms and constituting about 60% of the entire solvent-accessible surface area; i.e., the major part of the solvent-accessible surface area is hydrophobic. Within this region, there are a large number of pockets or islands corresponding to solvent-accessible polar atoms. Depending on the protein, a variable number of tiny hydrophobic patches (with size mostly well below 10 Å^2; i.e., less than the surface occupied by a single water molecules) may also exist.

Elsewhere (Eisenhaber & Argos, in prep.), it was shown that this estimate for the hydrophobic solvent-accessible surface contacting bulk water is too large. The following two theses have been proven.

(1) Definition of a hydrophobic surface region

From the view point of interaction with bulk water in a two-state hydration model, the protein can be considered as a structural entity together with its first hydration shell (defined as the set of waters bound to polar atoms). The hydrogen bonded water molecules cover not only polar area, but also part of the solvent-accessible hydrophobic surface, bridge polar sites, and, thus, dissect the single large hydrophobic surface region into smaller patches. Thus, a hydrophobic surface region can be defined as a continuous piece of surface that is formed exclusively of hydrophobic atoms and that is not occupied by water molecules bound to polar atoms.

(2) Method of computation

The formation of hydrophobic surface regions owing to the structure of the first hydration shell can be determined computationally with explicit structural models of the first hydration shell [for example, with AUTOSOL (Vedani & Huhta, 1991) or AQUARIUS2 (Pitt et al., 1993)]. The same effect can be simulated by a small increase of the radii of solvent-accessible polar atoms (~0.4 Å), followed by calculation of the remaining exposed hydrophobic patches.

In this work, I want to show that the specific free surface energy can be derived from the area distribution of hydrophobic surface regions in tertiary structures in protein crystals. This deduction is based on three assumptions.

1. The native fold corresponds to the state with lowest hydrophobic solvent-accessible surface area.
2. Individual hydrophobic surface regions are mutually independent (with respect to conformational changes).
3. Instead of considering the ensemble of *many* conformations of a *single* protein, the ensemble of *single* low-energy conformations of *many* proteins can be used to compute the required statistics of hydrophobic surface regions.

For a given protein conformation χ, the overall solvent-accessible hydrophobic surface area $A(\chi)$ is the summed area of many topologically disconnected hydrophobic surface regions $A_i(\chi)$

$$A(\chi) = \sum_i A_i(\chi). \tag{2}$$

It should be emphasized that, in the case of a "naked" protein without any hydration water, the hydrophobic portion of the solvent-accessible surface is represented almost exclusively by a single large and interconnected surface region. Only the consideration of water molecules bound at polar protein atoms results in a considerably smaller apolar surface area and in the disintegration of the single hydrophobic surface region into many smaller patches.

The contribution G_{hs} of hydrophobic surface regions to the overall solvation free energy G_s can be calculated as

$$G_{hs} = \sum_i \sigma_{phob} A_i = \sigma_{phob} A, \tag{3}$$

where the summation involves all hydrophobic regions i contributing to the solvent-accessible surface and σ_{phob} is the specific hydrophobic surface solvation energy. The same σ_{phob} is assumed for all hydrophobic atoms. We postulate that the native conformation of the protein generally corresponds to an energetic state with the smallest total hydrophobic solvent-accessible surface area (first assumption). In other words, the distribution of the hydrophobic solvent-accessible area $A(\chi)$ within the ensemble Ω of all conformations χ for the given protein is Boltz-

mann-like. The probability P of occurrence of a conformation with solvent-accessible hydrophobic area A can be computed as

$$P(A) = C \cdot e^{-\lambda A}, \tag{4}$$

with $\lambda = \sigma_{phob}/(kT)$ where k and T are Boltzmann's constant and the absolute temperature, respectively, and C is a normalizing factor. Taking into account Equation 3, Equation 4 can be rewritten as

$$P(A) = C \cdot \prod_i e^{-\lambda A_j}. \tag{5}$$

The first assumption is obviously not fulfilled for proteins that have extensive surface contacts, not with water, but with other proteins (subunits of oligomeric proteins), lipids (membrane proteins), and the like. It is known that a significant number of the hydrophobic surface regions on subunits in multimeric proteins and complexes are buried from contact to solvent (Argos, 1988; Janin et al., 1988; Miller, 1989; Janin & Chothia, 1990). The typical surface buried by one partner in a subunit contact is about 600–1,000 Å^2 with 55–70% nonpolar (Janin & Chothia, 1990). Hence, the further conclusions apply only to water-soluble globular and monomeric proteins.

For the next step, it is important that the surfaces considered are those of single macromolecules with many degrees of conformational freedoms. Each individual hydrophobic surface region A_i is influenced only by some set of them. Local conformational changes will usually not change the entire surface organization, except a few nearby hydrophobic regions or even only a single one. In this sense, the individual hydrophobic surface regions can be considered as mutually independent (second assumption). Therefore, the criterion of minimal total hydrophobic surface area for fold stability can then be reformulated such that a folded protein will tend to use all local possibilities to reduce the size of an individual hydrophobic patch. The condition of mutual independence of hydrophobic patch size is written mathematically as product of probabilities of independent events

$$P(A) = P\left(\sum_i A_i\right) = \prod_i P(A_i). \tag{6}$$

A comparison with Equation 5 suggests the functional form of $P(A_i)$ as proportional to $\exp(-\lambda A_i)$. Both the surface organization of a protein as described above and the formation of the succinic crystal (Rees & Wolfe, 1993) can be considered from a similar point of view. Not all changes in packing parameters modify the hydrophobic area of all crystal faces, i.e., every crystal face is individually (within the freedom of packing) arranged in such a manner that the hydrophobic surface on each side is minimized. The condition of mutual patch independence is important because it relates the distribution P of the overall hydrophobic area A with that of a single hydrophobic surface region A_i.

The following thoughts are aimed at motivating a functional form for an area distribution $P(A^*)$ of any hydrophobic surface region with area A^* in an ensemble of conformations of a single protein. Most conformations in the ensemble Ω differ from the native structure only by small changes. Let us consider a subset Ω_i of protein conformations with only local differences altering the hydrophobic surface region A_i and letting (almost) unchanged all other patches. Applying Equations 2 and 4, we obtain for the subspace Ω_i

$$P(A) = C \cdot \exp\left(-\lambda \sum_{j \neq i} A_j\right) e^{-\lambda A_i}; \text{ i.e.,}$$

$$P(A) = C \cdot \exp\left(-\lambda \sum_{j \neq i} A_j\right) e^{-\lambda A_i} = C_i \cdot e^{-\lambda A_i}, \tag{7}$$

where C_i is a new constant. Many such distributions obtained from various subspaces Ω_i can be overlayed to give a general distribution P of patch size A^* representative for the whole conformation space Ω

$$P(A^*) = \left(\sum_i C_i\right) \cdot e^{-\lambda A^*} = C^* \cdot e^{-\lambda A^*}, \tag{8}$$

where C^* is a new constant. Equation 8 is in full agreement with Equations 4–6:

$$P(A) = C \cdot \prod_i e^{-\lambda A_i} = (C^* \cdot e^{-\lambda A^*})^n, \tag{9}$$

where n denotes the number of hydrophobic surface regions. Thus, considering the ensemble of all conformations of a protein in solution, the area distribution $P(A^*)$ of its hydrophobic surface regions A^* will follow the Boltzmann law. Thus, it is possible to determine σ_{phob} from the area distribution of hydrophobic surface regions in Ω.

