ELSEVIER

Biophysical Chemistry 51 (1994) 327–336

Biophysical Chemistry

Spatial and free energy distribution patterns of amino acid residues in water soluble proteins

Vladimir V. Nauchitel [a,*], Rajmund L. Somorjai [b]

[a] *Oklahoma University Health Sciences Center, Post Office Box 26901, CHB, Rm. 115, Oklahoma City, Oklahoma 73190, USA*
[b] *Institute for Biodiagnostics, National Research Council of Canada, 435 Ellice Ave., Winnipeg, Manitoba R3B 1Y6, Canada*

Received 20 October 1993; accepted 14 January 1994

Abstract

We have calculated, for the 20 common amino acid residues: probability density functions that characterize the residues' tendency to occupy different locations in proteins; a propensity scale for the residues to be exposed or buried; mean force potentials that characterize the residues' free energy dependence on their degree of exposure; the average composition of water soluble proteins, and the composition of their core and surface. The nature of differences between different hydrophobicity-related scales is discussed.

Key words: Proteins; Residues; Accessibility; Hydrophobicity

1. Introduction

We address a number of important problems with regard to proteins. The problems are all closely tied to a better understanding of the role of water as the solvent that influences structure formation. We use a single approach, simple in execution and compelling physically, to answer the following questions from a unified point of view: (1) Should water soluble proteins have any specificity in their amino acid residue composition [1]? (2) Are there any distinct features in the way residues are distributed between the surface and core of proteins [2,3]? (3) Is there a reliable way to determine the residues' propensities to occupy preferred locations in the core or on the surface [4–6]? (4) What is the relative importance of the major forces responsible for stabilization of the 3D structure of proteins [4,5,7–12]?

There are two general approaches used to determine the propensity of amino acid residues to have (or avoid) contact with water. The first is based on experimental measurements of some solubility-related properties of molecules that incorporate one of the 20 amino acids. In principle, these experimental approaches can provide real physico-chemical characterization of individual residues, but each technique has its own limitation when dealing with amino acids. The second is based on studying the probability of different amino acids to occupy specific locations in the 3D structure of globular proteins. The approaches

* Corresponding author.

SSDI 0301-4622(94)00053-M

differ in how they characterize and define the extent of a residue's burial/exposure. Cornette et al. [6] compiled a large set of different hydrophobicity scales, derived both from experiments and from statistical considerations.

Here we have calculated, for the 20 common residues: (1) probability density functions that characterize the residues' tendency to occupy different locations in proteins; (2) a propensity scale for the residues to be exposed or buried; (3) mean force potentials that characterize the residues' free energy dependence on their degree of exposure; (4) the average composition of water soluble proteins, and (5) the composition of their core and surface. We used 64 proteins to obtain statistically reliable results. All our results are obtained with a single, simple approach. We compute characteristics related to surface accessibility. (This approach [13] also happens to be two orders of magnitude more efficient than the ones based on conventional surface accessibility calculations.) We emphasize that the propensity of residues to be buried/exposed is not equivalent to hydro(phobicity/philicity).

2. The concept of Gaussian-neighborhood

To estimate the degree of exposure of an atom to water we calculate the surroundings of the atom, filtered by some discriminating function [13]. Our choice of the functional form for this function is dictated by the nature of hydrophobic forces. These forces act to diminish the contact area between nonpolar atoms and water molecules. The more shielded a nonpolar atom the lower its free energy. Clearly, more than one layer of neighbors would provide complete shielding from water and thus the lowest energy. However, the first layer of neighbors is the most important. Thus, to express the degree of exposure of an atom to water, we have to discriminate between nearest neighbors of the atom and those situated farther away. As the discriminating function we chose a 2-parameter Gaussian, $\exp[-(r-r_0)^2/2\sigma^2]$, where σ controls its width and r_0 positions its maximum. We set $r_0 = R_i^0 + R_w$ (R_i^0 is the van der Waals radius of atom i, and R_w is the radius of the water molecule), thus emphasizing the role of the first layer.

We define the Gaussian neighborhood (G-neighborhood) of atom i as the single, location-sensitive number

$$G_i^* = \sum_{j \neq i} \exp\left\{-\left[R_{ij} - \left(R_i^0 + R_w\right)\right]^2 / 2\sigma^2\right\}, \quad (1)$$

where R_{ij} is the separation between atoms i and j. The atomic radii for carbon, $R_C^0 = 1.70$ Å, hydrogen, $R_H^0 = 1.20$ Å, oxygen, $R_O^0 = 1.52$ Å, nitrogen, $R_N^0 = 1.55$ Å and sulphur, $R_S^0 = 1.80$ Å were taken from ref. [14]. Values of $R_w = 1.5$ Å and $\sigma = R_w$ were used for all the calculations. The Gaussian function can be viewed as a filter which emphasizes the essential area around the surface of a sphere defined by radius $(R_i^0 + R_w)$ (usually employed when computing the surface accessible to water molecules). The G_i^* complements the more conventional measure of accessible surface, i.e. the more buried atom i, the larger its G_i^* (and the smaller its surface accessibility) and conversely, the more exposed the atom, the smaller its G_i^* (the larger its surface accessibility).

