Surface complementarity of buried protein residues

Rahul Banerjee*,†, Subhankar Ray and Dhananjay Bhattacharya***

*Saha Institute of Nuclear Physics, Sector 1, Block AF, Bidhan Nagar, Calcutta 700 064, India

**Department of Physics, Jadavpur University, Calcutta 700 032, India

The surface fit of residues buried in protein interiors has been estimated using a surface complementarity function (S_m). For a perfect match between two surfaces, S_m gives a value of 1.0. When all dot surface points were considered irrespective of having been contributed by main-chain or side-chain atoms, buried residues gave an average S_m of 0.60, ranging from 0.57 to 0.63 for the 20 amino acids. This indicated the ability of all residues, i.e. polar, apolar, or charged to pack with the same measure of complementarity in protein interiors. Further, the same calculation performed with surface points from side-chain atoms alone gave a lower global average of 0.47, implying the importance of main-chain atoms in internal packing. The lower value of surface complementarity obtained in this case also showed that the side-chains of buried residues need not invariably be involved in a 'jigsaw puzzle'-like complementarity with other side-chains of the polypeptide chain and can occasionally show serious packing defects, though they play a predominant role in packing compared to the main-chain atoms. However, the combination of both side- and main-chain atoms ensures a high average S_m of 0.60 (calculated with all atoms), indicating high steric fit for all residues upon burial.

THE prediction of a protein's physiologically active three-dimensional structure starting from its amino acid sequence constitutes one aspect of the protein folding problem. Although the complete solution to this problem is still elusive, it has become evident that correctly folded proteins invariably exhibit common structural patterns. One of these is the absence of unbalanced charges, by satisfaction of most of the main-chain hydrogen bonding (N–H . . . O=C) potential[1]. Another is the packing of apolar amino acids in the protein interior to form hydrophobic cores. The collapse of the polypeptide chain due to the hydrophobic effect[2], thereby segregating the apolar amino acids from the surrounding aqueous environment and the concomitant exposure of polar or charged side-chains to solvent has been found to occur early in the protein folding pathway. It is now generally accepted that hydrophobicity does indeed play a dominant role in protein folding.

There is considerable redundancy in the amino acid sequence of a protein coding for a particular three-dimensional fold (sequence identity of about 40% gives homologous structures). In this context, the apolar residues constituting hydrophobic cores carry high informational content[3]. This is borne out by the fact that site-directed mutagenesis of these residues are generally destabilizing[4,5] for the protein. Further sequence and structure comparison studies between naturally occurring homologous proteins show buried residues to mutate only among apolar amino acids[5]. In contrast, individual residues exposed to the solvent do not show such strong mutational constraints, though the protein surface as a whole remains by and large polar. Nor do they contribute as significantly to protein stability.

Several investigations, both experimental and computational have been performed to determine which property of a buried core residue, (1) hydrophobicity, (2) volume or (3) steric complementarity serves as the principal information carrier linking sequence to structure. Of the three, hydrophobicity and volume are properties intrinsic to the amino acids whereas steric complementarity has to do with the packing of buried residues with other amino acids non-local to it in sequence. The 'jigsaw puzzle' model first proposed by Crick[6] gave primary importance to internal packing as directing the three-dimensional protein fold. The 'knobs into holes' arrangement of residues involved in inter-helical association coupled with the fact that mutation of these residues led to the relative re-arrangement of the helices was cited in favour of the hypothesis[7]. The packing density inside proteins was also found to be high comparable to crystals of small organic molecules than oil[8,9], and was attributed to the exquisite surface complementarity between side-chains interdigitating in the interior of the molecule, rather like pieces of a three-dimensional jigsaw puzzle. In a recent experiment, Dahiyat and Mayo[10] redesigned the core of the B1 domain of protein G and demonstrated the stability of the protein as a function of packing efficiency. The assessment of steric complementarity using the method of small probe contact dots[11], reconfirmed the excellent packing interactions within proteins, with side-chains fitting neatly into their surrounding environment.

†For correspondence. (e-mail: surfaces@cmb2.saha.ernet.in)

Other calculations however failed to reveal any preference for pairwise interactions amongst buried apolar residues (except cysteine–cysteine due to the disulphide bridge) or any form of specificity within the core[12]. Experiments designed to probe the relative importance of hydrophobicity versus steric fit of a buried residue demonstrated that both for T4 lysozyme[13] and λ repressor[14] it was possible to generate physiologically viable protein folds with alternative core sequences, selected from the set of apolar residues. This was remarkable for T4 lysozyme[13] where mutation of 7 core residues to methionine yielded a structure with 50% activity of the wild type, implying specific packing interactions to be less critical in determining the overall fold. However, it was noted that although the overall fold of these mutant proteins were similar, they differed widely in enzymatic activity and thermal stability. Similar results were also obtained in the *de novo* design attempts of ubiquitin[15], where although alternative core arrangements gave a wide range of thermal stabilities, all mutants achieved a unique overall fold. These experiments thus indicated that the initial acquisition of the global fold by the protein depended to a large extent on the pattern of hydrophobicity down the polypeptide chain. Final optimization of the resulting structure for stability and activity was critically determined by packing details within the protein. Lattice calculations[16,17] in two and three dimensions confirmed that residue strings composed of only two residue types (hydrophobic, H and polar, P) gave compact shapes with well-defined cores. Design of a four-helix bundle on the basis of the H–P binary code alone with no reference to packing considerations was also experimentally successful[18]. Thus there is increasing consensus that the pattern of hydrophobic and hydrophilic residues down the polypeptide chain gives the initial impetus, to collapse the protein into its unique fold, with packing interactions being critical for the final optimization of the structure in terms of thermal stability and catalytic efficiency. However, more detailed data in the future will shed more light on the continuing controversy.

