JOURNAL OF MOLECULAR RECOGNITION, VOL. 8, 334–344 (1995)

Occluded Molecular Surface: Analysis of Protein Packing

N. Pattabiraman* and K. B. Ward
Laboratory for the Structure of Matter, Code 6030, Naval Research Laboratory, Washington, DC 20375-5000, USA

P. J. Fleming
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520-8114, USA

We describe a novel method to calculate the packing interactions in protein structural models. The method calculates the interatomic occluded surface areas for each atom in the protein model. The identification of, and degree of interaction with, neighboring atoms is accomplished by extending surface normals from a dot surface of each atom to the point of intersection with neighboring atoms. The combined occluded and non-occluded surface areas may be normalized for the amino acid composition of the protein providing a single parameter, the normalized protein surface ratio, which is diagnostic for native-like structures. Individual residues in the model which are in infrequent occluded surface environments may be identified. The method provides a means to explicitly describe packing densities and packing environments of individual atoms in a protein model. Finally, the method allows estimation of the complementarity between any interacting molecules, for example a ligand binding to a receptor.

Keywords: **occluded surface; protein packing; misfolded protein structure.**

Introduction

An understanding of the folding and stability of proteins depends on an accurate evaluation of the interaction of buried groups with each other and the interaction of exposed groups with the solvent. The basic concepts of packing density, packing interaction, atomic volume, and solvent accessible surface have been applied to this question for many years (Richards, 1977; Rose *et al.*, 1985; Bere *et al.*, 1991; Nicholls *et al.*, 1991). A number of approaches to the calculation of amino acid residue packing and interaction in proteins have been devised. Atomic volume calculations have demonstrated that proteins are densely packed on the average (Richards, 1974; Harpaz *et al.*, 1994) and this parameter may be used to evaluate the quality of structural models of proteins (Ponder and Richards, 1987; Gregoret and Cohen, 1990). Specific interaction between buried groups in proteins has also been extensively studied, especially as a means for predicting protein folds (Wodak and Rooman, 1993). Many of the methods for calculating interaction between specific buried groups have used simple distance measurements between atomic coordinates or between virtual group coordinates.

Sippl and colleagues have calculated a pairwise distance distribution for amino acid pairs of varying sequence separation (Hendlich *et al.*, 1990). The relative frequencies of amino acid pair distances found in a database of protein structures was transformed to potentials of mean force. These potentials show distinct distributions for specific amino acid pairs in native proteins and thus may be used to identify correctly folded proteins with some exceptions. Casari and Sippl (1992) have extended this approach to identify the importance of hydrophobic contacts in native protein folds. Based on this criterion alone they were able to identify the native fold of 55 out of 68 proteins when each respective sequence was mounted on all conformations in the database used. Other methods have been reported which depend on pairwise distance or frequency distributions (Tanaka and Scheraga, 1976; Bryant and Amzel, 1986; Hendlich *et al.*, 1990; Casari and Sippl, 1992; Marorov and Crippen, 1992; Coloros and Yeates, 1993). Although some of these methods demonstrate the ability to identify most misfolded protein models, they do not explicitly consider the actual packing of residues in the protein or do not allow identification of specific residues that are in unusual packing environments.

The use of packing density as a criterion for evaluating model protein structures was developed explicitly by Gregoret and Cohen (1990). In their method each amino acid residue is represented by a sphere which is allowed to 'grow' in size until contact with its neighbors occurs. In this way the average allowed space (ideal radius) for each particular amino acid in a set of protein structures was calculated. Another method which specifically includes a consideration of residue packing environment is the three-dimensional profile method described by Luthy *et al.* (1992). Detailed investigation of atomic contact arrangements have been reported (Singh and Thornton, 1990; Vriend and Sander, 1993). In both of these methods the relative spatial orientation

* Author to whom correspondence should be addressed. Present address: Frederick Biomedical Supercomputing Center, SAIC, NCI-Frederick Cancer Research and Development Center, PO Box B, Frederick, MD 21702-1201, USA.

CCC 0952-3499/95/060334-11

Received 4 May 1995
Accepted (revised) 18 August 1995

of residue–residue interactions was explored. But both methods define interaction based on simple distance criteria without a determination of actual shared contact surface. An estimate of shared contact surface was made by Behe *et al.* (1991) by calculation of the solvent-accessible area buried between paris of proximal residues. Results of this analysis were interpreted to indicate that no preferred pair-wise interactions exist, at least in terms of preferred packing interactions.

Atomic packing and shape complementarity are also important parameters for calculating ligand–protein and protein–protein interaction. Jiang and Kim (1991) described a 'soft docking' technique which matches a ligand and target surface by finding areas of antiparallel surface normals. In a related approach Lawrence and Colman (1993) have defined a shape correlation statistic which uses molecular dot surfaces and associated unit normals to calculate complementarity. They used this method to demonstrate that antigen/antibody interfaces have relatively poor shape complementarity.

We have developed a novel method of analyzing atomic packing in protein structural models which explicitly determines the packing interactions between buried groups. This method uses molecular dot surfaces with associated normals of defined length to identify a molecular occluded surface. The information obtained allows one to estimate a variety of useful parameters for analysis of protein structure. Using this method one is able to estimate atomic packing density, areas of large and small interatomic space, atomic surface occluded by other atoms, and identify atoms with van der Walls interaction. Adaptations of the method will be useful for evaluating protein structural models, predicting mutation effects, predicting ligand–receptor interactions, and studying subtle changes in protein packing.

Experimental

Data sets

Fifty-five non-homologous protein structures of more than 50 residues, with a crystallographic resolution of at least 2.0 Å and *R* factor of 0.20 or less were chosen from two published lists of proteins with differing structures; one by Gregoret and Cohen (1990) and the other by Hobohm *et al.* (1992). This data set of protein structures included only monomeric proteins. For the calculations described in this work only the first of any alternate residue conformations were considered. In addition no heteroatom records were considered in the calculations described in this work. The Brookhaven Protein Data Bank entry code and the corresponding name of the protein for the data set are listed in columns 1 and 2, respectively, in Table 1.

Coordinate files for misfolded protein models were obtained from two sources. The 25 misfolded models generated by Holm and Sanders (1992) were obtained from the EMBL database (ftp.embl-heidelberg.de). These misfolded models were generated by swapping residue sequences of pairs of proteins having an equal number of residues. Backbone conformation was retained and side chain conformation was optimized using a Monte Carlo algorithm with a simple energy function. The models were energy minimized using the program GROMOS. The four misfolded structures described by Novotny and colleagues (1984, 1988) were provided by Dr J Novotny. These misfolded models were generated by swapping the amino acid residues between an IgG light chain domain and hemerythrin. For one pair of models the side chain conformations were substituted and energy minimized using the program CHARMM; for a second pair, favorable side chain conformations were identified using the program CONGEN and the models were energy minimized with the version of CHARMM incorporated into CONGEN.