To characterize the 20 amino acid residues we calculated their G-neighborhood, G_k:

$$G_k = \sum_i G_{ik}^* / N_k, \quad (2)$$

where ΣG_{ik}^* is the sum over all N_k atoms i of residue k. G_k is a measure of the surroundings of residue k (averaged over all the atoms of the residue) by other atoms. We calculated the G_k for all residues of 64 protein molecules. All information on the atomic coordinates was taken from the Protein Data Bank [15,16] (entries: 1abp, 1acx, 1ccr, 1cc5, 1cy3, 1cyc, 1fdx, 1fx1, 1gcr, 1hds, 1hip, 1hmq, 1lz1, 1lzt, 1mbd, 1pfc, 1ppd, 1p2p, 1rei, 1rhd, 1rn3, 1tim, 2ccy, 2cdv, 2cga, 2cna, 2cyp, 2c2c, 2dhb, 2fd2, 2fxb, 2gbp, 2gcr, 2hhb, 2hla, 2kai, 2ldx, 2lhb, 2lh1, 2lyz, 2lzm, 2mcp, 2ptc, 2rhe, 2sod, 2sbt, 2sga, 2sns, 2taa, 256b, 3adk, 3app, 3est, 3fxc, 3fxn, 3gpd, 3icb, 3pcy, 3pgk, 3rp2, 3tln, 4cpv, 5cpa, 5cyt). We added hydrogens to all X-ray structures of the above proteins. The terminal residues were excluded from our analysis because they differ from the regular residues.

(More details about the approach and related references are given in ref. [13].)

3. G-neighborhood probability density profiles and mean force potentials of amino acid residues

We used the G-neighborhood approach to calculate probability density functions (histograms) of G_k for the 20 regular residues. The histogram heights are scaled numbers of those residues that have their G_k values inside the bins of the histogram. The scaling factor is the sum of all residues of a given type. A bin width of $\Delta G = 5$ was used for all our calculations. The heights divided by the bin width are average probability density values for the bin. To get more data we used two sets of histograms that were shifted by 2.5. These two sets of histograms allowed us to construct more precise probability density functions, Fig. 1(a,b). The populated G values (see Figs. 1a and 1b) range approximately from 20 to 60. The limits are imposed both by the intrinsic geometry of proteins and by the parameters of the Gaussian filter. To populate the range below $G = 20$, chemical bonds would have to be broken. (An isolated atom would have $G = 0$.) To noticeably populate the range above $G = 60$ the atoms would have to interpenetrate. Thus, residues with large G values approaching 50–60 are maximally buried inside the proteins, whereas those with $G \approx 20$–30 are maximally exposed to the solvent.

Based on the calculated probability profile patterns the standard residues are readily divided into two groups, with ten residues in each. One group comprises residues with their most proba-