In the present context it would be of interest to calculate the average surface complementarity of the 20 amino acids buried in the interior of proteins thereby estimating the informational value of steric fit. The jigsaw puzzle model proposes high surface complementarity between the inter-digitating side-chains of interior residues. However, a quantitative analysis of the relative contributions of the side-chain and main-chain atoms to the steric fit has not been performed. Such a calculation would also shed light on the authenticity of the 'jigsaw puzzle' model. Removal of a buried residue would leave a cavity in the protein interior. The inner wall of this cavity would be composed of atoms contributed by the rest of the polypeptide chain. The present calculation estimates the steric fit of buried residues with the inner wall of their corresponding cavities. To this end a modified version of the function proposed by Lawrence and Colman[19] has been used.

Methods

A set of 50 protein crystal structures was selected from the SCOP[20] database and downloaded from the Brookhaven Protein Data Bank[21]. All the structures had a resolution better than or equal to 2.5 Å (with the exception of 1fps, res. 2.60 Å). The structures along with their respective resolutions and Protein Data Bank (PDB) codes are shown in Table 1. Care was taken to represent every class – all alpha, all beta, alpha and beta, alpha plus beta and multidomain, and to select structures with no irregularities or discontinuities in the polypeptide chain.

The surface complementarity calculation of the buried protein residues was performed in the following steps:

Determination of buried residues

Initially the solvent accessible area[9] of all the 20 amino acids in the Gly–Xaa–Gly fragment with extended conformation was calculated with a probe radius of 1.4 Å. The side-chain torsion angles (χ) were adjusted to their probable and extended values using the side-chain rotamer library[22]. The ratio of the solvent accessible area of a residue in the sample protein to that of the same residue in the tripeptide fragment was used to determine burial. A residue was taken to be completely buried if the value of the above ratio was less than or equal to 0.01. The surface complementarity of these buried residues has only been calculated. The need for such a stringent criterion to decide burial will be made clear when the surface complementarity function used in the calculation is discussed.

Surface generation

The objective of this study was to calculate the surface complementarity between the surface of a buried residue and the inner wall of the cavity enclosing it. Surface generation was performed using Connolly's algorithm[23] which samples any protein surface as a discrete set of dot surface points or area elements (Figure 1). Each area element has associated with it a coordinate defining its position and its normal. The dot surface points for every residue of the polypeptide chain of the sample protein were calculated along with its sequentially adjacent residues. As the N and C atoms of every residue (barring terminal residues) are covalently bonded to the adjacent residue additional surface points would be generated if the dot surface points for every residue are calculated in isolation. After surface calculation of the tripeptide, the surface points of the central target residue were stored and

Table 1. PDB codes for 50 protein structures used in S_m calculation along with their respective resolutions and crystallographic R-factors

PDB code	Resolution (Å)	R value	PDB code	Resolution (Å)	R value
2cnd(α)	1.45	0.161	1cnh(α)	2.10	0.197
1lmb(α)	1.80	0.189	2asr(α)	2.30	0.203
2gst(α)	1.80	0.160	2pgd(α)	2.00	0.198
1gln(α)	2.50	0.185	1axn(α)	1.78	0.177
1dsb(α)	2.00	0.169	1ytf(α)	2.50	0.235
1fps(α)	2.60	0.194	1tta(β)	1.70	0.168
2bbk(β)	1.75	0.167	4gcr(β)	1.47	0.181
2sil(β)	1.60	0.166	2pec(β)	2.20	0.180
1dyn(β)	2.20	0.200	1thw(β)	1.75	0.181
1dif(β)	1.70	0.198	1eur(β)	1.82	0.173
1htp(β)	2.20	0.185	2rma(β)	2.10	0.170
1tml($\alpha\|\beta$)	1.80	0.184	1esc($\alpha\|\beta$)	2.10	0.166
6xia($\alpha\|\beta$)	1.65	0.141	1xyz($\alpha\|\beta$)	1.40	0.183
1lau($\alpha\|\beta$)	1.80	0.174	1dea($\alpha\|\beta$)	2.10	0.174
1phr($\alpha\|\beta$)	2.10	0.167	1vhr($\alpha\|\beta$)	2.10	0.176
1rva($\alpha\|\beta$)	2.00	0.162	3eca($\alpha\|\beta$)	2.40	0.149
2dri($\alpha\|\beta$)	1.60	0.194	1tph($\alpha\|\beta$)	1.80	0.185
1rnb($\alpha\|\beta$)	1.90	0.214	6lyz($\alpha + \beta$)	2.00	–
6rat($\alpha + \beta$)	1.50	0.152	3sic($\alpha + \beta$)	1.80	0.178
2cmd($\alpha + \beta$)	1.87	0.188	1ubi($\alpha + \beta$)	1.80	0.165
1kpa($\alpha + \beta$)	2.00	0.192	2aak($\alpha + \beta$)	2.40	0.221
2chs($\alpha + \beta$)	1.90	0.194	1sry($\alpha + \beta$)	2.50	0.184
1pne($\alpha + \beta$)	2.00	0.165	1lit($\alpha + \beta$)	1.55	0.180
1mkb($\alpha + \beta$	2.00	0.183	1rpl(md)	2.30	0.220
1jud(md)	2.50	0.193	1fin(md)	2.30	0.208

The class to which the protein belongs is enclosed in parenthesis (all alpha, α; all beta, β; alpha and beta, $\alpha|\beta$; alpha plus beta, $\alpha + \beta$; multidomain, md).