Calculation of occluded surfaces and packing parameters

Molecular surfaces are calculated for each residue in a protein model using the program MS (a program to generate molecular dotted surfaces written by Connolly (1983), which is part of the MidasPlus program (Ferrin *et al.*, 1988). The probe radius was 1.4 Å and the following atomic radii (in Å) were used: C, 1.9; N, 1.5; O, 1.4; S, 1.85. The output from MS consists of the coordinates of dots which define the molecular surface of a molecule. This surface consists of the contact and reentrant surfaces as defined by Richards (1977). The output also contains the associated surface area and unit normal vector pointing outward from the surface for each dot. These surface areas and vectors are used as described below to determine occluded surface area.

For the surface dot calculation the coordinates of each residue, without the presence of other atoms in the protein model, are used. When calculating the molecular surface of each residue, we include the atom C of the previous residue and the atom N of the following residue. The surfaces of these additional atoms are not considered in subsequent calculations of residue surface area.

Our aim is to determine which of the atomic surface dots are occluded by neighboring atoms when the residue with the molecular surface dots is placed back in the protein. For each atom in a residue the neighboring atoms within a center to center distance of 6.4 Å are identified. This value was chosen on the assumption that if an atom were farther away than 6.4 Å a water molecule could intervene and no part of the surfaces would be occluded. In contrast, an atom closer than 6.4 Å might lie in a position where it could occlude a portion of the molecular surface.

For each dot associated with the atoms we identify the intersections of the surface normal with the van der Waals sphere of the set of atoms that were identified by the distance criterion. If this surface normal intersects any such van der Waals surface, the respective dot is labeled as occluded, otherwise it is labeled as non-

Table 1. Normalized protein surface ratios[a]

Brookhaven Protein Data Bank entry[b]	Protein	$\overline{P_{sr}}$
4pti	Trypsin inhibitor bovine	0.60
2ovo	Ovomucoid third domain	0.61
2cdv	Cytochrome *c3 D. vulgraris*	0.61
1bp2	Phospholipase A2 bovine	0.63
2act	Actinidin kiwifruit	0.65
2sns	Staphylococcal nuclease *S. aureus*	0.65
4rxn	Rubredoxin *C. pasteurian*	0.65
6ldh	Apo M4 lactate dehydrogenase dogfish	0.65
45lc	Cytochrome c551 *P. aeruginosa*	*0.65*
1ton	*Tonin rat*	*0.66*
2cna	*Concanavalin A jack bean*	*0.66*
3est	*Elastase porcine*	*0.66*
3grs	*Gluthathione reductase human*	*0.66*
3cla	*Cloramphenicol acetyltransferase E. coli*	0.66
1hoe	Alpha-amylase inhibitor *S. tendae*	0.66
2ilb	Interleukin-1 beta human	0.66
1acx	Actinoxanthin *A. globisporus*	0.67
4dfr	Dihydrofolate reductase *E. coli*	0.67
9pap	Papain payaya	0.67
3b5c	Cytochrome *b5* soluble domain bovine	0.67
2mhr	Myohemerythrin *T. aostericola*	0.67
1paz	Pseudoazurin *A. faecalis*	0.67
2cpp	Cytochrome P450cam *P. putida*	0.68
4tnc	Troponin C chicken	0.68
1ccr	Cytochrome *c O. sativa*	0.68
1gox	Glycolate oxidase *S. oleracea*	*0.68*
1r69	*434 Repressor N-term domain phage 434*	*0.68*
1ctf	*L7/L12 50S ribosomal protein C-term domain E. coli*	0.69
1lh1	Leghemoglobin *L. leteus*	0.69
2alp	Alpha-lytic protease *L. enzymogenes*	0.69
3tln	Thermolysin *B. thermoproteol.*	*0.69*
4tpt	*Beta-trypsin bovine*	*0.69*
5cpa	*Carboxypeptidase A bovine*	*0.69*
5cyt	*Cytochrome c* tuna	0.69
1rnt	Ribonuclease T1 A. oryzae	0.70
1ubq	Ubiquitin human	0.70
2cab	Carbonic anhydrase form B human	0.70
2pcy	Plastocyanin poplar	0.70
3apr	Acid proteinase *R. chinensis*	*0.70*
3rn3	*Ribonuclease A bovine*	*0.70*
4fxn	*Lactate dehydrogenase Clostridium mp*	0.70
6lyz	Lysozyme hen egg-white	0.70
2gbp	Galactose binding protein *E. coli*	0.70
1eco	Erythrocruorin *C. thummi*	0.71
1gcr	Gamma-II crystallin bovine	0.71
1lzl	Lyozyme human	0.71
2cyp	Cytochrome *c* peroxidase *S. cerevisiae*	*0.71*
2lzm	*Lysosyme bacteriophage T4*	*0.71*
2prk	*Proteinase K*	*0.71*
2sga	*Porteinase A S. griseus*	0.71
3blm	Beta-lactamase *B. licheniformis*	0.71
4pep	Pepsin pig	0.71
1fx1	Flavodoxin *D. vulgaris*	0.72
5cpv	Ca-binding paravalbumin carp	0.72
3app	Acid proteinase *P. janthinellum*	*0.74*

[a] *Listed in ascending order of* $\overline{P_{sr}}$.
[b] Bernstein *et al.* (1977) and Abola *et al.* (1977).

occluded. A two-dimensional diagram of this analysis is shown in Fig. 1. By summing the surface area associated with each class of surface dot, we determine the total occluded and non-occluded surface areas for the residue. The occluded surface ratio, Q_{sr}, for a residue is defined as the ratio of the occluded surface over the

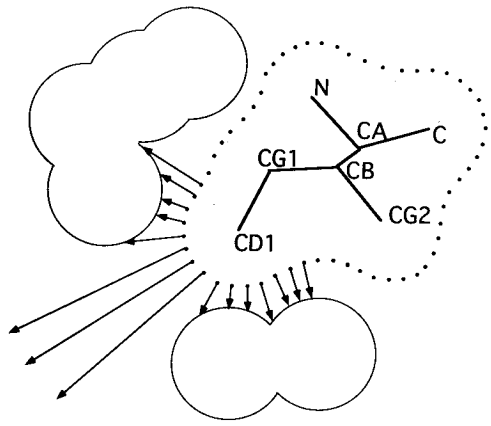

Figure 1. Two-dimensional diagram of the surface normals used to calculate occluded surface. The diagram depicts a slice through an isoleucine residue and the surface normals associated with the CD1 atom are drawn. A surface normal was extended outward from each surface dot. If the extended normal intersected the van der Waals surface of another atom within an atom center to center distance of 6.4 Å that dot was labeled as being occluded, otherwise the dot was labeled as representing a non-occluded surface patch. In this two-dimensional slice there are 12 occluded surface dots and 3 non-occluded surface dots associated with atom CD1.

total surface area. This procedure is repeated for all residues in each protein structure.

