PROTEINS: Structure, Function, and Genetics 20:264–278 (1994)

Polar and Nonpolar Atomic Environments in the Protein Core: Implications for Folding and Binding

Patrice Koehl[1] and Marc Delarue[2]
[1]UPR 9003 Cancérogénèse et Mutagénèse Moléculaire et Structurale du CNRS, ESBS, 67400 Illkirch Graffenstaden, France; and [2]Laboratoire d'Immunologie Structurale, Institut Pasteur, 75015 Paris Cedex, France

ABSTRACT **Hydrophobic interactions are believed to play an important role in protein folding and stability. Semi-empirical attempts to estimate these interactions are usually based on a model of solvation, whose contribution to the stability of proteins is assumed to be proportional to the surface area buried upon folding. Here we propose an extension of this idea by defining an environment free energy that characterizes the environment of each atom of the protein, including solvent, polar or nonpolar atoms of the same protein or of another molecule that interacts with the protein. In our model, the difference of this environment free energy between the folded state and the unfolded (extended) state of a protein is shown to be proportional to the area buried by *nonpolar* atoms upon folding. General properties of this environment free energy are derived from statistical studies on a database of 82 well-refined protein structures. This free energy is shown to be able to discriminate misfolded from correct structural models, to provide an estimate of the stabilization due to oligomerization, and to predict the stability of mutants in which hydrophobic residues have been substituted by site-directed mutagenesis, provided that no large structural modifications occur.**

Key words: hydrophobic interactions, protein stability, hydrophobicity scale, protein mutant stability

INTRODUCTION

Hydrophobic interactions are believed to be important determinants of the stability of proteins, multiunit protein complexes, and protein/ligand complexes.[1,2] Unfortunately, although they are easy to observe and describe qualitatively, their quantitative evaluation remains difficult.

Hydrophobicity can be defined as the low solubility of nonpolar molecules in water. A semiquantitative estimate of the hydrophobic energy has been obtained by assuming that it is proportional to the surface area buried upon folding.[3–7] Eisenberg and McLachlan[6] extended this idea by adding the hypothesis that the solvation free energy of an amino acid is the sum of the energies related to each of its atoms. They defined the total free energy of solvation of a protein as follows:

$$G_S = \sum_i ASP_i \, ASA_i \qquad (1)$$

where the summation extends over all atoms of the protein and ASP_i and ASA_i are the atomic solvation parameter and accessible surface area of atom i, respectively. Five adjustable parameters (ASP values for five types of atoms) were shown to be sufficient to reproduce with an excellent correlation coefficient the 20 distribution coefficients of amino acids between octanol and water.[8] The solvation contribution to the free energy of protein folding is obtained as the difference of the free energy of hydration described by equation (1) between the native and a reference denatured state, usually taken to be the fully extended structure. The denatured "state" of a protein is known to be a distribution of different molecular conformations, and the extended conformation is certainly a poor model of this state.[9] However, in the usual way in which equation (1) is used, the reference state has little influence because the energies of the same protein are compared in two conformations having the same reference state. Energy calculations based on equation (1) have been successful in identifying misfolded from correctly folded structures in cases in which intramolecular energy terms alone (i.e., van der Waals and electrostatic terms) showed no alarmingly high energy for the misfolded structures.[6] Recently, a good correlation has bee observed between the destabilization of Leu → Ala mutants in T4 lysozyme and the size (surface) of the cavity created by the mutation.[10]

Abbreviations: ASA, accessible surface area; ASP, atomic solvation parameter; M.W., molecular weight; NMR, nuclear magnetic resonance; NPCA, non-polar contact area; PCA, polar contact area; TASA, total accessible surface area; VdW, Van der Waals.

Received April 14, 1994; revision accepted June 17, 1994.
Address reprint requests to Patrice Koehl, UPR 9003 Cancérogénèse et Mutagénèse Moleculaire et Structurale du CNRS, ESBS, Boulevard Sébastien Brant, 67400 Illkirch Graffenstaden, France.

This hypothesis of linear correlation of the free energy with the accessible surface area has been recently questioned[11] since it cannot reproduce correctly some hydrophobic pair interactions, especially interactions between nonpolar groups that are found at the surface of a protein, in opposition with bulk hydrophobic interactions, which involve internal clusters of nonpolar groups.

We believe that the stability of proteins does not come uniquely from shielding hydrophobic atoms from solvent atoms. Indeed, some of the nonpolar atoms are exposed to the surface of the molecule or to interior polar atoms, for instance from nonexposed polar side chains. Conversely, polar atoms are not always exposed to the solvent, or in contact with hydrophobic atoms from the core, but, as in secondary structures for instance, they sometimes tend to maintain a polar environment inside the protein.

To take these effects into account, two routes are possible. The first involves performing statistical studies on the database of protein structures to highlight biases and preferences in atom–atom contacts, as observed experimentally in high-resolution structures. This can easily be made quantitative by considering the number of standard deviations from which an event deviates from the mean value of the database.

Estimation of hydrophobic interactions using analysis of atom–atom contacts in proteins have indeed been done.[12–17] Also, preferences of amino acids to be buried or exposed as well as atomic contacts have been included in the design of the so-called 1D-3D profiles[18] that have been successfully applied to the identification of misfolded structures[19] and, with some success, to the identification of sequences that fit a known 3D structure, i.e., the reverse folding problem.[18,20] The atom–atom contacts derived by Vriend and Sander[17] have also been applied to the evaluation of misfolded structures. However, their approach involves 57 different types of atoms. They also use 80 different types of residue fragments, which are further refined according to the helical or nonhelical nature of the residue; the index they derive is then dependent on the residue accessibility. This method undoubtedly allows the inclusion of such interactions as, for instance, the stabilization of a charged nitrogen by aromatic rings stacked above and below the charge, as observed in some proteins like acetylcholine esterase and in the Cambridge Data Base of small organic molecules. However, we felt that another approach, involving fewer parameters, should be possible.

It is worth mentioning that pair potentials need not be taken between atoms, but that it can be done between residues, which also reduces the number of parameters. Of particular interest on this subject is the work of Miyazawa and Jernigan,[21] which takes solvent interactions into account, and of the group of M. Sippl.[22,23] This has proved to be another powerful tool for the reverse folding problem.[20,24,25] However, we feel that an atomic description of the amino acids is needed, because it is not correct to try to derive a hydrophobic scale for amino acids: in the side chain of lysine, for instance, both charged and hydrophobic atoms (methylene groups) are found.

A second route is to try to generalize the approach of Eisenberg and McLachlan[6] [see Eq. (1)] to include atom-atom contacts effects, even crudely: in this work, only solvent, polar, and nonpolar atoms will be considered. We thus define another free energy, which we refer to as the environment free energy, G_e, given in Equation (2):

$$G_e = \sum_i A_i(ASA_i + PCA_i) + B_i\,NPCA_i \qquad (2)$$

where the summation extends over all atoms of the protein. ASA_i, PCA_i, and $NPCA_i$ are the accessible surface area, the surface area occluded by polar atoms, and the surface area occluded by nonpolar atoms, of atom i, respectively. A_i and B_i are parameters that will be described explicitly below.

In the first part of this study, we will re-examine the interior and surface of both monomeric and multimeric proteins from a set of 82 well-refined structures that have been determined to 2.5 Å resolution or better from the Brookhaven Data Bank,[26] focusing on the occluded surface of each atom. We analyze both $NPCA_i$ and ASA_i as a function of the molecular weight of the protein and also in terms of the nature of occluding atoms (polar, nonpolar, or charged). This study confirms and extends previous work.[27–31] We then derive an expression for the contribution of the environment free energy to the stability of the protein. Parameters A and B defined in equation (2) are estimated from data of the partition of amino acids between water and octanol[8] from which only the energy of contact is considered.[32,33] A systematic study of this environment free energy is presented, including its application to detect misfolded structures, its contribution to the free energy for protein-protein and protein–ligand interactions, and its use to predict the stability of mutants in which hydrophobic residues have been changed.

MATERIALS AND METHOD

Database

Recently, a sample of 103 high-resolution structures was selected to represent nonredundant protein folds in the Brookhaven Protein Data Bank (PDB).[34] We chose 82 proteins out of these 103 structures as a database to validate our approach (only 102 were actually given in Table II of Boberg et al.[34]; the 20 discarded structures either had missing residues or missing atoms). The corresponding entries in the PDB[26] are: 156b, 1acx, 1bp2, 1ccr, 1cpv, 1ctf, 1ctx, 1eca, 1gcr, 1hoe, 1ilb, 1ldm, 1lzl, 1mbd, 1nxb, 1pcy, 1phh, 1r69, 1rhd, 1rn3, 1sn3, 1snc, 1ubq, 2app, 2b5c, 2cab, 2cdv, 2ci2, 2cpp, 2gbp, 2lbp, 2lh2,

2mhr, 2ovo, 2paz, 2rnt, 2sga, 2sns, 2taa, 3adk, 3fxc, 3fxn, 3lzm, 3pgk, 3tln, 451c, 4fdl, 4pep, 5cpa, 5pti, 5rxn, 6acn, 8adh, 8dfr, and 9pap as monomeric proteins, and 1cse, 1fc2, 1gdl, 1lrd, 1prc, 1pyp, 1tim, 1tnf, 1utg, 2aat, 2aza, 2ccy, 2cts, 2fb4, 2gn5, 2hhb, 2hla, 2pab, 2sod, 2wrp, 3grs, 3ins, 3wga, 4atc, 4hvp, 4xia, and 5api as multimeric proteins. Unless otherwise stated, multimeric proteins are considered as such, rather than as a collection of separate protomers.