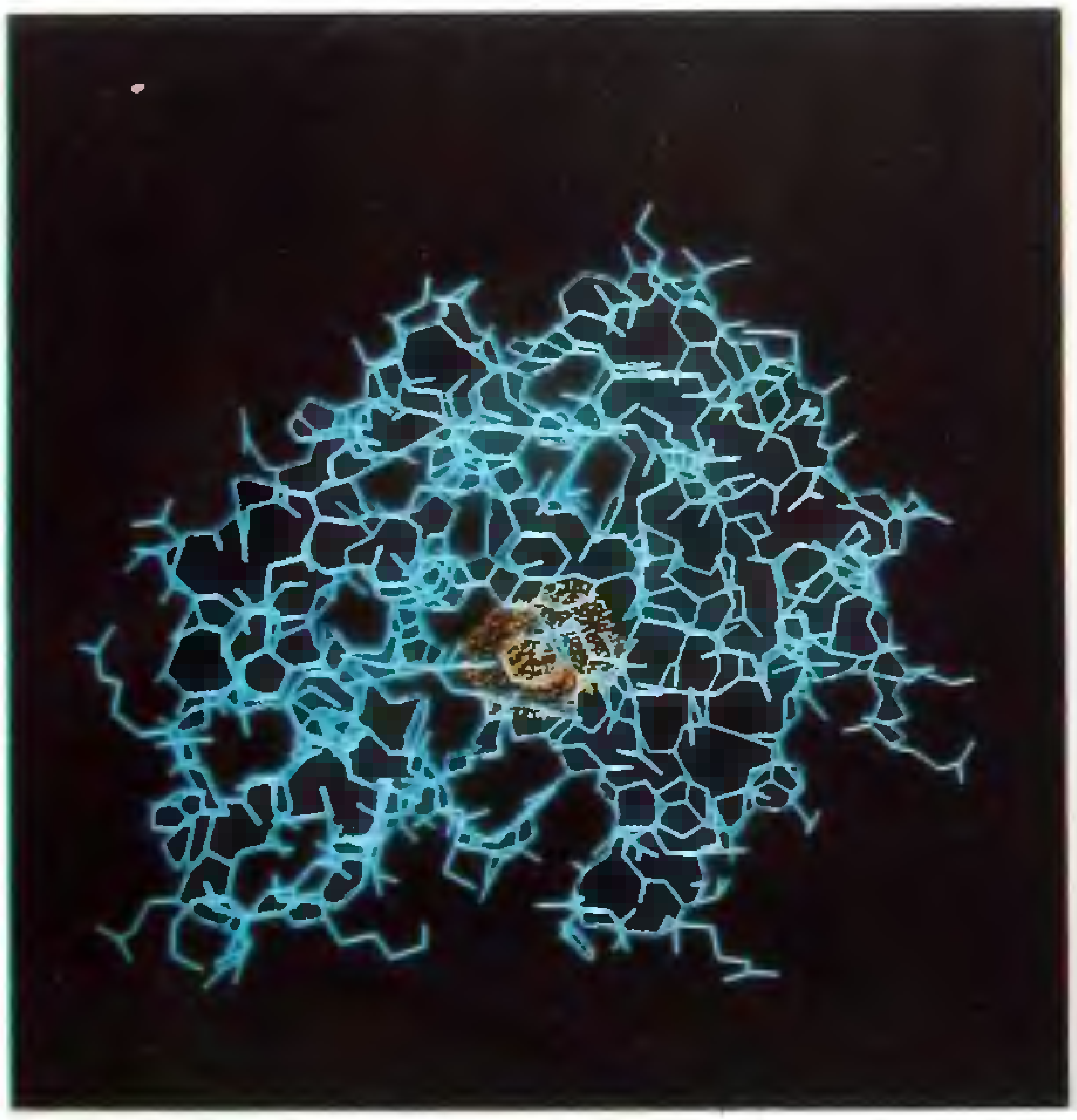

Figure 1. Dot surface points of a residue (Phe 112) buried in the interior of a protein (PDB code 2rma, ref. 34) displayed in FRODO[35].

those of the adjacent residues rejected. The entire dot surface for the polypeptide was thus generated in triplets (doublets for terminal residues) moving down the chain, with the protein surface being sampled at 10 dots/Å^2.

Calculation of surface complementarity for buried residues

The surface complementarity statistic S_m used in this calculation is a modified version of the function S_c proposed by Lawrence and Colman[19]. Consider P_A to be the set of dot surface points of a residue buried according to the criterion given earlier. Let x_A be the coordinates of a dot surface point selected from P_A (Figure 2). The surface point nearest in distance to x_A was searched from the set of points belonging to the rest of the polypeptide chain excluding the residue in question. For residues adjacent in sequence to the buried residue corresponding to P_A only surface points of the side-chain atoms were considered provided x_A was itself a side-chain atom surface point. Otherwise dot surface points of sequentially adjacent residues were excluded in a search to find the surface point nearest in distance to x_A. Thus for every buried residue, the surface of its immediate neighbourhood enclosing it is defined by the set of points nearest in distance to all the surface points in P_A, contributed by the rest of the polypeptide chain. Then following Lawrence and Colman[19],

$$S(x_A) = n_A \cdot n'_A \exp(-w|x_A - x'_A|^2), \qquad (1)$$

where x'_A is the surface point nearest in distance to x_A and n_A, n'_A are the surface normals at x_A and x'_A respectively. The scalar function (1) can then be defined for every dot surface point in P_A. S_m was then defined as $S_m = \{S\}$, where the curly brackets denote the median of the distribution of S over P_A. S_m is thus a truncated form of the functon proposed by Lawrence and Colman[19]. Here w was set to 0.5. In the calculation of S_c (the original form of the function proposed by Lawrence and Colman) a 1.5 Å band of surface points located at the periphery of the two interacting surfaces was removed from the calculation. In the present case no surface point has been removed from the calculation as complete burial of the residue has been ensured with a stringent criterion.