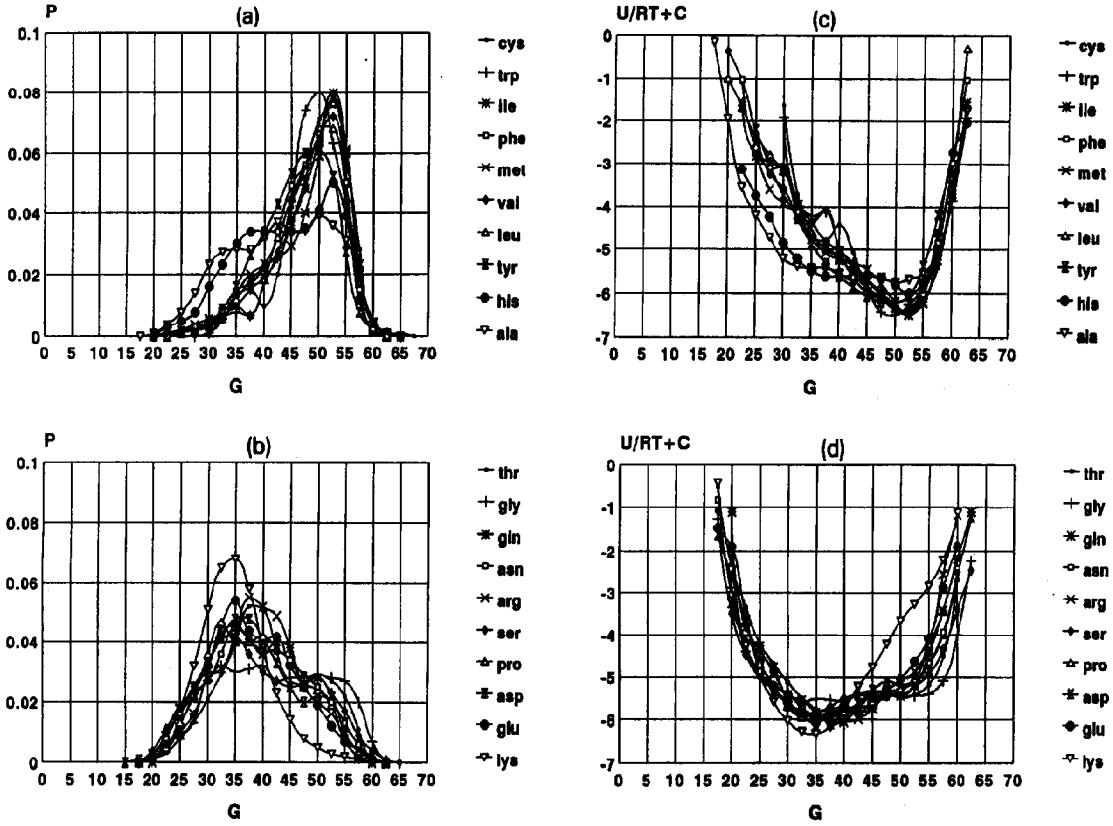

Fig. 1. (a,b) Probability density functions for the tendency of the twenty amino acid residues to be buried/exposed. (c,d) Free energy of burial for amino acid residues.

ble G values between 50 and 55. The left part of their G-profiles is longer tailed than the right one, Fig. 1a. These residues' probability to be shielded by other residues exceeds their probability to be in the outer shell of a molecule. The residues of the second group have their most probable G values grouped in the area 35 to 40 and prefer outer shell locations over the buried ones, Fig. 1b. We think that this very distinctive pattern of the residues' dividing into two groups is an interesting phenomenon.

Each of the two groups comprises residues that have different chemical nature. The first group includes residues with nonpolar side-chains ILE, LEU, VAL, ALA, MET, CYS as well as the ones with aromatic side-chains PHE, TRP, TYR, HIS. The residues with aromatic side-chains (PHE, TRP, TYR and HIS) have π-electrons and can interact strongly with the protons of water molecules. Furthermore, TYR and HIS have polar groups in their side-chains. Nevertheless, all the four residues belong to the group that is formed mostly by residues with nonpolar aliphatic side-chains which cannot interact strongly with water molecules. In contrast, the ring structure of PRO has no π-electrons, but contains 3 nonpolar CH_2 groups, yet it belongs to the group of polar and charged residues. The polar and charged residues (SER, THR, ASN, GLN, ASP, GLU, ARG, LYS) as well as GLY and PRO form the second group. The above difference between the two groups is expected for residues with aliphatic side-chains on one hand and residues with polar or charged side-chains on the other. The small nonpolar residues GLY, ALA, and the polar residue SER have broader G-profiles than do the bigger residues. They can occupy both inner and outer layers of a globule because they do not have spatial restrictions such as the big and bulky residues do and it is much easier to create for them a compromise surroundings than for big residues. The residues with small side-chains allow more freedom to structurally accommodate flanking bulky residues. Therefore, the small residues can participate in different hinge-like structures as well as in regular secondary structures. In general, small residues have broader profiles, Fig. 1, than the big ones, indicating their greater freedom in adopting different locations. CYS is an exception. This residue is more selective in its preference to be buried than almost all other residues. It is known that S–S bridging is responsible for this behavior.

We do not pretend to be the first who noticed that aromatic residues prefer to be buried inside of water soluble proteins, but we want to stress their chemical difference and we expect them to have some special functional role in the proteins. We believe that the aromatic residues' tendency to occupy inner locations is based on a different physico-chemical nature than that one of the nonpolar residues.

The G-neighborhood probability profiles for each residue, Figs. 1a and 1b, carry information on the free energy of transfer between locations with different degree of exposure. The probabilities and energies are related by

$$P(G) = \exp(C^* - U/RT), \tag{3}$$

where C^* is a constant, and U/RT is the dimensionless potential of mean force [17]. This mean force potential describes the dependence of the free energy of a residue on its degree of exposure. The potentials for the twenty residues are depicted in Figs. 1c and 1d. We do not need to know U to estimate the free energy difference ΔF_{ab}:

$$\Delta F_{ab} = (U_b - U_a)/RT = \ln(P_a/P_b) \tag{4}$$

is the free energy change on transferring a residue from location a to location b.

The energy change for polar and charged residues (except LYS), Fig. 1d, when they are moved from their most preferable positions (outer layers, $G \leqslant 35$) to the interior of the molecule, is noticeably smaller than the change for residues such as VAL, LEU, ILE, PHE, TRP, TYR, MET, CYS, Fig. 1c, when they are moved from their most preferable positions (inner layers, $G \geqslant 50$) close to the surface of the molecule. GLY bridges these two groups. Its energy changes very little from $G = 25$ to 55.