Surface Areas of Proteins

The area of an atom in a protein in contact with solvent is called the solvent accessible surface area. We will refer to it as the *TASA*, or total accessible surface area, if the protein is considered unfolded, in an extended state, and as *ASA* in a folded protein. By homology, we define the polar contact surface area, or *PCA*, and the nonpolar contact area, or *NPCA*, of an atom as its areas in contact with (or occluded by) polar and nonpolar atoms, respectively. *PCA* and *NPCA* should not be confused with the polar surface area and nonpolar surface area, which commonly correspond to the accessible surface area of polar and nonpolar atoms, respectively. All surfaces mentioned above (i.e., *TASA*, *ASA*, *PCA*, and *NPCA*) were computed based on a modified version of the method developed by Shrake and Rupley.[28] Hydrogen atoms were not considered explicitly. In another set of calculations, hydrogen atoms (polar or nonpolar) were explicitly considered (data not shown); all qualitative conclusions presented below were still valid. Carbon and sulphur atoms were classified as nonpolar atoms, while nitrogen and oxygen (neutral or charged) were classified as polar atoms. For each protein atom i, evenly distributed test points were placed on the solvation sphere of radius $R_i + R_W$, centered at the atomic position, at a density of 3 points/Å^2. R_i is the van der Waals radius of atom i (vdW radii used in the following computations are: nitrogen, 1.5 Å; oxygen, 1.4 Å; sulphur, 1.85 Å; carbonyl carbon, 1.5 Å; other carbon, 2.0 Å) and R_W is the effective radius of the solvent molecule used as a probe, chosen to be equal to 1.4 Å.[7,28,30] Atoms of the molecule other than atom i were considered as test atoms, belonging to two categories. They were short-range ("near," in the terminology of Shrake and Rupley[28]) test atoms if they belonged to the same residue as i, or to the backbone of the two neighboring residues. All other atoms were long-range ("long," in the terminology of Shrake and Rupley[28]) test atoms. A test point on the solvation sphere of i was occluded if its distance to the atomic position of a test atom j was smaller than the solvated radius of j (i.e., $R_j + R_W$). First, the list of test atoms occluding a given test point k was established. If this list was empty, the fraction of the surface of atom i related to k, S_k, was accounted for as accessible to solvent. If this list contained at least one short-range test atom, S_k was attributed to the short-range surface of i. Otherwise, each atom j of the list was given a weight that depends on its distance r_{ij} to the center of test atom i. This weight is obtained from a linear square well, as proposed by Holm and Sander,[35] equal to 1 from $r_{ij} = 0$ to $r_{ij} = R_i + R_j$, and then decreasing linearly down to 0, which is reached at $r_{ij} = R_i + R_j + 2\,R_W$. The fractions of polar and nonpolar atoms occluding k were then computed based on these weights, and the test point k contributed to the polar and nonpolar contact area of i (PCA and NPCA, respectively) according to these fractions. This procedure was repeated for each test point of the solvated sphere of i, yielding the ASA, PCA, and NPCA of atom i.

The main difference in our approach compared with that of Shrake and Rupley[28] resides in the definition of the short-range surface. This surface accounts for the stereochemistry of the residue to which i belongs and is not included in the subsequent calculations; however, it affects the absolute values of the different contact areas. In their method, the area assigned to short-range atoms in calculations for a folded structure is generally less than that occluded by the same short-range atoms in the unfolded state, since they give the same weight to short- and long-range test atoms that occlude a test point.[28] In our approach, the short-range surface remains approximately constant in the same conditions, since short-range atoms are privileged. Differences can occur due to changes in the conformation of the residues upon folding; these differences were found to have little effect on the close surface (data not shown).

The *TASA*, which is the total accessible surface of atom i in the unfolded, extended state of the protein, corresponds in fact to the total surface of the solvated sphere of i less the short-range surface of i. Under the assumption that the short-range surface area of an atom is the same in the folded and unfolded state of the protein, the following relation holds upon folding:

$$TASA = ASA + PCA + NPCA. \qquad (3)$$

The reason for our definition of the different surfaces will become apparent below.

The Environment Free Energy

Eisenberg and McLachlan[6] developed an atomic description of the free energy of solvation, in which the sign and strength of the water-solvent interaction are specified by the atomic solvation parameter (*ASP*) of each atom, and by its accessible surface area [Eq. (1)]. Its contribution to the stability of the protein is obtained as the difference of free energy between the native state and the unfolded state:

$$\Delta G_S = \sum_i ASP_i(ASA_i - TASA_i). \quad (4)$$

This approach, however, does not take into account the environment of each atom of the protein upon folding. For instance, a polar atom occluded by polar atoms will still be in a favorable situation, although buried. To quantify this effect, we generalize the free energy of solvation of Eisenberg and McLachlan[6] to the notion of free energy of environment, given by Equation (2). The environment free energy of a protein in the unfolded state is given by:

$$G_e(\text{unfold}) = \sum_i A_i\, TASA_i. \quad (5)$$

Hence the contribution of the free energy of environment to the stability of a protein, which we will term ΔG_e, is given by:

$$\begin{aligned}\Delta G_e &= G_e - G_e(\text{unfold}) \\ &= \sum_i [A_i(ASA_i + PCA_i) + B_i\, NPCA_i] \\ &\quad - \sum_i A_i\, TASA_i. \end{aligned} \quad (6)$$

From Equation (3), it follows that:

$$\Delta G_e = \sum_i (B_i - A_i)\, NPCA_i. \quad (7)$$

We now show that $B_i - A_i$ can be derived from experimental data on free energies of transfer of amino acid residue analogues from a nonpolar solvent to water. The environment free energy of an amino acid X in water is:

$$G_e(X)_{\text{water}} = \sum_i A_i\, TASA_i. \quad (8)$$

The environment free energy of the same amino acid in a nonpolar solvent is:

$$G_e(X)_{\text{octanol}} = \sum_i B_i\, TASA_i. \quad (9)$$

The change in environment free energy of amino acid X upon transfer from hydrophobic to hydrophilic conditions is:

$$\Delta G_e(X) = \sum_i (A_i - B_i)\, TASA_i. \quad (10)$$

$A_i - B_i$ can be deduced by identifying this change in environment free energy to the experimentally determined free energy of transfer of an amino acid analogue of X.

Comparison of Equation (10) with Equation (3) of Eisenberg and McLachlan[6] yields:

$$A_i - B_i = \Delta\sigma(i) = ASP_i. \quad (11)$$

Then the contribution of the environment free energy to the stability of a protein can be expressed as:

$$\Delta G_e = -\sum_i ASP_i\, NPCA_i. \quad (12)$$

Hence the contribution of the environment free energy to the stability of a protein is proportional to the area buried *by nonpolar atoms* upon folding.

Equation (12) is valid under the following underlying assumptions:

1. Equation (3) is valid (see above).
2. Polar and nonpolar atoms buried in a nonpolar environment in the protein core have the same environment free energy as when they are exposed to an nonpolar solvent, such as n-octanol. This is the basic assumption that allows estimation of the atomic solvation parameters from the modified experimental data on transfer of amino acid analogues from nonpolar to polar solvent[8,32]; the same assumption was implicitly used in the formalism of Eisenberg and McLachlan.[6]
3. Polar and nonpolar atoms buried in a nonpolar environment in the protein core have the same environment free energy as when they are exposed to solvent. This is just a generalization of the preceding assumption. Qualitatively, it is based on the fact that polar atoms are more favorably buried in a polar environment, while the environment of nonpolar atoms is preferentially hydrophobic. Precise quantification of these effects presents some problems, such as:

   a. Hydrogen-bonded pairs have a desolvation penalty if buried.[36–38]

   b. Polar-polar contacts may not be of the correct type or in the correct orientation for an hydrogen bond.

   c. There is no clear experimental evidence that a nonpolar atom in contact with polar atoms is in an unfavorable state compared with a situation in which it only sees nonpolar atoms. However, this latter hypothesis makes physical sense; it could be ascribed to the desolvation of the polar atoms.[38]

In a first approximation, these modulations were not included in equation (12).