In addition, the length of the extended surface normal from the surface dot to the point of intersection with the neighboring van der Waals surface is computed. The value of this length is divided by 2.8 to give a normalized packing parameter, *PP*, equal to zero for atoms in contact, and equal to 1.0 or greater for atoms separated by at least the diameter of a water molecule. Only these normals which intersect another atom surface are considered in this calculation.

Finally, the identity of atoms intersected by the extended surface normals is tabulated for a description of the packing environment for each residue.

Calculation of normalized surface ratios

To incorporate the observed frequency distribution of Q_{sr} for the 20 amino acids (Fig. 2) and also to account for the size of each amino acid, we use a weighted surface area scheme. For each of the 20 standard amino acids we define the weighted surface area $S^a_w(Q_{sr})$ as follows:

$$S^a_w(Q_{sr}) = \frac{S^a_t \times I^a_{sr}(Q_{sr})}{I^a_{sr}(\max)} \quad (1)$$

where S^a_t is the total surface area for amino acid, a; $I^a_{sr}(Q_{sr})$ is the *incidence* of the occluded surface area ratio, Q_{sr}, for the amino acid, a; and $I^s_{sr}(\max)$ is the maximum observed *incidence*, at any value of Q_{sr} for the given amino acid as shown in Fig. 2. In Eqn (1), for $I^a_{sr}(Q_{sr}) = I^a_{sr}(\max)$, $S^a_w(Q_{sr})$ will be equal to the total surface area of the amino acid, S^a_t. Table 2 shows the $S^a_w(Q_{sr})$ values for each amino acid for a given occluded surface ratio, Q_{sr}. These data reflect the distribution of occluded surface environments for each amino acid (see Results and Discussion) weighted by the total surface area of the amino acid.

From the weighted surface areas of the 20 amino acids we calculate a parameter called the weighted protein occluded surface, P_w, for the entire protein structure as follows:

$$P_w = \sum_{i=1}^{N} S^a_w(Q^i_{sr}) \quad (2)$$

where $S^a_w(Q^i_{sr})$ is the actual weighted surface area for residue i, having an occluded surface ratio Q_{sr}, and N is the total number of residues in the protein. The value of P_w depends upon the value of Q^i_{sr} for each residue and also the total number of residues in the protein. In order to normalize the weighted protein occluded surface, we calculated the ideal weighted protein occluded surface, P_I, for an 'ideal' protein structure consisting of residues which are all in their most frequent occluded surface environments (i.e., each residue has a Q_{sr} with maximum incidence in Fig. 2). It may be noted that for a residue in its most frequent occluded surface environment, $S^a_w(Q^i_{sr})$ is equal to the total residue surface, S^a_t, for that type of amino acid. Thus the ideal weighted protein occluded surface is defined as

$$P_I = \sum_{i=1}^{N} S^i_t \quad (3)$$

where S^i_t is the total surface area for amino acid residue, i. Then the normalized protein surface ratio, $\overline{P_{sr}}$, is defined as

$$\overline{P_{sr}} = \frac{P_w}{P_I} \quad (4)$$

For analysis of individual residues in protein model structures we define a term called the normalized residue surface ratio, $\overline{R_i}$ as follows:

$$\overline{R_i} = \frac{S^a_w(Q^i_{sr})}{S^i_t} \quad (5)$$

The value of $\overline{R_i}$ will equal 1.0 when the residue is in its most frequent occluded surface environment and less than 1.0 for all other occluded surface environments.

Calculation of protein volumes and packing densities

The Molecular Surface Package of programs written by Connolly was used to calculate solvent-excluded volumes. A probe size of 1.4 was used in the PQMS program. From the packing parameter, *PP*, for each dot, we calculated the residue packing parameter, *PPr*, by averaging the *PP* of all the dots associated with each residue. The average protein packing parameter *PPp*, was calculated as the mean of the residue packing parameters. The packing density is defined as $(1 - PPp)$.