The complete ensemble Ω of protein conformations in solution is difficult to generate. In a similar situation, other researchers have studied sets of crystallographic structures of different proteins and have observed an occurrence versus energy dependence for their parameter of interest that resembles Boltzmann statistics as theoretically deduced for Ω (third assumption). This has been the case for the occurrence of backbone and side-chain angles (Pohl, 1971) and *cis/trans* prolines (MacArthur & Thornton, 1991), for the derivation of an empirical scale for side-chain conformational entropy changes (Pickett & Sternberg, 1993), for the distribution of residues between the interior and the surface of proteins (Miller et al., 1987), for secondary structure propensities (Finkelstein et al., 1977) and residue occurrences at certain points in secondary structures (Serrano et al., 1992), and for the distribution of ion pairs (Bryant & Lawrence, 1991) and cavities (Rashin et al., 1985). The statistics of distances between residue centers in crystallographic structures have even been used to derive pairwise potentials (Tanaka & Scheraga, 1976; Crippen & Viswanadhan, 1985; Miyazawa & Jernigan, 1985, 1996; Casari & Sippl, 1992). Although not explicitly declared, essentially the same assumption of substituting the ensemble of conformations of a single protein by the ensemble of native structures has been used for the derivation of atomic environment energies (Delarue & Koehl, 1995). Gutin, Finkelstein, and Badretdinov (Gutin et al., 1992; Finkelstein et al., 1995a) have proven that, under the assumption of the "Random Energy Model," the occurrence of a structural element in native protein structures is indeed exponentially dependent on the own energy of the element

$$\text{Occurrence} \sim \exp(-\text{energy}/(kT^*)), \tag{10}$$

where T^* is the so-called conformational temperature. Simply, elements stabilizing the fold for many amino acid sequences are typical for globular proteins in contrast to structural features, which can be fitted only by a minor number of sequences. The typical temperature T^* was shown to be below the melting point and near 300 K under the condition of a minimal random deviation between real and fitted energies (Finkelstein et al., 1995b).

The exponential form of the distribution as predicted theoretically in Equation 8 poses strong demands on the number of hydrophobic regions and, therefore, on the number of proteins in the Protein Data Bank (PDB) subset. If the number is too small, only a weak signal over a mostly uniform random distribution of patch size will be observed. Generally, exponential curve behavior at smaller patch sizes is expected to be followed by an almost constant region (random distribution). At even larger areas, the patches will occur simply very infrequently.

Our next goal is the derivation of an estimate for the necessary number of observed hydrophobic surface regions to allow a reasonable fit by Equation 8 in the area interval $[A_b,A_e]$. With a bin width ΔA, the bars of a histogram are centered at $A_k = A_0 + k\Delta A$ with $A_b = A_0$ and $A_e = A_0 + K\Delta A$, where $K + 1$ denotes the number of bins. If Equation 8 holds, the overall number N of observations is calculated as

$$N = \sum_{k=0}^{K} C^*\Delta A \cdot \exp[-\lambda(A_0 + k\Delta A)]. \quad (11)$$

This sum can be treated as a geometric row. After a few transformations, we obtain

$$N = N_K \cdot \frac{q^{K+1} - 1}{q - 1} \text{ where}$$

$$N_K = C^*\Delta A \cdot \exp[-\lambda(A_0 + K\Delta A)] \quad \text{and} \quad q = e^{\lambda\Delta A}. \quad (12)$$

The constant N_K is in the occurrence of patches with an area between $A_0 + K\Delta A - 0.5\Delta A$ and $A_0 + K\Delta A + 0.5\Delta A$; i.e., the value for the last bin considered (onset of a constant region). Generally, greater ΔA (bin width) will enlarge N_K (for small $\lambda\Delta A$). Both growing $K + 1$ (number of bins) and ΔA will increase N relative to N_K as shown by Equation 12. Obviously, there is an upper limit for both values ΔA and $K + 1$ for a given number of proteins (and, subsequently, hydrophobic surface regions) in the sample. For example, the factor between N and N_K in Equation 12 is 21.2, 47.1, and 104.8 for $\lambda\Delta A = 0.1$ and $K + 1 = 8$, 16, and 24, respectively. For $\lambda\Delta A = 0.2$ or 0.3 and $K + 1 = 16$, the factor is equal to 106.3 and 344.5; i.e., the overall number of patches N should be larger than N_K by several orders of magnitude. In the case of $\lambda\Delta A = 0.2$ and $K + 1 = 16$, a level of about $N_K \sim 20$ patches in the last bins requires more than 2,000 hydrophobic regions in the interval $[A_b,A_e]$ for a fit to Equation 8 being reasonable. It is also possible to estimate the length of the interval $[A_b,A_e]$. Assuming a value of 20 cal/(mol·Å^2) for σ_{phob} as in studies on small organic molecules, the bin width would be 3, 6, and 9 Å^2 for $\lambda\Delta A = 0.1$, 0.2, and 0.3, respectively. For example, 16 bins each 6 Å^2 span only 96 Å^2. Even smaller bin widths will result in many non-occupied bins and large bin occupancy fluctuations due to the limited number of hydrophobic regions. The use of only a few large bins may lead to statistically insignificant curve fits.

Results and discussion

Area distribution of solvent-accessible hydrophobic surface regions

In Figure 1, the area distribution of solvent-accessible hydrophobic surface regions computed with several polar expansions is shown for a representative PDB subset of 127 nonhomologous and monomeric globular proteins. The bin size is 5 Å^2. It is a problem to select hydrophobic surface regions from the protein crystal structures that have actual contact with solvent. Even in the case of monomeric proteins, not all hydrophobic surface regions are in direct contact with solvent. On one side, there are many very small patches. This might be also an artefact of the spherical atomic model used. A water molecule occupies in the order of 10 Å^2 of solvent-accessible surface (given a volume of about 30 Å^3). Therefore, we will not consider hydrophobic patches below 10 Å^2. On the other side, larger hydrophobic regions may be involved in binding co-factors, substrates, and other types of ligands, as well as in crystal contacts (Janin & Rodier, 1995). Some patches may belong to empty cavities without water molecules, albeit rare (Hubbard et al., 1994). It is very