RESULTS AND DISCUSSION

Contact Areas as a Function of Molecular Weight

The accessible surface area (ASA) values of proteins have been found to correlate with their molecular weights (M.W.), both for monomeric and multimeric proteins.[3,30,31] This result was confirmed on our extended data set containing 55 monomeric and 27 oligomeric structures. Assuming a relationship between ASA and M.W. of the form:

$$ASA = a\, M.W.^b \quad (13)$$

we obtain the equation:

$$ASA = 4.9\, M.W.^{0.76} \quad (14)$$

which fits the observed ASA values for monomeric

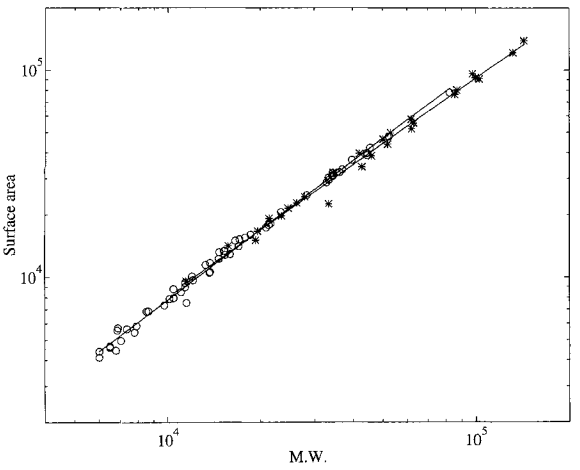

Fig. 1. Nonpolar contact area (in Å^2) versus molecular weight (in Da) for the 55 monomeric proteins (○) and 27 multimeric proteins (*) of our database. Straight lines show the best linear fit of log(Surface) to log(M.W.).

proteins to within 5% on average, with a correlation coefficient of 0.99, and equation

$$ASA = 2.6\ M.W.^{0.82} \quad (15)$$

which fits the observed ASA values for multimeric proteins to within 8% on average, with a correlation coefficient of 0.98. In equations (14) and (15) ASA is in Å^2 and M.W. in Da. These parameters differ slightly from those previously described.[30,31] These differences are not statistically significant if an 8% error on the determination of the ASA is accepted, in which case errors on b and a [defined in Eq. (13)] at 2 standard deviations are 0.04 and 2, respectively, for both monomeric and oligomeric proteins.

Similarly, we observe that the nonpolar contact area (NPCA) of proteins are correlated with their molecular weight (M.W.) (Fig. 1). Equation

$$NPCA = 0.27\ M.W.^{1.12} \quad (16)$$

fits the observed values (in Å^2 for NPCA and Da for M.W.) for monomeric proteins to within 4% on average, with a correlation coefficient of 0.999, and equation

$$NPCA = 0.50\ M.W.^{1.05} \quad (17)$$

fits the values for oligomeric proteins to within 5%, with a correlation coefficient of 0.996.

The accessible surface area ASA is found to be larger than the area of the molecular surface of the protein, since it is measured one probe radius (1.4 Å) away from that surface.[30] Similarly, the nonpolar contact area NPCA is found to be proportional to a quantity larger than the molecular weight or molecular volume of the protein.

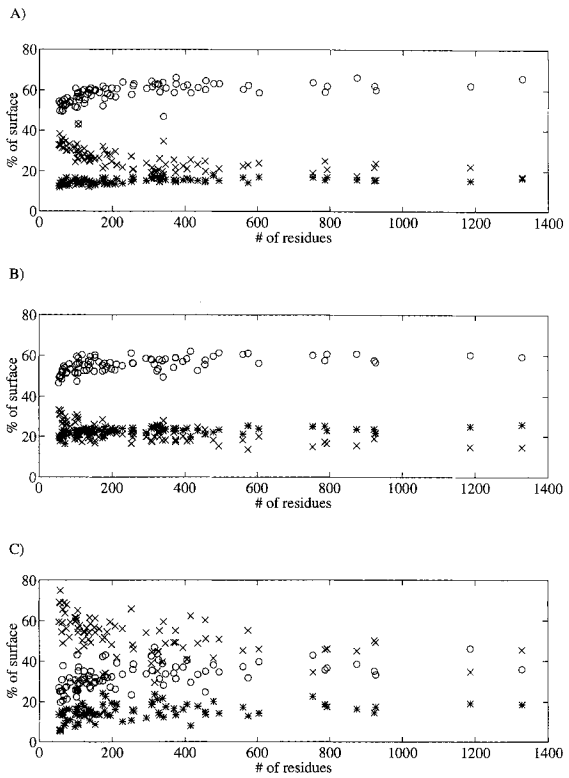

Fig. 2. Average partitioning of the total accessible surface area in the folded state of nonpolar atoms **(A)**, neutral polar atoms **(B)**, and charged atoms **(C)** versus the size of the protein. x, surface accessible to solvent (ASA); ○, surface occluded by nonpolar atoms (NPCA); *, surface occluded by polar atoms (PCA).

Contact Surfaces as a Function of Occluding Atoms

Each protein was divided into nonpolar, neutral polar, and charged atoms, and the accessible surface areas, the surfaces occluded by polar and nonpolar atoms of each component, were evaluated. Carbon and sulphur are taken to be nonpolar, and nitrogen and oxygen to be polar, neutral when they carry no charge, and charged in carboxylate, amino, and guanidinium groups. Partition of the atomic surface of the three classes of atoms defined here are plotted on Figure 2 for the native proteins. No systematic dependence on protein size is observed, except that small proteins (less than 100 amino acids) tend to have a higher fraction of accessible surface area than large proteins.

The environment of each class of atoms was further characterized by collecting data over the full set of 82 proteins, corresponding to 104,854 nonpolar atoms, 58,909 neutral polar atoms, and 5,793 charged polar atoms. Results are shown in Table I. On average, nonpolar atoms have a large fraction of their surface buried (77% of the total accessible surface area). This buried surface is divided in 79% of nonpolar contact area and 21% of polar contact

TABLE I. Average Contact Surfaces of Polar and Nonpolar Atoms in Proteins

Atom type	No. of atoms*	ASA (%)†	PCA (%)†	NPCA (%)†
Neutral polar atoms				
Mainchain	45,852	14	25	61
Sidechain	13,057	38	18	44
Total	58,909	20	23	57
Charged polar atoms‡	5,793	50	15	35
Nonpolar atoms				
Mainchain	35,821	22	16	62
Sidechain	69,033	24	15	61
Total	104,854	23	16	61

*Compiled over the 82 proteins of our data bank.
†Expressed in percent (%) of the total accessible surface area (TASA).
‡No charged polar atoms for the mainchain were considered.

area. Neutral polar atoms show very similar behavior, in that 80% of their total accessible surface area is buried upon folding, on average. This is contrary to the common view that polar atoms are predominantly located at the surface of the protein. The fraction of the surface area of polar atoms occluded by polar atoms is larger than the corresponding *PCA* of nonpolar atoms (23% versus 16%), while the nonpolar contact areas of polar atoms are accordingly smaller than the *NPCA* of nonpolar atoms (57% versus 61%). Sidechain polar atoms have a larger accessible surface area than mainchain polar atoms (Table I). The *PCA* of mainchain polar atoms, which is larger than the *PCA* of sidechain polar atoms, is related to the peptide groups that hydrogen-bond to form secondary structures in the interior of the protein (Table I). Interestingly, the average environments of neutral polar and nonpolar atoms are not drastically different. A Student's t test, however, shows that the differences in the mean values for *ASA*, *PCA*, and *NPCA* are significant, with a probability higher than 0.999.

In contrast to nonpolar and neutral atoms, charged atoms remain preferentially accessible to solvent (50% on average).

The *NPCA* of nonpolar, neutral polar, and charged polar atoms are 61%, 57%, and 35% of the total accessible surface area, respectively. This provides an empirical hydrophobic scale for atoms in a protein. it is noteworthy that the *PCA* represents 21%, 29%, and 31% of the *buried* surface of nonpolar, neutral polar, and charged polar atoms, respectively.

This analysis motivated us to define the environment free energy, since the difference kinds of environments of interior (polar or nonpolar) atoms are not taken into account in the solvation free energy of Eisenberg and McLachlan.[6]

Atomic Solvation Parameters

The atomic solvation parameters ASP_i were determined by a linear least-squares fit of equation (10) to observed values for the free energy of transfer from n-octanol to water, ΔG_{exp}, of the amino acids. The measured ΔG_{exp} were those determined by Fauchère and Pliska,[8] modified by Sharp et al.[32] to retain only the energy of contact.[33] The fit is carried out by considering five classes of atoms: carbon, neutral oxygen, and nitrogen (N/O), charged oxygen (O^-), charged nitrogen (N^+), and sulphur. This classification was introduced by Eisenberg and McLachlan.[6] The total accessible surface areas *TASA* of the five classes of atoms have been calculated for amino acid *X* in Gly-*X*-Gly sequences, averaged over the conformation this tripeptide adopts in lysozyme (PDB code 3lzm), thermolysin (PDB code 3tln), reduced cytochrome c (PDB code 5cyt), and flavodoxin (PDB code 4fxn). Similar results to those previously reported[28,30,39] were found (data not shown).