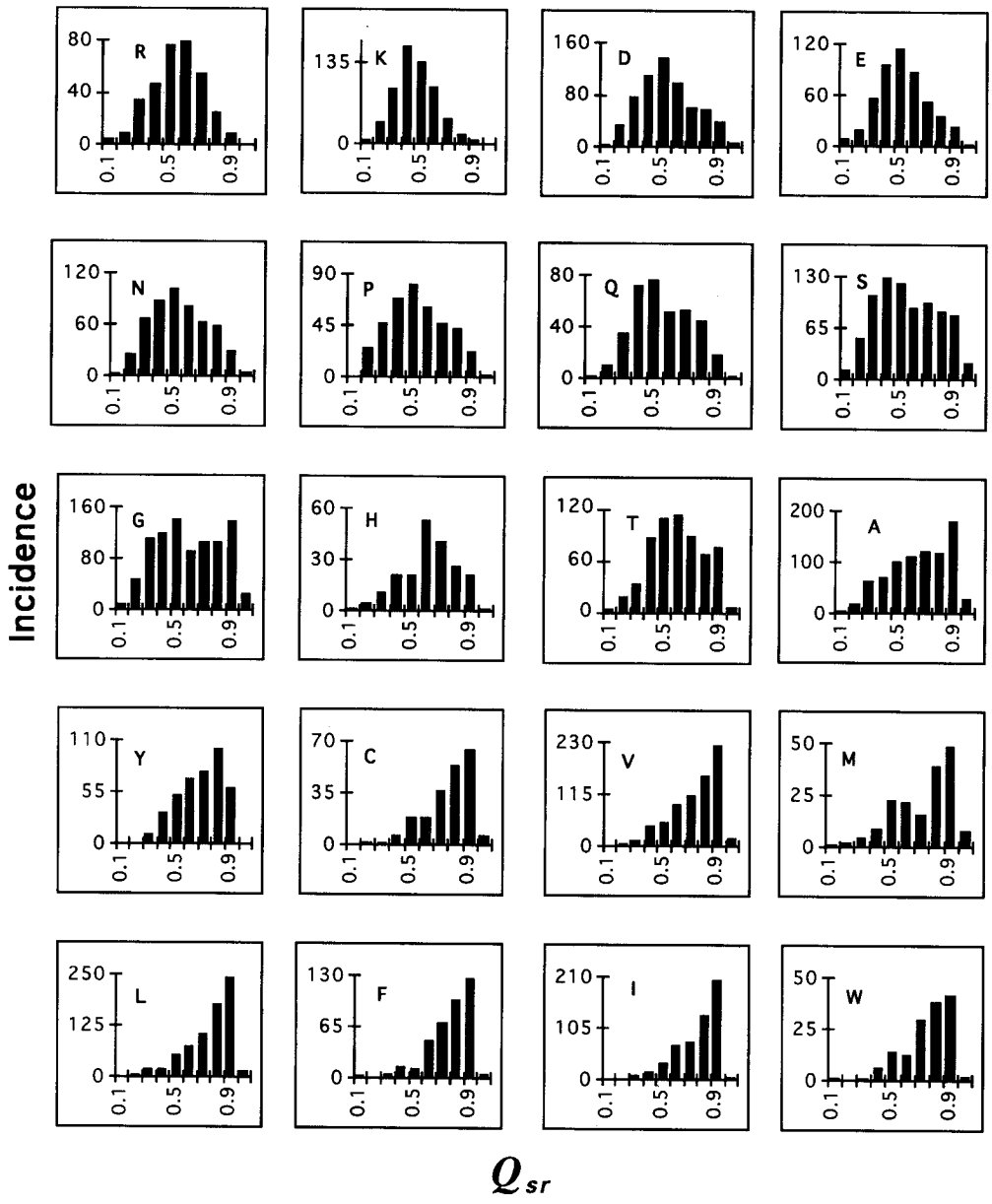

Figure 2. Frequency distributions of occluded surface ratio values, Q_{sr}, for each of the 20 standard amino acids in a dataset of 55 protein structures. Each histogram displays the incidence (frequency of occurrence) of a range of occluded surface ratios: zero to 0.1, 0.11 to 0.2, 0.21 to 0.3, etc., for each amino acid. Single letter codes identify the histogram for each amino acid. The total number of residues included in each plot is listed in Table 3. First, last and incomplete residues in each protein structure were not included in these plots. The individual histograms are presented according to a grouping based on the nature of the frequency distribution as described in the text.

Results and Discussion

Computation of parameters

In this work a molecular surface is used as a basis for comparative evaluation of amino acid environments. The molecular surfaces are represented by a set of dots distributed over the surface. This surface is composed of the contact and reentrant surfaces as defined by Richards (1977), and there are several methods to calculate the area of this surface (Richmond, 1984; Connolly, 1985). We have used a simple summation of the area associated with each dot describing this surface to calculate the total surface area of an amino acid residue. Although this method has the drawback that the distribution of dots in the reentrant surface may not be uniformly distributed, the ability to characterize the area associated with each dot is useful. The mean molecular surface areas for a large number of isolated amino acid residues, calculated by summing the dot-associated areas, are shown in Table 3. These values are averaged over the rotomer conformations present in 55 protein structural models (see Experimental). In agreement with previous calculations using accessible surface area (Rose *et al.*, 1985) the molecular surface areas of amino acid residues in this data set of proteins

Table 2. Weighted surface areas for amino acid residues[a]

Amino acid	Q_{sr} bin[b] 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ALA	1.13	7.18	23.80	26.44	37.77	42.69	46.09	44.20	68.75	10.20
ARG	7.16	16.10	62.63	82.31	137.78	141.36	98.42	44.73	16.10	1.79
ASN	1.82	22.75	60.97	79.17	93.73	73.71	57.33	52.78	27.30	3.64
ASP	1.97	21.63	50.47	72.09	89.79	65.54	40.63	38.01	24.91	3.28
CYS	0.00	1.35	2.69	8.08	25.58	25.58	49.81	72.70	86.16	8.08
GLN	2.86	14.31	50.07	103.01	110.16	72.96	75.82	62.95	27.18	2.86
GLU	8.34	17.61	51.89	88.03	106.56	80.61	49.11	33.36	21.31	1.85
GLY	3.41	18.58	41.70	45.49	53.83	34.50	40.18	39.80	52.31	10.24
HIS	2.27	9.09	25.00	47.73	47.73	120.45	93.18	59.09	47.73	2.27
ILE	0.00	0.56	5.00	8.34	17.79	39.47	42.80	72.26	113.40	3.34
LEU	1.47	1.96	8.81	9.30	26.92	36.23	53.85	91.05	114.06	5.38
LYS	6.16	26.97	70.89	124.06	104.03	73.20	33.13	13.10	5.39	0.00
MET	2.44	4.88	12.21	21.98	56.18	53.73	39.08	95.26	119.68	19.54
PHE	2.25	0.00	6.74	15.72	14.60	52.79	79.74	111.19	141.51	5.62
PRO	1.12	28.05	52.73	77.41	92.00	68.44	52.73	47.12	25.80	2.24
SER	6.93	30.05	61.25	73.96	69.92	52.58	56.05	50.27	46.80	12.13
THR	3.95	15.02	26.87	69.55	87.73	90.10	70.34	55.32	61.65	5.53
TRP	3.99	0.00	3.99	23.92	55.81	47.84	119.59	151.48	167.43	7.97
TYR	1.45	1.45	15.96	47.88	74.00	100.12	110.28	146.55	87.06	1.45
VAL	0.44	2.63	5.26	19.72	22.79	39.88	49.09	68.37	97.74	7.45

[a] The weighted surface area, $S^a_w(Q_{sr})$ was calculated from the data shown in Fig. 2 as described in Experimental.
[b] Each column represents the bin containing the weighted surface for occluded-surface/total-surface ratios. Q_{sr}, of 0–0.10, 0.11–0.20, 0.21–0.3, etc.