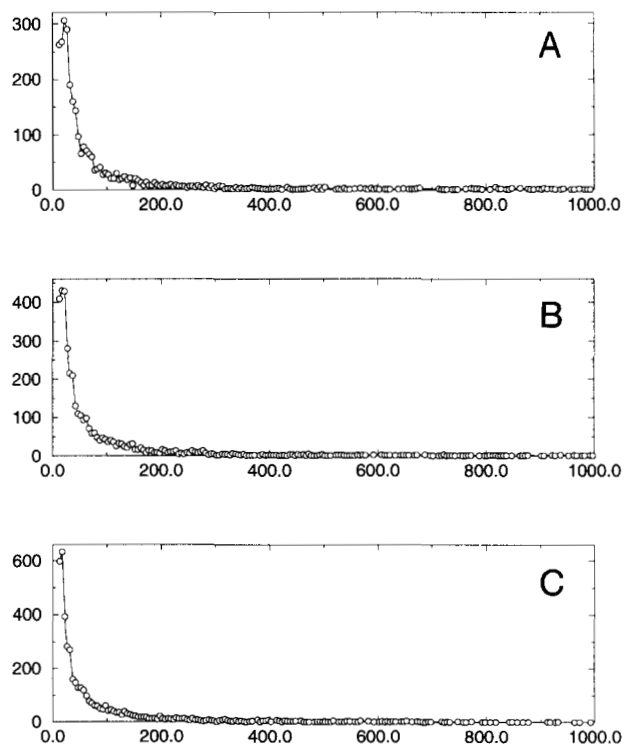

Fig. 1. Area distribution of hydrophobic surface regions. The area distribution of hydrophobic solvent-accessible surface regions is presented in the range of 10–1,000 Å^2 for the PDB subset of 127 proteins for various polar expansions. The abscissa and the ordinate show the patch area (in Å^2) and the patch occurrence, respectively. The bin size is 5 Å^2. Each open circle corresponds to a bin, the solid line connecting subsequent circles is shown to guide the eye. Bins with zero occupancy are not included. The dependency for various radial increments of solvent-accessible polar atoms is shown: **(A)** 0.4 Å^2; **(B)** 0.5 Å^2; and **(C)** 0.6 Å^2. In all cases, three intervals of different curve behaviors can be distinguished. First, the region of fast decrease (up to 100 Å^2); second, an almost constant region (up to about 150 Å^2); followed by a region of rare patch occurrence.

difficult to unselect automatically all these patches which, therefore, disturb the overall distribution, albeit probably not much because the water content in protein crystals is, in general, high and comparable with the protein fraction.

It has been demonstrated in the another article (Eisenhaber & Argos, in prep.) that a radial increment between 0.35 and 0.6 Å for solvent-accessible polar atoms is a suitable expansion for simulating the effect of bound water on the organization of the solvent-accessible surface and that 0.4 Å is good compromise for different proteins. Because the overall form of the dependencies is identical for various polar expansions (compare Fig. 1A,B,C), I will discuss only the data for the radial increment 0.4 Å in more detail. It can well be seen in Figure 1A that the statistics become scarce for areas above 100 Å^2. Hydrophobic surface regions with areas above 1,000 Å^2 are singular events due to the limited number of proteins in the PDB subset. Therefore, this data is not illustrated in Figure 1. Of 3,608 hydrophobic surface regions with more than 10 Å^2, 2,232 are smaller than 100 Å^2 (Table 1) and only 1,376 exceed this value. If bin sizes up to 10 Å^2 are used, the distribution experiences oscillating behavior at larger patch sizes (many consecutive bins with zero and small non-zero occupancy). Patches less than 10 Å^2 in size occupy generally less than 2% of the overall hydrophobic area.

Table 1. *Area distribution of hydrophobic surface regions in the range 10–200 Å^2* [a]

Patch area (Å^2)	Patch occurrence
12.5	263
17.5	268
22.5	306
27.5	290
32.5	190
37.5	160
42.5	144
47.5	97
52.5	66
57.5	78
62.5	71
67.5	65
72.5	60
77.5	36
82.5	38
87.5	41
92.5	28
97.5	31
102.5	28
107.5	21
112.5	21
117.5	30
122.5	19
127.5	21
132.5	24
137.5	19
142.5	22
147.5	8
152.5	21
157.5	17
162.5	13
167.5	9
172.5	14
177.5	9
182.5	9
187.5	13
192.5	9
197.5	7

[a] Area values at the center of a bin (5 Å^2 bin width) and the occurrences of hydrophobic surface regions in a set of 127 nonhomologous monomeric proteins are presented.

Determination of σ_{phob} from the area statistics of solvent-accessible hydrophobic regions

Before starting the regression analysis in accordance with Equation 8, a suitable area range $[A_b, A_e]$ for which sufficient data has been accumulated should be determined. For a bin size of 5 Å^2, the threshold 100 Å^2 of patch area is the onset of an almost constant region (up to 145 Å^2) with occurrences between 19 and 28 patches (Fig. 1A; Table 1). A linear regression analysis in the range from 102.5 to 142.5 Å^2 shows that the slope of the best regression line would be only -0.11 Å^{-2}. The Student's t-value of the slope is 1.12 and clearly below the critical value $t_{9\text{-}2,0.05/2} = 2.365$ of the two-sided criterion at the $\alpha = 5\%$ significance level (and at any smaller α); i.e., the assumption of constant distribution in the area range considered cannot be rejected. The same constant region is also visible for other bin widths (Fig. 2).

With $N = 2{,}232$ hydrophobic surface regions in the range [10 Å^2, 100 Å^2], a bin occupancy of about $N_K \sim 20$ is desirable for the bin with the largest area value (see Equation 12). This condition is fulfilled with a bin width of 5 Å^2 and 18 bins (N_K = 28, the quotient of N and N_K is 79.7). Thus, a fit with Equation 8 appears possible.

In the range 10–100 Å^2, the linear-logarithmic plot of occurrence versus area in Figure 3 is almost linear. The exponential fits (long-dashed line) are shown with together with the original data (solid line connecting open circles). The numerical data describing the parameters λ and C^* as well as the statistical significance of the fit for various polar expansions, are presented in Table 2. Surprisingly, the regression parameters λ and C^* do not depend on the value of the radial increment of solvent-accessible polar atoms in the critical range 0.35–0.8 Å to any significant extent. Such behavior is in agreement with the conclusion that larger polar expansions lead mainly to the proportional decrease in area of existing hydrophobic surface regions and not to their subdivision (Eisenhaber & Argos, in prep.). This result is very important because it proves that the value σ_{phob} derived from the area distribution is not influenced by the specific polar expansion used for delineating the hydrophobic surface regions. At the same time, it is notable that the standard deviations of σ_{phob} are minimal for the polar expansions 0.4 Å and 0.5 Å; i.e., at values that have been found optimal for modeling the hydration effect of bound water.

As shown in Table 2, the specific hydrophobic surface energy σ_{phob} has a value of about 18 cal/(mol·Å^2). Because its standard deviation is about 1 cal/(mol·Å^2), the confidence interval for σ_{phob} is 16–20 cal/(mol·Å^2) for the significance level of 0.05 (the Student's t-value $t_{16,0.05} = 2.120$ and the confidence interval is $(18 \pm t_{16,0.05} \cdot 1)$ cal/(mol·Å^2)). This error describes the statistical uncertainty due to non-ideal data fitting by the regression function. An additional, systematic source of error might be the sample size, which does not allow curve fitting on a larger

