Protein Engineering vol.15 no.8 pp.659–667, 2002

Quantifying the accessible surface area of protein residues in their local environment

Uttamkumar Samanta[1,2], Ranjit P.Bahadur[2,3] and Pinak Chakrabarti[1,3,4]

[1]Bioinformatics Centre and [3]Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Calcutta 700 054, India

[2]U.Samanta and R.Bahadur contributed equally to this work

[4]To whom correspondence should be addressed.
E-mail: pinak@boseinst.ernet.in

The quantification of the packing of residues in proteins and docking of ligands to macromolecules is important in understanding protein stability and drug design. The number of atoms in contact (within a distance of 4.5 Å) can be used to describe the local environment of a residue. As this number increases, the accessible surface area (ASA) of the residue decreases exponentially and the variation can be described in terms of an exponential equation of the form $y = a_1\exp(-x/a_2)$, each residue having its own set of parameters a_1 and a_2, which also depend on whether the whole residue or just the side chain is considered. Hydrophobic and hydrophilic residues can be distinguished on the basis of both the average number of surrounding atoms and the variation of ASA. For a given number of partner atoms, a comparison of the observed ASA with the expected value obtained from the equation provides a method of assessing the goodness of packing of the residue in a protein structure or its importance in the binding of a ligand. The equation provides a method to estimate the ASA of a protein molecule and the average relative accessibilities of different residues, the latter being inversely correlated with hydrophobicity values.

Keywords: accessible surface area/binding efficiency/hydrophobicity/packing of residues/residue partner number

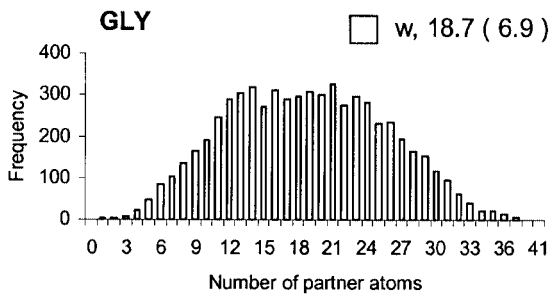
Introduction

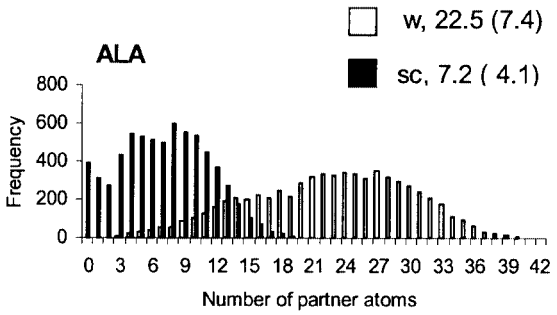
One notable feature of protein structures is their compactness (Chothia, 1975; Richards, 1977). The hydrophobic effect is thought to provide the bulk of the free energy needed to fold a protein into a compact state (Kauzmann, 1959; Dill, 1990). There is a linear relationship between the surface areas of amino acid residues (in a standard state) and the free energy changes associated with the transfer of the amino acids from water to an organic solvent (Chothia, 1974; Rose *et al.*, 1985; Sharp *et al.*, 1991). The packing of the interior residues was quantitatively evaluated by Richards (Richards, 1974) and Finney (Finney, 1975) in terms of packing density, which is the ratio of the volume of the van der Waals envelope of the molecule to the volume of space that it actually occupies. In crystals of small organic molecules (with a variety of molecular shapes), and also in proteins, this number is usually in the range 0.70–0.78, suggesting that in the folded state of proteins the groups are as closely packed as they can be. In spite of the average picture that the packing density conveys, it is not sufficient to analyze the efficiency of packing of individual residues in the structure (whether on the surface or the interior). Two factors are needed to address this issue: the optimum number of atoms (or residues) around a given residue and its accessible surface area. Samanta *et al.* (Samanta *et al.*, 2000) analyzed the environment of tryptophan residues in proteins and found that there is an exponential relationship between the number of residues (partners) in contact with the Trp residue (considering only the indole part or the whole residue) and the accessible surface area. Such a relationship is useful in assessing not only the efficiency of packing of a Trp residue in its local protein milieu, but also its role in binding a non-proteinous molecule (such as the substrate or a cofactor) and in protein–protein complexes (Samanta and Chakrabarti, 2001). However, it was realized that to be applicable to within proteins, as well as between a protein and a small molecule, it is necessary to consider partners in terms of atoms (and not residues), and in this paper we analyze all known protein structures to derive the number of partner atoms in contact with all the 20 amino acid residues (taking the whole residue and the side chain separately) and the relationship with the accessible surface area. The equations derived can be used to calculate parameters that reflect the hydrophobic character of the residues and to estimate the expected surface areas which can be compared with the observed values to assess the efficiency of packing of individual residues.

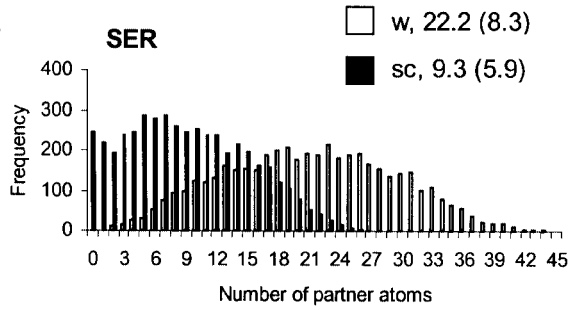
Materials and methods

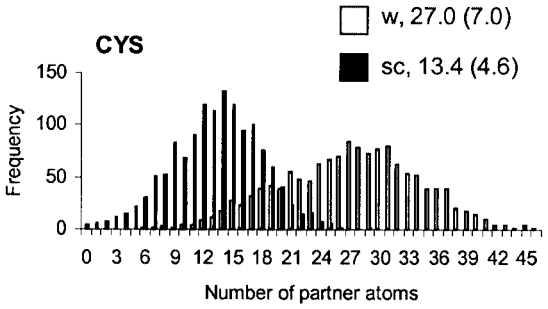
Atom coordinates were obtained from the Protein Data Bank (PDB), now operated by the Research Collaboratory for Structural Bioinformatics (Berman *et al.*, 2000); 432 chains (in 418 files) were selected using PDB_SELECT (Hobohm and Sander, 1994) from PDB files (as of March 2000) with an *R*-factor ≤20%, a resolution ≤2.0 Å and sequence identity <25%. To restrict the analysis to well-ordered residues, all polypeptide chains with >40% of atoms with temperature factor (*B*-factor) >30 Å^2 were excluded. Moreover, even at the level of residues, those with >40% atoms with *B*-factor >30 Å^2 were also not considered as the central residue (for which partners were to be found). All protein atoms (in the PDB file) in contact with any atom of the central residue (or its side chain) were first found provided they satisfied the following conditions: (i) distance >2.0 Å, but ≤4.5 Å (the lower limit was to exclude bonded atoms or those with unreasonably short contact), (ii) the occupancy factor is 1.0 and the *B*-factor ≤30 Å^2 and (iii) the atoms do not belong to the same residue or the main-chain atoms of the flanking residues. Then only the unique atoms in the list were retained and their count gave the number of partner atoms for the central residue.