Table 3. Occluded, non-occluded, and total molecular surface areas for the standard 20 amino acids[a]

		Occluded		Non-occluded		Total		
Residue	Num.[b]	Side chain[c]	Whole residue	Side chain[c]	Whole residue	Side chain[c]	Whole residue	SD[d]
ALA	816	16.56	42.24	13.22	26.52	29.78	68.75	(1.23)
ARG	340	45.89	70.45	58.98	70.91	104.88	141.36	(2.88)
ASN	520	23.76	46.34	33.63	47.39	57.39	93.73	(2.08)
ASP	623	22.00	43.76	31.79	46.03	53.78	89.79	(2.03)
CYS	208	34.66	60.98	13.86	25.19	48.52	86.16	(1.46)
GLN	365	32.18	55.86	41.69	54.30	73.87	110.16	(2.32)
GLU	495	27.89	50.52	42.55	56.05	70.44	106.56	(2.33)
GLY	897	0.00	29.27	0.00	24.56	0.00	53.83	(1.21)
HIS	200	44.79	69.57	39.31	50.88	84.10	120.45	(2.25)
ILE	545	55.71	81.18	23.89	32.22	79.60	113.40	(2.08)
LEU	713	54.02	80.20	24.73	33.86	78.75	114.06	(1.96)
LYS	593	29.56	51.71	58.21	72.35	87.76	124.06	(2.51)
MET	174	54.36	80.31	29.34	39.36	83.69	119.68	(2.80)
PHE	383	74.43	100.79	31.46	40.72	105.89	141.51	(2.50)
PRO	399	25.95	45.11	34.08	46.89	60.03	92.00	(1.53)
SER	796	15.08	37.73	20.57	36.23	35.64	73.96	(1.33)
THR	615	27.00	50.23	26.95	39.87	53.95	90.10	(1.68)
TRP	146	90.57	115.81	41.92	51.62	132.49	167.43	(3.33)
TYR	404	66.85	92.13	43.87	54.42	110.72	146.55	(2.47)
VAL	715	42.20	67.23	21.17	30.51	63.36	97.74	(1.56)

[a] Values are the mean surface areas, in Å^2, for the respective amino acid. The areas were obtained by summing the areas associated with each moelcular surface dot as described in the text.
[b] The number of residues of each type included in determination of surface areas.
[c] Atoms CB and beyond are included in the side chain.
[d] Numbers in parentheses are the standard deviations of the whole residue total surface area.

have small deviations about the mean; the distribution of rotomers does not significantly change the average residue surface area. The values are approximately 60% as large as the accessible surface area values except for cysteine which we have calculated as the free sulfhydryl rather than the half-cystine. Therefore the use of a mean value is appropriate for statistical comparisons such as presented here.

Our method of computation of interatomic occluded and non-occluded surface areas of amino acid residues in proteins is unique in the use of explicit vectors but we obtain information which is similar to that computed by the methods of either Rashin *et al.* (1986) or Holm and Sander (1992). Our focus is on the interatomic occluded area rather than solvated area. For this purpose we identify each patch of surface area that is occluded by neighboring atoms. The calculated ratio of occluded surface area (column 4 in Table 3) to total residue area (column 8), which is the mean occluded surface ratio, $\langle Q_{sr} \rangle$, is conceptually equivalent to the mean fractional area lost on transfer from the unfolded to the folded state as defined by Rose *et al.*, (1985). The two parameters are highly correlated as expected from the above discussion. However, our method of computing occluded and non-occluded surface area extracts characteristics of amino acids in proteins which were not obtained in the distribution functions of accessibility reported by Rose *et al.*, (1985) as explained below.

The frequency distributions of occluded surface ratios, Q_{sr}, for the standard amino acids are plotted in Fig. 2. The 20 amino acids may be separated into three categories based on the Q_{sr} distribution occurrences. One category (R,K,D,E,N,P) has an almost normal distribution around $Q_{sr}=0.5$. This category includes the polar residues with solvation free energies (ΔGo_{solv}) (Eisenberg *et al.*, 1989) of -1.37 to -0.82 and proline. Although proline is relatively non-polar with a ΔGo_{solv} of 0.98, it is frequently found on the surface of proteins as part of turn regions. Another category (Q,S,G,H,T) has a maximum *incidence* of Q_{sr} between 0.4 and 0.6 but also shows a significant number of occurrences with greater values. This intermediate category has ΔGo_{solv} values of -0.03 to 0.35. The third category (A,Y,C,V,M,L,F,I,W) has a skewed distribution heavily weighted toward $Q_{sr}=0.9$. This category is composed of the non-polar residues ($\Delta Go_{solv}=0.42$ to 3.07) and cysteine which frequently forms disulfide bonds.

Our results differ from those of Rose *et al.* in several respects in spite of the fact that the two types of average surface ratios $(A^\circ - \langle A \rangle)/A^\circ$ and $\langle Q_{sr} \rangle$ correlate well. For example, in Fig. 2 histidine is clearly within the intermediate category of occluded surface ratios whereas it was categorized as largely buried in the previous treatment. Importantly, the absolute values of occluded surface ratios are much more widely distributed than a similar measure of solvent exposure [cf. Fig. 1 in Rose *et al.* (1985)]. Atomic surface area which is defined as buried by the criterion of accessibility may not be occluded as defined here. However, our results are consistent with the interpretation of Rose *et al.* that, in spite of much non-polar surface being exposed, there is a strong correlation between hydrophobicity and buried surface. The correlation between hydrophobicity and occluded surface ratio is shown in Fig. 3 for the standard amino acids. Although these two parameters show some correlation it is clear that the occluded surface ratio is something other than an empirical hydrophobicity scale (Casari and Sippl, 1992).

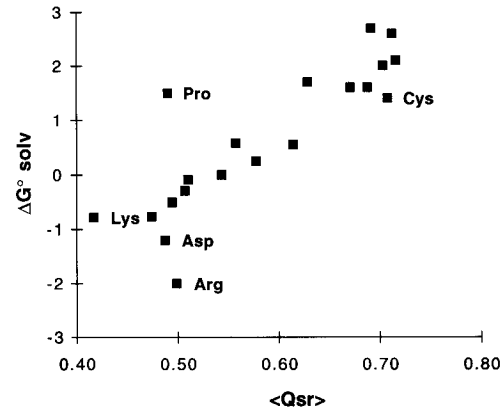

Figure 3. Plot of free energy of solvation Eisenberg *et al.* (1989) vs the mean occluded surface ratio $\langle Q_{sr} \rangle$ for the standard 20 amino acids as calculated in this work. Residues showing poor correlation are labeled.

Normalized protein surface ratios

The normalized protein surface ratio values, $\overline{P_{sr}}$, for a data set of crystallographic protein structures are listed in Table 1 and shown graphically in Fig. 4. The range of values for P_{sr} is remarkably narrow (0.60–0.74). Within this small range we found no correlation of the P_{sr} with molecular weight. Thus, although larger proteins, with a greater volume of buried residues, have the potential to obtain more occluded surface area they maintain the same relative occluded surface area per residue. This is because both P_w and P_I increase as a linear function of molecular weight (data not shown). Similar results

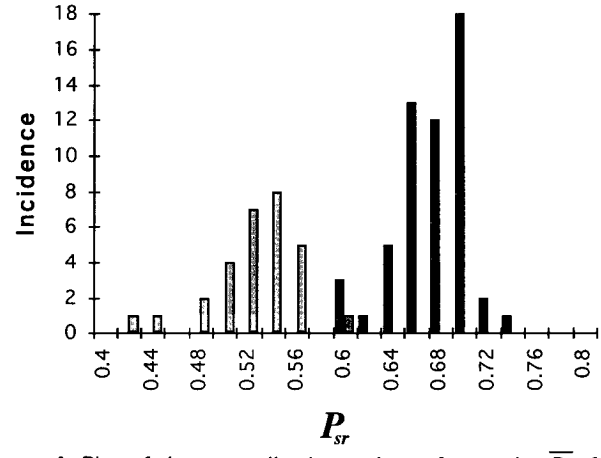

Figure 4. Plot of the normalized protein surface ratio, $\overline{P_{sr}}$, for protein models. The solid bars represent the incidence of $\overline{P_{sr}}$ in native proteins in a dataset of 55 structural models. The shaded bars represent the incidence of $\overline{P_{sr}}$ in a dataset of 25 misfolded models built as described by Holm and Sander (1992).

were obtained when only the side chain atoms of amino acid residues were included in the calculation.