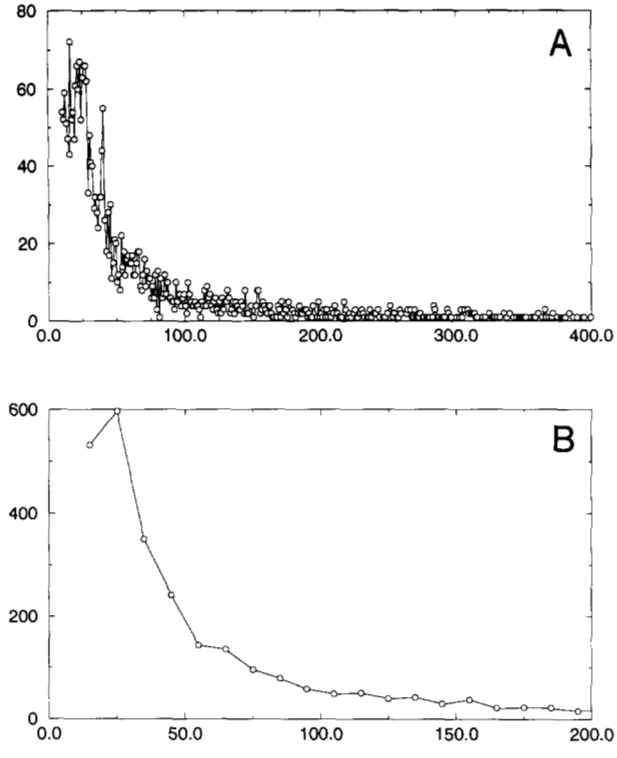

Fig. 2. Influence of bin size. The area distribution of hydrophobic solvent-accessible surface regions computed for the polar expansion 0.4 Å is presented for the bin sizes: **(A)** 1 Å^2 and **(B)** 10 Å^2. The abscissa and the ordinate show the patch area (in Å^2) and the patch occurrence, respectively. Each open circle corresponds to a bin with non-zero occupancy. In the first case, heavy fluctuations are observed. The second graph shows a smoothed curve but with only a few data points.

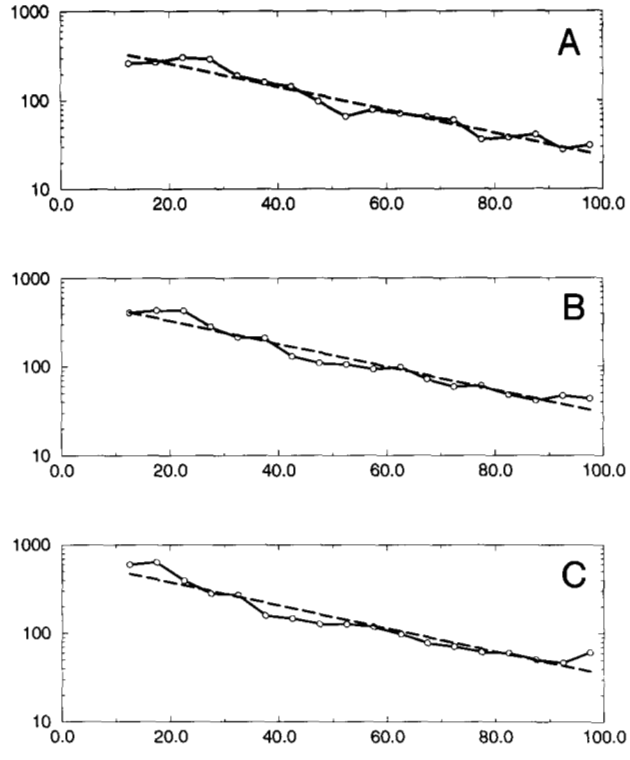

Fig. 3. Exponential fit to the area distribution in the range 10–100 Å^2. The half-logarithmic plots of occurrence versus area for solvent-accessible hydrophobic surface regions in the range of 10–100 Å^2 have been calculated for various polar expansions: **(A)** 0.4 Å^2; **(B)** 0.5 Å^2; and **(C)** 0.6 Å^2. The abscissa and the ordinate show the patch area (in Å^2) and the patch occurrence, respectively. Each open circle corresponds to a bin. The fit with respect to Equation 8 is a long-dashed line. The parameters of the fit are described in Table 2.

area range than [10 Å^2, 100 Å^2]. Because the tail of the area distribution in the large area region has the tendency to be overestimated in small protein samples (uniform distribution of patches with respect to the area), the regression produces too small slopes if the fit interval $[A_b, A_e]$ is taken too long. Therefore, the value of 18 cal/(mol·Å^2) for σ_{phob} obtained in this work can also be considered as lower limit for σ_{phob}.

Comparison with σ_{phob} values obtained with other approaches

In Table 3A and B, a literature review of atomic surface parameters σ_{phob} for hydrophobic atoms obtained from experimental data on small organic compounds, from the thermodynamics of mutated proteins and from neural network classifications of native and non-native protein conformations, is presented. The most known attempts for deriving hydrophobic surface energies (entries S1–3 in Table 3A) exploited the changes in solubility of homologous series of hydrocarbons without and with functional groups (Herrmann, 1972, 1977; Reynolds et al., 1974; Amidon et al., 1975; Valvani et al., 1976). The approaches vary in the selection of organic compounds, the method to calculate the atomic surface (atomic and solvent radii, choice of extended conformations with the largest possible area or the use of surface areas averaged over canonical ensembles), and also in the issue whether the surface area is only linearly related with or directly proportional to the energy of transfer. In the latter case, the regression line passes through the origin (a cavity of zero area corresponds to zero transfer energy). Independent of these choices, the σ_{phob} is in the range 20–33 cal/(mol·Å^2) (Reynolds et al., 1974; Herrmann, 1977; Tanford, 1979). In a concurrent optimization of hydrophobic and hydrophylic group parameters (Amidon et al., 1975; Valvani et al., 1976), σ_{phob} is in the range of 16.6–20.1 cal/(mol·Å^2) in various homologous series with a statistical error of about ±2 cal/(mol·Å^2) at the 5% significance level. The value yielded for pure hydrocarbon is 19.5 cal/(mol·Å^2). A specific apolar surface energy of 17–33 cal/(mol·Å^2) was derived from the geometry of succinic acid crystals (Rees & Wolfe, 1993) assuming a packing optimization aimed at a minimal hydrophobic surface (S10 in Table 3A).

The results for macroscopic surface tension (S4 in Table 3A) and the microscopic surface energy obtained with the solvent-accessible surface [~20 cal/(mol·Å^2)] differ drastically. A recent correction of the transfer energies with a mixing term taking into account volume differences in solute and solvent size resulted in an increased microscopic hydrophobic effect of 47 cal/(mol·Å^2) (De Young & Dill, 1990; Sharp et al., 1991). Further correction terms for surface curvature, surface roughness, and thermal fluctuations lead to 67 cal/(mol·Å^2) (Sharp et al., 1990; Nicholls et al., 1991). With the same assumption but without curvature and roughness terms and relying on the molecular

Table 2. *Exponential fit of the patch area distribution and its statistical significance*[a]

Polar expansion (in Å)	0.35	0.40	0.50	0.60	0.80
$\lambda \cdot 1{,}000$	31.10	30.00	30.00	30.00	29.80
$sd(\lambda) \cdot 1{,}000$	1.84	1.72	1.66	2.08	1.79
t-Test for λ	16.92	17.47	18.04	14.31	16.62
σ_{phob} (cal/(mol·Å^2))	18.52	17.87	17.87	17.87	17.74
$sd(\sigma_{phob})$ (cal/(mol·Å^2))	1.10	1.02	0.99	1.24	1.07
C^*	6.03	6.16	6.40	6.52	6.64
$sd(C^*)$	0.11	0.10	0.10	0.13	0.11
t-Test for C^*	53.98	59.03	63.27	51.52	60.93
F-test	143.18	152.67	162.72	102.38	138.11