Explicit hydrogen atoms and atomic van der Waal's radii

Lawrence and Colman have noted that the value of S_c is relatively insensitive to the van der Waal's radii of the atoms. Calculation for the set of 50 proteins was performed without explicitly fixing hydrogens and the atomic van der Waal's radii used were from the general molecular mechanics force field[24]. For a select subset of 20 highly resolved structures the calculations were repeated after fixing all the hydrogen atom positions using X-PLOR[25]. Consequently, the atomic radii were also changed in accordance with the new force field[26]. Barring minor details, the results of both sets of calculations with and without explicit hydrogens were similar. However, the average S_m values for the 20 amino acids showed a maximum rise of about 5% when compared with the non-hydrogen set of calculations, due to denser packing on including additional hydrogen atoms.

Side-chain and main-chain atoms

For every buried residue the S_m calculation was performed thrice, each time with a different set of points. For Case 1, S_m was calculated between the surface points of the side-chain atoms of the buried residue with respect to its corresponding set of nearest neighbour points contributed by the side-chain atoms alone. For Case 2, S_m was calculated between the surface points of the side-chain atoms alone of the buried residue and its corresponding set of nearest neighbour points irrespective of having been contributed by side-chain or main-chain atoms. Lastly for Case 3, all the dot surface points both for the buried residue and its set of nearest neighbour area elements were considered in the S_m calculation with no discrimination between side-chain or main-chain atoms. In summary, Case 1 estimates side-chain versus side-chain steric fit, Case 2 side-chain versus all atoms and Case 3 all atoms of the buried residue versus all atoms of its environment.

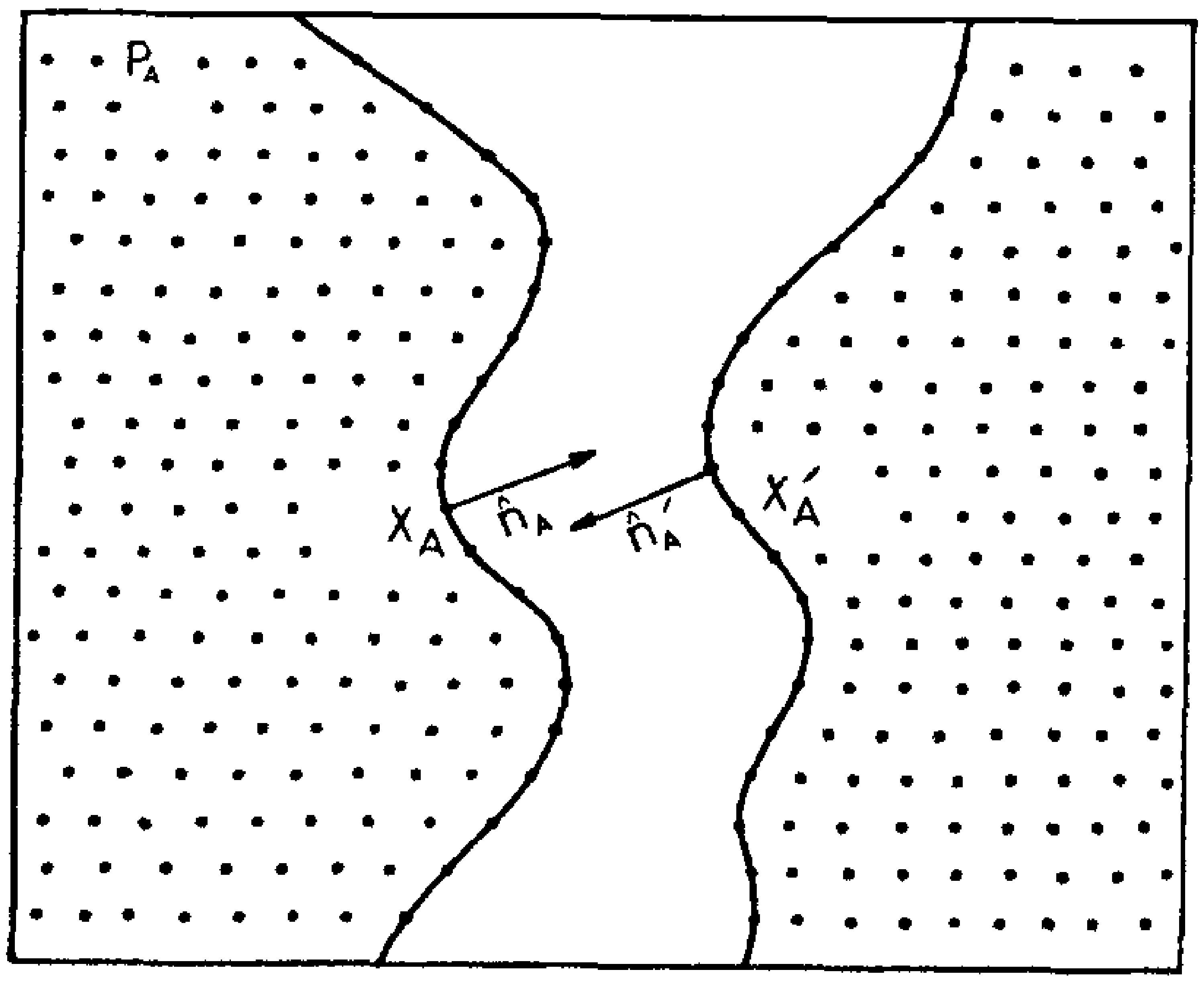

Figure 2. The surface complementarity function[19] defined as the median of the distribution $S(x_A) = n_A n'_A \exp(-w|x_A-x'_A|^2)$, over all the points of the buried residue (P_A), where x_A is a dot surface point of the buried residue corresponding to P_A and x'_A is its nearest neighbour contributed by the rest of the protein. Their normals are n_A, n'_A respectively.