Three proteins in the data set (2cdf, 2ovo, and 4pti) have P_{sr} values of 0.60–0.61 and appear to be exceptions to the normal distribution. The protein cytochrome c_3 (2cdv) is a small protein with four heme prosthetic groups and these were not included in the calculations of occluded surface ratios. If the four heme groups are included, 2cdv has a $\overline{P_{sr}}$ value of 0.69, placing it well within the normal range of native proteins. The protease inhibitors, ovomucoid third domain (2ovo) and trypsin inhibitor (4pti), have loop regions which are not tightly packed and which fit into the active site of the respective protease. These loop regions exhibit very low normalized residue surface ratios, R_i, and a low P_{sr} value is therefore obtained for the whole protein. In addition, both of these small proteins are stabilized by internal disulfide bonds and do not depend solely on non-bonded interactions for their stability.

The importance of the narrow range of $\overline{P_{sr}}$ values is that this range defines properly folded globular proteins. In contrast to native or properly folded proteins, most misfolded proteins have P_{sr} values which are significantly lower. The P_{sr} values for 25 misfolded proteins are also shown in Fig. 4. These misfolded porteins are essentially indistinguishable from the correctly folded models in terms of potential energy using the program GROMOS (Holm and Sander, 1992). The method described here accurately identifies these models as misfolded with the exception of 1rn3on1p2p (ribonuclease sequence on phospholipase A2 backbone) which has a P_{sr} value of 0.60.

However, the relatively large $\overline{P_{sr}}$ value for 1rn3on1p2p is due to an interesting structural analogy. In both ribonuclease and phospholipase A2, the first 17 residues are in a similar *alpha*-helix/turn secondary structure combination. In both proteins this helix is on the surface of the respective protein. Thus, mounting the ribonuclease amino acid sequence on the phospholipase A2 backbone permits the ribonuclease residues to have a similar environment as they do when in the ribonuclease structure. The first 17 residues of 1rn3on1p2p have a pattern of individual R_i values that is similar to that of native proteins. This result is expected because these 17 residues are in a similar environment compared to the properly folded structure. This segment of apparently normal residue occluded surface ratios gives the protein a relatively large P_{sr} value for a misfolded protein. In retrospect it is clear that only a portion of this structure is misfolded. Our method correctly identified the first 17 residues as being in a native-like structural environment.

Thus, inspection of the four models with intermediate values of P_{sr} (≈ 0.6) has identified interesting structural characteristics in each case. Protein models that fall into this category should be investigated with regard to prosthetic group binding sites and/or loosely structured loop regions.

All four of the misfolded protein models generated by Novotny and colleagues are easily distinguished from the correctly folded models by low P_{sr} values (0.51–0.54). We find no difference between the misfolded models constructed using CONGEN and those using CHARMM despite the fact that the former are significantly improved models in terms of empirical potential energies (Novotny *et al.*, 1988).

These results show that P_{sr} can usefully distinguish deliberately misfolded, but energy minimized protein model structures. Other algorithms have been reported which will also accomplish this end (Novotny *et al.*, 1988; Holm and Sander, 1992; Vriend and Sander, 1993). However, an advantage to our method is that the distribution of native surface environments for each amino acid is considered discretely; the distribution function of occluded surface is not smoothed by regression fitting and no windowing of residue segments is required to calculate P_{sr}. Our treatment is especially useful for a consideration of individual residues in a protein structure. Those residues which are in unusual occluded surface environments may be easily distinguished.

Normalized residue surface ratios

A plot of $\overline{R_i}$, the normalized residue surface ratio values for the well characterized protein hen egg lysozyme is shown in Fig. 5. Other native proteins have similar R_i distribution profiles. In this display a score of 1.0 is obtained by residues with occluded surface ratios equal to the peak value in the histograms shown in Fig. 2, i.e., these residues are in their respective most frequent occluded surface environment. Approximately 25% of the residues are in this category. A low R_i value is obtained if the residue has an infrequent occluded surface ratio. Residues with less frequent occluded surface ratios ($R_i < 0.5$) comprise less than 20% of the protein and are distributed individually throughout the amino acid sequence rather than in clusters. Visual inspection of the location of these residues in lysozyme shows that they are also widely distributed throughout the tertiary structure.

Further analysis of the occluded surface characteristics that contribute to the relatively low R_i values for these residues is in progress. Those residues with R_i values less than 0.2 have been found frequently in intramolecular contact due to the crystal lattice. Such crystal contacts explain the low values for Arg14, Thr47 and Arg128 in hen egg lysozyme (Fig. 5). Because of the crystal contacts, these surface residues have more occluded surface than they would have for lysozyme in solution. Not all crystal lattice contacts cause a residue to have a low R_i value however. For instance, both the OE1 and NE2 atoms of Gln41 are involved in crystal contacts but this residue has a R_i value of 0.94.