[a] The parameters λ and C^* and their standard deviations (denoted as sd) for an exponential fit (Equation 8) to the area distribution of hydrophobic surface regions in a set of 127 nonhomologous monomeric proteins are presented. The regression was obtained for the area range 10–100 Å^2 with 18 bins of 5 Å^2 width. The standard two-sided t-test (Student's t-test) proves the significance of the parameter with respect to the hypothesis, H0, assuming that the parameter is zero. The t-values for the regression $\ln(P) = \ln(C^*) - \lambda \ln(A^*)$ (compare Equation 8) are calculated as:

$$t - \text{value}(\lambda) = \lambda \sqrt{\frac{(m-1) \cdot s_{m-1}(A^*)}{R}} \text{ and}$$

$$t - \text{value}(\ln(C^*)) = \lambda \sqrt{\frac{(m-1)m \cdot s_{m-1}(A^*)}{R \sum_{l=0}^{m-1} A_l^{*2}}},$$

where m is the number of bins ($m = 18$); the summation is executed over the squares of all argument areas (centers of bins) and the expressions $s_k(A)$ and R are:

$$s_k(A) = \frac{1}{k} \sum_{l=0}^{m-1} (A_i - \bar{A})^2 \text{ and}$$

$$R = \frac{1}{m-2} \sum_{l=0}^{m-1} [\ln(P_l) - \ln(\hat{P}_l)]^2.$$

The summation in the first equation is made over all differences between values of area in data pairs and the mean value. R is the sum of squared distances between the natural logarithms of actual and fitted P-values (weighted residual of the regression). In the case of our regressions, the test is always passed because the t-values are significantly larger than 2.120 and 3.922 for the significance levels 0.05 and even 0.001, respectively (16 degrees of freedom).

The standard F-test (Fisher's test) checks the suitability of the regression with respect to the hypothesis, H0, assuming that the average function value is a similarly good approximation. The F-test value is calculated as:

$$F = \frac{s_2(P) - \frac{m-2}{2} R}{R}.$$

Because $F_{2,16} = 3.63$ (significance level 0.01) and $F_{2,16} = 6.23$ (significance level 0.05) (i.e., both are always significantly smaller) the hypothesis H0 is not acceptable and the regression is statistically significant. Surprisingly, the polar expansion does not have any important influence on the regression parameters in the studied range 0.35–0.8 Å.

The value σ_{phob} and its standard deviation have been calculated by multiplying with kT the parameter λ and its standard deviation respectively. We assumed $T = 300$ K, in agreement with Finkelstein et al. (1995b).

surface (without incrementation of atomic radii by the probe radius as in the case of the solvent-accessible surface), Tuñón et al. (1992) arrived at 69 cal/(mol·Å^2) with additional 4 cal/(mol·Å^2) for thermal fluctuations. Later criticism revealed that any concentration as well as solute and solvent size dependencies are already included in the standard transfer energy (Holtzer, 1992, 1994; Ben-Naim & Mazo, 1993; Juffer et al., 1995) and only corrections to solute-solvent interactions expressed by activity coefficients may be applied (see Equation 2 in Juffer et al., 1995).

Already Tanford (1979) supposed that the discrepancy between microscopic and macroscopic hydrophobic effect might be resolved by considering a type of surface differing from the solvent-accessible surface by reduced atomic radii. Indeed, on the time-averaged positions at the interface, the molecules are expected to be packed in a plane and, consequently, molecular and solvent-accessible surfaces do not differ in area. This is not the case for individual organic molecules. Jackson and Sternberg (1994) derived the surface energy for the molecular surface within the scaled particle theory and showed that the microscopic hydrophobic effect does not differ in magnitude from the macroscopic one. Therefore, both the microscopic and the macroscopic hydrophobic effects are related mainly by the scaling factor between solvent-accessible and molecular surfaces areas if a possible shape dependency is neglected.

In a second group of approaches, many atomic solvation parameters have been optimized concurrently to fit transfer energies in accordance with Equation 1. In this case, the parameter for hydrophobic atoms may be influenced seriously by side effects as a result of polar interactions with solvent that contribute the major share to the transfer energy of polar compounds. This happened obviously in the case of the parameter for carbonyl carbons in the set of Ooi et al. (1987), which is 427 cal/(mol·Å^2), compared with only 8 cal/(mol·Å^2) for any apolar carbon (S6 in Table 3A). Similarly, optimization of the surface energy parameters within the CHARMm force field (Schiffer et al., 1993) results in a large hydrophobic energy (S9 in Table 3A). Our value of σ_{phob} compares favorably with atomic solvation parameters for carbon such as 16–18 cal/(mol·Å^2) (Eisenberg & McLachlan, 1986; Eisenberg et al., 1989) or 12 cal/(mol·Å^2) (Wesson & Eisenberg, 1992) obtained from transfer energies of small organic compounds from a hydrophobic environment to water (entries S5, S7, and S9 in Table 3A). Surprisingly, both different transfer energies (octanol → water compared with vacuum → water, without and with "correction with entropy of mixing) and small differences in molecular geometry change little the hydrophobic atom parameter, whereas polar parameters change drastically even in the order of magnitude.

Koehl and Delarue (1994) repeated the calculation of Eisenberg and McLachlan (1986) with the same assumptions and obtained surprisingly 36 cal/(mol·Å^2) (S11 in Table 3A). Perhaps the discrepancy is due to their amino acid conformations taken from only four native protein structures (compared with 201 amino acid conformations in Eisenberg et al., 1989) biased to lower accessibilities of apolar atoms.

To summarize, the interpretation of small organic molecule experimental data is consistent with a specific hydrophobic surface energy in the ranges of 19.5–33 cal/(mol·Å^2) (hydrocarbon solubility), 16.6–20.1 and 17–33 cal/(mol·Å^2) (solubility of alkanes with functional groups and succinic crystal geometry, respectively), and 12–18 cal/(mol·Å^2) (carbon in atomic solva-

Table 3. *Specific free surface energy of hydrophobic solvent-accessible surface*[a]

No.	σ_{phob} (cal/(mol·Å^2))	Method (reference)
A. Derivation from small molecule data		
S1	28–33	Solubility of series of alkyl derivatives (transfer pure organic compound → water), probe radius 1.5 Å (Herrmann, 1972, 1977)
S2	20–25	Solubility of series of alkyl derivatives (transfer pure organic compound → water), probe radius 1.5 Å (Reynolds et al., 1974)
S3	16.6–20.1 (±2)	Solubility of series of alkyl derivatives (transfer pure organic compound → water), probe radius 1.5 Å (Amidon et al., 1975; Valvani et al., 1976)
S4	73	Interfacial free energy between pure hydrocarbon and water (Tanford, 1979)
S5	16 (±4)	Amino acid transfer energies octanol → water (Eisenberg & McLachlan, 1986)
S6	8	Transfer of small organic solutes gas phase → water (Ooi et al., 1987)
S7	18 (±2)	Amino acid transfer energies octanol → water (Eisenberg et al., 1989)
S8	12 (±6)	Amino acid transfer energies vacuum → water, correction for entropy of mixing (Wesson & Eisenberg, 1992)
S9	32.5	Combined structure optimization with CHARMm and surface energy (Schiffer et al., 1993)
S10	17–33[b]	Geometry of succinic acid crystal (Rees & Wolfe, 1993)
S11	36 (±1)	Amino acid transfer energies octanol → water, correction for entropy of mixing (Koehl & Delarue, 1994)
B. Derivation from protein data		
P1	20 (±10)[c]	Protein stability change due to cavity-creating mutation in T4 lysozyme (Eriksson et al., 1992, 1993)
P2	19[d]	Protein stability change due to mutation at solvent-accessible site 44 in T4 lysozyme (Blaber et al., 1993, 1994)
P3	16 (±6)[e]	Protein stability change due to cavity-creating mutation in sperm-whale myoglobin (Pinker et al., 1993)
P4	8–18 (1–27)[f]	Classification of native and non-native protein conformations with a neural network as estimated from their Figure 3 (Wang et al., 1995)
P5	18 (±2)	Area distribution of hydrophobic surface regions (this work)

[a] Values or ranges for σ_{phob} as described in the literature are listed. Part A of the table comprises approaches based on small organic compounds (series S), the methods of part B use experimental data for proteins (series P). If a standard deviations was reported, the error margin for the 0.05 significance level of σ_{phob} is given in parentheses (this statistical error is about twice the standard deviation).