The energy plots of the first group of residues, Fig. 1c, show that the compact structures of proteins are stabilized to a great degree by burying these residues. It appears that the buried residues

of second group do not contribute to energetical stability, Fig. 1d. For polar and charged residues inside proteins, the energy lost on breaking contact with water molecules is not compensated completely by inner interactions between polar and charged groups, Fig. 1d. The sheltering of these residues requires some energy, but because such sheltered residues help to bury more residues of the first group, ultimately the energy balance becomes negative. Furthermore, the interaction between buried polar groups is selective because any charge tends to be surrounded by charges of opposite sign, or to form a salt bridge with an atom that has opposite charge, or to participate in an H-bond if it is a donor or an acceptor of the proton. This property of the coulombic interaction is largely responsible for the uniqueness of protein structure, stabilizing it against other compact structures which cannot provide optimal coulombic interaction inside the molecule. (We also include hydrogen bonds when we refer to interaction between polar groups. A similar role for hydrogen bonds was suggested earlier [18].) Another selective factor is the solvent (water). All water soluble proteins should have sufficiently uniform distribution of polar groups at their surface. Otherwise, the molecules would adhere to each other by their large nonpolar patches. The limitations caused by the nature of polar groups on one hand and by water on the other impose severe restrictions on the possible 3D structures of proteins. A review of some aspects of distributions of different groups on the surface of proteins has been published recently by Rashin [19].

4. Propensity scales of amino acid residues to be buried / exposed and sources of their discrepancy

The probability density functions of G contain information on the residues' probability to occupy different locations (to have different extent of exposure to a solvent) and, therefore, the functions are superior to any kind of one number characteristics. Nevertheless, it is convenient to have a single characteristics number to describe these preferences and many scientists use such characteristics as hydrophobicity. It is interesting to compare relative characteristics for the twenty standard residues that can be produced from the probability profiles, Fig. 1, with other scales. For this one number characteristics average values $\langle G_r \rangle$ ($r = 1, 2, \ldots, 20$) can be used. The propensity-to-be-buried (PB) values for the twenty standard residues can be defined by

$$\langle g_r \rangle = \left[2\langle G_r \rangle - \left(\langle G_r \rangle_{\max} + \langle G_r \rangle_{\min}\right)\right] / \left(\langle G_r \rangle_{\max} - \langle G_r \rangle_{\min}\right), \quad (5)$$

where $\langle G_r \rangle_{\max}$, $\langle G_r \rangle_{\min}$ are the largest and smallest values of $\langle G_r \rangle$ for the twenty residues, respectively. With this definition, the $\langle g_r \rangle$ are scaled between $+1$ (the highest propensity to be buried) and -1 (the highest propensity to be exposed). The $\langle g \rangle$ plot, Fig. 2a, $\langle g \rangle$ values, displays the PB values for the twenty amino acid residues.

Many different scales have been proposed to characterize the extent by which amino acid residues avoid contact with water [6,20–33]. Here we compare a number of different scales and discuss possible sources of differences between them. To make this comparison easier we used the maximum and minimum values of each scale to convert the data into $+1$, -1 range, as we did it in (5). The normalized propensity scales (here we use propensity as a generic term) are presented in Figs. 2a and 2b. We employ the notation of ref. [6] for the different scales. The different scales are named after their authors.

The scales in Fig. 2a are based on statistical studies of real 3D structures of proteins. The experiment-based group, Fig. 2b, deals mostly with side-chain analogs (small molecules that incorporate side-chains of the amino acids). The most noticeable differences between the two groups are for CYS and PRO residues. One would expect CYS to be less hydrophobic than VAL, ILE, LEU because the nonpolar side-chain of CYS is smaller, and it is the case for the scales in Fig. 2b. On the other hand, the propensity of CYS to be buried is amongst the highest ones when based on crystal structures, Fig. 2a. Opposite situation takes place for PRO which prefers to stay on the surface of water soluble proteins, Fig. 2a, but shows much higher hydrophobicity in experiments on small molecules, Fig. 2b. There

