Protein Science (1995), *4*:1516–1527. Cambridge University Press. Printed in the USA.

The optimization of protein–solvent interactions: Thermostability and the role of hydrophobic and electrostatic interactions

VELIN Z. SPASSOV,[1,2] ANDREJ D. KARSHIKOFF,[1] AND RUDOLF LADENSTEIN[1]

[1] Centre for Structural Biochemistry, Karolinska Institutet, NOVUM, S-14157 Huddinge, Stockholm, Sweden
[2] Institute of Biophysics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

(RECEIVED April 4, 1995; ACCEPTED May 9, 1995)

Abstract

Protein–solvent interactions were analyzed using an optimization parameter based on the ratio of the solvent-accessible area in the native and the unfolded protein structure. The calculations were performed for a set of 183 nonhomologous proteins with known three-dimensional structure available in the Protein Data Bank. The dependence of the total solvent-accessible surface area on the protein molecular mass was analyzed. It was shown that there is no difference between the monomeric and oligomeric proteins with respect to the solvent-accessible area. The results also suggested that for proteins with molecular mass above some critical mass, which is about 28 kDa, a formation of domain structure or subunit aggregation into oligomers is preferred rather than a further enlargement of a single domain structure. An analysis of the optimization of both protein–solvent and charge–charge interactions was performed for 14 proteins from thermophilic organisms. The comparison of the optimization parameters calculated for proteins from thermophiles and mesophiles showed that the former are generally characterized by a high degree of optimization of the hydrophobic interactions or, in cases where the optimization of the hydrophobic interactions is not sufficiently high, by highly optimized charge–charge interactions.

Keywords: electrostatic interactions; hydrophobic interactions; solvent accessibility; thermostability

In recent years a certain success has been achieved in understanding the molecular basis of protein thermostability, mainly due to the considerable increase in the number of available amino acid sequences and three-dimensional structures and the opportunity for comparison of the structures of related proteins from thermophilic and mesophilic organisms (Zuber, 1988). However, the main question, namely, which properties are responsible for the shift in the denaturation temperature of thermostable proteins, is still to be answered (Rehaber & Jaenicke, 1992). It was shown that small differences in amino acid sequence can cause changes in thermostability of proteins. Thus, for example, the thermostable ferredoxin from *Clostridium thermosaccharolyticum* differs from its less stable relative from *Clostridium tartarivorum* in only two positions: glutamines 31 and 44 are replaced by glutamates (Perutz & Raidt, 1975). Also, the enhanced thermal stability of hemoglobin A_2 is determined by one extra hydrogen bond and two additional nonpolar contacts in comparison with adult human hemoglobin (Perutz & Raidt, 1975). Amino acid point mutations can change the thermal stability of proteins, as well (Eijsink et al., 1992; Goward et al., 1994). Theoretical and experimental analysis of proteins has convincingly shown that the three-dimensional structure of a folded protein and the resulting properties, including thermostability, are a result of the delicate balance of different types of interactions: van der Waals interactions, hydrogen bonds, charge–charge interactions, hydrophobic effect. Amino acid sequences and structure of proteins from thermophilic and mesophilic organisms have been compared by means of statistical analysis of amino acid exchanges (Argos et al., 1979). The pattern of exchanges has suggested that thermal stability is largely achieved by an additive series of small improvements at many locations in the molecule without significant changes in the tertiary structure. Their overall effect is primarily to increase the internal and decrease external hydrophobicity, as well as to favor helix-stabilizing residues in α-helices and strand-stabilizing residues in β-strands. Stellwagen and Wilgus (1978) have proposed that the ratio of surface area to volume for a given protein, domain, or subunit is a critical factor of thermal stability.

According to the thermodynamic hypothesis of Anfinsen (1973), the unique native structure of a globular protein corresponds to the minimum of its free energy. It is usually represented as a sum of the individual contributions of the different

Reprint requests to: Andrej D. Karshikoff, Centre for Structural Biochemistry, Karolinska Institutet, NOVUM, S-14157 Huddinge, Stockholm, Sweden; e-mail: aka@barra.csb.ki.se.

types of interactions. However, the minimum of the free energy does not require a minimum of each of its components. In other words, the native protein structure is not necessarily characterized, say, by a minimum of the electrostatic term of the free energy. Thus, for instance, the electrostatic energy of the trimer-trimer interaction in phycocyanin from *Fremyella diplosiphon* as a function of the mutual subunit orientation has a minimum close, but not directly at the observed crystallographic structure (Karshikov et al., 1991). In this case, the dominant factor stabilizing the phycocyanin hexamer was the steric complementarity, i.e., the protein-solvent contact area is minimized. Obviously, the role of the different interactions in folding or stability at given conditions must be analyzed in the context of all other interactions. The solution of this task is rather laborious, mainly due to the complex coupling between the different types of interactions. A common way to avoid this difficulty is to analyze the protein behavior under conditions where only one of the factors is changed, whereas all others are kept constant or are neglected. In the present study, we followed this concept; however, our analysis was focused on the divergence of a given type of interactions from its minimum in the real structure, which would be achieved if all other interactions are neglected. To this difference we refer the *optimization* of the corresponding type of interactions for a given real protein structure: the smaller the difference, the higher the optimization. Let us assume that there is a structure that corresponds to the energy minimum of a certain type of noncovalent interactions. We call this hypothetical structure an *optimal structure* with regard to this type of interaction. This structure can then be used as a reference for evaluation of the optimization of a given type of interactions. For example, if we consider only the local main-chain hydrogen bonds, the optimal structure would correspond to a 100% saturation of hydrogen bonding and would then be an α-helix. By this way, the myoglobin molecule (84% α-helix) is better optimized with respect to these interactions than cytochrome *c* (45% α-helix) (Spassov et al., 1994). In our previous work (Spassov & Atanasov, 1994; Spassov et al., 1994), we have used this approach for an analysis of the optimization of the charge-charge interactions in globular proteins. A random distribution of the charges over the protein surface was used for the reference state because both most optimized and deoptimized structures cannot be defined in this case. Some common features related to the optimization of the charge-charge interactions were detected: the enzymes are generally better optimized than the proteins without enzymatic functions; the proteins that belong to the mixed $\alpha\beta$ folding type are electrostatically better optimized than pure α-helical or β-strand structures; proteins with a low degree of electrostatic optimization are covalently stabilized by disulfide bonds. It was also found that the electrostatic interactions in a native protein are effectively optimized by rejection of the conformers that lead to repulsive charge-charge interactions.

Our results obtained by the analysis of optimization of the electrostatic interactions in proteins (Spassov et al., 1994) have shown that the approach of separating the different types of interactions by the way described above is able to detect some general rules governing stabilization of native protein structures. This encouraged us to apply this approach for an analysis of protein-solvent interactions.

The protein-solvent interactions are considered as an important factor responsible for protein stability. Their significance, especially the hydrophobic interactions, are widely discussed in the literature (Ponnuswamy, 1993; Rose & Wolfenden, 1993). There are different and sometimes controversial opinions about the molecular nature of the hydrophobic interactions (Dill, 1990; Ponnuswamy, 1993). The most common definition of hydrophobic interactions is related to the process in which an apolar group is transferred from a polar phase to an apolar phase (Rose & Wolfenden, 1993). This reflects in the folding process a certain tendency of apolar side chains to leave the water and thus to form the hydrophobic core of a protein. On the basis of thermodynamic data of transfer of hydrocarbons from the liquid state to water and protein unfolding, Privalov and Makhatadze (Makhatadze & Privalov, 1993; Privalov & Makhatadze, 1993) have defined the hydrophobic interactions as a sum of the van der Waals interactions between the apolar atoms and the hydration of these atoms; the latter destabilizes the native structure. This definition shows clearly that the hydrophobic interactions are a complex phenomenon including different type of interactions. An experimental fact is that protein stability depends on the exposure of apolar groups to the solvent (Hirono et al., 1991; Tuñon et al., 1992). Thus, the solvent accessibility of the apolar atoms appears to be an appropriate tool for the analysis of the contribution of the hydrophobic interactions as a part of the protein-solvent interactions and their role in protein stability. A decisive advantage is that this parameter is linearly related to the energetics of the protein (Chothia, 1974) and can be easily calculated if the three-dimensional protein structure is available.