[b] The range is 40–80 cal/(mol·Å^2) for the molecular surface and has been converted to that for solvent-accessible surface with factor 2.4^{-1} (Rees & Wolfe, 1993).

[c] Statistically not significant regression. The error margins have been calculated in this work (for methodology, see legend to Table 2).

[d] Regression is based on eight residue types, data points for other amino acids are very distant from the regression line and have been excluded. Values of area burial are given only in graphic form. Statistical significance of regression has not been calculated.

[e] The error margins have been calculated in this work (for methodology, see legend to Table 2).

[f] The first range is given for aliphatic carbons and aromatic carbons of phenylalanine and the 6-carbon ring of tryptophane. The range in parentheses includes any type of carbon.

tion parameter sets). These results are in good agreement with the value for σ_{phob} = 18 (±2) cal/(mol·Å^2) obtained in this work.

Only a few recently published reports describe techniques to derive ASP using data on proteins (entries P1–4 in Table 3B). Wang et al. (1995) optimize the ASP with a neural network under the condition of maximal solvation energy difference between ensembles of native and generated non-native conformations. Unfortunately, it appeared difficult to converge the neural network reproducibly for various selections of protein structures. Therefore, only a probable range for atomic solvation parameters could be published. The value for hydrophobic carbon (see atom types C_{al}, C_W, and C_F in Figure 3 of Wang et al., 1995) is approximately between 8 and 18 cal/(mol·Å^2). The range for any type of carbon is 1–27 cal/(mol·Å^2). This technique has some other unsolved methodological problems, such as the description of the denaturated state by generated non-native conformations and the issue of concurrent optimization of all ASPs [the inadequacy of Equation 1 for highly polar and charged atoms (Juffer et al., 1995) may lead to bias for the hydrophobic parameters].

Eriksson et al. (1992, 1993) and Pinker et al. (1993) derived σ_{phob} from stability changes in mutated proteins. Both claim to obtain a linear correlation between the change in free energy $\Delta\Delta G$ resulting from a cavity-creating mutation and the surface of the cavity created, both in T4 lysozyme (Eriksson et al., 1992, 1993) and for sperm whale myoglobin (Pinker et al., 1993), with a slope of about 20 cal/(mol·Å^2). We repeated their regression

analysis and found that the assumed linear dependency of change in thermodynamic stability $\Delta\Delta G$ and in cavity surface is of poor statistical significance due to considerable scatter of the data. For example, Eriksson et al. (1992) published six data pairs (P1 in Table 3B). The slope is really 19.9 cal/(mol·Å^2), but with a confidence interval of ±9.8 cal/(mol·Å^2) at the α = 5% significance level. The t-values for the slope and intercept are 4.08 and 4.4, respectively, i.e., both slope and intercept are significantly different from zero at $\alpha = 5\%$ ($t_{4,0.05} = 2.776$, statistical significance), but not at the stronger $\alpha = 1\%$ ($t_{4,0.01}$ = 4.604, no statistical significance). The F-test of the regression results in 8.34, which is smaller than the critical value $F_{2,4,0.01}$ = 16.69; i.e., the assumption of linear dependency can be rejected. That means that the entry P1 in Table 3B does not represent a scientific result in the closer meaning the word. In another set of mutation data (stability of mutations at the solvent-accessible site 44 in an α-helix of T4 lysozyme, entry P2 in Table 3B), a slope of about 19 cal/(mol·Å^2) has been obtained for eight amino acid types (Blaber et al., 1993, 1994). The significance of this regression is not clear because data points for many other amino acid types that are very distant from the regression line have been excluded from the analysis. In addition, it is even more difficult than for the cavity-creating mutations to assess the contribution of the hydrophobic effect among other energy terms.

Pinker et al. (1993) published the regression of cavity area (ordinate) versus energy (abscissa) for data from nine mutation experiments and obtain a slope of 52 Å^2 per kcal/mol (P3 in Table 3B). Because the data points scatter considerably, it is not correct to take the reciprocal of the slope to get the specific hydrophobic surface energy, as was done in the article because the regression is not commutative. Instead, a new regression of energy (ordinate) versus cavity area (abscissa) has to be computed. In this case, the slope is only 15.7 cal/(mol·Å^2). The error for $\alpha = 5\%$ is ±5.6 cal/(mol·Å^2).

In addition to the issue of statistical significance of the regressions, systematic errors have to be considered. Both Eriksson et al. (1992, 1993) and Pinker et al. (1993) rely on the assumption that the difference in free energy between various data points can completely be assigned to the change in surface energy of the cavity created. Other factors are also influential, such as conformational adjustments, changes in van der Waals contacts, and electrostatic interactions inside the protein and with the solvent (Eriksson et al., 1993; Jackson & Sternberg, 1994). To conclude, previous attempts to derive a specific hydrophobic surface energy from protein data (Table 3B) yielded only estimates with large confidence ranges [Wang et al., 1995, 8–18 cal/(mol·Å^2); Eriksson et al., 1992, 10–30 cal/(mol·Å) with a no statistically significant regression; Pinker et al., 1993, 10–22 cal/(mol·Å)].

The agreement of our result for σ_{phob} with estimates of other authors obtained independently allows also another interpretation. The conformational temperature of 300 K assumed in our derivation of σ_{phob} (as suggested by Finkelstein et al., 1995b) and, therefore, the hypothesis of the REM model appear to be correct. This conclusion is in contrast with the result of Thomas and Dill (1996), who related fractional accessibilities of residue types in proteins and transfer energies and obtained conformational temperatures of 640 K and 1,800 K for different sets of protein structures. I suggest that it is not correct to assume the validity of Equation 1 also for polar atoms and to translate the full transfer energy into surface area without explicit electrostatics as Thomas and Dill did. In this work, I considered only hydrophobic surface area, for which Equation 1 is a good approximation.

Conclusion

In this work, it was shown that it is possible to derive an atomic solvation parameter for solvent-accessible hydrophobic atoms directly from macromolecular structural data using the area distribution of hydrophobic surface regions. The value of 18 ± 2 cal/(mol·Å^2) for apolar atoms is in agreement with the results from studies on low-molecular compounds. The transferability of solvation parameters derived from small molecules on macromolecules has been proven for the case of hydrophobic atoms. The results also suggests that the conformational temperature as defined by Finkelstein (1995b) is really near 300 K.