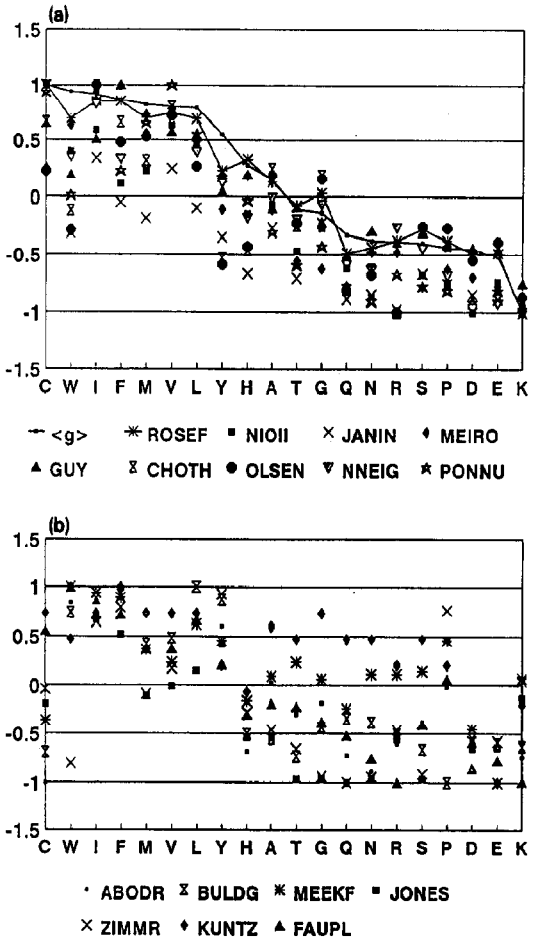

Fig. 2. (a) PB scales of amino acid residues based on theoretical treatment of X-ray data: $\langle g \rangle$ stands for the scale obtained in this paper; ROSEF [5]; NIOII [29]; JANIN [26]; MEIRO [27]; GUY [10]; CHOTH [3]; OLSEN [30]; NNEIG [6]; PONNU [4]. (b) PB scales of amino acid residues based on experiments: ABODR [22]; BULDG [23]; MEEK [28]; JONES [24]; ZIMMR [20]; KUNTZ [21]; FAUPL [31].

are also noticeable differences in dispersion of points belonging to different scales for some residues. Thus propensities of TRP and PHE in Fig. 2a are more dispersed than these in Fig. 2b, while propensities of THR, ASN, ARG, SER, and LYS are more dispersed in Fig. 2b.

The differences between different scales can be attributed to the particular techniques that were used to obtain the scales. The differences among statistically derived scales, Fig. 2a, result partially from different definitions of what is buried and exposed, and partially from the choice of which proteins comprised the particular set. Some of the early scales were calculated from small sets of molecules that made the results less reliable. Some of the small sets included water insoluble proteins, such as crambin, that could distort the derived scales.

Note that experimental propensities, Fig. 2b (KUNTZ), of nonpolar residues and CYS, based on the measurements of the amount of water that does not freeze when an aqueous macromolecular solution is rapidly frozen and then equilibrated at -20 to -40°C, differ from those based on small molecules (other scales of Fig. 2b). The corresponding scale [21] singles out three charged residues (ASP, GLU, LYS), however, it differentiates very slightly (with respect to other scales) between polar and nonpolar residues. Different experimental techniques are used to obtain information on the solubility of the various residues, and it is difficult to bring them to the same standard condition. Some residues have very low solubility in nonpolar solvents, while others are almost insoluble in water. Often the experimental measurements are corrected by additional calculations, especially for CYS and PRO, e.g. [31]. The use of different techniques, together with different interpretations of the results are responsible for some of the differences between the experimental scales, Fig. 2b.

It is feasible to derive some scale based on X-ray data without any additional corrections (as long as we understand that there are other influencing factors in addition to hydrophobicity). Among the scales we analyze, the one proposed by Rose et al. [5], Fig. 2a (ROSEF), is the closest to our scale of propensity, Figs. 2a ($\langle g \rangle$). Points of each of the two scales are connected by lines in Fig. 2a. These scales are close because they are based on conceptually similar approaches. Rose et al. deal with surface accessibilities of amino acid residues, while we calculate complementary characteristics using the notion of Gaussian neighborhood [13]. We, as Rose et al., do not make any assumptions about which residues should belong to the surface and which to the core of proteins. Initial assumptions such as that

residues are considered exposed if they have more than 80% (or sometimes 95%) of their surface exposed, may cause additional discrepancies.

Because the scales give information related to solubility in water, the major differences between the two groups may be attributed to differences between the behavior of proteins and the small molecules in water. Therefore, the lesser propensity of CYS to be shielded, obtained in many such experiments, Fig. 2b, is reasonable and can be attributed to hydrophobicity alone. The experiments mostly measure the true hydrophobic characteristics of CYS. It is known that the hydrophobic behavior of CYS in proteins is enhanced by S–S bridging between these residues. PRO is more hydrophobic than polar and charged residues, Fig. 2b; its high accessibility in proteins (and lower propensity to be buried, Fig. 2a) is known to be caused by its rigid structure which fits easily into turns of globular proteins. We call scales based on X-ray structures PB scales to emphasize that the spatial distribution of the residues in proteins is influenced by other factors beside hydrophobicity.

We want to emphasize that hydrophobicity is a manifestation of certain properties of both the solute molecule and water as the solvent. All the experiments measure hydrophobicity-related characteristics that always include other factors (e.g. the interaction with a reference solvent when the partition coefficients are measured). The interaction with solvent differs for different residues, especially for residues that belong to different groups (aliphatic, aromatic, polar, charged).