In this paper, a method for an analysis of protein-solvent interactions is described. It is based on the calculation of an optimization parameter using atomic solvent accessibilities. The analysis was performed on a set of 183 nonhomologous proteins with known three-dimensional structure available in the last issue (1994) of the Protein Data Bank (PDB) (Bernstein et al., 1977). As expected, the proteins with less than about 1,000 atoms are characterized by a lower optimization, due to an incomplete hydrophobic core. This limit coincides with the smallest known single-domain enzymes. It was shown that once a hydrophobic core of sufficient size is formed, further increase in the molecular mass of a protein is realized by the formation of domains and/or subunit aggregation into oligomers, but not by continuous enlargement of a single hydrophobic core. The analysis also showed that the thermostable proteins are generally characterized by a high degree of optimization of the hydrophobic interactions, or in cases where the optimization of the hydrophobic interactions is not sufficiently high, by highly optimized charge-charge interactions.

Results and discussion

In the following analysis we will use the number of the nonhydrogen atoms in the protein, N_t, as a measure of its molecular mass. The reason for this is that the relation between the molecular weight and N_t is practically linear and that the solvent accessibilities as well as the optimization parameters are directly related to the different type of atoms, not to the amino acid type.

Solvent-accessible area of completely unfolded proteins

Figure 1 represents a plot of the the solvent-accessible areas of all atoms, SA_t^u, and of the hydrophobic atoms, SA_h^u, for all

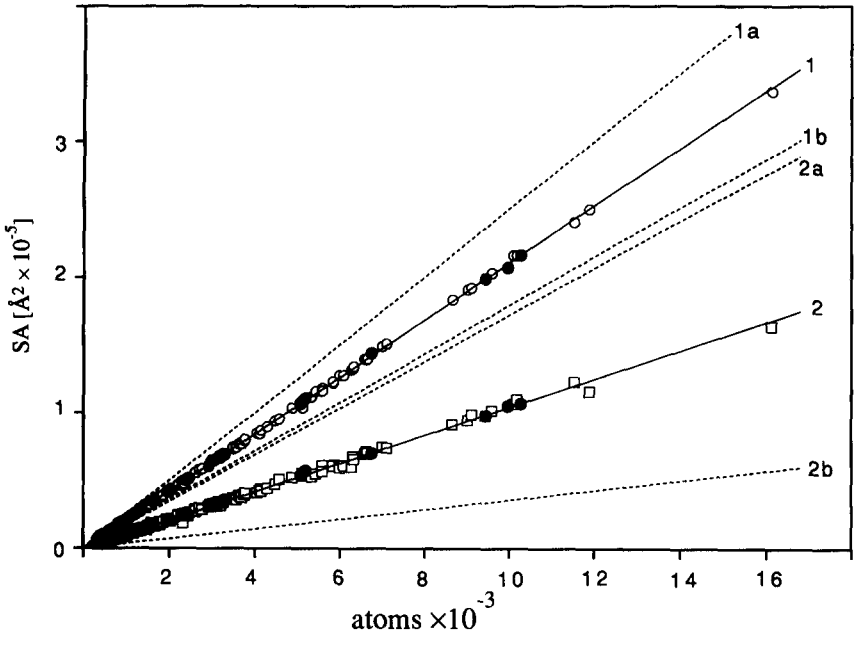

Fig. 1. Solvent-accessible area versus the number of atoms, N_t, of completely unfolded proteins. ○, SA_t^u, total solvent-accessible area of the individual proteins. □, SA_h^u, hydrophobic solvent-accessible area of the individual proteins; ●, SA_t^u and SA_h^u for thermostable proteins. Line 1, linear fit of SA_t^u, the flanking lines 1a and 1b correspond to SA_t^u for the extreme cases of poly-Met ($\alpha_t^{Met} = 25.0$ Å^2/atom) and poly-Trp ($\alpha_t^{Trp} = 17.9$ Å^2/atom), respectively. Line 2, linear fit of SA_h^u, lines 2a and 2b correspond to SA_h^u for poly-Ile ($\alpha_h^{Ile} = 17.2$ Å^2/atom) and poly-Asp ($\alpha_h^{Asp} = 3.5$ Å^2/atom), respectively.

proteins in the representative data set (Table 3) in the reference state versus N_t. It can be seen that both SA_t^u and SA_h^u have a strong linear dependence on N_t:

$$SA_t^u(N_t) = \alpha_t^u \cdot N_t, \qquad \alpha_t^u = 21.05 \text{ Å}^2/\text{atom}$$

$$SA_h^u(N_t) = \alpha_h^u \cdot N_t, \qquad \alpha_h^u = 10.45 \text{ Å}^2/\text{atom},$$

with correlation coefficients $r_t = 0.999$ and $r_h = 0.998$, respectively. It is clear that the solvent-accessible area of the polar atoms, $SA_p^u(N_t)$, is also linear, because $SA_p^u = SA_t^u - SA_h^u$. These linear dependencies were first noticed by Chothia (1976, 1975) for an essentially smaller data set. The values $\alpha_t^u = 1.44M$ ($M \approx 14$ converts α from Å^2/atom to Å^2/molecular mass as given in the papers quoted below) obtained by Chothia (1975) and $\alpha_t^u = 1.48M$ obtained by Miller et al. (1987b) are very close to those obtained in this work ($\alpha_t^u \approx 1.5M$). This good agreement between the results obtained on the basis of very different data sets shows that the linear relationship found by Chothia is a fundamental characteristic of proteins. Furthermore, the slopes α_t^u and α_h^u are approximately equal to the solvent-accessible area per atom averaged over all 20 amino acids in model tripeptides, Ala-X-Ala: $\langle sa_t \rangle_x = 20.96$ Å^2/atom and $\langle sa_h \rangle_x = 10.16$ Å^2/atom. Some mean values, sa_t^x, calculated for the individual amino acid, X, in the model tripeptides, differ significantly from the average value $\langle sa_t \rangle_x$ (see also Lee & Richards, 1971). The aromatic residues and prolines are characterized by a lower solvent accessibility per atom, whereas the linear side chains tend to have higher sa_t^x values. The range of possible values of SA_t^x is determined by lines 1a (poly-Met) and 1b (poly-Trp) in Figure 1. In the case of the hydrophobic solvent-accessible surface, the limiting lines of poly-Ile and poly-Asp (2a and 2b in Figure 1, respectively) cover a rather large range; however, the deviation from the average (line 2 in Fig. 1) is negligibly small. This result shows clearly that the values of SA_h^u (also SA_t^u), characterizing the solvent-accessible surface area of the atoms of given type in the reference state, depend on the number of atoms (or the molecular mass) and do not depend on the amino acid composition.

The quantities sa_h^x and sa_t^x are intrinsic for a particular amino acid, X, and reflect its ability to form contacts with the solvent, as well as to participate in intermolecular contacts. Thus, one could expect that the predominance or the reduction of the number of residues of a given type may reflect in certain properties regulating the packing of proteins. It has been presumed, for instance, that thermostable proteins are characterized by an increased hydrophobic index and by an increased ratio (Arg + Lys) to the total number of amino acids in comparison with the proteins from mesophiles from the same genus (Merkler et al., 1981). As seen from Figure 1, SA_h^u for thermostable proteins coincides very well with the average; hence the increased hydrophobicity index is not due to a massive increase of amino acids with a larger hydrophobic surface. It seems that the possibility for variation of SA_h^u and SA_t^u is not realized in the natural evolution process of creating the amino acid composition of proteins.