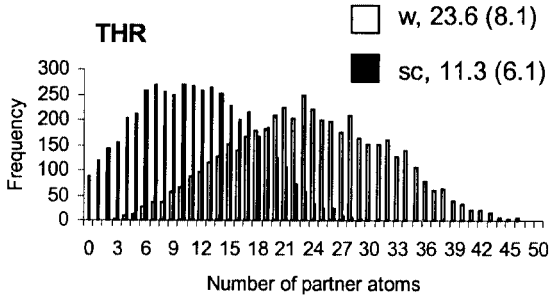
The solvent accessible surface area (ASA) was computed using the program ACCESS (Hubbard, 1992), which is an implementation of the Lee and Richards (1971) algorithm.

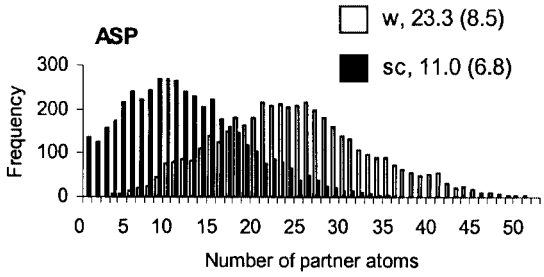


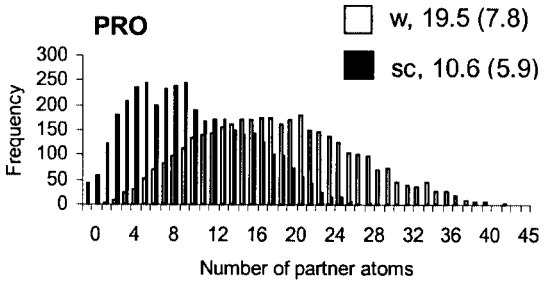


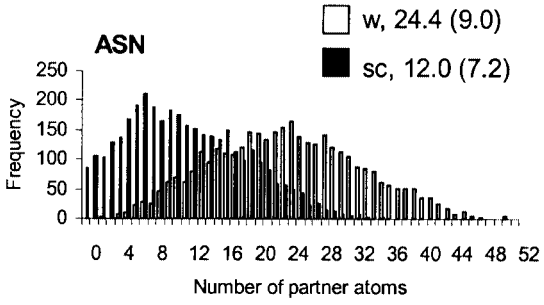


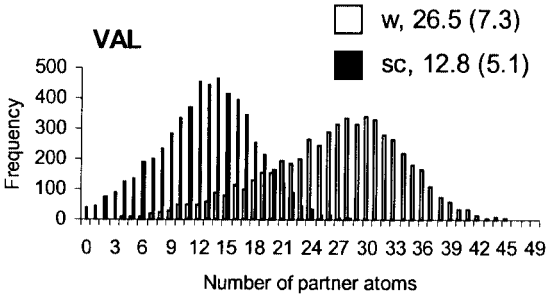


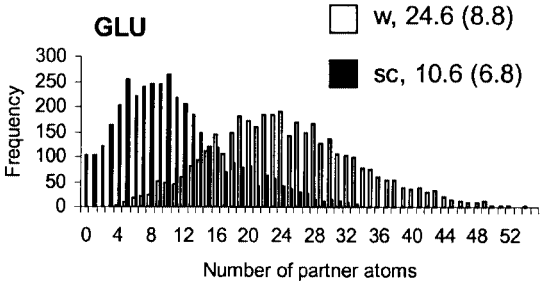












continued

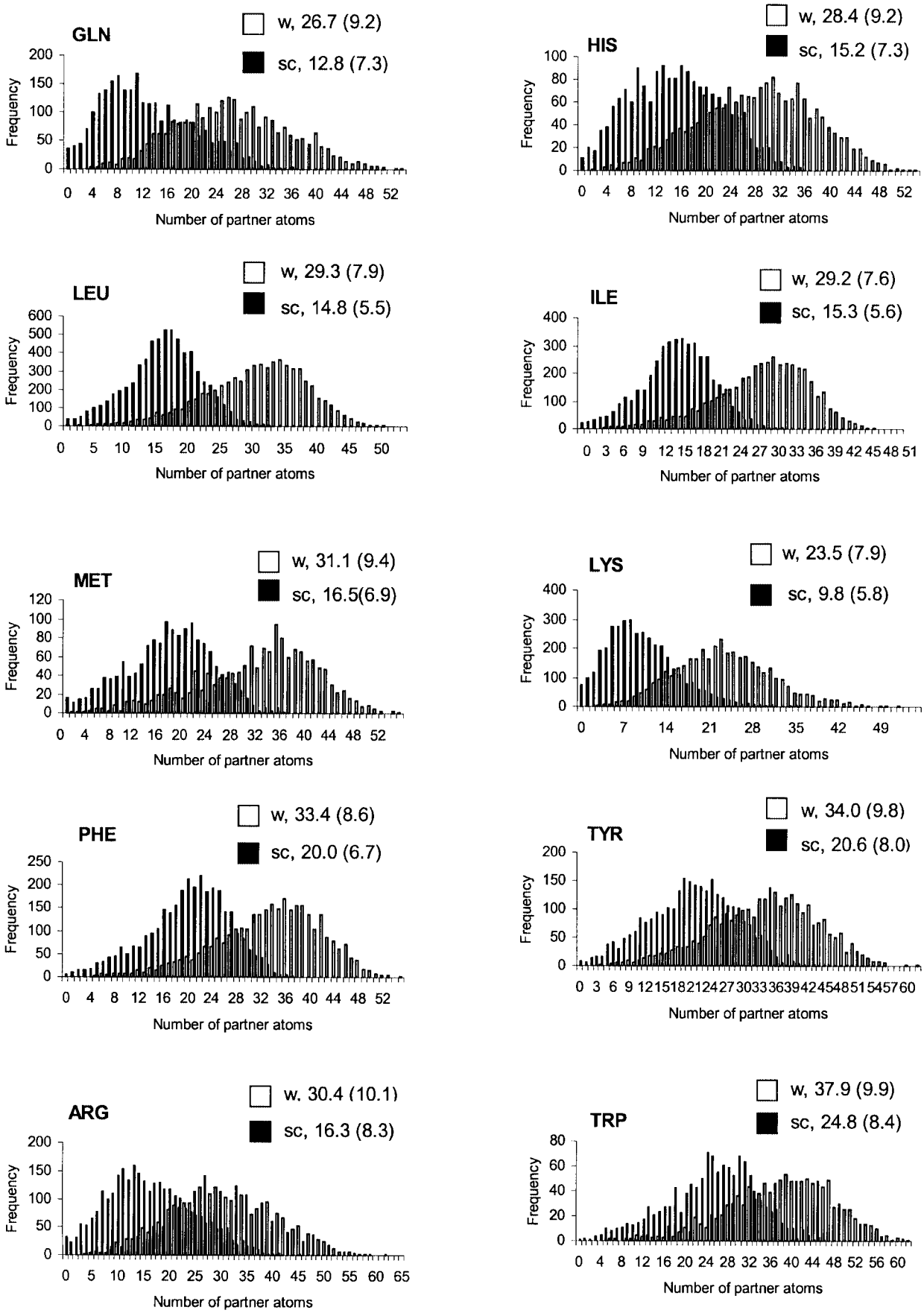


Fig. 1. Histograms showing the distribution of the number of partner atoms in contact with amino acid residues in proteins. The bars corresponding to the side chain (sc) and the whole residue (w) appear grouped in the left and the right sides of the plot, respectively, and the average value (with standard deviation) of partner number in each case is shown. The residues are arranged according to their volume (Chothia, 1975).