Since all residue types exhibit distributions of occluded surface (Fig. 2), an individual residue with an unusual R_i value may not have structural significance. In addition, we have not included crystallographically defined water or other heteroatoms in this first analysis. A more extensive analysis of a larger dataset, including heteroatoms, is in progress and this analysis may demonstrate improved discrimination of significant

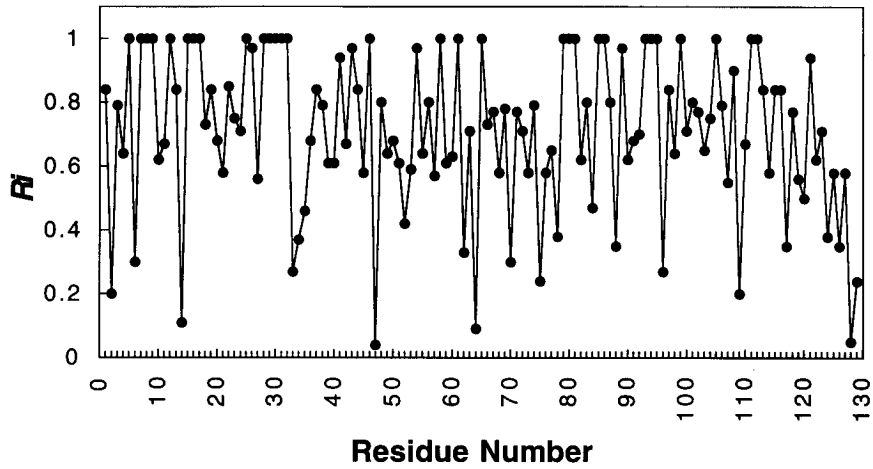

Figure 5. Plot of normalized residue surface ratio, $\overline{R}_i$, vs residue number for hen egg lysozyme (6lyz).

deviations from the norm. However, we have found that inspection of R_i for crystallographic protein models undergoing refinement is a useful tool to identify residues which have been incorrectly modeled.

Packing density calculations

The ensemble of extended normals associated with a buried residue describes the interatomic space surrounding that residue. Surface areas with a mean *PP* approaching zero will be in close van der Waals contact with a neighboring atom. Whereas surface areas with a mean *PP* of 1.0 will be at least 2.8 Å distant from any other atom and are defined as not occluded. The mean length of extended normals for each residue provides an estimate of the packing proximity for that particular residue. And similarly, the mean length of extended normals for all residues in a protein structure estimates the overall packing of that protein. The value of $(1 - PPp)$ is operationally similar to a packing density value. As an example of the effectiveness of this calculation we have calculated the mean value of $(1 - PPp)$ for each of nine ribonuclease-A structures in which the packing was known to be slightly different. Crystallographic determination of ribonuclease at temperatures from 98 to 320 K was reported by Tilton *et al.* (1992). Figure 6 shows the variation of $(1 - PPr)$ with temperature for these nine structures. The packing varies in a reciprocal relationship to the solvent-excluded volume for these nine structures.

Packing environments

The calculation of packing, or contact, environments has been an integral aspect of many protein fold prediction algorithms (Singh and Thornton, 1990; Behe *et al.*, 1991; Bowie *et al.*, 1991; Overington *et al.*, 1992). In most cases the packing environment is determined from

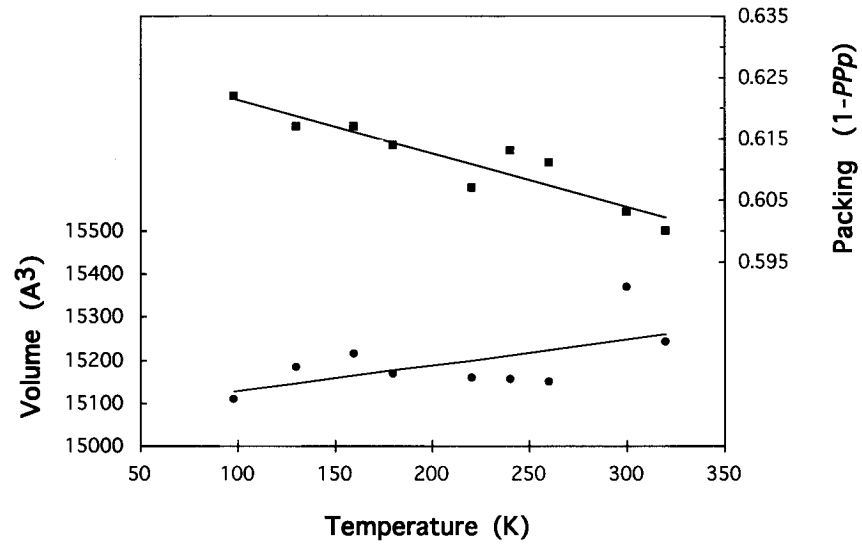

Figure 6. Plot of residue packing and solvent excluded volume vs temperature for ribonuclease-A. The nine crystallographic structures reported by Tilton *et al.* (1992) were used to determine packing parameter, *PPp*, as described in the text. Solvent excluded volume was calculated using a probe size of 1.4 A with the program PQMS.

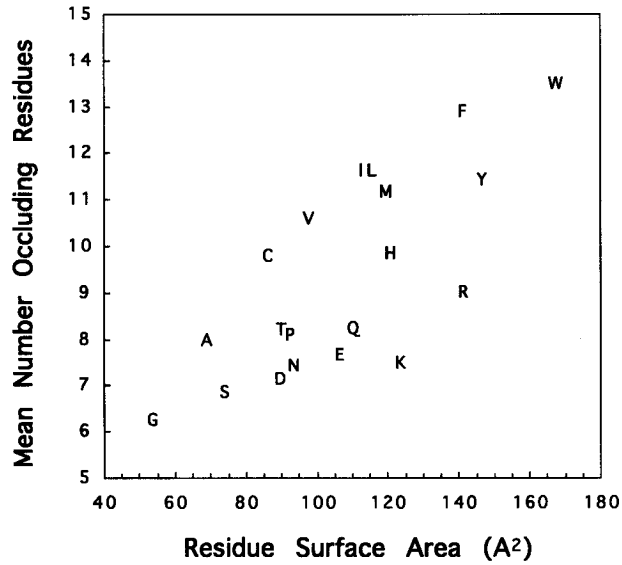

Figure 7. Plot of the mean number of occluding residues vs residue surface area. The residue surface area data were taken from Table 3 and the mean number of occluding residues was calculated from the list of residues intersected by extended normals from each residue in the data set. Single letter codes for the 20 common amino acids are used for the respective data points.

pairwise distance calculations without an explicit determination of whether or not the two residues in a pair actually may be in van der Waals contact with each other. Our method directly determines the identity and degree of interaction between two atoms and provides information about residue–residue interaction similar to the area buried between contacting residues (Bene *et al.*, 1991). We have calculated the packing environment for each residue in the protein structures listed in Table 1. Figure 7 shows the mean number of residues comprising the total occluding surface of each type of amino acid plotted vs the surface area of the amino acid. Although some correlation of the number of occluding residues with surface area is apparent it is surprising that A, T, P, Q, E, and K all have approximately the same mean number of occluding residues. Qualitatively, the data in Fig. 7 confirm that the packing environments of hydrophobic residues is different from that of polor and charged residues. The results shown in Fig. 7 are in contrast to results obtained when strict distance criteria are used. Calonna-Cesari and Sander (1990) used a distance cut-off similar to the one used here in their derivation of protein contact counts. In this latter case the contact counts are almost exactly correlated with surface area. Because the results in Fig. 7 do not show an exact correlation with surface area it is apparent that atoms within a distance separation of 6.4 Å may not be occluded by each other, although simple distance calculations would categorize them as in contact.