Our estimated *ASP* values are given in Table II. They are in agreement with recently published *ASP* values, calculated under the same assumptions.[40] It should be noted that our *ASP* values follow the hydrophobic scale of the five classes of atoms given by Eisenberg and McLachlan,[6] though the sign difference between polar and nonpolar atoms is not respected any more. The contribution of nonpolar atoms (C and S) is still positive and that of charged polar atoms (N^+ and O^-) negative. In contrast, neutral polar atoms (N and O) have now a small positive contribution. The meaning of the sign of the *ASP* parameters is unclear. A second example that also contradicts the results of Eisenberg and McLachlan[6] can be derived by calculating the *ASP* values from the original data of Fauchère and Pliska[8] but with separate values for neutral nitrogen and oxygen. In that case also, a positive *ASP* is found for neutral nitrogen.

Recently, a great deal of effort has been devoted to trying to reconcile the magnitude of the "microscopic" hydrophobic effect, measured with the surface area dependence, and the magnitude of the "macroscopic" hydrophobic effect, derived from surface tensions of hydrocarbon-water interfaces.[41,42]

TABLE II. Atomic Solvation Parameters

Atoms	ASP values (cal mol^{-1} Å^{-2})	Error*	Contribution (%)†
N/O	8.1	4	15.4
C	36	1	69.0
O^-	−5	8	4.6
N^+	−46	8.5	5.4
S	44	6.5	5.6

*Estimated errors on the ASP values at 2 standard deviations, assuming an error of 0.25 kcal/mol on each modified free energy of transfer from octanol to water (see text).
†Relative contribution of each atom class to the total accessible surface of the 20 amino acids.

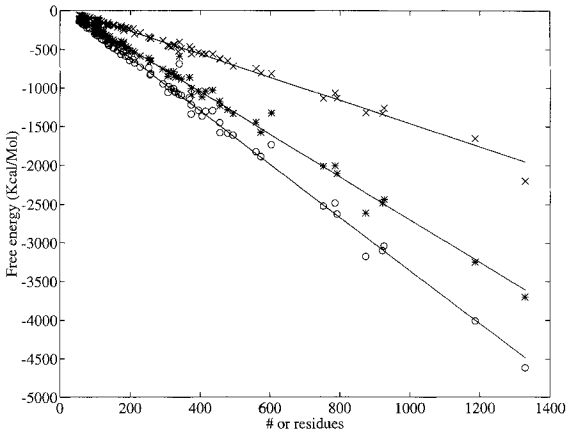

Fig. 3. Calculated contributions of the solvation free energy ΔG_s (○) and environment free energy ΔG_e(*) versus the number of residues, N, of the 82 proteins of our database. On the same plot, contribution of the mainchain only of the 82 proteins to ΔG_e is shown by x. Linear least-square fits to the data are shown.

The macroscopic value is approximately 72 cal mol^{-1} Å^{-2},[43] while values for the microscopic effect ranging from 20 to 73 cal mol^{-1} Å^{-2} have been reported.[3,6,32,41,42,44] These variations are essentially due to the use of different solvent for the determination of the experimental free energies of transfer, different values for the total accessible surface area of atoms of a residue in the extended state of the protein,[45] different measures of the surface area,[42] and different ways of expressing the free energy of transfer of a solute between two solvents.[32] It is noteworthy that recent experimental studies have shown a good linear correlation between the change in free energy induced by a cavity-creating substitution and the surface of the cavity created, both in T4 lysozyme[46] and sperm whale myoglobin,[47] with a slope of 20 cal mol^{-1} Å^{-2}.

There is not yet any clear indication as to which values should be used in a calculation in which the free energy is taken to be proportional to a surface.[33] Here we have decided to rely on the Flory-Huggins theory for the analysis of the transfer of free energy data since it provides the contact part of the difference in chemical potential between the two solvents.[33] This seems most appropriate for defining the environment free energy which analyzes interatomic contacts in a protein. The corresponding atomic solvation parameter value for carbon atom (the predominant atom of a protein when hydrogen is not taken into account explicitly) we report here is 36 cal mol^{-1} Å^{-2}

The Environment Free Energy

The contribution of the free energy of solvation [Eq. (4)] and of the environment free energy [Eq. (12)] to the stability of the 82 proteins of our database is plotted in Figure 3 versus the number of residues. Both ΔG_s and ΔG_e vary linearly with N, with the same correlation coefficient of 0.997. Linear least-squares fits to the data yield equations (18) and (19):

$$\Delta G_s = 70.2 - 3.43\,N \qquad (18)$$

and

$$\Delta G_e = 51.1 - 2.75\,N. \qquad (19)$$

Equation (18) fits the solvation free energy difference to within 6% on average, while equation (19) fits the difference in environment free energy to within 6.4% on average. A linear relationship between ΔG_s and N has already been described for monomeric proteins,[48] in which case the fit was:

$$\Delta G_s = 15.3 - 1.13\,N \qquad (20)$$

[Eq. (1) of Chiche et al.[48]]. Here it is established for oligomeric proteins also. Equation (20) was obtained with the *ASP* values determined by Eisenberg and McLachlan,[6] calculated from the unmodified transfer energy data of Fauchère and Pliska.[8] It yields a contribution of the free energy of solvation to the stability of a protein approximately 3 times lower than the estimates given by equation (18), which was computed with *ASP* values based on the modified free energy of transfer. Such a difference was expected.[32]

We believe that equation (18) provides an overestimate of the contribution of the hydrophobic effects to the stability of the folded state of a protein compared with its denatured state, since it does not take into account the type of contacts induced in the core of the protein upon folding, which may not all be stabilizing. Indeed, our free energy of environment ΔG_e, which tries to account for these contacts, provides a lower estimate of the hydrophobic effects [Fig. 3, and Eq. (19) compared with Eq. (18)].

ΔG_e is proportional to the nonpolar contact area [Eq. (12)], which in turn has been found to be approximately proportional to the molecular weight, both for monomeric and oligomeric proteins (this

work). This explains the linear relationship between ΔG_e and N.

ΔG_s linearly depends on the total accessible surface *TASA* and on the accessible surface area [equation (4)]. *TASA* is proportional to the *M.W.* (for example, see ref. 30), while *ASA* is found to be proportional to the *M.W.* to the power 0.76 for monomeric proteins and 0.83 for oligomeric proteins (this work). The contribution of the unfolded state (through *TASA*) is predominant, which explains the linear relationship between ΔG_s and N (results not shown).

The contribution of the mainchain interactions in a protein to ΔG_e is shown in Figure 3. To estimate this contribution, chimeric proteins corresponding to the 82 proteins of the data base were modeled, in which all residues except glycines were substituted to alanine, keeping the same fold for the backbone (practically, all sidechains were truncated after the CB atom). It appears also to vary linearly with N, with a correlation coefficient of 0.995. A least-square fit to the data yields:

$$\Delta G_e \text{ (backbone)} = 34.5 - 1.5\,N. \qquad (21)$$

This equation fits the values to within 8% on average. From the comparison of equations (19) and (21), it results that backbone-backbone interactions, and consequently sidechain-sidechain contacts, both represent approximately half of the total hydrophobic interactions that stabilize the folded form of a protein.

A Database-Derived Hydrophobic Scale of the 20 Amino Acid Residues

Equation (12), which provides an estimate of the contribution of the free energy of environment ΔG_e to the stability of a protein, can be seen as the sum over all atoms or over all residues of the protein. The specific contributions of the sidechains of each type of amino acids to ΔG_e were calculated based on the 82 proteins in our database, and Figures 4 and 5 show histograms of these contributions for all residues, except glycine. Distributions for hydrophobic residues are highly asymmetric, with a long one-sided tail that extends towards a zero contribution of the residue to ΔG_e. The position of the mean of the distribution shifts towards 0 as the hydrophobicity of the residue decreases. Polar residues have broader distributions, whose means are small, with tails that extend in both directions. For two large residues (arginine and lysine) (Fig. 4) cases can be found in which the contribution of the residue is positive, i.e., the environment of the residue in the folded protein is less favorable than a complete solvent environment in the theoretical unfolded state. Distributions for residues such as Ala, Ser, Thr, and Pro show two maxima. These two maxima reflect the distribution of the corresponding residue in the core and at the surface of the protein. This is illustrated in Figure 5 for alanine: the contributions of alanines that are mainly accessible (i.e., such that $> 40\%$ of their total surface is accessible to solvent) and of alanine that are mainly buried (such that $> 60\%$ of the total surface is occluded) have been computed separately and were found to correspond to the two maxima observed in the global distribution.