The ASA for each residue (and its side chain) were calculated without considering the non-proteinous constituents of the PDB files. To represent the standard state, the ASA of residue X in both the tripeptides, Gly–X–Gly and Ala–X–Ala, was calculated, the models having being made in the extended conformation ($\phi = \psi = 180°$) using the program InsightII (Molecular Simulations, San Diego, CA). The relative accessibility of a residue is the ASA normalized by the standard ASA and in this work, for reasons discussed below, the Gly-based model tripeptide provided a more suitable standard. In addition to the 'observed' value, an 'estimate' of the relative accessibility was also obtained when the ASA of the residue was derived from the equation (discussed below) relating ASA to its partner number.

For a given residue, the average value of the ASA (and the standard deviation) at a given partner number were calculated. For all residues other than Gly the partners were defined for the side chain also and, consequently, there were 39 sets of data. When the average ASA values (y) were plotted against the partner numbers (x), the points could be fitted into an exponential equation of the form $y = a_1\exp(-x/a_2)$. When the plot was extrapolated to $x = 0$, the value obtained was closer to that obtained for the residue in the tripeptide Gly–X–Gly, which was thus taken as the standard state (A_0) (the expected value when the protein is in the unfolded state with the minimum contact with the rest of the molecule). However, in general for the whole residue, the calculated value (at $x = 0$) was about 20% greater than A_0 and hence it was decided to fix a_1 at the value of the standard state and fit only the parameter a_2 against the observed data. Though the quality of fit was slightly worse (judged from the R^2 value), this set of parameters was used for further calculations as they gave the expected value of A_0.

To find the optimum value of the cut-off distance, the following calculations were performed at different values (4.0, 4.5, 5.0, 6.0 and 7.0 Å). (i) The average partner numbers for all residue types (whole) in the database were determined. (ii) For each residue the average accessible surface area <ASA> at each partner number was found. Then an exponential fitting was performed to relate <ASA> to the whole range of partner numbers. Hence for each cut-off distance we had a set of 20 equations for all residue types. (iii) For a given PDB file, we took each residue in turn, calculated its partner number in the structure and from the equations (found above) calculated its ASA. We did this for all the residues in the file and summed to obtain the total 'calculated' ASA for the molecule. This was compared with the 'observed' ASA (obtained using ACCESS) to give the parameter R_A = ASA(calc)/ASA(obs). (iv) R_A values were computed for all the PDB chains (139 in total) (for which not more than 5% of the residues were rejected based on the selection criteria enumerated at the beginning) and their average, $<R_A>$, was found. The cut-off distance giving a value (0.93) closest to 1 was 4.5 Å.

A similar experiment was carried out to decide if the bonded atoms were to be included in the calculation of partner numbers. The calculations were repeated (only at the cut-off distance of 4.5 Å) by considering even the backbone atoms of the neighbouring residues as partners. When these partner numbers were fitted to ASA, the R^2 value of fitting was considerably poorer than that obtained when these atoms were excluded when calculating partner numbers. Also, when this set of equations was used, $<R_A>$ was 1.3 ($\pm$2), which differed by a greater amount from the ideal value of 1.

The codes for the PDB files used are given in the Appendix.

Results and discussion

Number of partner atoms around each residue

The local environment of a residue can be characterized by the number of atoms in contact with it, within a distance of 4.5 Å (the reason behind the use of this limit is elaborated in a separate section).

The distributions of the numbers of atoms surrounding each residue and just the side chain are presented in Figure 1, where the residues are ordered according to increasing volume as given by Chothia (Chothia, 1975). The largest residue, Trp, has the highest average number (38) of contacts and the smallest, Gly, the least (19). For all non-Gly residues, the difference in numbers for the whole residue and the side chain (i.e. the number due to the main-chain atoms) is ~13. Pro, with a pyrrolidine ring encompassing both the side chain and the main chain, has a smaller number of partners than suggested by its size; Pro and Gly have similar values. Pro residues are generally restricted to loop regions which are more frequently on the surface of a protein, thus leading to a lower than expected atom neighbour count. Of the three hydrophobic residues having nearly identical volumes, Leu, Ile and Met, the last residue has a slightly higher number of contacts, possibly indicating a greater inclination of the Met S, as compared with an aliphatic carbon atom, to interact with other groups (Pal and Chakrabarti, 2001). Between two residues, Met and Lys, which are also of comparable volume, the latter has a smaller number of contacts, indicating that a hydrophilic residue has a smaller number of protein atoms around it than a hydrophobic residue of equal size. However, although more hydrophilic than Leu and Ile, which follow it, His has about the same number of contacts. This, together with the large values observed for Phe and Tyr, may suggest that the aromatic residues are better packed than the aliphatic residues.

Although the shape of the distribution is symmetric for all the residues when the whole residue is considered, the hydrophilic residues can be distinguished by the asymmetric (skewed) nature of the distributions for their side chains. Taking two residues, Ser and Lys, as examples, it can be seen that the distributions have positive skewness, with the peaks shifting towards lower partner numbers. Pro behaves more like a hydrophilic residue, whereas Arg, which has a planar group in its side chain has a rather symmetric distribution, like other planar aromatic residues. Thus, a hydrophilic residue can be distinguished from a hydrophobic residue of comparable

Table I. Correlation coefficients between the average partner numbers and other contact numbers

	BJ	PBV	KZB(w)	KZB(sc)	ZK	W	SC
NO	−0.43	0.94	0.63	0.53	0.23	0.56	0.51
BJ		−0.26	−0.76	−0.84	0.49	−0.73	−0.77
PBV			0.57	0.41	0.39	0.51	0.41
KZB(w)				0.96	−0.38	0.99	0.96
KZB(sc)					−0.49	0.92	0.98
ZK						−0.42	−0.54
W							0.94

The references and the conditions [cut-off distance (Å), whether the parameters were obtained using the whole residue (w), the side-chain (sc) or only the C^α atoms (CA)] used to derive the different sets of values are as follows: NO, Nishikawa and Ooi (1980) [8, CA]; BJ, Bahar and Jernigan (1997) [6.4, CA]; PBV, Panjikar *et al.* (1997) [6.5, CA]; KZB(w/sc), Karlin *et al.* (1999) [5, w/sc]; ZK, Zhang and Kim (2000) [6.5, CA, only helical residues]; W/SC, this work [4.5, w/sc].

size on the basis of both the number of surrounding atoms and their distribution.