Solvent-accessible area of folded proteins

The solvent-accessible surface area of proteins has been analyzed in detail in a number of publications (Chothia, 1975, 1976; Miller et al., 1987a, 1987b; Janin et al., 1988; Koehl & Delarue, 1994b). The results obtained in this work on the basis of an extended set of protein structures do not differ essentially from those reported previously. The dependence of the solvent-accessible area on the molecular mass obeys the relation

$$SA_t^n = aN_t^x, \tag{1}$$

proposed by Miller et al. (1987a, 1987b) with the parameters $a = 45.7$, $x = 0.732$ for monomers and $a = 36.9$, $x = 0.773$ for oligomers (Fig. 2, curves 1 and 2, respectively). The values of these parameters are in excellent agreement with those obtained by Miller et al. (1987a, 1987b) and later by Koehl and Delarue

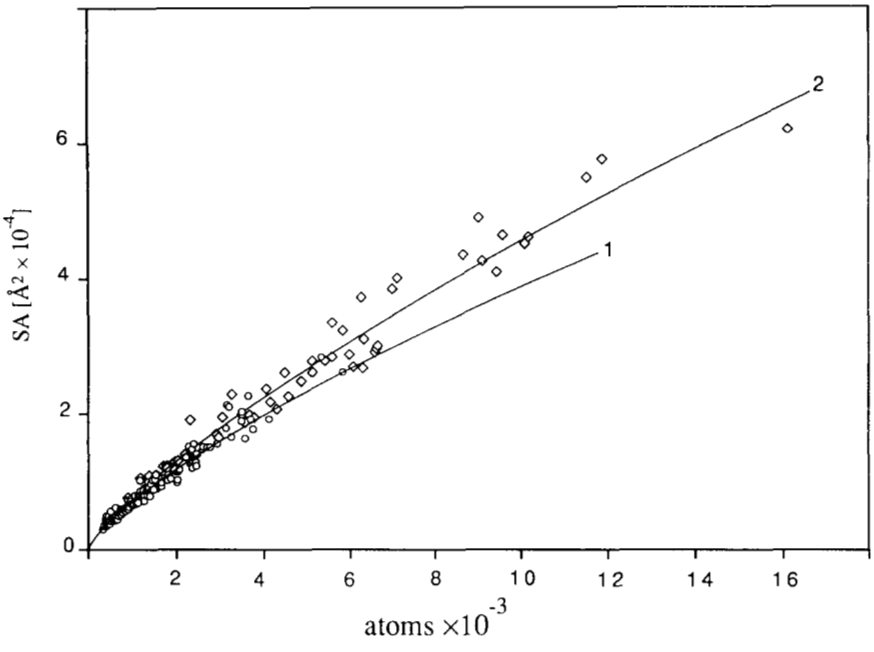

Fig. 2. Solvent-accessible area versus the number of atoms of folded proteins: ○, monomers; ◇, oligomers. Curves 1 (correlation coefficient, $r = 0.981$) and 2 ($r = 0.985$) represent the best fit of Equation 1 for monomers and oligomers, respectively. Regression curve obtained for the small proteins (see text) coincides with curve 1 in the region $N_t < 2{,}000$. Regression curve for proteins with $N_t > 2{,}000$ coincides with curve 2 in the region $2{,}000 < N_t < 8{,}000$.

(1994b). However, the average deviation from the fit obtained here is 8% and in some cases approaches 30%. These values are essentially larger than those reported by Miller et al. (1987a, 1987b): 4% and 5% for monomers and oligomers, respectively, with a maximum deviation of about 12%. Obviously, this discrepancy originates from the different data sets: a three times larger data set was used in this work. The larger deviations obtained here suggest that $\boldsymbol{SA}_t^n$ does not depend only on the molecular mass of the proteins as had been concluded on the basis of a smaller set of proteins (Miller et al., 1987b). It probably reflects the differences in the amino acid sequence and in the packing of the individual proteins.

The distribution of the monomeric and oligomeric proteins in the molecular mass scale is not uniform: the proteins with lower molecular mass are predominantly monomers, whereas the large proteins are more frequently oligomers. Thus, the classification monomers/oligomers can formally be replaced by small/large proteins. Although the definition of small and large proteins is quite arbitrary, we fitted the data to Equation 1 using the classifications: "small" proteins (N_t less than a certain critical number, N_c) and "large" proteins ($N_t > N_c$). It was found that the regression curves obtained for the subset of the small and large proteins when $N_c \approx 2{,}000$ coincide very well with those for monomers and oligomers, respectively. These curves are not shown due to essential overlapping with the curves 1 and 2 in Figure 2. It is notable that for the proteins with $N_t < 2{,}000$ the parameter x (Equation 1) is smaller than that for the large proteins. This tendency maintains when N_c increases. Thus, for example, the nonlinear fit of the data for proteins with $N_t > 6{,}000$ (only oligomers), gives $\boldsymbol{SA}_t^n = 16.3N_t^{0.863}$. Very similar results have been obtained for a set of 30 mainly large proteins (Chan et al., 1995). As far as x reflects the compactness of the packing, it shows that the proteins of different size are packed by a different way. The small proteins, which are predominantly monomers, are characterized by a parameter x close to that corresponding to the most compact arrangement, $x = 2/3$. The deviation of x from this value when N_c is shifted to larger N_t shows that the larger proteins are less compact. This can be related to formation of domains, which are weakly bound to each other, as well as to subunit aggregation and formation of oligomers. It follows that formation of single-domain proteins, characterized by a relatively compact packing, is preferred to a certain molecular mass; according to the evaluation made here, it is $N_t \approx 2{,}000$ (≈28 kDa). With a further increase in the molecular mass, formation of domains or oligomers is preferred rather than a continuous enlargement of a single domain.

The values of $\boldsymbol{SA}_t^n$ for the majority of proteins with $N_t > 8{,}000$ are above the regression lines for oligomers and for both proteins with $N_t > 2{,}000$ and $N_t > 6{,}000$. It seems then that Equation 1 is a rather rough approximation and does not describe the mass-dependence of $\boldsymbol{SA}_t^n$ with sufficient accuracy. However, the lack of structural data for proteins with $N_t > 8{,}000$ does not allow any better assumption to be made.

Optimization of protein–solvent interactions

As it was mentioned in the introduction, the optimization of a given type of interaction can be used as a measure of the significance of such interactions with respect to the variety of protein properties in the context of the other noncovalent interactions. Applied to a set of nonhomologous proteins, optimization can also reveal some common principles or tendencies that are responsible for these properties. In this paper, we report our results of the analysis of the optimization of protein–solvent interactions. We distinguish two types of protein–solvent interactions in this study: hydrophobic interactions, with an optimization parameter ξ_h, and hydrophilic interactions, with an optimization parameter ξ_p. To a certain degree, the optimization parameters introduced here are similar to the partitioning of the total accessible surface area into apolar, polar, and charged atoms presented in the recent paper of Koehl and

Delarue (1994b). In the following discussion, however, we will analyze these quantities in a different aspect.

The values of ξ_h and ξ_p versus the number of atoms, N_t, of the proteins in the representative set are plotted in Figure 3. The two parameters decrease when N_t increases. This dependence can be described by the relations:

$$\xi_h(N_t) = 117.3/N_t + 0.202, \quad (2A)$$

$$\xi_p(N_t) = 83.9/N_t + 0.261, \quad (2B)$$

with correlation coefficients $r = 0.835$ and $r = 0.834$, respectively. It is seen from the above relations that the optimization of the two types of interactions has a compensatory character: the efficiency of the hydrophobic interactions increases in a way similar to decreasing the hydrophilic interactions. However, the curves $\xi_h(N_t)$ and $\xi_p(N_t)$ cross at N_t about 600, which is evidence that for larger N_t the hydrophobic interactions are more effectively optimized than the hydrophilic interactions are deoptimized. For proteins with $N_t > 1{,}000$, this difference reaches 30%. This effect is due to the fact that the two forces expressed by the parameters ξ_h and ξ_p are opposite but not symmetric. As noted in Method of calculation, one structure (the reference state, $\xi_p = \xi_h = 1$) corresponds both to the most optimized hydrophilic and most deoptimized hydrophobic interactions, whereas different structures correspond to the most deoptimized hydrophilic ($\xi_p = 0$) and most optimized hydrophobic ($\xi_h = 0$) interactions. The predominance of the optimization of hydro-