The propensity P (ref. 27), for a residue to get buried was calculated as the proportion of a particular amino acid buried divided by the proportion of all amino acids buried.

$$P = (N_b/T_b)/(N_p/T_p),$$

where N_b is the number of residues of a particular amino acid, b, buried in the interior, T_b is the total number of residues of that specific amino acid in the data base, N_p is the total number of residues which are buried and T_p is the total number of residues in the database. P thus gives a value of 1 if there is no preference, whereas values greater than 1 favour the amino acid in the buried environment and those lesser than 1 do not.

Results and discussion

Analysis of the 50 proteins yielded 2457 buried residues out of a total of 11,312 amino acids. All the hydrophobic residues alanine, valine, leucine, isoleucine, phenylalanine and methionine show burial propensity (Table 2) higher than 1.00. Cysteine exhibits the highest propensity of 2.21. Tryptophan, by virtue of its large size and unusual shape, recorded a comparatively low count, thereby indicating its limited acceptability in cores. Tyrosine, serine and threonine have the highest propensities of 0.88, 0.90 and 0.82, respectively amongst the polar amino acids. The relatively high incidence of serine and threonine in protein interiors is perhaps due to their ability to form a hydrogen bond back to the main-chain[28]. The burial of fully charged residues aspartic acid, glutamic acid, histidine, arginine and lysine occurs 116 times in terms of numbers. The lowest frequency is for lysine which occurs thrice. Cysteine was treated as an outlier in the entire analysis as it forms disulphide bridges. Comparison of the distribution patterns of the 20 amino acids buried in the interior of the protein on one hand, and those found in subunit interfaces reveal significant differences[29,30]. Although the interface is more hydrophobic than the rest of the surface it is not as pronounced as protein interiors, with a higher proportion of charged or polar amino acids being buried at the interface[29]. Notable among them is arginine which occurs with high frequency in interfaces[31] contrary to interiors where it has the lowest propensity greater than only lysine. Again, amongst the buried hydrophobic residues in subunit interfaces aromatic residues are favoured over aliphatic amino acids with the case being the reverse for protein interiors. Methionine maintains high frequency of occurrence in both cases[32].

Mathematically, the highest value obtainable by the presently used surface complementarity function S_m is 1.00 which enjoins an absolute matching of two surfaces. Previous discussions of steric complementarity in the literature involved terms like 'exquisite' which gave the qualitative impression of extremely high surface fit. However, in the present case the initial task would be to associate the numerical range of the function S_m with a more physically meaningful description. There is always an arbitrary element associated in correlating qualitative terms with numerical values of a function, in this case S_m. The highest possible value (Table 3) of S_m obtainable by any residue saturates to about 0.75 for those residues with side chains and to 0.80 for glycine. The value of S_c obtained by Lawrence and Colman[19] between the surfaces of interacting subunits of an oligomeric protein was 0.72. However, differences in the definitions of S_c and S_m do not warrant a direct comparison. Minimum S_m falls to 0.36 for Case 3 (with all atoms considered) but is much lower for Case 1. Comparison with another measure of packing density proposed by Gregoret and Cohen[33] shows S_m to be greater than 0.60 for most residues with dense packing (according to the statistics proposed by Gregoret and Cohen).

Thus it may be reasonable to assume that residues with S_m greater than 0.60 have excellent fit (jigsaw puzzle class) with their surrounding surface, approaching the best possible complementarity of surfaces in the protein interiors. Residues with S_m in the range 0.50–0.60 would have good complementarity, those in the range 0.40–0.50 mediocre and less than 0.40 would definitely be classified as poor.

The distribution of S_m values for the twenty residues appears quite uniform (Tables 3 and 4 *a–c*). For Case 3 (Tables 3 and 4 *c*, Figure 3 *c*) when all surface points were considered irrespective of having been contributed by side- or main-chain atoms, the frequency of occurrence of most of the residues of the 20 amino acids fell in the range 0.50–0.70, with a relatively smaller fraction less than 0.40 on one hand, and greater than 0.70 on the other. Almost all the buried residues exhibit high complementarity with roughly half in the jigsaw puzzle class (> 0.60) and the rest lying between 0.50 and 0.60.

Global average of all the residues in Case 3 is 0.60 indicative of the high goodness-of-fit upon burial. Comparison of the average S_m values of the 20 amino acids shows good agreement within the limits of error (Figure 4). The highest $\langle S_m \rangle$ is maintained by tryptophan in all the three cases. This may be due to the fact that the large vol-

Table 2. Propensity for amino acid burial

Residue	Total (T_b)	Buried (N_b)	P
GLY	927	201	1.00
ALA	954	317	1.53
VAL	778	327	1.94
LEU	968	402	1.91
ILE	604	284	2.16
MET	223	91	1.88
PHE	454	149	1.51
TYR	417	80	0.88
TRP	165	35	0.98
SER	667	131	0.90
THR	635	113	0.82
ASN	512	55	0.49
GLN	427	22	0.24
ASP	700	41	0.27
GLU	666	17	0.12
HIS	230	33	0.66
LYS	674	3	0.02
ARG	617	22	0.16
PRO	527	54	0.47
CYS	167	80	2.21

*For a particular residue, N_b is the number buried and T_b is the total number of this residue in the database. $P = (N_b/T_b)/(N_p/T_p)$, where N_p is the total number of amino acid residues found buried and T_p is the total number of residues in the database.