The number of partner atoms delineated here is similar to the number of ligand atoms in the first coordination sphere of a metal ion in inorganic chemistry. Starting with Nishikawa and Ooi (Nishikawa and Ooi, 1980), many workers have analysed spatial neighbours in terms of contact numbers (Panjikar *et al.*, 1997; Karlin *et al.*, 1999; Zhang and Kim, 2000). These have also been used in deriving potentials of mean force for interactions among residues, for application in threading sequences into the correct fold (Miyazawa and Jernigan, 1996; Bahar and Jernigan, 1997). Correlation coefficients between mean partner numbers and some of these earlier publications are given in Table I. As could be expected, the two sets of values calculated by us are highly correlated between themselves, as well as with the values of Karlin *et al.* (Karlin *et al.*, 1999), who also considered all surrounding atoms around a residue, but at a slightly longer distance of 5 Å. The correlation is rather poor (and inverse) with the values of Bahar and Jernigan (Bahar and Jernigan, 1997) and Zhang and Kim (Zhang and Kim, 2000), who employed a low-resolution model in which a residue is represented by a single interaction site located at the C^{α} position. Irrespective of the type of residue, on average about six non-bonded residues are found within a sphere of radius 6.5 Å centred on it and, as such, these values are less discriminating. An all-atom model, on the other hand, using a cut-off length larger than the van der Waals contact distance, provides coordination numbers which reflect the nature of the residue in a much more realistic manner.

Variation of the accessible surface area with the number of partner atoms

The mean values of the accessible surface areas (ASAs) of the residues (considering either the whole residue or only the side chain) at different values of the partner number are plotted for two representative residues in Figure 2. It is seen that the standard deviations of the mean values decrease with increasing number of partners and the variation of ASA can be adequately represented by an exponential form (Table II). In the majority of cases, extrapolation to $x = 0$ (i.e. no partner) leads to a value which is close to the value for the residue (X) obtained in the fragment Gly–X–Gly in an extended conformation. As a result, the ASA value of the residue flanked by Gly residues (and not Ala) on either side can be taken as the standard state (A_0) for the residue. This is also justified by the fact that in our methodology the side-chain atoms (starting at C^{β}) of the flanking residues are legitimate contenders to be counted as partners for the central residue X.

The decrease in the observed ASA with partner number has distinct features typical of the hydrophobicity of the residue. Taking residues of comparable volume, Met and Lys, the decrease is slower with the more hydrophilic residue (Lys). A slower fall is also noticed when the whole residue is considered, as compared with the corresponding side chain alone. The fitted parameter a_2 embodies the rate of decrease, with a larger value indicating a slower decay. Thus between two equal-sized residues, the more hydrophilic one has a larger a_2, as does the value for the whole residue in relation to that for the side chain (see note added in proof). The fitting to the exponential

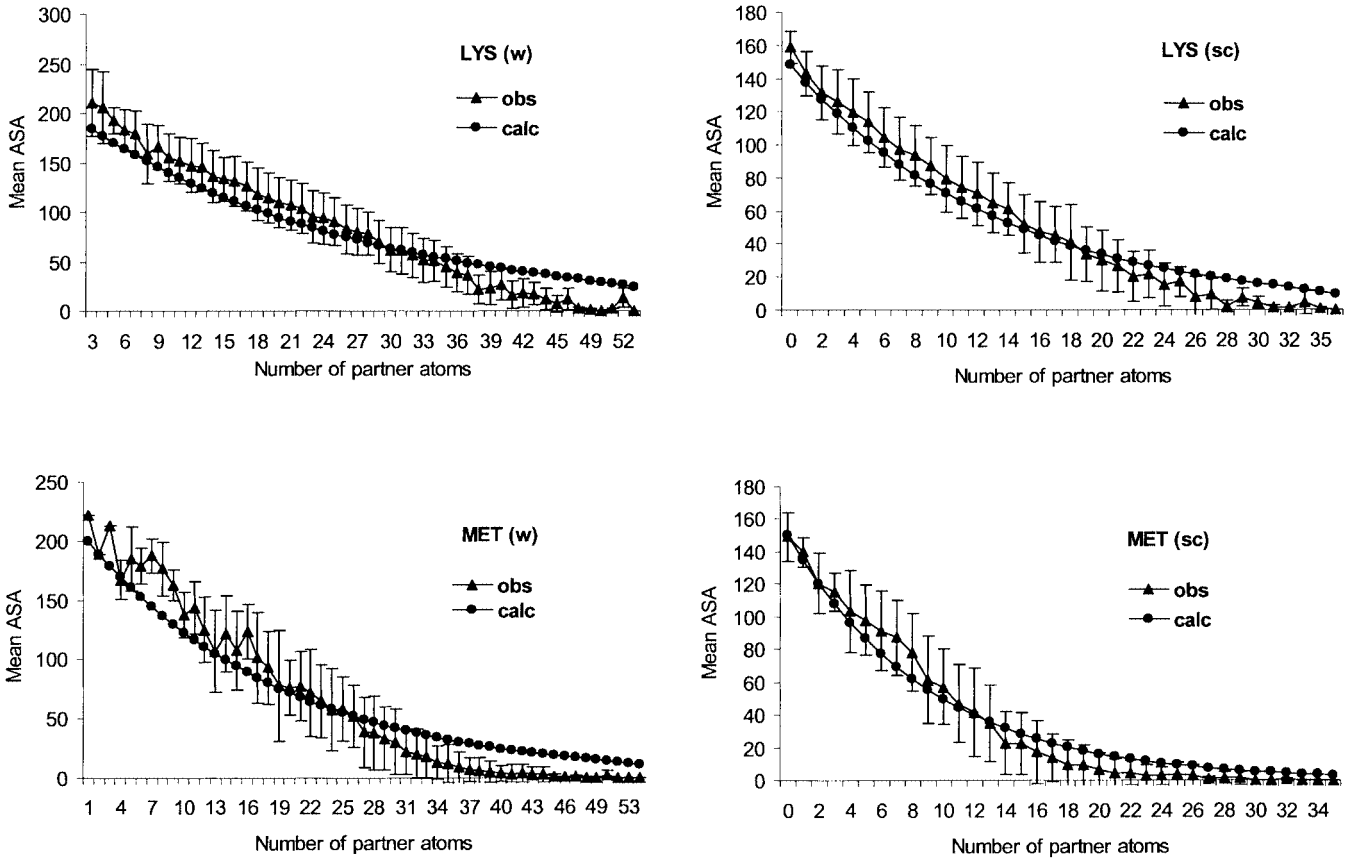


Fig. 2. Variation of the mean accessible surface areas, ASA (Å^2) (vertical bars representing standard deviations) with the number of partner atoms for two typical residues (whole or w) and their side chains (sc). The curve corresponding to the best least-squares fit, $y = a_1\exp(-x/a2)$ (with a_1 and a_2 given in Table II) is also shown.