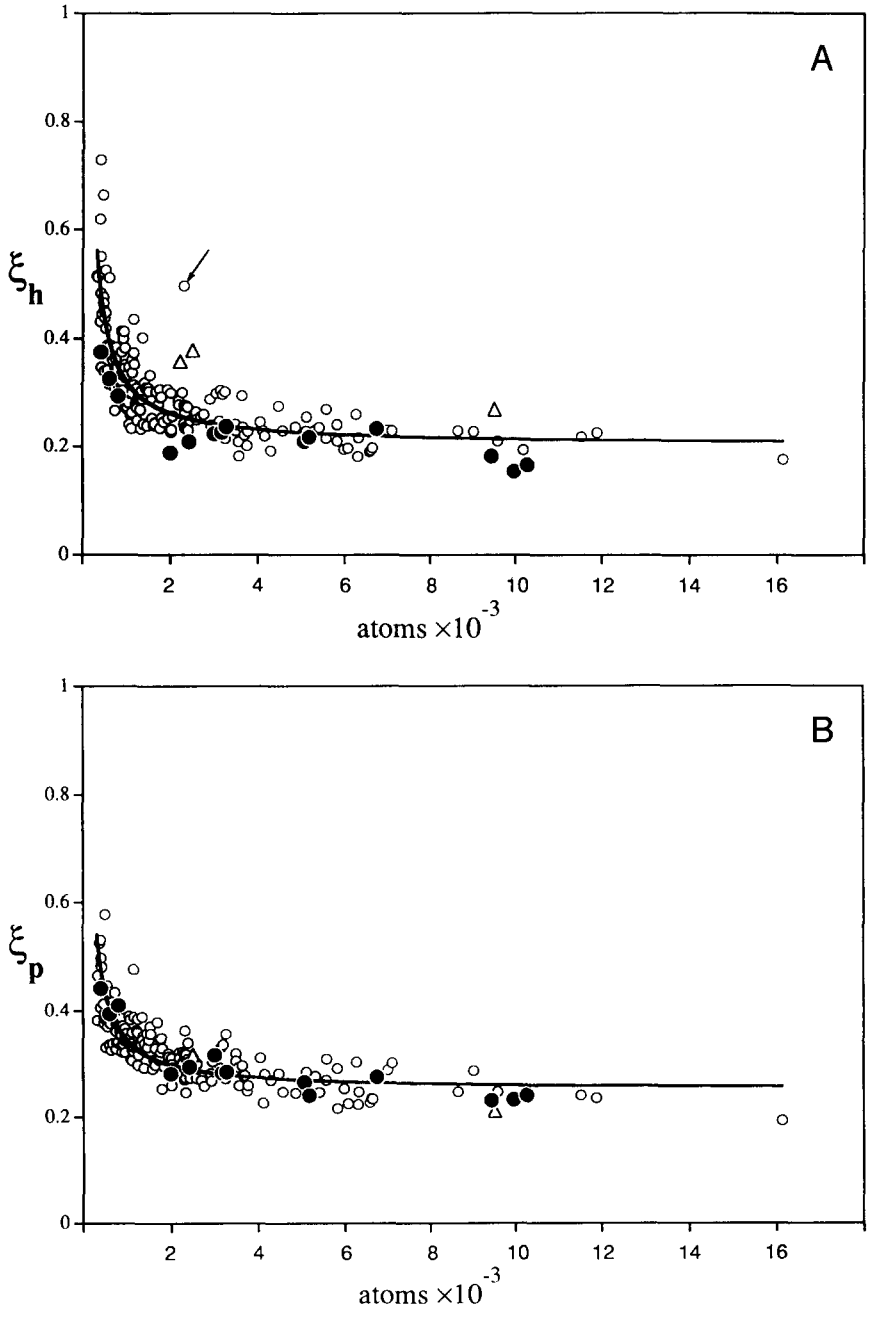

Fig. 3. Optimization parameters versus nonhydrogen atoms in proteins: **A:** ξ_h. **B:** ξ_p. Values for the representative data set are presented by open circles. Data for thermostable proteins analyzed in this study (see Tables 2, 3) are given in filled circles. In addition to the representative data set, three membrane proteins, porin (2POR), photosynthetic reaction center (1PRC), and bacteriochlorophyll A protein (3BCL) are given (△).

phobic interactions over the hydrophilic is another evidence that the former are a driving force in the process of protein folding.

The optimization parameters calculated separately for main-chain polar atoms, ξ_{mc}, side-chain polar atoms, ξ_{sc}, and charged atoms, ξ_{ch}, are shown in Figure 4A, 4B, and 4C, respectively. The side-chain polar and charged atoms are characterized by a relatively high optimization and a significant dispersion of the optimization parameters. Conversely, the main-chain polar atoms exhibit a very high deoptimization. It follows that the main contribution to the deoptimization of the hydrophilic interactions comes from the main-chain polar atoms. Obviously, their reduced accessibility to the solvent, pointed out also by Koehl and Delarue (1994b), is due to the hydrogen bond network of the protein secondary structure. Thus, in spite of the almost symmetric relationship of $\xi_h(N_t)$ and $\xi_p(N_t)$, the deoptimization of the hydrophilic protein-solvent interactions is compensated rather by the hydrogen bonding in the interior of the protein molecule than by the opposite hydrophobic interactions.

The dependence of $\xi_h(N_t)$ diminishes when $N_t > 1{,}000$ and reaches asymptotically a certain value (see Equation 2). This value, as well as the asymptote of the hydrophilic interactions is close to the average fractional apolar and polar areas calculated by Koehl and Delarue (1994b), but essentially higher than the apolar fractional area given by Lesser and Rose (1990). The optimization of the hydrophobic interactions dramatically decreases when N_t falls below 1,000. This can be related to an "insufficiency of material" and an unfavorable surface-to-volume relation for the formation of a hydrophobic core. Taking into account that $\xi_h(N_t)$ and $\xi_p(N_t)$ cross close to this point, one can conclude that the region $N_t \approx 1{,}000$ corresponds to the critical mass necessary for formation of the hydrophobic core of proteins. Once the hydrophobic core is formed, no significant increase of the optimization of the hydrophobic interactions takes place with increasing molecular mass. Interestingly enough, $N_t \approx 1{,}000$ coincides with the lowest limit of the molecular mass of enzymes. The smallest enzyme in the representative data set, acylphosphatase (PDB code 1ASP), has 775 non-hydrogen atoms—also very close to the $\xi_h(N_t)/\xi_p(N_t)$ cross point. It follows that the enzymes in general are characterized by a well-developed hydrophobic core and hence by a well-optimized structure with respect to the hydrophobic interactions. The enzymes also show a higher optimization of the charge-charge interactions (Spassov et al., 1994) and seem to exhibit an enhanced requirement for optimization of the factors stabilizing the native structure.

In Figure 5 the parameter ξ_h for the largest monomeric protein in the data set, aconitase (PDB code 6ACN), is presented as a function of the chain length. The partial values of ξ_h were calculated by consecutive addition of amino acids according to the sequence and the three-dimensional data, beginning with the N-terminal amino acid. According to this, the value of ξ_h for the first amino acid (the N-terminus) is very close to that one for the reference state whereas ξ_h reaches the value for the complete protein after addition of the last amino acid (the C-terminus). The pattern of ξ_h follows the dependence found for all proteins in the data set. As is seen from Figure 5, the distances between the minima of the curve are between 1,000 and 2,000 atoms, and they coincide exactly with the domain formation in this protein. Moreover, the values of ξ_h at the minima are very close to the curve defined by Equation 2A, showing that the individual domains are characterized by an optimization of