Table 3. Average and standard deviation of S_m for buried amino acid residues for Cases 1–3

Buried residues	Average $S_m(\sigma)$		
	Case 1	Case2	Case3
GLY	–	–	0.60 (0.08)
ALA	0.44 (0.12)	0.58 (0.07)	0.60 (0.07)
VAL	0.50 (0.09)	0.58 (0.05)	0.59 (0.05)
LEU	0.49 (0.08)	0.57 (0.06)	0.58 (0.05)
ILE	0.50 (0.08)	0.58 (0.06)	0.59 (0.05)
MET	0.49 (0.10)	0.59 (0.06)	0.60 (0.05)
PHE	0.52 (0.07)	0.60 (0.05)	0.61 (0.05)
TYR	0.49 (0.07)	0.59 (0.05)	0.61 (0.05)
TRP	0.50 (0.05)	0.62 (0.05)	0.63 (0.05)
SER	0.38 (0.13)	0.58 (0.08)	0.61 (0.07)
THR	0.44 (0.11)	0.58 (0.06)	0.60 (0.06)
ASN	0.42 (0.11)	0.59 (0.06)	0.60 (0.07)
GLN	0.43 (0.09)	0.59 (0.06)	0.59 (0.06)
ASP	0.44 (0.11)	0.57 (0.07)	0.59 (0.06)
GLU	0.43 (0.11)	0.59 (0.06)	0.59 (0.05)
HIS	0.45 (0.09)	0.58 (0.05)	0.59 (0.05)
LYS	0.27 (0.12)	0.55 (0.02)	0.57 (0.02)
ARG	0.40 (0.08)	0.55 (0.06)	0.56 (0.06)
PRO	0.45 (0.12)	0.59 (0.06)	0.59 (0.06)
CYS	0.42 (0.15)	0.53 (0.12)	0.56 (0.10)

Case 1, Surface complementarity S_m between buried residues and their corresponding inner cavities. Here the surface points contributed by the side-chain atoms alone have been considered; Case 2, S_m has been calculated between dot points of buried residues contributed by side-chain atoms alone and all the dot points of the inner cavity wall; Case 3, S_m between buried residue and cavity with all points included irrespective of whether contributed by side- or main-chain atoms.

ume of tryptophan would have a singularly disruptive influence on the molecule unless its compact and efficient packing was ensured. The charged residues with extended side-chains lysine and arginine showed the lowest complementarity. The uniformity in $\langle S_m \rangle$ for every type of residue indicates that upon burial all residues regardless of shape, size or charge exhibit the same measure of complementarity with their immediate neighbourhood. Case 2 (Figure 3 *b*) resembles Case 3 in the pattern of S_m distribution. However, in Case 2 residues occur with double the frequency in the range 0.50–0.60 than in the range 0.6–0.7, in contrast to Case 3 where both the ranges are more or less evenly populated. The global average of S_m in Case 2 also falls marginally to 0.58 compared to Case 3.

Examination of Case 1 (Tables 3 and 4 *a* and Figure 3 *a*) where surface points contributed by side-chain atoms alone have been considered exhibits an altered distribution of S_m. The standard deviation in S_m for most of the 20 amino acids in Case 1 is approximately double that in Cases 2 and 3. The bulk of the residues fall between 0.5 and 0.6, though a significant fraction of residues appear in the range 0.40–0.50. This is in contrast to Case 3 where

Table 4*a*. Frequency of occurrence of the buried residues with respect to S_m for Case 1*

S_m →	0.0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8
GLY	–	–	–	–	–	–	–	–
ALA	5	6	31	65	102	87	21	0
VAL	0	1	10	28	121	142	25	0
LEU	2	1	4	32	170	171	22	0
ILE	0	0	7	25	85	149	17	1
MET	0	1	3	9	32	34	12	0
PHE	0	0	2	8	35	92	12	0
TYR	0	0	2	8	28	39	3	0
TRP	0	0	0	2	15	18	0	0
ASN	1	2	7	9	25	10	1	0
GLN	0	0	1	7	10	3	1	0
SER	5	10	20	32	39	23	2	0
THR	1	2	13	14	42	39	1	1
ASP	0	1	5	7	13	14	1	0
GLU	0	1	2	2	8	4	0	0
HIS	0	1	2	2	20	8	0	0
LYS	0	1	0	2	0	0	0	0
ARG	0	0	3	4	13	2	0	0
PRO	0	2	4	8	21	16	3	0
CYS	1	8	10	12	19	26	4	0

*Surface points contributed by side-chain atoms alone have been considered both for the residue and its inner cavity. Ranges of S_m not included in the tables are not populated by any residue.

Table 4*b*. Frequency of occurrence of the buried residues with respect to S_m for Case 2*

S_m →	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8
GLY	–	–	–	–	–	–
ALA	0	5	46	130	130	6
VAL	0	0	30	172	124	1
LEU	0	2	46	236	118	0
ILE	0	0	26	162	93	3
MET	0	0	8	42	39	2
PHE	0	0	5	73	70	1
TYR	0	0	3	43	34	0
TRP	0	0	0	10	23	2
ASN	0	0	5	28	20	2
GLN	0	0	1	12	9	0
SER	0	3	23	49	49	7
THR	0	1	10	67	32	3
ASP	0	0	8	16	17	0
GLU	0	0	1	9	7	0
HIS	0	0	1	21	11	0
LYS	0	0	0	3	0	0
ARG	0	0	3	13	6	0
PRO	0	0	4	28	21	1
CYS	4	13	8	24	30	1

*S_m has been calculated between dot points of buried residues contributed by side-chain atoms alone and all the dot points of the inner cavity wall. Ranges of S_m not included in the tables are not populated by any residue.

the overwhelming number of residues are greater than 0.50. The global average of S_m for all the 20 amino acids in Case 1 is 0.47. Average S_m values for each of the 20 amino acids rarely cross 0.50. There is a drop in $\langle S_m \rangle$ from 0.49 for apolar to about 0.43 for polar and charged amino acids (Case 1). The lowest value of 0.27 was recorded for lysine (although the statistics for it are bad due to the low count of lysine as buried). The differences in $\langle S_m \rangle$ between apolar, polar and charged residues in Case 1 are levelled out in Cases 2 and 3 (Table 3). Presence of buried residues with S_m less than 0.40 in Case 1 shows that packing among side-chain atoms alone can at times be attended with serious defects.