Interestingly, we found that the average environment free energies of the side chain of the 19 amino acids $\overline{\Delta G_e}$ are very well correlated with the modified experimental transfer energy from octanol to water of Fauchère and Pliska,[8] with a correlation coefficient of 0.98 (Fig. 6). The relationship

$$\overline{\Delta G_e}(X) = -0.82\,\Delta G_{tr}(X) - 0.9 \qquad (22)$$

fits the data to within 7% on average. Hence ΔG_e provides a database-deduced hydrophobic scale that correlates well with any hydrophobic scales deduced from mean are lost upon folding,[30,39] as well as with hydrophobic scales deduced from the data of Fauchère and Pliska.[6,8] However, it does not correlate well with another database-deduced hydrophobic scale based on the hydration of the sidechains of the residues, defined as the binding of crystallographic water molecules directly to protein surface.[49] This is probably because crystallographic water molecules represent only a fraction of actual water molecules in contact with the surface of the protein.[50]

Stability of Oligomeric Structures

The environment free energy of both oligomeric and monomeric proteins have been found to vary linearly with N (equation 19), with no noticeable difference. However, the energy of one subunit of a multimeric protein varies significantly depending on whether it is considered free in solution or as part of the oligomer (Fig. 7). Stabilization is always observed upon oligomerization, with changes in environment free energy varying between 1% and 108% (14% on average). These variations are protein size dependent, i.e., the stabilization of a small subunit is usually found to be large, while stabilization of subunits larger than 50 amino acids only varies between 1% and 30%. No differences have been observed for subunits belonging to dimers, trimers, or tetramers. The order of magnitude of the interaction energy between monomers is 30–40 kcal/mol, which compares well with other estimates. For example, the present estimates of the hydrophobic energy gains are 24 kcal/mol for the trypsin-inhibitor complex (file 2PTC in the PDB databank) and 32 kcal/mol for the hemoglobin α-β dimer (file 2MHB in the PDB databank). Chothia and Janin[51] estimated the same energies to be about 35 and 43 kcal/mol, respectively, while Miyazawa and Jernigan[21] calculated total interprotein contact energies of 61 kcal/

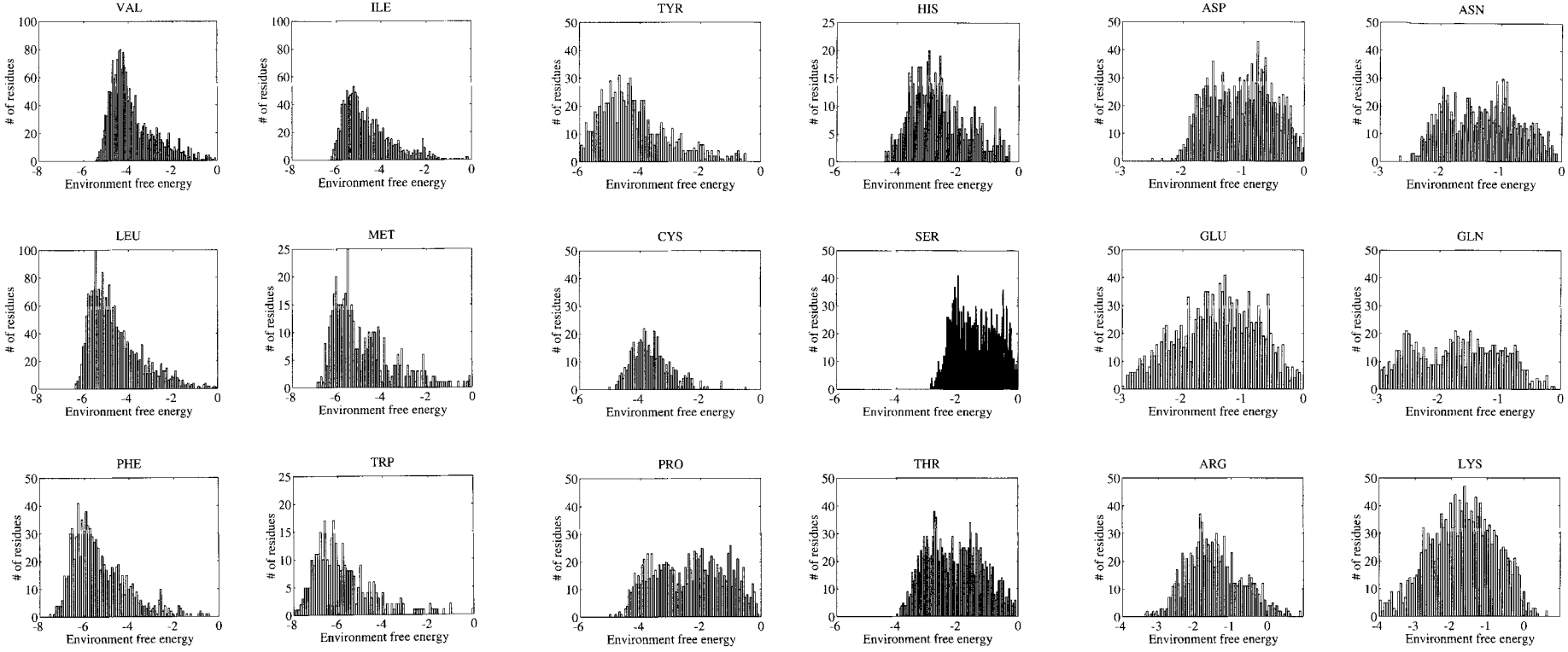

Fig. 4. Distributions of the contribution of the sidechains of 18 amino acids to the environment free energies of the 82 proteins of the database. The two missing amino acids are Gly, which, when hydrogens are not considered explicitly, does not have a sidechain, and Ala, which is treated explicitly in Figure 5.

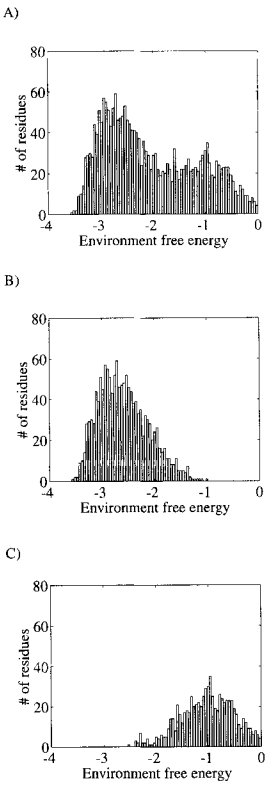

Fig. 5. Distribution of the contribution of the side chain of alanine to the environment free energy of the 82 proteins of our database. **A:** All alanines. **B:** Alanines whose solvent accessibility (defined as the ratio of its ASA to its total accessible surface TASA) is lower than 30%. **C:** Alanines whose solvent accessibility is greater than 30%.

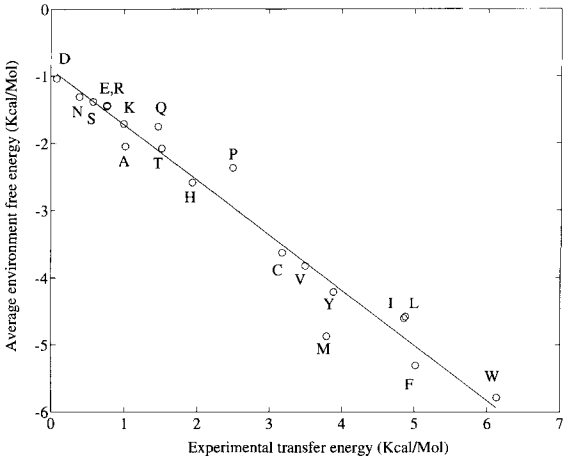

Fig. 6. Average contribution of amino acid side chains to the environment free energy versus the experimental free energy of transfer of the same amino acids from octanol to water (see text). Least-squares fit to the data is shown.

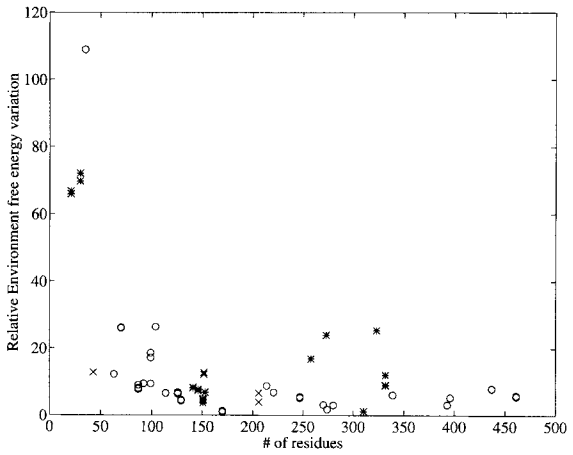

Fig. 7. Stabilization of protein subunits upon oligomerization, as observed by the change in environment free energy of each subunit considered isolated or in the oligomer (shown as 100 $[\Delta G_e(\text{oligomer}) - \Delta G_e(\text{isolated})]/\Delta G_e(\text{isolated})$, versus N, the number of residues. Values for dimeric proteins are indicated by ○, those for trimers by x, and those for tetramers by *.

mol for the trypsin inhibitor and 66 cal/mol for the hemoglobin dimer with a totally different method. A direct comparison with experimental data of association is hindered by the difficulty of getting a correct estimate of translational and rotational entropy losses upon association.