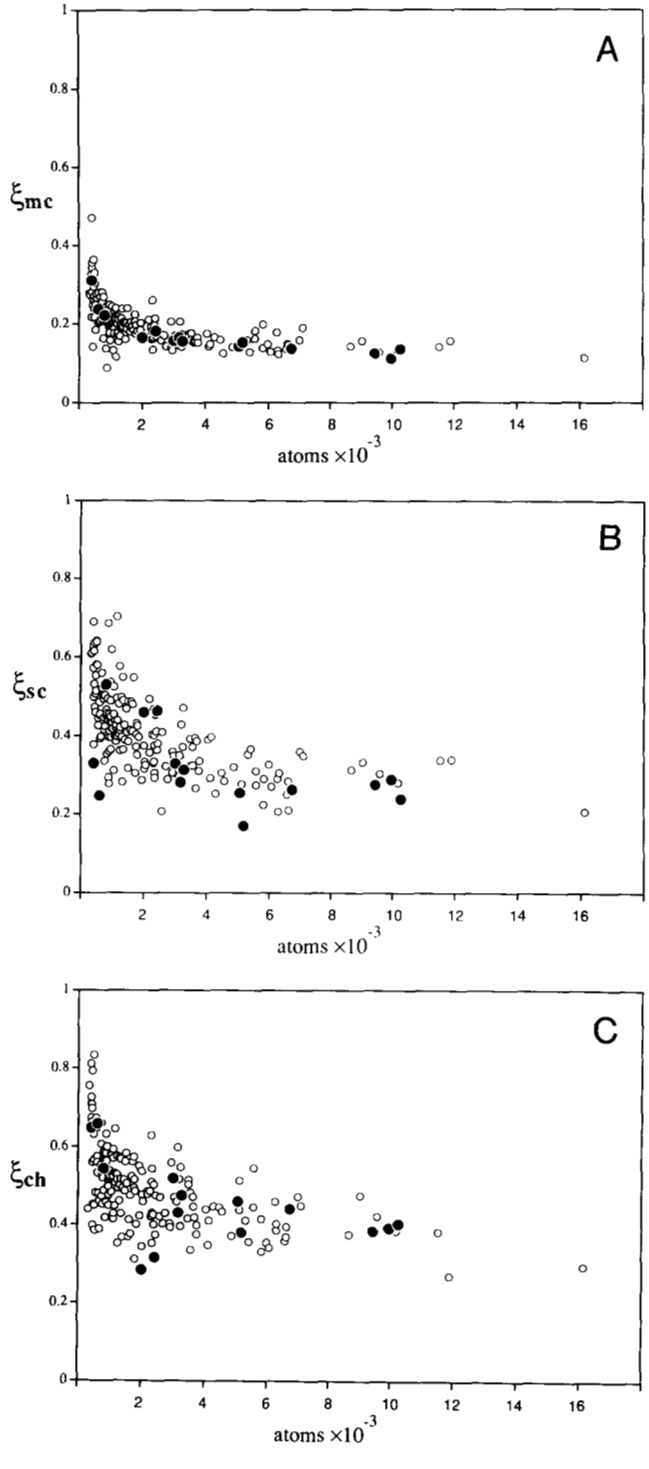

Fig. 4. Optimization parameters of hydrophilic interactions calculated for different types of polar and charged atoms. **A:** Main-chain polar atoms (ξ_{mc}). **B:** Side-chain polar atoms (ξ_{sc}). **C:** Charged atoms (ξ_{ch}). Values obtained for the proteins from the representative data set are given with open circles. Values calculated for the thermostable proteins (see Tables 2, 3) are given in filled circles.

hydrophobic interactions that is typical for a single protein. These results suggest that the average size of a domain is between 1,000 and 2,000 atoms. Thus, one can distinguish two critical masses: the first one ($N_t \approx 1{,}000$) needed for formation of a hy-

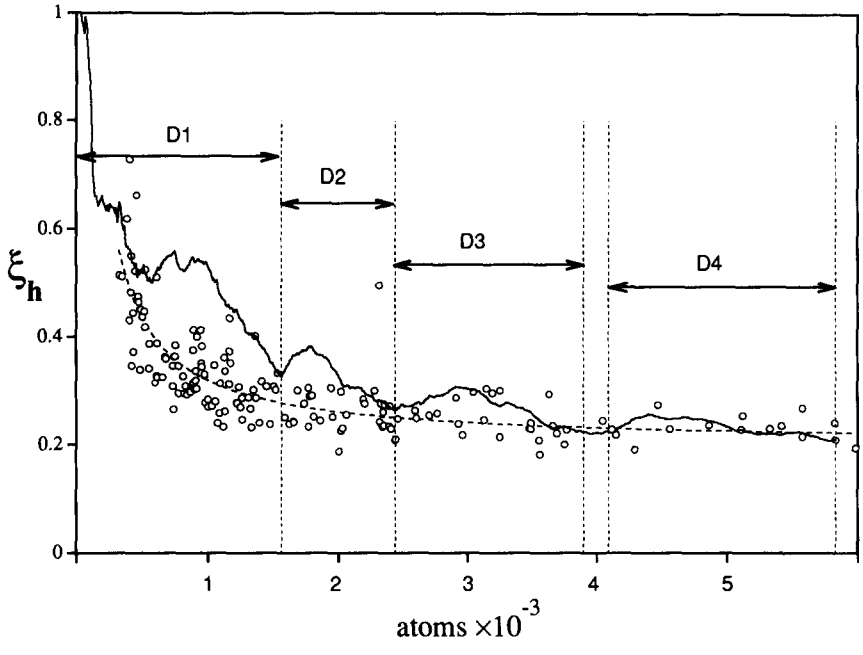

Fig. 5. Optimization parameter ξ_h for aconitase (6ACN) calculated for different chain lengths of the molecule, beginning from the N-terminal amino acid (see text). Regions of the four domains D1 (residues 1–201), D2 (202–319), D3 (320–512), and D4 (537–574) are given according to the description given in PDB; see also Robbins and Stout (1989). For comparison, data for the proteins from the representative set are also shown as in Figure 3A.

drophobic core; and the second one ($N_t \approx 2{,}000$), which determines the upper limit for the mass of a domain. This is in accord with the conclusion made above on the basis of the analysis of Equation 1. On the basis of a thermodynamic analysis, Privalov (1989) has shown that a domain must include at least 50 amino acids to be stable at room temperatures and that it usually does not have more than 200 amino acids. These two values are very close to the $\xi_h(N_t)/\xi_p(N_t)$ cross point and $N_t \approx 2{,}000$ obtained here.

The calculations performed for the subunits of oligomeric proteins showed that the individual subunits are less optimized than the corresponding oligomers. As a rule, the subunit aggregation is accompanied by an increase in the optimization of the hydrophobic interactions. For example, $\xi_h = 0.313$ when calculated for the separate subunit of tetrameric beef liver catalase (PDB code 8CAT), whereas a significantly lower value ($\xi_h = 0.175$) is obtained for the tetramer. Thus, the separated subunits differ from the monomeric proteins with regard to the optimization of the hydrophobic interactions, whereas there is no systematic difference between the optimization parameters for monomers and oligomers. This is more evidence that variation of the parameters in Equation 1 is rather due to a mass effect than due to differences between monomers and oligomers.

One exception from the general tendency of $\xi_h(N_t)$ was obtained: the structure of wheat germ agglutinin (9WGA, in Fig. 3A pointed by an arrow) is characterized by extremely low optimization of the hydrophobic interactions. The optimization of the charge–charge interactions of this protein has been found to be extremely low too (Spassov et al., 1994). This deviation from the common tendency can be explained by the presence of 16 disulfide bridges, which reduce the necessity of optimizing the noncovalent interactions.

Sensitivity of the optimization parameters

The sensitivity of the optimization parameters was analyzed by comparing point-mutated structure, as well as using artificial substitution of a given type of amino acid. The effect of point mutations was estimated for 20 mutant structures of lysozyme T4 (PDB files 1L01 to 1L20). The values of ξ_h and ξ_p calculated for the mutants differ from the wild-type protein (PDB file 2LZM) by about 0.3%. This is an expected result because the optimization parameter is a ratio between two large numbers. Detectable changes in the optimization parameters can take place after a massive replacement of amino acids. Thus, for example, the substitution of all 14 isoleucines to alanine or all 22 valines to threonine in the structure of thermitase (1THM) reduces the optimization of the hydrophobic interactions by 12% and 4%, respectively. Also, the substitution of all 13 leucines to alanine in proteinase K (2PRK) changes ξ_h by 10%. These examples demonstrate that the optimization parameters are insensitive to point mutations; only essential changes in the amino acid composition and sequence can be detected. As an illustration, the optimization parameters of three membrane-bound proteins are plotted in Figures 3 and 4. The optimization of the hydrophilic interactions for these proteins follows the general trend found for the proteins from the representative data set. However, they show an essential increase of ξ_h, which demonstrates that the parameters defined here are sensitive to structural peculiarities of proteins, which correspond to a specific environment.

Proteins from thermophilic organisms

The proteins from thermophilic organisms are active at temperatures where the proteins from mesophiles usually denature. The analysis of the molecular basis of this phenomenon is a possible way for approaching the most general problem of protein stability. It is clear that the environmental factors, such as temperature, force an adaptation of the structure. However, the rules governing this response remain unclear. It was widely demonstrated that small changes in the structure may reflect a significant shift in the thermal stability of proteins (Perutz & Raidt, 1975; Eijsink et al., 1992; Goward et al., 1994). The analysis of

these changes showed that both electrostatic and hydrophobic interactions are involved in the increase in the thermal stability of proteins. Indeed, differences in a few salt bridges, hydrogen bonds, or apolar contacts occur throughout the structures of homologous proteins; however, they do not necessarily point to a functionally relevant increase in the thermal stability. It follows that the changes leading to an increase in the thermal stability probably obey some common principles. These speculations are supported by a number of comparative studies, where structural trends, which are characteristic for the thermostable proteins, have been revealed (Zuber, 1981; Adams, 1993; Andreotti et al., 1994). In the following discussion, we will analyze the optimization parameters for the protein-solvent as well as for the charge-charge interactions in light of the stability of proteins from thermophilic organisms.