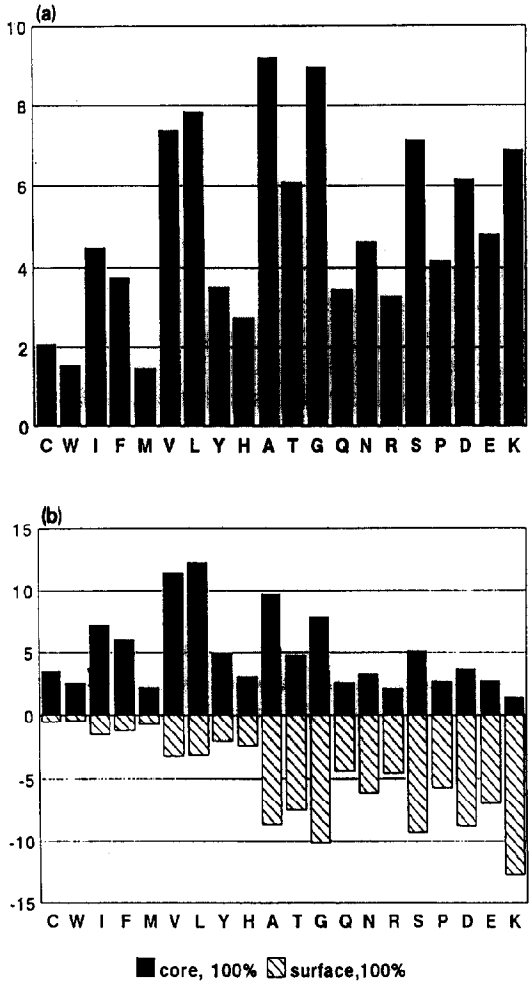

Fig. 3. (a) Average residue composition of proteins. The numbers (in %) are frequencies of occurrence of amino acid residues in water soluble proteins. (b) Average residue compositions of surface and core portions of proteins. Frequencies of occurrence (in %) near the surface and in the core. The histograms with positive numbers are for core residues; he sum of the numbers is 100%. The histograms with negative numbers are for the surface residues; the sum is 100%.

5. Residue compositions of core and surface of water soluble proteins

The probability to have a G value above or below some specific value can be used to find the percentage of each residue populating the surface and core portions of protein molecules. $G = 42.5$ is approximately the average over all the residues of proteins we dealt with, and it is equidistant between the two most probable values for the two groups of the residues, Figs. 1a and 1b.

The twenty standard residues are not present in proteins with the same frequency. The percentage of each residue, based on the 64 molecules, is shown in Fig. 3a. These frequencies of occurrence of the residues are typical for any large set of water soluble globular proteins, compare [1]. The scaled product of the frequency of occurrence f_i of residue i, Fig. 3a, and the proba-

bility p_i of that residue to have its $G < 42.5$ gives its surface proportion among other surface residues; the product of f_i with the probability to have $G \geqslant 42.5$, i.e. $(1 - p_i)$ gives the corresponding core proportion. Thus, the data of Fig. 3a present the composition of an average water soluble protein, whereas Fig. 3b shows the compositions of the surface and the core of this average 'molecule'. The values are scaled, such that both the sum of all surface residues, and the sum of all core residues is 100%. (We plotted the proper scale for core residues and the negative of the scale for surface residues to better visualize the two compositions.) This average protein has 48.9% of its residues on the surface and 51.3% in the core. To obtain the distribution ratio between surface and core for each residue type the surface fractions, in Fig. 3b, should be multiplied by 0.489 and the core fractions by 0.513. The relative fractions (all buried residues versus all exposed ones) of real proteins depend on their size. For smaller proteins the ratio of exposed to buried residues is greater than for the larger ones, but both the surface and the core composition of any water soluble protein should reflect the main features of their average characteristics (Fig. 3b). It is worth emphasizing that the % surface compositions of different residues, Fig. 3b, do not correlate directly with their propensities, Figs. 2a and 2b). (The same is true for core residues.) Thus, there are more ALA, GLY residues in the surface layer than ASN, ARG, GLU, even though the latter have higher propensities to be exposed; similarly there are more ALA, GLY in the core than TRP, MET which have higher propensities to be buried. This apparent paradox results from two different factors. The propensity to be buried is defined as the averaged ratio of any of the 20 residues, say ALA, to be buried to all the ALA residues, whereas the percentage of this residue on the surface (or in the core) depends on both its PB value and its frequency of occurrence in proteins in general. GLY and ALA are essential components of both core and surface portions of proteins.

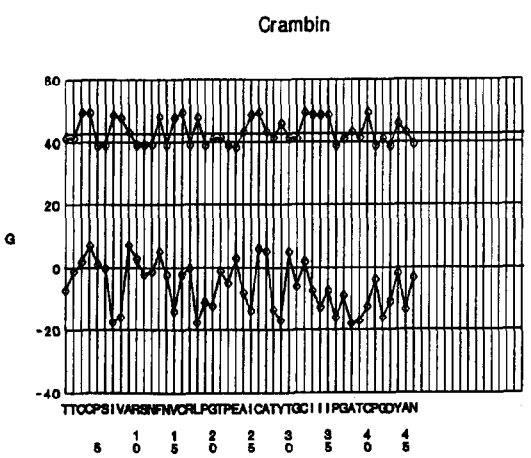

Fig. 4. Deviations of G-neighborhood values for residues of Crambin from average values for the residues (lower plot); average G-neighborhood values plotted against sequence of crambin (upper plot).