A word of caution is in order here. The data presented can be interpreted as association energy only under the assumption that the free subunits adopt the same conformation as in the complexes. This hypothesis may not be always valid. In these cases $\Delta\Delta G_e$ only provide an estimate of the contact energies between the different subunits. The case of Trp repressor circumvents this problem, since both the structures of the apo form (file 3WRP in the Brookhaven databank[52]) and the holo form (file 2WRP[53]) are known. The environment free energy of the apo-Trp repressor dimer is −521 kcal/mol, while the environment free energy of the holo-Trp repressor dimer in the presence of tryptophan is −526 kcal/mol. Hence tryptophan is estimated to enhance the stability of the Trp repressor dimer by 5 kcal/mol. Interestingly, the environment free energy of the holo-Trp dimer without considering the presence of tryptophan was found to be −520 kcal/mol, i.e., very similar to the energy of the apo-Trp repressor. Similar calculations based on the solvation free energy ΔG_s yielded a stabilization of the Trp repressor dimer due to the tryptophan of 2 kcal/mol only. The experimental value, which includes entropy losses, is 6 kcal/mol (J. Carey, personal communications).

Assessment of Protein Models

With the tremendous flow of new protein structures determined either by X-ray crystallography or nuclear magnetic resonance (NMR), and the devel-

opment of tools for predicting the three-dimensional conformation of a protein from its sequence and a related structure, a new challenge has appeared that deals with the design of criteria that can discriminate a good model or structure from a bad or misfolded one. The revision of several protein structures solved both by X-ray crystallography and by NMR (for review, see Branden and Jones[54]) and the design of deliberately misfolded structures[55,56] have shown that simple criteria based on the quality of the fit to the experimental data (R factor in X-ray crystallography, usually target functions in NMR), on the internal energy of the protein (including bonded interactions and VdW and electrostatics interactions), or on Ramachandran plots were not sufficient to identify a misfolded structure. As a consequence, several new methods have been developed that either improve these standard criteria[57–59] or rely on the evaluation of hydrophobic effects. These latter are based on the definition of the solvation free energy,[6,47,60] on statistical studies on residue distribution between the interior and surface of the protein and residue-residue interactions,[55,61–63] or on statistical studies on atomic solvation preference[35] and atomic contacts.[17] Correct structures were also identified from misfolded structures using the so-called 1D-3D profile,[19] which includes interactions with the solvent, interactions within the protein itself, and information on the secondary structures.[18]

Since our environment free energy accounts for both the interactions of the protein with the solvent and the environment of each atom in the interior of the protein, we decided to test its ability to assess protein models in cases in which the correct structure in known and incorrect folds have been modeled. The proteins considered include the ferredoxin of *A. vinelandii* (whose correct structure is entry 4FD1 in the PDB, and for which a misfolded structure has been reported as entry 2FD1 in the PDB), the HIV protease, the rubisco small subunit, and the Ig k VL region and hemerythrin. In the latter cases, misfolded structures were generated by modeling the side chains of the α-helical hemerythrin on the β-sheet backbone of Ig k VL domain, and vice versa. Two sets of misfolded structures were generated either by modeling the incorrect sidechains in all trans conformation[55] using the program CHARMM,[64] or by optimization of the position of the sidechains,[56] using a conformational space sampling program, CONGEN.[65] Our environment free energy and the solvation free energy (evaluated with the modified ASP values given in this work) were compared and the results are summarized in Table III. The environment free energy ΔG_e is found to be able to identify misfolded structures at least as well as the solvation free energy ΔG_s. The difference in energy between the two misfolded hemerythrins and the correct hemerythrin is proportionally 2 times larger than the same difference in ΔG_s, while ΔG_s provides a better identification of the misfolded Ig VL domains. It is noteworthy that the environment free energy shows a clear difference between the CONGEN-generated and the CHARMM-generated misfolded structures, which is expected because of the better packing provided by CONGEN.[56] The same difference is not as clear with ΔG_s. Both the environment free energy and the solvation free energy are found to be less discriminative than 1D-3D profiles[18] in identifying misfolded structures (Table III).

Another set of misfolded structures was considered previously, in which poly(Ala) were modeled with the same backbone folds as the 82 proteins of our database. Energies for these poly(Ala) were found to be half the energies of the correctly folded protein (Fig. 3).

Prediction of the Stability Effects of Site-Directed Mutagenesis on a Protein

Typical experimental efforts to understand the strength of the hydrophobic effect in stabilizing a protein are usually based on mutation studies in which a hydrophobic residue within the core of a protein is substituted by a smaller residue. The resulting change in the stability of the folded versus the denatured state of the protein is taken to be a measure of the difference between the hydrophobic stabilization of the two residues. The limits of any semiempirical model that attempts to quantify these hydrophobic effects can then be defined by testing its ability to predict these changes in stability upon hydrophobic residue substitution. Other attempts to compute the energies of these mutants have recently been described.[66,67]

One obvious difficulty in trying to predict the effect of a mutation is that the possible structural changes in the protein induced by the substitution are usually not known. For that reason, we chose, as a first test of our environment free energy, 17 mutants of T4 lysozyme that have recently been characterized both in terms of stability and in terms of structure.[10,46,68] Seven of these mutants involve substitution of one or two residues to alanine, for which only local structural changes around the mutated residues were found and such that a cavity always remained. The change in stability of these mutants was found to correlate with the size of this cavity.[10] The other mutants involve two bulky amino acids (a leucine and a phenylalanine), which have been replaced by a series of hydrophobic residues. Mutations of the leucine residues were found to induce few structural changes, while Phe mutants showed adjustment around the mutation.[46] To predict the change of stability induced by these mutations, the contributions ΔG_e and ΔG_s of the environment free energy and solvation free energy, respectively, to the stability of the wild-type protein

TABLE III. Assessment of Protein Models

Protein	Source: Correct	Source: Misfold	ΔG_e (correct) (kcal/mol)	ΔG_e (misfold) (kcal/mol)	Error ΔG_e* (%)	Error ΔG_s† (%)	Error profile‡ (%)
Hiv protease	3hvp	—**	−237	−216	9	7.4	37.5
Ferredox in hemerythrin	4fd1	2fd1	−275.8	−192.9	30.1	27.5	76.5
	—	—¶	−292	−235	20	9.0	60
		—§	−292	−241	17.5	9.0	
Ig k VL	—	—¶	−249	−215	14	17	85
		—§	−249	−231	8	14	
Rubisco small subunit	3rub	—¶¶	−296	−260	12.1	8	72

*Relative difference in environment free energy between the misfolded and the correct structures, evaluated as: $[\Delta G_e(\text{correct}) - \Delta G_e(\text{misfolded})]/\Delta G_e(\text{correct}) \times 100$.
†Relative difference in solvation free energy between the misfolded and the correct structures, evaluated as: $[\Delta G_s(\text{correct}) - \Delta G_s(\text{misfolded})]/\Delta G_s(\text{correct}) \times 100$; ΔG_s were computed by equation (4) (see text).
‡Relative difference in Z-score deduced from 3D-1D profiles; scores have been taken from Lüthy et al.[19] and the relative error is evaluated as: [score(correct) − score(misfold)]/score(correct) × 100.
**Ref. 71.
¶Ref. 55.
§Ref. 56.
¶¶Ref. 72.

and of models of the 17 mutants were computed. For a mutation to alanine, the mutant was constructed by truncating the sidechain of the residue to be changed after the CB atom. For the other mutations, all rotamers of the mutant residues were tested, with the rest of the molecule fixed, and the rotamer with the lowest environment free energy was kept. We chose the rotamer library of Tuffery et al.[69] The calculated differences in ΔG_e and ΔG_s were found to correlate well with the observed destabilization, with a correlation coefficient of 0.90. Interestingly, even the absolute values of the predicted differences in stability were found to be correct. Results for ΔG_e are shown in Figure 8. $\Delta\Delta G_e$ and $\Delta\Delta G_s$ were found to be very similar in amplitude (result not shown). An explanation for this small difference is given in Figure 9: since no significant variations in the polar contact area (*PCA*) were found between the mutants and the wild-type T4 lysozyme and since the environment free energy differs from the solvation free energy in the way it handles polar environment in the protein interior, the two methods should provide the same results. At this point, it is worth mentioning that modeling the denatured state as a fully extended polypeptide chain is probably wrong: this state probably contains some contacts.[9] However, in the evaluation of the energy of mutants, this is no longer a problem, since what we evaluate is now a difference of differences. Discrepancies between the predicted values and the experimental values are larger when the mutant residue is not an alanine (Table IV). It is noteworthy, for example, that for what should be cavity-filling mutations like L99F and L99M, the environment free energy predicts an increase in hydrophobic stabilization in the mutant compared with the wild-type protein, while a destabilization is observed experimentally. Similar results were found for two other cavity-filling mutations (L133F and A129V), in which case we predict a stabilization (Table IV), while the true mutants were found to be less stable.[68] The most plausible explanation for these discrepancies is that the more bulky hydrophobic residues introduce strains in the core of the protein.[46,68]

We also tested our program on the 83 mutants of staphylococcal nuclease involving hydrophobic residues for which stability data are available.[70] These mutants involve single alanine and glycine substitutions for each of the leucine, valine, tyrosine, isoleucine, methionine, and phenylalanine residues, as well as each isoleucine mutated to valine. Prediction of the possible destabilization induced by these mutations based on our environment free energy did not show a reasonable correlation with the experimental values (correlation coefficient 0.52; Fig. 10). The lack of structural data for these mutants of staphylococcal nuclease makes it difficult to relate this poor correlation with the globally correct results obtained for T4 lysozyme. These results draw the limit of validity of our method.