Table II. Accessible surface areas (ASA) for the whole residue (w) or the side chain (sc) for different residues (X) in model peptides and the parameters of the equation that describes the variation of mean ASA with the number of partner atoms

Residue	ASA (Å^2)[a]				Fitted parameters					
	Gly–X–Gly		Ala–X–Ala		Whole (w)			Side chain (sc)		
	w	sc	w	sc	a_1	a_2	R^2	a_1	a_2	R^2
Gly	83.91	–	70.27	–	83.91	15.10	0.67			
					101.75	*12.67*	*0.89*			
Ala	116.40	55.40	102.68	55.28	116.40	15.80	0.84	55.40	4.97	0.93
					147.20	*12.76*	*0.94*	*65.61*	*4.31*	*0.97*
Ser	125.68	69.08	111.97	68.97	125.68	17.36	0.88	69.08	7.68	0.96
					154.58	*14.34*	*0.96*	*76.22*	*7.03*	*0.98*
Cys	141.48	82.07	127.72	81.91	141.48	11.67	0.84	82.07	6.09	0.96
					121.17	*13.26*	*0.82*	*75.22*	*6.58*	*0.96*
Thr	148.06	88.62	134.28	88.45	148.06	18.20	0.85	88.62	9.80	0.92
					193.59	*14.21*	*0.96*	*104.79*	*8.42*	*0.97*
Asp	155.37	97.80	141.61	97.66	155.37	20.61	0.88	97.80	11.51	0.93
					186.00	*17.38*	*0.95*	*113.22*	*10.07*	*0.97*
Pro	144.80	106.44	126.78	98.23	144.80	17.07	0.89	106.44	9.32	0.94
					173.32	*14.44*	*0.95*	*115.25*	*8.68*	*0.96*
Asn	168.87	109.92	155.22	109.87	168.87	20.22	0.89	109.92	11.12	0.95
					202.70	*17.04*	*0.96*	*112.19*	*10.10*	*0.97*
Val	162.24	103.12	148.51	103.00	162.24	14.61	0.91	103.12	7.01	0.96
					188.38	*12.81*	*0.94*	*118.68*	*6.20*	*0.98*
Glu	187.16	132.53	173.46	132.42	187.16	22.46	0.86	132.53	11.36	0.97
					238.50	*17.90*	*0.96*	*139.96*	*10.79*	*0.98*
Gln	189.17	129.68	175.42	129.52	189.17	23.51	0.84	129.68	12.76	0.94
					228.02	*19.67*	*0.92*	*146.75*	*11.38*	*0.97*
His	198.51	141.27	184.79	141.17	198.51	20.73	0.88	141.27	11.46	0.96
					237.37	*17.59*	*0.94*	*156.70*	*10.40*	*0.98*
Leu	197.99	141.52	184.33	141.47	197.99	16.26	0.90	141.52	7.60	0.98
					249.87	*13.17*	*0.97*	*151.58*	*7.14*	*0.98*
Ile	189.95	130.71	176.22	130.58	189.95	15.72	0.86	130.71	8.01	0.96
					185.16	*16.08*	*0.86*	*147.08*	*7.22*	*0.98*
Met	210.55	150.39	196.87	150.32	210.55	18.59	0.89	150.39	9.02	0.96
					254.13	*15.66*	*0.95*	*161.66*	*8.45*	*0.97*
Lys	207.49	147.99	193.73	147.83	207.49	25.44	0.86	147.99	13.51	0.95
					259.16	*20.32*	*0.96*	*164.83*	*12.17*	*0.97*
Phe	223.29	164.18	209.64	164.14	223.29	18.50	0.91	164.18	10.61	0.96
					244.15	*17.06*	*0.93*	*182.97*	*9.62*	*0.98*
Tyr	238.30	180.03	224.68	180.01	238.30	20.52	0.91	180.03	12.40	0.97
					248.23	*19.78*	*0.92*	*191.72*	*11.70*	*0.98*
Arg	249.26	190.24	235.45	190.04	249.26	25.67	0.84	190.24	14.82	0.97
					270.27	*23.68*	*0.87*	*203.44*	*13.91*	*0.98*
Trp	265.42	209.62	251.78	209.57	265.42	22.09	0.85	209.62	14.10	0.95
					329.01	*18.43*	*0.91*	*232.89*	*12.80*	*0.97*

The equation is of the form $y = a_1\exp(-x/a_2)$. Two sets of values are given (see Materials and methods): in the second set (in italics) both a_1 and a_2 have been fitted, whereas in the first (which has been used for further calculations) a_1 was fixed at A_0.
[a]The value for the Gly-based peptide is taken as the standard state (A_0).

curve is much better for the side chain than the whole residue, as can be seen from the R^2 values. Generally for the latter, the calculated ASA values are lower than the observed values at lower partner numbers and higher at higher partner numbers, with the crossover occurring at or slightly beyond the average partner numbers.

Estimates of average relative accessibilities of residues

Using the equation relating ASA and partner number, it is possible to calculate the ASA ($<A>_{calc}$) corresponding to the average number of partners for the whole residue and the side chain. These estimates of average ASA compare very well with the means of the observed values ($<A>_{obs}$) obtained using the standard algorithm of Lee and Richards (Lee and Richards, 1971) (Table III); the hydrophobic residues show better agreement than the hydrophilic residues and the whole residue compared with the side chain. On dividing $<A>_{calc}$ by the standard value (taken as the ASA of the residue X in the peptide Gly–X–Gly, as discussed earlier) one obtains an estimate of the average relative accessibility for the residue in a protein structure. Comparison of the values for the whole residue with those for the side chain reveals an interesting feature. For polar residues the side chain shows higher accessibility than the whole residue (with Lys being the most prominent). This is along the expected line, as for these residues the more hydrophilic part is in the side chain, which is thus more exposed than the rest of the residue. For hydrophobic residues the values are nearly identical, with the side chain in some cases having a slightly lower value than that for the corresponding residue taken as the whole. (However, this trend becomes much clearer if one does a similar calculation with a smaller cut-off distance of 4.0 Å; data not shown.)

Table III. Estimates of accessible surface area (ASA) and relative accessibility when the whole residue (w) and the side chain (sc) are surrounded by the average number of partner atoms

Residue	Average partner number[a]		$<A>$, average ASA (Å^2)				$<A>_{calc}/A_0$[d]	
	w	sc	$<A>_{obs}$[b]		$<A>_{calc}$[c]		w	sc
			w	sc	w	sc		
Gly	18.7		27 (25)		24.30		29.0	
Ala	22.5	7.2	28 (31)	18 (21)	28.04	13.03	24.1	23.5
Ser	22.2	9.3	39 (33)	28 (24)	35.05	20.69	27.9	30.0
Cys	27.0	13.4	17 (21)	10 (16)	13.97	9.14	9.9	11.1
Thr	23.6	11.3	44 (36)	36 (30)	40.42	28.06	27.3	31.7
Asp	23.3	11.0	58 (37)	48 (31)	50.07	37.61	32.2	38.5
Pro	19.5	10.6	54 (40)	43 (33)	46.12	34.02	31.9	32.0
Asn	24.4	12.0	58 (41)	48 (34)	50.60	37.43	30.0	34.1
Val	26.5	12.8	24 (33)	19 (27)	26.40	16.68	16.3	16.2
Glu	24.6	10.6	73 (42)	64 (36)	62.73	51.95	33.5	39.2
Gln	26.7	12.8	69 (43)	60 (38)	60.76	47.63	32.1	36.7
His	28.4	15.2	54 (45)	47 (39)	50.49	37.63	25.4	26.6
Leu	29.3	14.8	29 (38)	23 (33)	32.70	20.29	16.5	14.3
Ile	29.2	15.3	25 (35)	21 (31)	29.68	19.33	15.6	14.8
Met	31.1	16.5	36 (46)	29 (38)	39.43	24.14	18.7	16.1
Lys	23.5	9.8	96 (43)	85 (37)	82.48	71.91	39.8	48.6
Phe	33.4	20.0	31 (40)	25 (35)	36.69	24.97	16.4	15.2
Tyr	34.0	20.6	46 (45)	40 (40)	45.52	34.13	19.1	19.0
Arg	30.4	16.3	86 (53)	77 (48)	76.30	63.25	30.6	33.2
Trp	37.9	24.8	44 (48)	38 (44)	47.78	36.21	18.0	17.3

[a]From Figure 1.
[b]Observed. Standard deviations are in parentheses.
[c]Calculated at the average partner number using the equations given in Table II.
[d]Estimates of % relative accessibility, at average partner number. A_0 in the standard state is the value corresponding to the peptide Gly–X–Gly in Table II.