Materials and methods

The solvent-accessible surface area as defined by Lee and Richards (1971) was computed with the fast and accurate analytical routine ASC (Eisenhaber & Argos, 1993; Eisenhaber et al., 1995a, see http://www.embl-heidelberg.de/~eisenhab/). The atomic radii from Table 1 of Juffer et al. (1995) were used. The probe (water) radius was set equal to 1.4 Å. The complete topological description of the surface, which is a side product of the analytical area computation, was used to calculate the solvent-accessible faces of atoms. The cycle-face assignment problem, which appears in the case of several cycles of solvent-accessible arcs on one and the same atom, was solved with stereographic projection (Connolly, 1983). Hydrophobic regions were determined as sets of spherical faces on hydrophobic atoms (carbon and sulphur) being connected through common solvent-accessible circular arcs.

Solvent-accessible surface computations were performed under various conditions using: (1) only protein atoms and (2) only protein atoms, but with enlarged radii for polar atoms accessible in case (1). A polar expansion of 0.4 Å is suitable to model the effect of the first hydration shell on the hydrophobic portion of the protein surface (Eisenhaber & Argos, in prep.).

The selection of polypeptide chains with nonhomologous tertiary structures (residue identity ≤35%, resolution ≤2.0 Å) taken from the PDB (Bernstein et al., 1977; Abola et al., 1987) was performed automatically with the program OBSTRUCT (Heringa et al., 1992). Protein structures determined by NMR techniques as well as structures with incomplete backbone or side chains (except for the first or the last residue of the polypeptide chain) were excluded. From the resulting set of 229 proteins, 100 polypeptide chains that were subunits of multimeric proteins as recognized by a chain identifier or an appropriate commentary in a REMARK record were removed. I also excluded two membrane proteins. The remaining list of 127 entries was constituted by monomeric proteins whose atomic coordinates are given in PDB files identified as: 153l, 1aaj, 1acf, 1acx, 1adl, 1ahc,1alc, 1ald, 1amp, 1arb, 1ast, 1bcx, 1btc, 1btl, 1cbs, 1cdp, 1cfb, 1cgt, 1cie, 1cnr, 1cot, 1cp4, 1crm, 1ctf, 1cyo, 1eas, 1edb, 1enj, 1esl, 1fas, 1fdx, 1fkb, 1flp, 1frd, 1fxd, 1gcd, 1gcs, 1gdi, 1gky, 1glg, 2glt, 1gof, 1gox, 1gpr, 1hbg, 1hfc, 1hoe, 1hpi, 1huw, 1hyp, 1iag, 1icn, 1igd, 1kdb, 1knt, 1l99, 1lis, 1mba, 1mjc, 1nar, 1npc, 1ofv, 1oyb, 1paz, 1pbe, 1php, 1pii, 1poa, 1poc, 1poh, 1ppa, 1ppn, 1ptx, 1r69, 1rbp, 1rbu, 1ris, 1sbp, 1sgt, 1shg, 1srp,

1st3, 1tag, 1tca, 1tdy, 1ten, 1thw, 1tib, 1tml, 1top, 1ubi, 1ukz, 1yma, 21bi, 2acs, 2apr, 2bat, 2che, 2cna, 2ctb, 2cut, 2cy3, 2dri, 2ebn, 2exo, 2fcr, 2fx2, 2hts, 2lhb, 2mcm, 2pia, 2ran, 2sil, 2sn3, 351c, 3bcl, 3blm, 3cms, 3fxn, 3il8, 4enl, 4icb, 5fd1, 6abp, 6ldh, 6pti, and 7pcy.

Acknowledgments

I thank Birgit Eisenhaber for advice in regression analysis and in the use of statistical criteria and Jaap Heringa as well as Dirk Walther for help in constructing a list of monomeric proteins. I am grateful to Patrick Argos for discussion and, together with Cornelius Frömmel, for constant support. I have received financial help from the Fund "Wisssenschaftler-Integrationsprogramm" (grant 020386/B) administered jointly by the East-German Länder and the Federal Republic of Germany.

References

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information content, software systems, scientific applications.* Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107-132.

Amidon GL, Yalkowsky SH, Anik ST, Valvani SC. 1975. Solubility of nonelectrolytes in polar solvents. V. Estimation of the solubility of aliphatic monofunctional compounds in water using a molecular surface area approach. *J Phys Chem 79*:2239-2245.

Argos P. 1988. An investigation of protein subunit and domain interfaces. *Protein Eng 2*:101-113.

Ben-Naim A, Mazo RM. 1993. Size dependence of the solvation free energy of large solutes. *J Phys Chem 97*:10829-10834.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535-542.

Blaber M, Zhang XJ, Lindstrom JL, Pepiot SD, Baase WA, Matthews BW. 1994. Determination of alpha-helix propensities within the context of a folded protein: Sites 44 and 131 in bacteriophage T4 lysozyme. *J Mol Biol 235*:600-624.

Blaber M, Zhang XJ, Matthews BW. 1993. Structural basis of amino acid alpha helix propensities. *Science 269*:1637-1640.

Bryant SH, Lawrence CE. 1991. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. *Proteins Struct Funct Genet 9*:108-119.

Casari G, Sippl MJ. 1992. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins are able to identify native folds. *J Mol Biol 224*:725-732.

Chothia C. 1974. Hydrophobic bonding and accessible surface areas in proteins. *Nature 248*:338-339.

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

Connolly ML. 1983. Analytical molecular surface calculation. *J Appl Crystallogr 16*:548-558.

Covell DG, Smythers GW, Gronenborn AM, Clore MG. 1994. Analysis of hydrophobicity in the a and b chemokine families and its relevance to dimerization. *Protein Sci 3*:2064-2072.

Crippen GM, Viswanadhan VN. 1985. Sidechain and backbone potential function for conformational analysis of proteins. *Int J Peptide Protein Res 25*:487-509.

Delarue M, Koehl P. 1995. Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J Mol Biol 249*: 675-690.

De Young LR, Dill KA. 1990. Partitioning on nonpolar solutes into bilayers and amorphous *n*-alkanes. *J Phys Chem 94*:801-808.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry 29*:7133-7155.

Eisenberg D, McLachlan AD. 1986. Solvation energy in protein folding and binding. *Nature 319*:199-203.

Eisenberg D, Wesson M, Yamashita M. 1989. Interpreting of protein folding and binding with atomic solvation parameters. *Chemica Scripta 29 A*: 217-221.

Eisenhaber F, Argos P. 1993. Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. *J Comput Chem 14*:1272-1280.

Eisenhaber F, Argos P. 1996. Hydrophobic regions on protein surfaces. I. Definition with the structure of the first hydration shell and a quick method for region computation. Forthcoming.

Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M. 1995a. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and for generating dot surfaces of molecular assemblies. *J Comp Chem 16*:273-284.

Eisenhaber F, Persson B, Argos P. 1995b. Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. *CRC Crit Rev Biochem Mol Biol 30*:1-94.

Eriksson AE, Baase WA, Matthews BW. 1993. Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences. *J Mol Biol 229*:747-769.

Eriksson AE, Baase WA, Zhang XJ, Heinz DW, Blaber M, Baldwin EP, Matthews BW. 1992. Response of a protein structure to cavity creating mutations and its relation to the hydrophobic effect. *Science 255*:178-183.

Finkelstein AV, Badretdinov AY, Gutin AM. 1995a. Why do protein architectures have Boltzmann-like statistics? *Proteins Struct Funct Genet 23*:142-150.

Finkelstein AV, Gutin AM, Badretdinov AY. 1995b. Perfect temperature for protein structure prediction and folding. *Proteins Struct Funct Genet 23*:151-162.