CONCLUDING REMARKS

Recent studies of the stability consequences of altering amino acid sequences of small proteins strongly support the hypothesis that hydrophobic interactions play a very important role in protein stability and folding. Attempts to model these hydrophobic interactions usually assume that they are proportional to the area buried upon folding. In this work we extend this idea by also paying attention to the interatomic contacts in the core of the protein and we define an energy that characterizes the *en-*

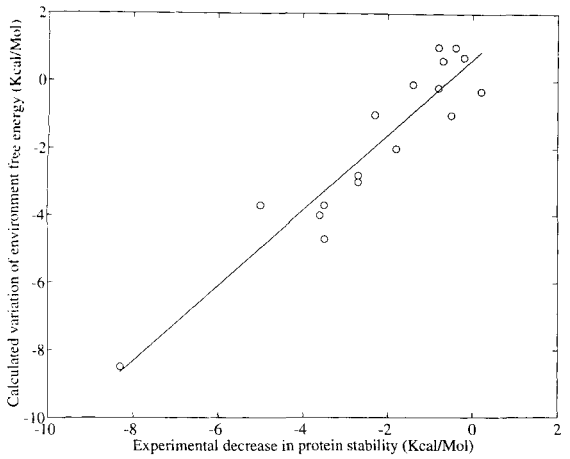

Fig. 8. Predicted change in the environment free energy of folding of mutant T4 lysozymes relative to wild type, versus the experimentally determined change in the free energy of unfolding (from Table IV). The full line shows the first diagonal.

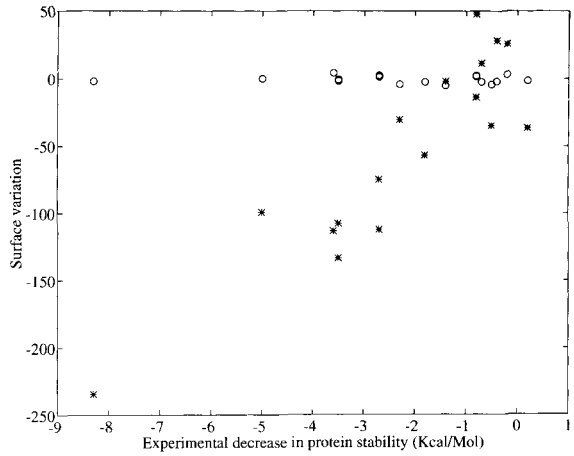

Fig. 9. Calculated change in the total polar contact area (○) and nonpolar contact area (*) of mutant T4 lysozymes relative to wild type, versus the experimentally determined change in the free energy of unfolding.

vironment of each atom, be it solvent, polar, or nonpolar. The contribution of this environment free energy to the stability of the folded state versus the unfolded state of a protein was found to be proportional to the area buried by *nonpolar atoms* upon folding.

This environment free energy provided a basis for a database-deduced hydrophobic scale, which was found to correlate well with the experimental data of the energy of transfer of amino acid analogues from octanol to water of Fauchèere and Pliska[8] from which only the contact energies have been extracted.[32,33]

The environment free energy has also been shown to be useful in assessing the quality of a experimentally determined structure or a theoretical model. It should be mentioned that it provides a relative test,

TABLE IV. Prediction of Change in Stability Induced by "Cavity-Creating" or "Cavity-Filling" Substitutions in 17 Mutants of T4 Lysozyme*

Mutant	$\Delta\Delta G$ experimental† (kcal/mol) (pH 3.0)	$\Delta\Delta G_e$ (calculated)
L46A	−2.7	−2.8
L118A	−3.5	−3.7
L121A	−2.7	−3.0
L133A*	−3.6	−4
L99A	−5.0	−3.7
F153A	−3.5	−4.7
L99A/F153A	−8.3	−8.5
L99I	−1.4	−0.1
L99M	−0.7	+0.6
L99F	−0.4	+1.0
L99V	−2.3	−1.
F153I	−0.5	−1.
F153L	+0.2	−0.3
F153M	−0.8	−0.2
F153V	−1.8	−2.0
L133F*	−0.2	+0.7
A129V*	−0.8	+1.0

All mutants were constructed from and compared with the WT T4 lysozyme, in which the two cysteines have been substituted (C54T/C91A), except mutants indicated with an asterisk, which were constructed from, and compared with the WT T4 lysozyme.
†Data from Karpusas et al.[68] and Eriksson et al.[10,46]

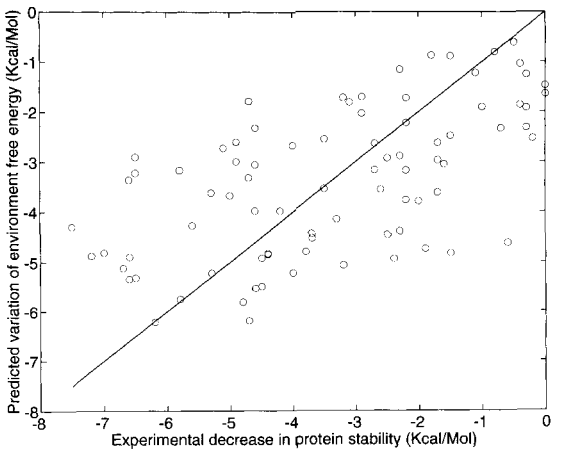

Fig. 10. Predicted change in the environment free energy of folding of staphylococcal nuclease mutants, versus the experimental change in the free energy of unfolding. The line shows the first diagonal.

in that it only differentiates a misfolded structure from a known correct structure. Truly incorrect structures can be identified on the basis that they do not fulfill the linear relationship of the environment free energy with the number of residues of the protein. In some cases, however, the result is less clear-cut.

Theoretical prediction of the stability of protein mutants is an important yet unsolved problem in

biophysics. Here we show that the environment free energy provides a correct estimate of the change in stability induced by a mutation in cases in which only local structural changes close to the mutation have been observed. If these results are confirmed, then in cases in which the predicted change in stability does not agree with the experimental thermodynamic data, this discrepancy could be interpreted as a sign that larger structural changes have occurred. Clearly, more work is needed before any definite and general statement can be made in this regard.

The model we have developed represents one step towards the ultimate goal of correctly modeling the energy of burying polar or nonpolar atoms. We have shown that, in a crude approximation, it was possible to account for favorable polar–polar contacts and unfavorable polar–nonpolar contacts in the interior of a protein within the framework of Eisenberg and McLachlan.[6] The advantage of this model is that it relies on a small number of parameters derived from experimental data taken from a field totally different from the one in which the effects it is trying to explain are observed. The cases in which the environment free energy performs better than the solvation free energy are clear: they are cases in which favorable polar–polar contacts, or unfavorable polar-nonpolar contacts occur, for example due to insufficient secondary structure formation such as in wrong models built in poor electron density maps. This is a reason for which ΔG_e was usually found to provide better discrimination between incorrectly folded and correctly folded models. This will not occur, however, in experiments in which only nonpolar atoms are involved, such as the lysozyme T4 mutants mentioned above.

The environment free energy described here is implemented in a program called ENVIRON. This program is available from one of the authors (P.K.) upon request to koehl@bali.u-strasbg.fr.

ACKNOWLEDGMENTS

We thank Dr. Bruccoleri, Dr. Eisenberg, and Dr. Fitzgerald for making available to us the coordinates of undeposited structures, and Dr. Jannette Carey and Dr. Thomas Simonson for helpful comments. We also acknowledge encouragement and support from J.-F. Lefèvre and D. Moras, in whose laboratories this work was carried out.