Table IV. Correlation coefficients between the average relative accessibilities of different residues and some representative hydrophobicity data

Hydrophobicity scale of	Average relative accessibility	
	Whole residue	Side chain
Fauchère and Pliska (1983)	−0.89 (−0.96)[a]	−0.91 (−0.96)[a]
Kyte and Doolittle (1982)	−0.83	−0.85
Miller *et al.* (1987)	−0.90 (−0.93)[b]	−0.94 (−0.96)[b]
Ponnuswamy *et al.* (1980)	−0.93	−0.92
Wolfenden *et al.* (1981)	−0.64 (−0.82)[c]	−0.72 (−0.82)[c]
Eisenberg *et al.* (1982)	−0.77 (−0.89)[d]	−0.82 (−0.92)[d]

Average relative accessibilities are the estimated values from Table III. Gly is excluded when the side chain is considered and Wolfenden *et al.* (Wolfenden *et al.*, 1981) do not have the value for Pro. In parentheses are the correlation coefficients on excluding [a]Arg, Pro and Trp, [b]Arg and Gly, [c]Ala, Arg and Gly and [d]Arg, Gly and Pro.

It may be mentioned that Rose *et al.* (Rose *et al.*, 1985) directly calculated the mean fractional area buried (which is $1 - <A>/A_0$). The relative accessibility (%) ($100<A>/A_0$) is estimated here indirectly using an estimated value of $<A>$ from consideration of the average number of partner atoms. The two sets of values are in good agreement (with a correlation coefficient of –0.97).

Reduction of surface area on folding and correlation with hydrophobicity

The accessible area of a fully extended polypeptide chain is reduced by a factor of about three on folding into the native structure (Chothia, 1975). Considering individual residues (Table III), it is found that in general the estimates of the relative accessibilities of hydrophilic residues are ~30 (Lys being the most exposed), whereas for hydrophobic residues the values are in the range 10–20. This clear demarcation between the two types of residues led us to examine whether there is any relationship between relative accessibility and a few hydrophobicity scales taken from the literature (Cornette *et al.*, 1987). Indeed, it was found that there is an inverse correlation (Table IV). Thus the fraction of a residue that is buried on folding is directly proportional to its hydrophobicity. Of the hydrophobicity scales that were tried, the match is poor for the one due to Wolfenden *et al.* (Wolfenden *et al.*, 1981), which measures the distribution of amino acid side chains between dilute aqueous solutions and the vapour phase. However, the correlation improves on the exclusion of Gly, Ala and Arg. The other experimental scale of Fauchère and Pliska (Fauchère and Pliska, 1983) using octanol–water distribution measurements is in excellent agreement (Figure 3), as also is the statistical scale of Miller *et al.* (Miller *et al.*, 1987) based on the distribution of residues between the surface and interior of proteins; the match with the latter scale shows further improvement on exclusion of Arg and Gly, two residues almost from the two ends of the size spectrum.

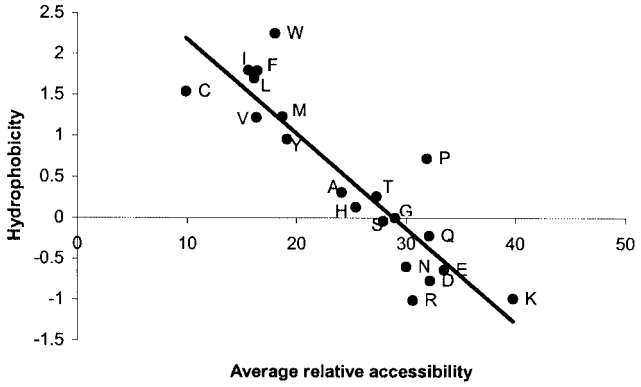


Fig. 3. Plot of the estimates of the average relative accessibilities of different residues (whole) against the hydrophobicity values (in kcal/mol) due to Fauchère and Pliska (Fauchère and Pliska, 1983). A least-squares line, $y = -0.11x + 3.19$ ($R^2 = 0.91$), can be obtained by excluding Arg, Pro and Trp from the regression analysis.

Table V. Average R_A values (and the standard deviations) at different cut-off distances

Distance (Å)	$<R_A>$
4.0	0.76 (±0.09)
4.5	0.93 (±0.11)
5.0	1.32 (±0.19)
6.0	1.82 (±0.30)
7.0	2.17 (±0.40)

$<R_A>$ is defined in Materials and methods.

The optimum cut-off distance and the estimation of the accessible surface area of a protein molecule

Two factors went into consideration in the choice of the cut-off distance. First, we wanted to include those atoms which have van der Waals and other specific non-covalent interactions with a given residue. For delineating hydrogen bonds a distance of <3.9 Å is generally used (McDonald and Thornton, 1994) and weak interactions (such as the C–H···O hydrogen bond) are known to extend to about 4.0 Å (Desiraju, 1996). When the closest contact distance between any two atoms of a pair of aromatic rings is within 4.5 Å, there is a binding interaction between the rings (McGaughey *et al.*, 1998). As such, a limiting distance of 4.5 Å was deemed to be reasonable. On the other hand, as we were attempting to correlate the partner number with the accessible surface area and the latter can be affected owing to the screening of solvent by other atoms not in immediate contact (i.e., at a longer distance), we tried a number of distances (4.0, 4.5, 5.0, 6.0 and 7.0 Å). As discussed in Materials and methods, the optimum value was selected by finding out the partner numbers (at different limiting ranges) of all the residues in a polypeptide chain and then converting these to the corresponding ASA values (using the appropriate exponential equations); the distance which provided the best match (the perfect match would give a value of 1) of the 'calculated' ASA of the molecule to its 'observed' ASA was found to be 4.5 Å (Table V). An $<R_A>$ of 0.93 suggests that we have an alternative procedure to compute the ASA of a protein molecule that is within 7% the true value.