It was shown above that the total solvent accessibility for the reference state does not depend on the amino acid composition, but only on the molecular mass. The value of solvent accessibility for the native protein, however, depends on the amino acid composition via the folded structure "encoded" in the sequence. As far as the optimization parameters ξ_h and ξ_p for the individual protein are defined as the ratio between these two quantities, they depend on the molecular mass (N_t) and on the three-dimensional structure (defined by the atomic coordinates, R^n). If one assumes that the N_t dependence of the two optimization parameters follows Equation 2, the deviation of a given value from Equation 2 must include the influence of the three-dimensional structure and sequence of the individual protein. We have shown that in the case of the membrane proteins this difference reflects the peculiarities of the protein structure corresponding to their specific environment. Thus, if thermal stability is reflected by the protein-solvent interactions, the corresponding optimization parameters are expected to differ significantly from Equation 2.

We found 14 crystal structures of proteins from thermophiles in PDB that have a complete set of atomic coordinates (see Table 2). The optimization parameters calculated for these proteins are represented in Figure 3. It can be seen from Figure 3A that ξ_h for the majority of thermostable proteins falls under the curve defined by Equation 2 and forms the bottom limit of the distribution determined by the proteins from the representative set. Such a tendency is not observed for ξ_p: the points corresponding to ξ_p for the thermostable proteins are distributed around the regression line similarly to those of the mesophilic proteins (Fig. 3A). No peculiarities were found for ξ_{mc}, ξ_{sc}, and ξ_{ch} calculated for the set of thermostable proteins (Fig. 4). As far as ξ_{mc} reflects the secondary structure of proteins (see the discussion above, as well as Koehl & Delarue [1994a]), we can conclude that no unusual secondary structure elements or arrangements may be expected for thermostable proteins. These results suggest that, in terms of the optimization of protein-solvent interactions, the main contributor to the enhanced thermal stability is the increased optimization of the hydrophobic interactions. In other words, a tendency is observed for the thermostable proteins of reduction of the hydration of apolar groups, a factor destabilizing the native structure.

As was mentioned above, the deviation of ξ_h from Equation 2 reflects the specificity of the amino acid composition and three-dimensional structure of proteins. A quantitative estimation of this deviation for the individual proteins can be given by the ratio $\delta\xi_h = [\xi_{h,I} - \xi_h(N_t)]/\xi_h(N_t)$, where $\xi_{h,I}$ is the optimization parameter calculated for the individual thermostable proteins. The values of $\delta\xi_h$ are listed in Table 1. The evaluation of the sensitivity of the optimization parameters showed that a deviation of ξ_h from Equation 2 of about 10% can be associated with changes in the amino acid composition and sequence leading to a significant change in the protein-solvent interactions. Therefore, proteins with $\delta\xi_h \leq 0.1$ are assumed to be substantially optimized. As seen from Table 1, a considerable number of the thermostable proteins are characterized by this feature.

In the third column of Table 1 the optimization parameter, $S_{opt} = (\Delta G_{ei,ntv} - \langle \Delta G_{ei,rnd} \rangle)/\sigma$, for charge-charge interactions introduced by us previously (Spassov et al., 1994) is presented. This parameter represents the degree of deviation of the electrostatic interaction energy, $\Delta G_{ei,ntv}$, of the native structure from the mean electrostatic energy, $\langle \Delta G_{ei,rnd} \rangle$, of randomly distributed charges on the protein surface with standard deviation σ. The mean value of S_{opt} is about -2 and values of $S_{opt} \leq 3$ correspond to high optimization of charge-charge interactions (Spassov et al., 1994).

The comparison of $\delta\xi_h$ and S_{opt} for thermostable proteins (see Table 1) shows complementarity of the two optimization parameters. Proteins with a high $\delta\xi_h$ are characterized by a moderate optimization of the charge-charge interactions and vice versa; proteins with moderate $\delta\xi_h$ have $S_{opt} < 3$. Interestingly, there are no proteins with high optimization of both hydrophobic and charge-charge interactions. On the basis of these results, one can conclude that a common principle governing the enhanced thermal stability of proteins of thermophiles is the considerable increase in either the optimization of the hydrophobic

Table 1. *Comparison of the hydrophobic and electrostatic optimization in proteins from thermophiles*[a]

PDB code	$\delta\xi_h$	S_{opt}	Hydrophobic	Electrostatic
1LDN	−0.28	−2.50	O	M
1THM	−0.28	−2.88	O	M
1GD1	−0.23	−2.39	O	M
1CAA	−0.23	−1.66	O	M
2FXB	−0.17	−2.91	O	M
3TLN[b]	−0.17	–	O	–
1RIS	−0.16	−2.44	O	M
3PFK	−0.15	−0.58	O	L
1BMD[c]	−0.07	−2.09	M	M
1PHP	−0.07	−4.39	M	O
1EFT	−0.05	−3.62	M	O
1IPD	−0.03	−4.56	M	O
3MDS[c]	−0.01	−2.02	M	M
1SET	+0.06	−3.78	M	O

[a] In the last two columns an evaluation of the optimization parameters is given as follows: O, high optimization: $\delta\xi_h < -0.1$, $S_{opt} < -3.0$; M, moderate optimization: $-0.1 < \delta\xi_h < 0.1$; $-3.0 < S_{opt} < -1.0$; L, low optimization: $\delta\xi_h > 0.1$; $S_{opt} > -1.0$.

[b] The electrostatic interactions for this protein are strongly influenced by specific calcium binding, therefore S_{opt} was not calculated (Spassov et al., 1994).

[c] Superoxide dismutase and malate dehydrogenase are characterized by moderate optimization of both hydrophobic and electrostatic interactions, i.e., they cannot formally be distinguished from the nonthermostable proteins. However, among the homologous proteins from mesophilic organisms, they are the best optimized structures (see Table 2).

interactions (the reduction of the hydration) or the optimization of the charge-charge interactions.

The principle stated above is illustrated in Table 2, where the optimization parameters ξ_h and S_{opt} for the thermostable proteins are compared with those for functionally homologous proteins from mesophilic organisms. It is seen that among the homologous proteins, those from thermophiles are characterized by an increased optimization of at least one of the two criteria (the corresponding values are given in bold in Table 2). For rubredoxin and ferredoxin, the optimization of the electrostatic interactions is predominant. This can be related to the fact that these proteins are small, with a molecular mass below the critical value $N_t \approx 1{,}000$, needed for formation of a hydrophobic core of sufficient size. The electrostatic interactions are relevant for the thermostability of the larger proteins, as well. As seen from Table 2, the number of salt bridges for the majority of thermostable proteins is larger than the statistically expected number.

Method of calculation

Optimization criteria for protein-solvent interactions

Our theoretical approach is based on the approximation introduced by Eisenberg and McLachlan (1986) in which the solvation energy is represented as the sum of the contributions of each protein atom. The individual contributions are calculated as a product of the corresponding solvent accessibility and the atomic solvation parameter. The solvation free energy term, ΔG_s^n, of a native protein can then be expressed by the equation:

$$\Delta G_s^n = \sum_i \Delta f_\tau sa_i(R^n)$$

where the sum is taken over all protein non-hydrogen atoms. The quantity Δf_τ is the atomic solvation parameter for atoms of type τ and has the dimension [energy/Å^2]. It is positive for the apolar atoms and negative for the polar atoms. The solvent accessibility, $sa_i(R^n)$, depends on the coordinates, R^n, of the individual atom, i, and thus reflects the conformation of a folded protein. It depends also on τ via the corresponding atomic radius. The equation above can also be applied to a completely unfolded protein:

$$\Delta G_s^u = \sum_i \Delta f_\tau \langle sa_i(R^u)\rangle$$

where $\langle sa_i(R^u)\rangle$ is the solvent-accessible area averaged over all possible random coil conformations. Here, R^u corresponds to the coordinates of atom i in a given random coil conformation. The average can be estimated on the basis of the assumption that the most populated random coil conformations are characterized by a maximum solvent accessibility of all atoms. In this way, $\langle sa_i(R^u)\rangle$ can be replaced by the atomic solvent accessibility, sa_i^u, calculated for model peptides Ala-X-Ala in extended conformation, where X is the amino acid the atom i belongs to. The contribution of the protein-solvent interactions to the free energy of folding, $\Delta G_s^{(u-n)}$, can be written as

$$\Delta G_s^{(u-n)} = \sum_i \Delta f_\tau [sa_i(R^n) - \langle sa_i(R^u)\rangle]$$

$$= \sum_i \Delta f_\tau [sa_i(R^n) - sa_i^u].$$

For our further consideration, it is more convenient to rewrite the above equation as a sum over different atom types τ:

$$\Delta G_s^{(u-n)} = \sum_\tau \Delta f_\tau \sum_i^{N_\tau} [sa_i(R^n) - sa_i^u]$$

$$= \sum_\tau \Delta f_\tau \boldsymbol{SA}_\tau^u (\xi_\tau - 1)$$

where N_τ is the set of the atoms of type τ and

$$\xi_\tau = \frac{\sum sa_i(R^n)}{\sum sa_i^u} = \frac{\boldsymbol{SA}_\tau^n}{\boldsymbol{SA}_\tau^u}. \quad (3)$$