Therefore all residues irrespective of shape, size and charge bury with the same measure of complementarity (within the limits of error) in protein interiors, which is indeed high when all the atoms have been considered in the calculation (Case 3). The differences in the global averages between Cases 1 and 3 indicate the non-trivial role played by the main-chain atoms in internal packing over and above side-chain interactions. Even lysine which recorded the lowest count in Case 1 rises to about 0.57 in Case 3. Further, the packing between side-chains may not be invariably high (jigsaw puzzle category), and can occasionally be attended with considerable packing defects.

Identical calculations performed with explicit hydrogen atoms (Table 5) on a subset of 20 highly resolved (equal to or better than 2.00 Å) structures, gave an identical pattern of results, though with an increase in average S_m values by a maximum of about 5%. The present calculations

Table 4c. Frequency of occurrence of the buried residues with respect to S_m for Case 3

$S_m \rightarrow$	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8
GLY	3	20	64	95	19
ALA	3	24	115	163	12
VAL	0	13	155	156	3
LEU	0	22	227	153	0
ILE	0	10	141	130	3
MET	0	2	38	48	3
PHE	0	3	51	92	3
TYR	0	0	36	44	0
TRP	0	1	8	24	2
ASN	0	5	18	29	3
GLN	0	2	10	10	0
SER	0	9	50	62	10
THR	1	4	57	49	2
ASP	0	3	19	18	1
GLU	0	0	9	8	0
HIS	0	3	15	14	1
LYS	0	0	3	0	0
ARG	0	3	13	6	0
PRO	0	3	27	23	1
CYS	8	12	23	36	1

*All surface points irrespective of whether contributed by side- or main-chain atoms have been included in the S_m calculation. Ranges of S_m not included in the tables are not populated by any residue.

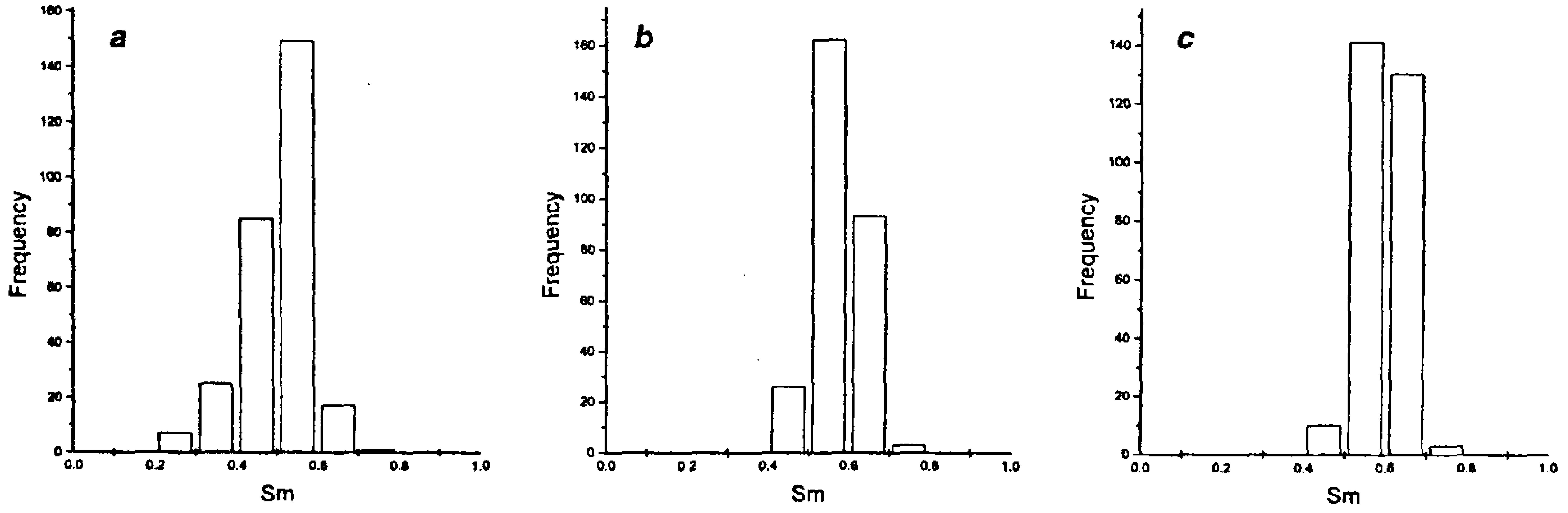

Figure 3. Frequency distribution of residue Ile with respect to S_m. ***a***, Case 1, where only the side-chain dot surface points have been considered for S_m; ***b***, Case 2, with S_m calculated between dot points of buried residues contributed by side-chain atoms alone and all the dot points of the inner cavity wall; ***c***, Case 3, S_m calculated with all points included irrespective of whether contributed by side- or main-chain atoms.

Table 5. Calculation of average S_m for the 20 amino acids with explicit hydrogens for Cases 1–3

Buried residues	⟨S_m⟩		
	Case 1	Case 2	Case 3
GLY	–	–	0.615
ALA	0.437	0.565	0.609
VAL	0.508	0.600	0.616
LEU	0.518	0.601	0.616
ILE	0.525	0.602	0.618
MET	0.533	0.640	0.640
PHE	0.554	0.644	0.656
TYR	0.491	0.616	0.633
TRP	0.527	0.652	0.656
ASN	0.435	0.598	0.613
GLN	0.479	0.622	0.603
SER	0.381	0.597	0.627
THR	0.459	0.592	0.619
ASP	0.431	0.604	0.622
GLU	0.505	0.616	0.630
HIS	0.493	0.629	0.620
LYS	0.130	0.557	0.569
ARG	0.441	0.571	0.572
PRO	0.479	0.637	0.645
CYS	0.454	0.562	0.579

PDB codes of the structures used are: 1dif, 1dsb, 1kpa, 1lau, 1lit, 1lmb, 1pne, 1rva, 1thw, 1tml, 1tta, 2chs, 2cmd, 2dri, 2end, 2gst, 3sic, 4gcr, 6lyz, 6rat.