Implications and summary

Although proteins are characterized by well-packed cores, it has been difficult to achieve a unique structure for proteins designed *de novo* or when the core residues are randomized (Betz *et al.*, 1993; Axe *et al.*, 1996). It is therefore important to have a detailed representation of the local packing quality (Word *et al.*, 1999). In this paper we have characterized two fundamental features of packing and their interrelationship, viz., the number of atoms (partners) in contact with a residue and how they cover the accessible surface area (ASA) of the residue. In addition to considering the whole residue, its side chain has also been taken into account separately. The hydrophobic residues can be distinguished from the hydrophilic residues using both the partner number and the accessible surface area. Between two residues of comparable volume, the hydrophilic residue has a smaller number of partner atoms around it and the histogram showing the distribution of partner numbers is more skewed than for the hydrophobic residues (Figure 1). The decrease in ASA with increase in partner number takes place more slowly for hydrophilic residues than for hydrophobic residues (Figure 2). This variation can be represented by an exponential equation, $y = a_1\exp(-x/a_2)$ where each residue has its own set of a_1 and a_2 depending on whether the whole residue or just the side chain is considered (Table II). When the equations are used to estimate the average relative accessibilities (Table III), the values compare very well with the observed values and are found to be inversely correlated with the hydrophobicities (Table IV and Figure 3).

The average (or expected, based on the equation derived here) ASA of a residue corresponding to a given number of partners provides a means to assess the efficiency of packing of a residue. If the observed ASA is more than the expected value, it suggests that the partners have been less efficient in covering the surface of the residue, whereas a smaller observed value indicates a tighter packing by the surrounding atoms. It is conceivable that the number of partners will depend on, in addition to the size and type of the residue, the location in the tertiary fold and as the ASA depends on this number it is likely that not all residues can be packed equally well in a given location. Hence the residue-specific exponential relationship between the partner number and ASA may offer a new algorithm for a threading procedure (to identify the possible fold for a given sequence) that is conceptually different from other methods of protein-fold recognition (Torda, 1997) and we are working on its development. Quantifying the steric fit of a ligand to a macromolecule is equivalent to quantifying the internal packing in protein and the aforementioned equations can be used to assess the importance of different residues in the binding site of a ligand, as has been attempted in the case of Trp (Samanta and Chakrabarti, 2001). Finally, we have developed a procedure for estimating the ASA of a protein chain, which is within 7% of the value obtained using the protocol of Lee and Richards (Lee and Richards 1971) (Table V).

Acknowledgements

We are grateful to an anonymous reviewer for many comments, especially that relating to the use of a number of cut-off distances. We thank the Department of Biotechnology for computational facilities and a research associateship (to U.S.) and the Indo-French Centre for the Promotion of Advanced Research for a fellowship (to R.P.B.).

References

Axe,D.D., Foster,N.W. and Fersht,A.R. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 5590–5594.

Bahar,I. and Jernigan,R.L. (1997) *J. Mol. Biol.*, **266**, 195–214.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res*., **28**, 235–242.
Betz,S.F., Raleigh,D.P. and Degrado,W.F. (1993) *Curr. Opin. Struct. Biol*., **3**, 601–610.
Chothia,C. (1974) *Nature*, **248**, 338–339.
Chothia,C. (1975) *Nature*, **254**, 304–308.
Cornette,J.L., Cease,K.B., Margalit,H., Spouge,J.L., Berzsofsky,J.A. and DeLisi,C. (1987) *J. Mol. Biol*., **195**, 659–685.
Desiraju,G.R. (1996) *Acc. Chem. Res.*, **29**, 441–449.
Dill,K.A. (1990) *Biochemistry*, **29**, 7133–7155.
Eisenberg,D., Weiss,R.M., Terwilliger,T.C. and Wilcox,W. (1982) *Faraday Symp. Chem. Soc*., **17**, 109–120.
Fauchère,J. and Pliska,V. (1983) *Eur. J. Med. Chem.*, **18**, 369–375.
Finney,J.L. (1975). *J. Mol. Biol*., **96**, 721–732.
Hobohm,U. and Sander,C. (1994) *Protein Sci.*, **3**, 522–524.
Hubbard,S. (1992) *ACCESS: a Program for Calculating Accessibilities.* Department of Biochemistry and Molecular Biology, University College London, London.
Karlin,S., Zhu,Z.-Y. and Baud,F. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 12500–12505.
Kauzmann,W. (1959) *Adv. Protein Chem.*, **14**, 1–63.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Lee,B. and Richards,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.
McDonald,I.K. and Thornton,J.M. (1994) *J. Mol. Biol*., **238**, 777–793.
McGaughey,G.B., Gagne,M. and Rappe,A.K. (1998) *J. Biol. Chem*., **273**, 15458–15463.
Miller,S., Janin J., Lesk,A.M. and Chothia,C. (1987) *J. Mol. Biol.*, **196**, 641–656.
Miyazawa,S. and Jernigan,R.L. (1996) *J. Mol. Biol*., **256**, 623–644.
Nishikawa,K. and Ooi,T. (1980) *Int. J. Pept. Protein Res*., **16**, 19–32.
Pal,D. and Chakrabarti,P. (2001) *J. Biomol. Struct. Dyn*. **19**, 115–128.
Panjikar,S.K., Biswas,M. and Vishveshwara,S. (1997) *Acta Crystallogr.*, **D53**, 627–637.
Ponnuswamy,P.K., Prabhakaran,M. and Manavalan,P. (1980) *Biochim. Biophys. Acta,* **623**, 301–316.
Richards,F.M. (1974) *J. Mol. Biol.*, **82**, 1–14.
Richards,F.M. (1977) *Annu. Rev. Biophys. Bioeng.*, **6**, 151–176.
Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) *Science*, **229**, 834–838.
Samanta,U. and Chakrabarti,P. (2001) *Protein Eng*., **14**, 7–15.
Samanta,U., Pal,D. and Chakrabarti,P. (2000) *Proteins: Struct. Funct. Genet.*, **38**, 288–300.
Torda,A.E. (1997) *Curr. Opin. Struct. Biol*., **7**, 200–205.
Sharp,K.A., Nicholls,A., Friedman,R. and Honig,B. (1991) *Biochemistry*, **30**, 9686–9697.
Wolfenden,R., Andersson,L., Cullis,P.M. and Southgate,C.C.B. (1981) *Biochemistry*, **20**, 849–855.
Word,J.M., Lovell,S.C., LaBean,T.H., Taylor,H.C., Zalis,M.E., Presley,B.K., Richardson,J.S. and Richardson,D.C. (1999) *J. Mol. Biol*., **285**, 1711–1733.
Zhang,C. and Kim,S.-H. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 2550–2555.