6. Crambin versus water soluble proteins

In this paragraph we apply information on $\langle G \rangle$ for the twenty standard amino acids to analyze possible difference between water soluble and insoluble proteins both on their 3D structures as well as their sequences. It was shown in [13] that nonpolar residues VAL, ILE, LEU of Crambin occupy positions mostly in the outer shell of the molecule. In contrast, in water soluble proteins they most frequently occupy inner core positions. We have calculated, based on X-ray data, G-neighborhood characteristics for the residues of the 3D structures of Crambin and Myoglobin. The plots of deviations of the residues' G-neighborhood values from their average values $\langle G \rangle$, calculated for water soluble proteins, are depicted on the lower parts of Figs. 4, 5. The plot of crambin is noticeably shifted below the zero (reference) line whereas the dots representing myoglobin are distributed more evenly above and below the reference. The large nonpolar residues of crambin have large negative deviations. This means that these residues form an essential part of the outer layer of this molecule, presumably typical for the structure of proteins that have stable, compact structures in alcohol solvents.

So far we dealt only with known 3D structures of proteins. This allowed us to obtain important information about spatial and free energy distributions of amino acid residues in water soluble proteins. Having the average G-neighborhood values, $\langle G \rangle$, of the twenty amino acid residues made it possible to look for anticipated differences at the sequence level between crambin and water soluble proteins (without accounting for 3D structures of the molecules). Thus plots for crambin and myoglobin sequences (upper curves in Figs. 4 and 5) show that the average values $\langle G \rangle$ for residues of myoglobin are distributed more evenly below and above the value 42.5 (the average over all residues in water soluble proteins), whereas the values for crambin are shifted upwards. The points that belong to the upper part of the plots are for residues that prefer to be buried inside water soluble proteins. Crambin has too many such residues and this makes it insoluble in water. It is impossible to rearrange this small molecule in such a way that it would shield enough nonpolar residues to make it soluble in water.

We believe that the approach of Gaussian neighborhood or some modification of it may provide a tool to look for the sections of membrane proteins that penetrate the membranes. This approach can also be used to study features of the molecules that form membrane channels.

7. Conclusion

G values characterizing residues' surroundings do not take into consideration the nature of the surroundings. Not any surrounding with the same G would bring the same free energy. Figs. 1a–1d do not tell anything about the nature of surrounding residues that make those G values. It is possible to get the same G surrounding a polar residue by exclusively nonpolar ones, or surrounding a nonpolar residue by polar ones. In real proteins the nonpolar residues face polar groups of the polar residue with –CO or –NH groups of their backbones not with their nonpolar groups; and polar residues face nonpolar sidechains with their nonpolar atoms or groups. Not any surrounding with the same G is described by the probability profiles, Figs. 1a and 1b, and potentials of mean force, Figs. 1c and 1d, because only equilibrium structures or conformations of stable protein molecules were used to obtain these potentials. Even for real protein situations the same residue with the same G value but different surroundings may have different free energies. The mean force potentials for the amino acid residues clarify (in a general sense) the role of different residues in protein structure formation and stabilization. However, these potentials cannot be used when searching for a stable structure of a protein. The mean force potentials comprise

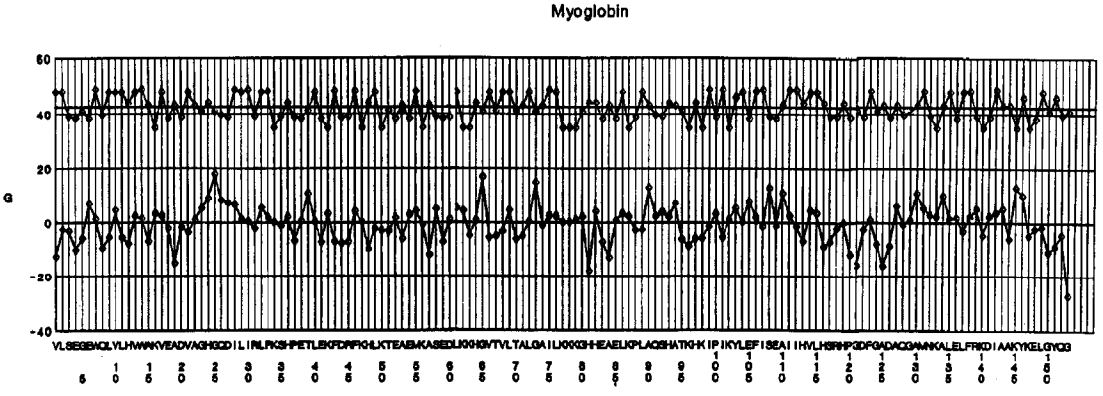

Fig. 5. Deviations of G-neighborhood values for residues of Myoglobin from average values for the residues (lower plot); average G-neighborhood values plotted against sequence of myoglobin (upper plot).