We define the parameter ξ_τ as the optimization criterion for protein-solvent interactions. The nominator on the right side of Equation 3, $\boldsymbol{SA}_\tau^n$, represents the solvent-accessible surface of the native protein formed by the atoms of type τ. The denominator, $\boldsymbol{SA}_\tau^u$, is the sum of the solvent accessibilities of all atoms of type τ in the unfolded protein. According to the approximation used here, this sum does not depend on the atomic coordinates and appears to be an intrinsic characteristic of all proteins. It was shown in the Results and discussion that it depends on the number of atoms only. It is convenient, therefore, to define the completely unfolded protein as the reference state. For the reference state $\xi_\tau = 1$ for all τs and their contribution to $\Delta G_s^{(u-n)}$ is zero. For the extreme case, where the solvent-accessible surface of the atoms of type τ is zero, ξ_τ is zero, too. In this study, we consider two basic types of atoms: hydrophobic atoms (optimization parameter ξ_h), characterized by $\Delta f_\tau > 0$; all other polar atoms with $\Delta f_\tau < 0$ are called hydrophilic (optimization parameter ξ_p). For the hydrophobic atoms, $\xi_h = 0$ corresponds to the most optimal, *but not to the native* protein structure. This is a micelle-like structure, where all hydrophobic atoms are isolated from the solvent by the shell of the hydrophilic atoms. In terms of Privalov's definition of hydrophobic interactions (Makhatadze & Privalov, 1993; Privalov & Makhatadze, 1993), the most optimized structure corresponds to a minimized hydration of the apolar atoms, which is a factor destabilizing the native structure, and maximized van der Waals interactions between these atoms. For the hydrophilic atoms, the most optimized structure is characterized by a maximum solvent-accessible area $\boldsymbol{SA}_p^n = \boldsymbol{SA}_p^u$, i.e., at $\xi_p = 1$, which in our model is the completely unfolded state. Each other protein structure, including the native one, is then less optimized. Because the completely unfolded state corresponds also to $\xi_h = 1$, it follows that the two forces are opposite and that the hydrophilic interactions, as defined in this model, are not a driving force in protein folding. It is worth noting that, although the completely unfolded state is common for both optimization parameters, there is no common structure corresponding to the opposite case. Thus, the two forces do not complement each other and determine different landscapes in the conformational space.

The optimization parameters, ξ_p and ξ_h, as defined above, have some advantages for the analysis of the protein-solvent interactions. First, the coefficient relating the free energy changes with the exposure of hydrophobic atoms to water ranges from 16-24 cal/mol/Å^2 (Chothia, 1974; Eisenberg & McLachlan, 1986) to 47 cal/mol/Å^2 (Sharp et al., 1991). This uncertainty is avoided here. Second, the optimization parameters are easily extractable from the three-dimensional structure of a given pro-

Table 2. *Optimization parameters ξ_h and S_{opt} calculated for proteins from thermophilic organisms (underlined) and corresponding structures from mesophiles*[a]

Protein & source	PDB code	N_{at}	ξ_h	S_{opt}	N_{slt}
Rubredoxin					
Pyrococcus furiosus	1CAA	413	0.374	**−1.66**	4 (4.2)
Desulfovibrio desulforicans	6RXN	358	0.369	−0.32	0 (1.1)
Desulfovibrio vulgaris	7RXN	389	0.380	−0.21	1 (3.4)
Desulfovibrio gigas	1RDG	398	0.370	−0.78	1 (3.9)
Clostridium pasteurianum	5RXN	422	0.347	−0.50	2 (3.6)
Ferredoxin					
Bacillus thermoproteolyticus	2FXB	612	0.325	**−2.91**	5 (3.1)
Clostridium acidiurici	1FDN	380	0.372	−2.17	2 (1.0)
Desulfovibrio gigas	1FXD	430	0.323	−2.19	1 (1.5)
Spirulina platensis	3FXC	732	0.346	−1.46	3 (2.6)
Azotobacter vinelandii	5FD1	1,841	0.294	−2.42	8 (6.6)
Thermitase					
Thermoactinomyces vulgaris	1THM	2,003	**0.188**	−2.88	13 (4.2)
Subtilisins					
Carlsberg (*Bacillus licheniformis*)	1SCA	1,920	0.213	−3.27	9 (3.8)
BL (*Bacillus lentus*)	1ST3	1,888	0.201	−2.70	8 (3.6)
BPN′ (*Bacillus amyloliquefaciens*)	2ST1	1,938	0.219	−3.17	10 (4.2)
Phosphoglycerate kinase					
Bacillus stearothermophilus	1PHP	3,008	**0.223**	−4.39	32 (30.4)
Saccharomyces cerevisiae	3PGK	3,148	0.304	−1.04	8 (20.8)
Superoxide dismutase (Mn, Fe)					
Thermus thermophilus	3MDS	3,282	**0.236**	**−2.02**	10 (10.1)
Pseudomonas ovalis	3SDP	2,910	0.287	−1.89	6 (6.2)
Escherichia coli	1ISB	3,006	0.245	−1.22	4 (6.6)
Phosphofructokinase					
Bacillus stearothermophilus	3PFK	9,444	**0.182**	−0.58	15 (15.2)
Escherichia coli	2PFK	9,024	0.288	−0.88	10 (15.8)
Malate dehydrogenase					
Thermus flavus	1BMD	4,976	**0.209**	−2.09	17 (16.7)
Porcine	4MDH	5,106	0.228	−2.57	15 (17.7)
Lactate dehydrogenase					
Bacillus stearothermophilus	1LDN	9,792	**0.153**	−2.50	21 (15.4)
Lactobacillus casei	1LLC	9,816	0.196	−0.64	12 (14.2)
Dogfish	6LDH	10,180	0.194	−2.63	19 (14.6)
Porcine	9LDB	10,272	0.173	−2.06	13 (13.0)
D-glyceraldehyde-3-phosphate dehydrogenase					
Bacillus stearothermophilus	1GD1	110,100	**0.165**	−2.39	16 (19.5)
Lobster	1GPD	10,040	0.244	−3.13	15 (13.3)
Ribosomal protein					
Thermus thermophilus	1RIS	817	0.292	−2.44	9 (6.8)
Thermolysin					
Bacillus thermoproteolyticus	3TLN	2,432	0.208	Not estimated	
Elongation factor Tu					
Thermus aquaticus	1EFT	3,175	0.226	−3.62	36 (28.3)
3-Isopropylmalate dehydrogenase					
Thermus thermophilus	1IPD	5,180	0.217	−4.56	29 (19.5)
Seryl-TRNA synthethase					
Thermus thermophilus	1SET	6,746	0.233	3.78	36 (28.6)

[a] The values in bold indicate the optimization parameter calculated for thermostable proteins, if they are best optimized for a group of functionally homologous proteins. N_s in the last column is the number of salt bridges in the native structure; the values given in parentheses are the number of salt bridges statistically expected for that protein structure (Spassov et al., 1994).