Finkelstein AV, Ptitsyn OB, Kozitsyn SA. 1977. Theory of protein molecule self-organization. II. A comparison of calculated thermodynamic parameters of local secondary structure with experiment. *Biopolymers 16*:641-656.

Gutin AM, Badretdinov AY, Finkelstein AV. 1992. Why is the statistics of protein structures Boltzmann-like? *Mol Biol 26*:94-102.

Harvey SC. 1989. Treatment of electrostatic effects in macromolecular modeling. *Proteins Struct Funct Genet 5*:78-92.

Hasel W, Hendrickson TF, Still WC. 1988. A rapid approximation to the solvent accessible surface area of atoms. *Tetrahedron Computer Methodology 1*:103-116.

Heringa J, Sommerfeldt H, Higgins D, Argos P. 1992. OBSTRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *Comput Appl Biosci 8*:599-600.

Herrmann RB. 1972. Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility with solvent cavity surface area. *J Phys Chem 76*:2754-2759.

Herrmann RB. 1977. Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions. *Proc Natl Acad Sci USA 74*:4144-4145.

Holtzer A. 1992. The use of Flory-Huggins theory in interpreting partitioning of solutes between organic liquids and water. *Biopolymers 32*:711-715.

Holtzer A. 1994. Does Flory-Huggins theory help in interpreting solute partitioning experiments? *Biopolymers 34*:315-320.

Honig B, Yang AS. 1995. Free energy balance in protein folding. *Adv Protein Chem 46*:27-58.

Hubbard SJ, Gross KH, Argos P. 1994. Intramolecular cavities in globular proteins. *Protein Eng 7*:613-626.

Jackson RM, Sternberg MJE. 1994. Application of scaled particle theory to model the hydrophobic effect: Implications for molecular associations and protein stability. *Protein Eng 7*:371-383.

Jackson RM, Sternberg MJE. 1995. A continuum model for protein-protein interactions: Application to the docking problem. *J Mol Biol 250*:258-275.

Janin J, Chothia C. 1990. The structure of protein-protein recognition sites. *J Biol Chem 265*:1627-1630.

Janin J, Miller S, Chothia C. 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol 204*:155-164.

Janin J, Rodier F. 1995. Protein-protein interactions at crystal contacts. *Proteins Struct Funct Genet 23*:580-587.

Jones G, Willet P, Glen RC. 1995. Molecular recognition of receptor sites using a genetic algorithm with a description of solvation. *J Mol Biol 245*:43-53.

Juffer A, Eisenhaber F, Hubbard SJ, Walther D, Argos P. 1995. Comparison of atomic solvation parameter sets: Applicability and limitations in protein folding and binding. *Protein Sci 4*:2499-2509.

Koehl P, Delarue M. 1994. Polar and nonpolar atomic environments in the protein core: Implications for folding and binding. *Proteins Struct Funct Genet 20*:264-278.

Korn AP, Burnett RM. 1991. Distribution and complementarity of hydropathy in multisubunit proteins. *Proteins Struct Funct Genet 9*:37-55.

Leckband DE, Schmitt FJ, Israelachvili JN, Knoll W. 1994. Direct force measurements of specific and nonspecific protein interactions. *Biochemistry 33*:4611-4624.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.
MacArthur MW, Thornton JM. 1991. Influence of proline residues on protein conformation. *J Mol Biol 218*:397–412.
Miller S. 1989. The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng 3*:77–83.
Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. *J Mol Biol 196*:641–656.
Miyazawa S, Jernigan RL. 1985. Estimation of interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.
Miyazawa S, Jernigan RL. 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol 256*:623–644.
Nicholls A, Sharp KA, Honig B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct Funct Genet 11*:281–296.
Ooi T, Oobatake M, Nemethy G, Scheraga HA. 1987. Accessible surface areas as measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci USA 84*:3086–3090.
Otting G, Liepinsh E. 1995. Protein hydration viewed by high-resolution NMR spectroscopy: Implications for magnetic resonance image contrast. *Acc Chem Res 28*:171–177.
Phillips GN Jr, Pettitt BM. 1995. Structure and dynamics of the water around myoglobin. *Protein Sci 4*:149–158.
Pickett SD, Sternberg MJE. 1993. Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol 231*:825–839.
Pinker RJ, Rose GD, Kallenbach NR. 1993. Effect of alanine substitutions in alpha-helices of sperm whale myoglobin on protein stability. *Protein Sci 2*:1099–1105.
Pitt WR, Murray-Rust J, Goodfellow JM. 1993. AQUARIUS2: Knowledge-based modeling of solvent sites around proteins. *J Comp Chem 14*: 1007–1018.
Pohl FM. 1971. Empirical protein energy maps. *Nature New Biol 23*:277–279.
Rashin AA, Ionif M, Honig B. 1985. Internal cavities and buried waters in globular proteins. *Biochemistry 25*:3619–3625.
Rees DC, Wolfe GM. 1993. Macromolecular solvation energies derived from small molecule crystal morphology. *Protein Sci 2*:1882–1889.
Reynolds JA, Gilbert DB, Tanford C. 1974. Empirical correlation between hydrophobic free energy and aqueous cavity surface area. *Proc Natl Acad Sci USA 71*:2925–2927.
Richards WG, King PM, Reynolds CA. 1989. Solvation effects. *Protein Eng 2*:319–327.
Rupley JA, Careri G. 1991. Protein hydration and function. *Adv Protein Chem 41*:37–172.
Schiffer CA, Caldwell JW, Kollman PA, Stroud RM. 1993. Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol Simul 10*:121–149.
Serrano L, Sancho J, Hirshberg M, Fersht AR. 1992. α-Helix stability in proteins. I. Empirical correlations concerning substitutions of side chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J Mol Biol 227*:544–559.
Sharp KA, Nicholls A, Fine RF, Honig B. 1990. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science 252*:106–109.
Sharp KA, Nicholls A, Friedman R, Honig B. 1991. Extracting hydrophobic free energies from experimental data: Relationship to protein folding and theoretical models. *Biochemistry 30*:9686–9697.
Still WC, Tempczyk A, Hawley RC, Hendrickson T. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc 112*:6127–6129.
Tanaka S, Scheraga HA. 1976. Medium and long-range interaction parameters between amino acids for predicting three-dimensional structure of proteins. *Macromolecules 9*:945–950.
Tanford C. 1979. Interfacial free energy and the hydrophobic effect. *Proc Natl Acad Sci USA 76*:4175–4176.
Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol 257*:457–469.
Tuñón I, Silla E, Pascual-Ahuir JL. 1992. Molecular surface area and hydrophobic effect. *Protein Eng 5*:715–716.
Valvani SC, Yalkowsky SH, Amidon GL. 1976. Solubility of nonelectrolytes in polar solvents. VI. Refinements in molecular surface area computations. *J Phys Chem 80*:829–835.
Vedani A, Huhta D. 1991. An algorithm for the systematic solvation of proteins based on the directionality of hydrogen bonds. *J Am Chem Soc 113*:5860–5862.
Wang Y, Zhang H, Scott RA. 1995. A new computational model for protein folding based on atomic solvation. *Protein Sci 4*:1402–1411.
Wesson L, Eisenberg D. 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci 1*:227–235.
Young L, Jernigan RL, Covell DG. 1994. A role for surface hydrophobicity in protein–protein recognition. *Protein Sci 3*:717–729.