averaged information on intramolecular interactions as well as on the interaction with the solvent. The probability density profiles, Figs. 1a and 1b, and free energy profiles, Figs. 1c and 1d), for the twenty amino acid residues provide information averaged over different sequence-structure situations. The equations (3) and (4) could provide absolutely correct assessments of free energy changes if residues forming proteins were independent of each other. The more residues are involved the better are conditions for applying our approach. When there is a thermodynamically equilibrium state with many residues involved, different local features balance each other making this approach more applicable. Avbelj [12] calculated the autocorrelation function of residue burial along a sequence and has shown that the correlation coefficient for the first neighbors is only 0.24. The author concludes that there is satisfactory independence of residue burial, although there are some residues that do not meet the requirement of independence, e.g. PRO.

The functional form of the mean force potential for the twenty amino acid residues is given in an accompaniing paper where it is used for research of antigen–antibody recognition.

To search for an unknown 3D structure of a protein a set of potentials that deal explicitly with specific interactions (H-bonding, electrostatic) between parts of the protein as well as between the solvent and the residues of the protein is required. We believe the G-neighborhood approach could provide a means for developing such potentials for the solvent–residue interaction, and work is in progress.

References

[1] P.Y. Chou, in: Prediction of protein structure and the principles of protein conformation. Prediction of protein structural classes from amino acid compositions, ed. G.D. Fasman (Plenum Press, New York, 1989) p. 549.
[2] B. Lee and F.M. Richards, J. Mol. Biol. 55 (1971) 379.
[3] C. Chothia, J. Mol. Biol. 105 (1976) 1.
[4] P.K. Ponnuswamy, M. Prabhakaran and P. Manavalan, Biochim. Biophys. Acta 623 (1980) 301.
[5] G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee and M.H. Zehfus, Science 229 (1985) 834.
[6] J.L. Cornette, K.B. Cease, H. Margalit, J.L. Spouge, J.A. Berzofsky and C. DeLisi, J. Mol. Biol. 195 (1987) 659.
[7] W. Kauzmann, Advan. Protein Chem. 14 (1959) 1.
[8] C. Tanford, The hydrophobic effect (Wiley, New York, 1973).
[9] F.M. Richards, Ann. Rev. Biophys. Bioeng. 6 (1977) 151.
[10] H.R. Guy, Biophys. J. 47 (1985) 61.
[11] K. Dill 29 (1990) 7133.
[12] F. Avbelj, Biochemistry 31 (1992) 6290.
[13] V.V. Nauchitel and R.L. Somorjai, Proteins Struct. Funct. Genet. 15 (1993) 50.
[14] A. Bondi, J. Phys. Chem. 68 (1964) 441.
[15] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F.Jr. Meyer, M.D. Brice J.R. Rogers, O. Kennard, T. Shimanouchi and M. Tasumi, J. Mol. Biol. 112 (1977) 535.
[16] E.E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle, J. Weng, "Protein Data Bank" in Crystallographic Databases-Information Content, Software Systems, Scientific Applications, eds. F.H. Allen, G. Bergerhoff and R. Sievers (Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987) p. 107.
[17] C.L. Brooks, M. Karplus and B.M. Pettitt, Proteins. A theoretical perspective of dynamics, structure and thermodynamics (Wiley, New York, 1988).
[18] C.H. Chothia and J. Janin, Nature 256 (1975) 705.
[19] A.A. Rashin, Progr. Biophys. Mol. Biol. (1993) 74.
[20] J.M. Zimmerman, J. Theoret. Biol. 21 (1968) 170.
[21] I.D. Kuntz, J. Am. Chem. Soc. 93 (1971) 514.
[22] A.A. Aboderin, Intern. J. Biochem. 2 (1971) 537.
[23] H.B. Bull and K. Breese, Arch. Biochem. Biophys. 161 (1974) 665.
[24] D.D.J. Jones, Theoret. Biol. 50 (1975) 167.
[25] D.H. Wertz and H.A. Scheraga, Macromolecules 11 (1978) 9.
[26] J. Janin, Nature 277 (1979) 491.
[27] H. Meirovitch, S. Rackovsky and H.A. Scheraga, Macromolecules 13 (1980) 1398.
[28] J.L. Meek, Proc. Nat. Acad. Sci. USA 77 (1980) 1632.
[29] K. Nishikawa and T. Ooi, Intern. J. Pept. Protein Res. 16 (1980) 19.
[30] K.W. Olsen, Biochim. Biophys. Acta 622 (1980) 259.
[31] J. Fauchere and V. Pliska, Eur. J. Med. Chem. 18 (1983) 369.
[32] R.M. Sweet and D. Eisenberg, J. Mol. Biol. 171 (1983) 479.
[33] S. Miyazawa and R.L. Jernigan, Macromolecules 18 (1985) 534.