Table 3. *PDB codes of the protein structures used in this work*

Monomeric proteins (126 entries)									
1CRN	5HIR	4TG1	9INS	3MT2	2OVO	1EPI	5RXN	1AAP	2GB1
1ROP	1PI2	1NXB	1DTX	1R69	1SN3	2CTX	2CI2	1C5A	2HIP
1HOE	1UBQ	1HO1	1CC5	2FXB	1HIP	351C	1TPK	1EG1	3FXC
3B5C	1PCY	1SAR	2MCM	1APS	8RNT	4CPV	2TRX	1FKB	4FD1
256B	1RNB	1TGI	1YCC	1CCR	1CD8	1C2R	1PAZ	2PF1	7RSA
1GMF	1BP2	2MHR	2CDV	3CHY	1LZT	3FGF	1LZ1	1CY3	1IF1
3FXN	1ECA	1F3G	2SNS	1MBA	1END	4CLN	2SNV	1LPE	2I1B
2LH4	2RN2	2SGA	1MBD	2LZM	5P21	2FCR	1CD4	1ETU	2ALP
1RBP	1GKY	1GCR	8DFR	3ADK	1SGT	9PAP	1HNE	1TON	1GCT
2FVW	1LTE	1THM	2PRK	3BLM	2CBA	1BIA	2REB	6ABP	1RHD
2CYP	2GBP	1FNR	4APE	1TRB	3APR	3TLN	5CPA	2REN	2LBP
1GOX	1ALD	1CPK	1PHH	3PGK	2BJL	2CPP	1PII	1NPX	1GLY
1PGD	2TAA	1COX	1ACE	1LFI	6ACN				
Oligomeric proteins (57 entries)									
6RLX	1MLP	1CDT	2UTG	1IL8	1FIA	2GN5	2SSI	5HVP	2WRP
2RSP	1MSB	2CCY	2AZA	1BBH	2SOD	9WGA	1SDH	3SDP	1COL
6XIA	3GAP	2PAB	1TNF	1GST	5TIM	1RVE	1AAI	2TSC	1PYP
1THB	1FBP	4MDH	3CLA	1VSG	8ADH	1BBP	1GP1	1LTS	1NSB
1HSA	7ICD	7AAT	4ENL	1CSC	5RUB	3GRS	8ATC	4CNA	2PMG
2PFK	1WSY	4GPD	6LDH	1OVA	1HGE	8CAT			

tein: one only needs to calculate the solvent accessibilities of the atoms of a given type.

Solvent-accessibility calculation

The following characteristics were calculated: (1) SA_t, the total solvent-accessible area. (2) SA_h, the hydrophobic solvent-accessible area. This area is formed by the carbon atoms only. The carbon atoms covalently bound to polar atoms are, in fact, polar; however, their solvent accessibility is rather small in both the native and the reference protein structures (Chothia, 1976). Therefore, their contribution to SA_h does not affect the results. (3) SA_p, the hydrophilic solvent-accessible area. This area is formed by the nitrogen, oxygen, and sulfur atoms.

The solvent-accessibility calculations were performed using a program based on the well-known algorithm of Lee and Richards (1971). The atomic radii used in the calculations were also taken from Lee and Richards (1971). In order to test the sensitivity of the optimization parameters on the atomic radii (Chothia, 1976), test calculations with different atomic radii were performed for 10 proteins with different molecular weight. The deviation in ξ_o and ξ_h was about 2%, which is insignificant for the results presented in this paper.

The atomic solvent accessibilities for the reference state were obtained from data on the tripeptide Ala-X-Ala in extended conformation, where X is the amino acid in the sequence of the corresponding protein. The molecular models of the tripeptide were built with the program Insight-II (Biosym Technologies, Inc., San Diego, California).

The optimization parameters, ξ_τ, were calculated for a set of 183 nonhomologous proteins. The protein entries were chosen from the new representative set (Boberg et al., 1995) of sequence-unbiased PDB structures. The PDB (version 1994) codes of proteins are listed in Table 3 in two subsets of proteins: monomeric (126 entries) and oligomeric (57 entries). Calculations of ξ_τ were performed for closely related proteins as well. These proteins are listed in Table 2.

Acknowledgments

We thank Dr. Mauno Vihinen (CSB, Novum, Karolinska Institutet) for help with construction of the representative data set. This work was partially supported by the European Community project "Biotechnology of Extremophiles" via a grant from NUTEK, Sweden.

References

Adams MWW. 1993. Enzymes and proteins from organisms that grow near above 100 °C. *Annu Rev Microbiol 47*:627–658.

Andreotti G, Tutino ML, Sannia G, Marino G, Cubellis MV. 1994. Indole-3-glycerol-phosphate synthase from *Sulfolobus solfataricus* as a model for studying thermostable TIM-barrel enzymes. *Biochim Biophys Acta 1208*:310–315.

Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science 181*:223–230.

Argos P, Rossman MG, Gran KM, Zuber H, Frank G, Tratschin JD. 1979. Thermal stability and protein structure. *Biochemistry 18*:5698–5703.

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

Boberg J, Salakoski T, Vihinen M. 1995. Representative selection of proteins based on nuclear families. *Protein Eng.* Forthcoming.

Chan MK, Mukund S, Kletzin A, Adams MWW, Rees DC. 1995. Structure of hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science 267*:1463–1469.

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

Chothia C. 1975. Structural invariants in protein folding. *Nature 254*: 304–308.

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

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

Eijsink VGH, Vriend G, van der Burg B, van der See JR, Venema G. 1992. Increasing the thermostability of a neutral protease by replacing positively charged amino acids in the N-terminal turn of α-helix. *Protein Eng 5*:165–170.

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

Goward CR, Miller J, Nicholls DJ, Irons LI, Scawen MD, O'Brien R, Chowdhry BZ. 1994. A single amino acid mutation enhances the thermal

stability of *Escherichia coli* malate dehydrogenase. *Eur J Biochem 224*: 249-255.
Hirono S, Liu Q, Moriguchi I. 1991. High correlation between hydrophobic free energy and molecular surface area characterized by electrostatic potential. *Chem Pharm Bull (Tokyo) 39*:3106-3109.
Janin J, Miller S, Chothia C. 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol 204*:155-164.
Karshikov A, Duerring M, Huber R. 1991. Role of electrostatic interactions in the stability of the hexamer of constitutive phycocyanin from *Fremyella diplosiphon*. *Protein Eng 4*:681-690.
Koehl P, Delarue M. 1994a. Application of a self-consistent mean field theory to predict protein side-chain conformation and estimate their conformational entropy. *J Mol Biol 239*:249-275.
Koehl P, Delarue M. 1994b. Polar and nonpolar atomic environments in the protein core: Implication for folding and binding. *Proteins Struct Funct Genet 20*:264-278.
Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379-400.
Lesser GJ, Rose GD. 1990. Hydrophobicity of amino acid subgroups in proteins. *Proteins Struct Funct Genet 8*:6-13.
Makhatadze GI, Privalov PL. 1993. Contribution of hydration to protein folding thermodynamics. I. The enthalpy of hydration. *J Mol Biol 232*: 639-659.
Merkler DJ, Farrington GK, Wedler FC. 1981. Protein thermostability. Correlations between calculated macroscopic parameters and growth temperature for closely related thermophilic and mesophilic bacilli. *Int J Pept Protein Res 18*:430-442.
Miller S, Janin J, Lesk AM, Chothia C. 1987a. Interior and surface of monomeric proteins. *J Mol Biol 196*:641-656.
Miller S, Lesk AM, Janin J, Chothia C. 1987b. The accessible surface area and stability of oligomeric proteins. *Nature 328*:834-836.
Perutz MF, Raidt H. 1975. Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature 255*:256-259.
Ponnuswamy PK. 1993. Hydrophobic characteristics of folded proteins. *Prog Biophys Mol Biol 59*:57-103.
Privalov PL. 1989. Thermodynamic problems of protein structure. *Annu Rev Biophys Biophys Chem 18*:47-69.
Privalov PL, Makhatadze GI. 1993. Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration. *J Mol Biol 232*:660-679.
Rehaber V, Jaenicke R. 1992. Stability and reconstitution of D-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic eubacteria *Thermotoga maritima*. *J Biol Chem 267*:10999-11006.
Robbins AH, Stout CD. 1989. The structure of aconitase. *Proteins Struct Funct Genet 5*:289-312.
Rose GD, Wolfenden R. 1993. Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu Rev Biophys Biomol Struct 22*:381-415.
Sharp KA, Nicholls A, Fine RF, Honig B. 1991. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effect. *Science 252*: 106-109.
Spassov VZ, Atanasov BP. 1994. Spatial optimisation of electrostatic interactions between the ionized groups in globular proteins. *Proteins Struct Funct Genet 19*:222-229.
Spassov VZ, Karshikoff AD, Ladenstein R. 1994. The optimization of the electrostatic interactions in proteins of different functional and folding type. *Protein Sci 3*:1556-1569.
Stellwagen E, Wilgus H. 1978. Relationship of protein thermostability to accessible surface area. *Nature 275*:342-343.
Tuñon I, Silla E, Pascual-Ahuir JL. 1992. Molecular surface area and hydrophobic interactions. *Protein Eng 5*:715-716.
Zuber H. 1981. Structure and function of thermophilic enzymes. In: Eggerer H, Huber R, eds. *Structural and functional aspects of enzyme catalysis*. Berlin/Heidelberg/New York: Springer-Verlag. pp 114-127.
Zuber H. 1988. Temperature adaptation of lactate dehydrogenase. Structural, functional and genetic aspects. *Biophys Chem 29*:171-179.