Finally, the method described here may be applied to the interaction between a ligand and receptor protein. Analysis of the extended surface normals associated with the bound ligand allows one to estimate the complimentary of ligand with receptor site. This type of analysis has proven useful for evaluating the pseudo energetics of drug and inhibitor binding (Baldwin *et al.*, 1995).

Conclusion

Several different representations of molecular surface previously have been found to be useful in the evaluation of protein structures. One of the most widely utilized types of surface is the accessible surface (Lee and Richards, 1971). Although this surface represents a projection some distance from the van der Waals surface it is a measure of solvent accessibility and therefore has been useful for calculations involving hydrophobicity and other interactions with water. Another useful defined surface, especially for graphical representations, is the molecular surface composed of contact and re-entrant segments (Richards, 1977; Connolly, 1983). We have defined a subset of the molecular surface, called the occluded surface, which represents that portion of the molecular surface in a macromolecule occluded by interatomic contact. The occluded surface is related to, but distinct from the buried surface of a folded protein.

Calculation of this occluded surface using extended surface normals permits a direct investigation of the packing environment of atoms in a macromolecule such as a protein. Our method does not require regression fitting of parameters, a variable window of residues in the amino acid sequence, determination of secondary structure or the use of solvation parameters. In this report we describe the information that may be obtained by a consideration of the ratio of occluded to total surface area for amino acid residues in protein strucutres, the calculation of atomic packing density, the specific environments of atoms in macromulecules, and the complimentarity of interacting molecules. We are in the process of developing a potential function for each amino acid based on the distribution of Q_{sr} (Fig. 2) and also a potential function for each amino acid based on the distribution of occluding amino acids. These potentials will be used to evaluate the folding and misfolding of protein structural models. The atomic interactions described by adjacent occluded surfaces will enable one to quantify the interaction developed for each functional group by quantitative structure–activity relationship studies. We believe that this method provides the basis for many useful applications in the field of macromolecular interaction.

Acknowledgements

PFJ would like to thank Dr Fred Richards for the encouragement and facilities to complete this project, Dr Joachim Jaeger for help with determining crystal lattice contact residues and critical comments, Dr Jiri Novotny for providing the coordinates for several of the misfolded models used in this study, and Dr Shaojian Sun for valuable discussion.

References

Abola, E. E., Bernstein, F. C., Bryant, S. F., Koetzle, T. E. and Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, eds. F. H. Allen, G. Bergerhoff, and R. Sievers, pp. 107–132. Data Commission of the International Union of Crystallography, Bonn.

Baldwin, E. T., Bhat, T. N., Liu, B., Pattabiraman, N. and Erickson, J. W. (1995). *Nature: Struct. Biol.* **2**, 244–249.

Behe, M. J., Lattman, E. E. and Rose, G. D. (1991). *Proc. Natl. Acad. Sci. USA* **88**, 4195–4199.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bowie, J. U., Luthy, R. and Eisenberg, D. (1991) *Science* **253**, 164–170.

Bryant, S. H. and Amzel, L. M. (1986) *Int. J. Peptide Protein Res.* **29**, 46–52.

Casari, G. and Sippl, M. J. (1992) *J. Mol. Biol.* **224**, 725–732.

Colonna-Cesari, F. and Saunder, C. (1990) *Biophys. J.* **57**, 1103–1107.

Colovos, C. and Yeates, T. O. (1993) *Protein Science* **2**, 1511–1519.

Connolly, M. L. (1983) *Science* **221**, 709–713.

Connolly, M. L. (1985) *J. Am. Chem. Soc.* **107**, 1118–1124.

Eisenberg, D., Wesson, M. and Yamashita, M. (1989) *Chem. Scripta* **29A**, 217–221.

Ferrin, T. E., Conrad, C. C., Jarvis, L. E. and Langridge, R. (1988) *J. Mol. Graphics* **6**, 13–27.

Gregoret, L. M. and Cohen, F. E. (1990) *J Mol. Biol.* **211**, 959–974.

Harpaz, Y., Gerstein, M. and Chothia, C. (1994) *Structure* **2**, 641–649.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Cesari, G. and Sippl, M. J. (1990) *J. Mol. Biol.* **216**, 167–180.

Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) *Protein Science* **1**, 409–417.

Holm, L. and Sander, C. (1992) *J. Mol. Biol.* **225**, 93–105.

Jiang, F. and Kim, S.-H. (1991) *J. Mol. Biol.* **219**, 79–102.

Lawrence, M. C. and Colman, P. M. (1993) *J Mol. Biol.* **234**, 946–950.

Lee, B. and Richards, F. M. (1971) *J. Mol. Biol.* **55**, 379–400.

Luthy R., Bowie, J. U. and Eisenberg, D. (1992) *Nature* **356**, 83–85.

Maiorov, V. N. and Crippen, G. M. (1992) *J. Mol. Biol.* **227**, 876–888.

Nicholls, A., Sharp, K. A. and Honig, B. (1991) *Proteins* **11**, 281–296.

Novotny, J., Bruccoleri, R. and Karplus, M. (1984) *J. Mol. Biol.* **177**, 787–818.

Novotny, J., Rashin, A. A. and Bruccoleri, R. E. (1988) *Proteins: Structure, Function, and Genetics* **4**, 19–30.

Overington, J., Donnelly, D., Johnson, M. S., Sali, A. and Blundell, T. L. (1992) *Protein Science* **1**, 216–226.

Ponder, J. A. and Richards F. M. (1987) *J Mol. Biol.* **193**, 775–791.

Richards, F. M. (1974) *J. Mol. Biol.* **82**, 1–14.

Rashin, A. A., Iofin, M. and Honig, B. (1986) *Biochemistry* **25**, 3619–3625.

Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Richmond, T. J. (1984) *J. Mol. Biol.* **178**, 63–89.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. and Zehfus, M. H. (1985) *Science* **229**, 834–838.

Singh, J. and Thornton, J. M. (1990) *J. Mol. Biol.* **211**, 595–615.

Tanaka S. and Scheraga, H. A. (1976) *Macromolecules* **9**, 945–950.

Tilton, R. F., Dewan, J. C. and Petsko, G. A. (1992) *Biochemistry* **31**, 2469–2481.

Vriend, G. and Sander, C. (1993) *J. Appl. Cryst.* **26**, 47–60.

Wodak, S. J. and Rooman, M. J. (1993) *Curr. Opin. Struct. Biol.* **3**, 247–259.