Received May22, 2001; revised April 26, 2002; accepted May 21, 2002

Appendix

Codes for the PDB files used

The subunit identifier, if present, is given as the fifth letter. 1A1IA, 1A1YI, 1A28B, 1A2PA, 1A2ZA, 1A34A, 1A3C_, 1A48_, 1A4IB, 1A6M_, 1A7S_, 1A8D_, 1A8E_, 1A9XB, 1ABA_, 1ADOA, 1ADS_, 1AE9B, 1AFWA, 1AGQA, 1AHO_, 1AIE_, 1ALVA, 1AMF_, 1AMM_, 1AMX_, 1AOCA, 1AOHB, 1APYA, 1AQB_, 1ARV_, 1ATLA, 1AUN_, 1AVWB, 1AXN_, 1AY7B, 1AYFA, 1AYL_, 1AYOA, 1AZO_, 1B0NA, 1B0NB, 1B0UA, 1B0YA, 1B16A, 1B2VA, 1B3AA, 1B4KB, 1B5EA, 1B65A, 1B67A, 1B6A_, 1B6G_, 1B7CA, 1B8OA, 1B93A, 1BA8A, 1BABB, 1BBHA, 1BBPA, 1BDO_, 1BE9A, 1BEA_, 1BEC_, 1BENB, 1BF6A, 1BFG_, 1BFTA, 1BG6_, 1BGF_, 1BI5A, 1BJ7_, 1BK0_, 1BK7A, 1BKRA, 1BQCA, 1BRT_, 1BS4A, 1BS9_, 1BSMA, 1BTN_, 1BU7A, 1BX4A, 1BX7_, 1BXAA, 1BXOA, 1BY2_, 1BYI_, 1BYQA, 1BYRA, 1C24A, 1C2AA, 1C3D_, 1C3MA, 1C3WA, 1C52_, 1CBN_, 1CC8A, 1CCZA, 1CEQA, 1CEWI, 1CEX_, 1CF9A, 1CFB_, 1CG6A, 1CKAA, 1CLEA, 1CMBA, 1CNV_, 1COZA, 1CPO_, 1CPQ_, 1CQYA, 1CS1A, 1CTJ_, 1CTQA, 1CV8_, 1CVL_, 1CXQA, 1CXYA, 1CY5A, 1CYDA, 1CYO_, 1CZFA, 1CZPA, 1D3VA, 1D7PM, 1D9CB, 1DBWB, 1DCIA, 1DCS_, 1DF4A, 1DFNA, 1DG9A, 1DGWY, 1DHN_, 1DI6A, 1DIN_, 1DLFH, 1DLFL, 1DOKA, 1DOSA, 1DOZA, 1DPSD, 1DPTA, 1DUN_, 1DXGA, 1ECD_, 1ECPA, 1EDG_, 1EDMB, 1EGPA, 1EUS_, 1EXTB, 1EZM_, 1FCE_, 1FIPA, 1FIT_, 1FLEI, 1FLTV, 1FLTY, 1FNA_, 1FRPA, 1FUS_, 1FVKA, 1G3P_, 1GCI_, 1GDOB, 1GOF_, 1GP1A, 1GPEA, 1GSA_, 1GUQA, 1HFC_, 1HFES, 1HKA_, 1HLEB, 1HOE_, 1HTRP, 1HUUA, 1HXN_, 1IAB_, 1ICFI, 1IDAA, 1IFC_, 1IIBA, 1ISUA, 1IXH_, 1JDW_, 1JER_, 1JHGA, 1KNB_, 1KOE_, 1KP6A, 1KPTA, 1KVEA, 1KVEB, 1LAM_, 1LATA, 1LBU_, 1LCL_, 1LKFA, 1LKKA, 1LOUA, 1LTSA, 1LTSC, 1LUCA, 1MAI_, 1MDC_, 1MFMA, 1MGTA, 1MKAA, 1MLA_, 1MML_, 1MOF_, 1MOLA, 1MOQ_, 1MPGA, 1MRJ_, 1MROA, 1MROB, 1MROC, 1MSI_, 1MSK_, 1MTYB, 1MTYG, 1MUGA, 1MUN_, 1NAR_, 1NBCA, 1NCOA, 1NIF_, 1NKD_, 1NKR_, 1NLS_, 1NOX_, 1NP4A, 1NPK_, 1NULB, 1OAA_, 1OBWA, 1OPD_, 1OPY_, 1ORC_, 1OTFA, 1PBE_, 1PCFA, 1PDO_, 1PGS_, 1PHF_, 1PLC_, 1PNE_, 1POA_, 1POC_, 1PPN_, 1PSRA, 1PTQ_, 1PTY_, 1PYMB, 1QB7A, 1QCXA, 1QCZA, 1QD1A, 1QDDA, 1QFMA, 1QFOA, 1QGIA, 1QGWB, 1QGWD, 1QH4A, 1QH5A, 1QH8A, 1QH8B, 1QHFA, 1QJ4A, 1QJ8A, 1QKSA, 1QMPD, 1QQ4A, 1QQ5A, 1QQP1, 1QQP2, 1QQP4, 1QREA, 1QRRA, 1QSGA, 1QTSA, 1QTWA, 1QU9A, 1RB9_, 1RCF_, 1REC_, 1REGY, 1RGEA, 1RHS_, 1RIE_, 1RZL_, 1SCJB, 1SFP_, 1SGPI, 1SLUA, 1SMD_, 1SMLA, 1SRA_, 1SUR_, 1SVFA, 1SVFB, 1SVPA, 1SVY_, 1SWUB, 1TAFA, 1TAXA, 1TC1A, 1TEN_, 1TGXA, 1TIB_, 1TIF_, 1TL2A, 1TML_, 1TOAA, 1TTBA, 1TVXB, 1U9AA, 1UBPA, 1UBPB, 1UNKA, 1UOX_, 1VCAA, 1VFRA, 1VFYA, 1VHH_, 1VID_, 1VIE_, 1VLS_, 1VNS_, 1VSRA, 1WAB_, 1WAPB, 1WDCA, 1WHI_, 1WHO_, 1WWCA, 1XNB_, 1YACA, 1YAGG, 1YCC_, 1YGE_, 1YTBA, 2A0B_, 2ABK_, 2ACY_, 2AHJC, 2ARCB, 2AYH_, 2BC2A, 2BOPA, 2BOSA, 2CBP_, 2CCYA, 2CHSA, 2CPGA, 2CTC_, 2DRI_, 2DTR_, 2EBN_, 2EBOA, 2END_, 2ERL_, 2FDN_, 2GAR_, 2GDM_, 2HBG_, 2HDDB, 2HFT_, 2HMZA, 2IGD_, 2ILK_, 2KNT_, 2LISA, 2MSBB, 2MYR_, 2NLRA, 2PII_, 2PSPA, 2PTH_, 2PVBA, 2QWC_, 2RN2_, 2SAK_, 2SICI, 2SN3_, 2SNS_, 2SPCA, 2TNFA, 2TPSA, 2TRXA, 2TYSB, 2UBPC, 3CHBD, 3CHY_, 3CLA_, 3CYR_, 3ENG_, 3EZMA, 3GRS_, 3LZT_, 3PTE_, 3PVIA, 3PYP_, 3SDHA, 3SEB_, 3SIL_, 3STDA, 3TDT_, 3TSS_, 3VUB_, 4EUGA, 4MT2_, 5HPGA, 5PTI_, 6CEL_, 6GSVA, 7A3HA, 7RSA_, 8ABP_, 8PRKA, 9WGAA, 16PK_, 19HCA, 153L_, 256BA, 451C_.

Note added in proof

There is a good inverse correlation between a_2 and the hydrophobicity scales used in Table IV, the correlation coefficients being:

	FAUPL	KYTDO	MILLER	PONNU	WOLF	EISEN
(w)	–0.61	–0.79	–0.76	–0.61	–0.84	–0.77
(sc)	–0.40	–0.75	–0.71	–0.50	–0.81	–0.63