REFERENCES

1. Dill, K.A. Dominant forces in protein folding. Biochemistry 29:7133–7155, 1990.
2. Oobatake, M., Ooi, T. Hydration and heat stability effects on protein unfolding. Prog. Biophys. Mol. Biol. 59:237–284, 1993.
3. Chothia, C. Structural invariants in protein folding. Nature 254:304–308, 1975.
4. Richards, F.M. Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977.
5. Janin, J. Surface and inside volumes in globular proteins. Nature 277:491–492, 1979.
6. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
7. Ooi, T., Oobatake, M., Nemethy, G., Sheraga, H.A. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc. Natl. Acad. Sci. U.S.A. 84:3086–3090, 1987.
8. Fauchère, J.-L., Pliska, V. Hydrophobic parameters π of amino acid side-chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. Chim. Ther. 18: 369–375, 1983.
9. Dill, K.A., Shortle, D. Denatured states of proteins. Annu. Rev. Biochem. 60:795–825, 1991.
10. Eriksson, A.E., Baase, W.A., Zhang, X.-J., Heinz, D.V., Blaber, M., Baldwin, E.P., Matthews, B.W. Response of a protein structure to cavity creating mutations and its relation to the hydrophobic effect. Science 255:178–183, 1992.
11. Wood, R.H., Thompson, P.T. Differences between pair and bulk hydrophobic interactions. Proc. Natl. Acad. Sci. U.S.A. 87:946–949, 1990.
12. Singh, J., Thornton, J.M. The interaction between phenylalanine rings in proteins. FEBS Lett. 191:1–16, 1985.
13. Singh, J., Thornton, J.M. Sirius, an automated method for the analysis of the preferred packing arrangements between protein groups. J. Mol. Biol. 211:595–615, 1990.
14. Burley, S.K., Petsko, G.A. Aromatic-aromatic interactions: A mechanism of protein structure stabilization. Science 229:23–28, 1985.
15. Burley, S.K., Petsko, G.A. Amino-aromatic interactions in proteins. FEBS Lett. 203:139–143, 1986.
16. DeLaCruz, X., Fita, I. Atomic accessible and contact surfaces as restraints in the Hendrickson and Konnert refinement program. J. Appl. Crystallogr. 24:941–946, 1991.
17. Vriend, G., Sander, C. Quality control of protein models: Directional atomic contact analysis. J. Appl. Crystallogr. 26:47–60, 1993.
18. Bowie, J.U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structures. Science 253:164–170, 1991.
19. Lüthy, R., Bowie, J.U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature 356: 83–85, 1992.
20. Wilmanns, M., Eisenberg, D. 3D profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. Proc. Natl. Acad. Sci. U.S.A. 90:1379–1383, 1992.
21. Miyazawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macrmolecules 18:534–552, 1985.
22. Sippl, M.J. Calculation of conformational ensembles from potential of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213:859–883, 1990.
23. Casari, G., Sippl, M.J. Structure-derived hydrophobic potential: Hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds. J. Mol. Biol. 224:725–732, 1992.
24. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 227:227–238, 1992.
25. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. nature 358:86–89, 1992.
26. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
27. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55: 379–400, 1971.
28. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms in lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.
29. Connolly, M.L. Solvent-accessible surfaces of proteins and nucleic acids. Science 221:709–713, 1983.
30. Miller, S., Janin, J., Lesk, A.M., Chothia, C. Interior and surface of monomeric proteins. J. Mol. Biol. 196:641–656, 1987.

31. Miller, S., Lesk, A.M., Janin, J., Chothia, C. The accessible surface area and stability of oligomeric proteins. Nature 328:834–836, 1987.
32. Sharp, K.A., Nicholls, A., Friedman, R., Honig, B. Extracting hydrophobic free energies from experimental data: Relationship to protein folding and theoretical models. Biochemistry 30:9686–9687, 1991.
33. Holtzer, A. The use of Flory-Huggins theory in interpreting partitioning of solutes between organic liquids and water. Biopolymers 32:711–715, 1992.
34. Boberg, J., Salakoski, T., Vihinen, M. Selection of a representative set of structures from Brookhaven protein data bank. Proteins 14:265–276, 1992.
35. Holm, L., Sander, C. Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 225:93–105, 1992.
36. Gilson, M., Honig, B. Destabilization of an α-helical bundle by helix dipoles. Proc. Natl. Acad. Sci. U.S.A. 86:1524–1528, 1989.
37. Ben-Naim, A. The role of hydrogen bonds in protein folding and protein association. J. Phys. Chem. 95:1437–1444, 1991.
38. Yang, A.-S., Sharp, K.A., Honig, B. Analysis of the heat capacity dependence of protein folding. J. Mol. Biol. 227: 889–900, 1992.
39. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
40. Pickett, S.D., Sternberg, M.J.E. Empirical scale of side-chain conformational entropy in protein folding. J. Mol. Biol. 231:825–839, 1993.
41. Sharp, K.A., Nicholls, A., Fine, R.F., Honig, B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. Science 252:106–109, 1991.
42. Tunon, I., Silla, E., Pascual-Ahuir, J.L. Molecular surface area and hydrophobic effect. Protein Eng. 5:715–716, 1992.
43. Aveyard, R., Haydon, D.A. Thermodynamic properties of aliphatic hydrocarbon/water interfaces. J. Colloid Interface Sci. 20:2255–2261, 1965.
44. Hermann, R.B. 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. U.S.A. 74:4144–4145, 1977.
45. Zielenkiewicz, P, Saenger, W. residue solvent accessibilities in the unfolded polypeptide chain. Biophys. J. 63: 1483–14, 1992.
46. Eriksson, A.E., Baase, W.A., Matthews, B.W. 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, 1993.
47. Pinker, R.J., Rose, G.D., Kallenbach, N.R. Effect of alanine substitutions in alpha-helices of sperm whale myoglobin on protein stability. Protein Sci. 2:1099–1105, 1993.
48. Chiche, L., Grigoret, L.M., Cohen, F.E., Kollman, P.A. Protein model structure evaluation using the solvation free energy of folding. Proc. Natl. Acad. Sci. U.S.A. 87:3240–3243, 1990.
49. Kuhn, L.A., Siani, M.A., Pique, M.E., Fisher, C.L., Getzoff, E.D., Tainer, J.A. The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. J. Mol. Biol. 228:13–22, 1992.
50. Otting, G., Liepinsh, E., Wuthrich, K. Protein hydration in aqueous solution. Science 254:974–980, 1991.
51. Chothia, C., Janin, J. Principles of protein-protein recognition Nature 256:705–708, 1975.
52. Zhang, R.G., Joachimiak, A., Lawson, C.L., Schevitz, R.W., Otwinowski, Z., Sigler, P.B. The crystal structure of TRP aporepressor at 1.8 angstroms shows how binding tryptophan enhances DNA affinity. Nature 327:591–597, 1987.
53. Schevitz, R.W., Otwinowski, Z., Joachimiak, A., Lawson, C.L., Sigler, P.B. The three dimensional structure of TRP repressor. Nature 317:782–786, 1985.
54. Branden, C.I., Jones, T.A. Between objectivity and subjectivity. Nature 343:687–689, 1990.
55. Novotny, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded models. Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
56. Novotny, J., Rashin, A.A., Bruccoleri, R.E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4:19–30, 1988.
57. Gonzalez, J.A., Rullmann, C., Bonvin, A.M.J.J., Boelens, R., Kaptein, R. Toward an NMR R-factor. J. Magn. Reson. 91:659–664, 1991.
58. Nilges, M., Habazettl, J., Brünger, A.T., Holak, T.A. Relaxation matrix refinement of the solution structure of squash trypsin inhibitor. J. Mol. Biol. 219:499–510, 1991.
59. Brünger, A.T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. Nature 355:472–475, 1992.
60. Vila, J., Williams, R.L., Vasquez, M., Sheraga, H.A. Empirical solvation models can be used to differentiate native from near-native conformations of bovine-pancreatic trypsin inhibitor. Proteins 10:199–218, 1991.
61. Baumann, G., Frömmel, C., Sander, C. Polarity as a criterion in protein design. Protein Eng. 2:329–334, 1989.
62. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M.J. Identification of native protein folds amongst a large number of incorrect models. J. Mol. Biol. 216:167–180, 1990.
63. Maiorov, V.N., Crippen, G.M. Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227:876–888, 1992.
64. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.
65. Bruccoleri, R.E., Karplus, M. Prediction of folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
66. Lee, C., Levitt, M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature 352:448–451, 1991.
67. Van Gunsteren, W.F., Mark, A.E. Prediction of the activity and stability effects of site-directed mutagenesis on a protein core. J. Mol. Biol. 227:389–395, 1992.
68. Karpusas, M., Baase, W.A., Matsumura, M., Matthews, B.W. Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. Proc. Natl. Acad. Sci. U.S.A. 86:8237–8241, 1989.
69. Tuffery, P., Etchebest, C., Hazout, S., Lavery, R. A new approach to the rapid determination of protein side chain conformations. J. Biomol. Struct. Dynam. 6:1267–1289, 1991.
70. Shortle, D., Stites, W.E., Meeker, A.K. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. Biochemistry 29:8033–8041, 1990.
71. Navia, M.A., Fitzgerald, P.M.D., McKeever, B.M., Leu, C.T., Heimbach, J.C., Herber, W.K., Sigal, I.S., Darke, P.L., Springer, J.P. Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. Nature 337:615–620, 1989.
72. Chapman, M.S., Suh, S.W., Ciermi, P.M.G., Cascio, D., Smith, W.W., Eisenberg, D.E. Tertiary structure of plant rubisco: Domains and their contacts. Science 241:71–74, 1988.