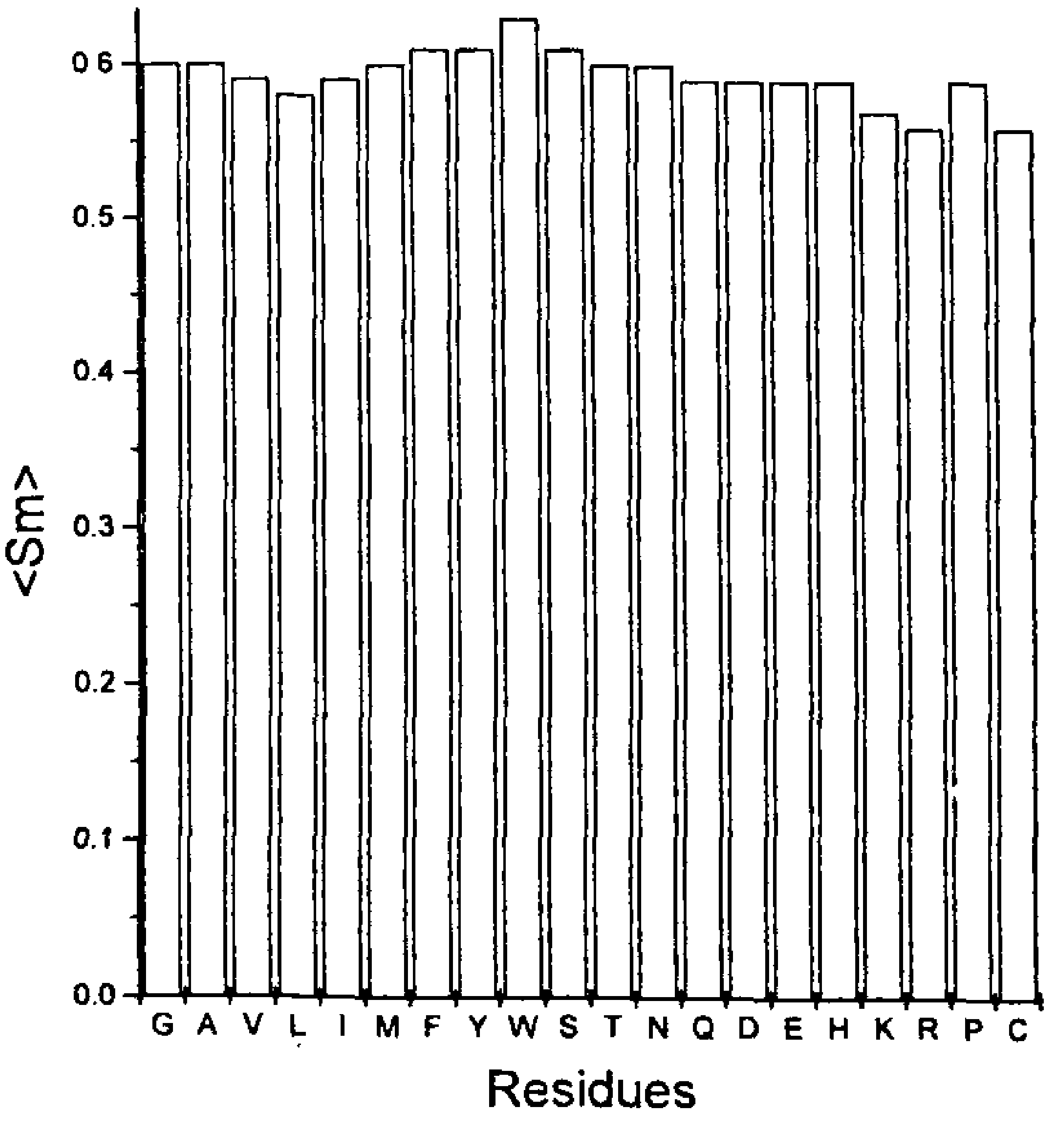

Figure 4. Average S_m for Case 3 for the 20 amino acids with value given in Table 3.

Table 6. Maximum and minimum S_m for buried amino acids for Case 1 (for structures in Table 1). For the same residue S_m has been calculated with the side-chain atoms of the buried residue against the main-chain atoms alone of its enclosing cavity (side-chain vs main-chain). The corresponding S_m for Case 2 for each residue (Case 2* for minimum and Case 2 ~ for maximum S_m in Case 1) is also tabulated

Buried residue	Case 1 (minimum)	Side-chain vs main-chain	Case 2*	Case 1 (maximum)	Side-chain vs main-chain	Case 2 ~
ALA	0.01	0.51	0.61	0.67	0.00	0.68
VAL	0.19	0.42	0.62	0.67	0.00	0.68
LUE	0.03	0.57	0.63	0.66	0.00	0.66
ILE	0.22	0.34	0.58	0.71	0.00	0.72
MET	0.15	0.16	0.49	0.69	0.08	0.74
PHE	0.29	0.24	0.56	0.66	0.01	0.69
TYR	0.24	0.16	0.52	0.64	0.01	0.65
TRP	0.36	0.25	0.69	0.59	0.00	0.60
SER	0.04	0.55	0.69	0.61	0.01	0.65
THR	0.09	0.41	0.56	0.70	0.01	0.71
ASN	0.09	0.38	0.43	0.61	0.03	0.66
GLN	0.26	0.34	0.59	0.62	0.00	0.70
ASP	0.16	0.23	0.47	0.61	0.10	0.65
GLU	0.17	0.15	0.51	0.60	0.04	0.63
HIS	0.17	0.38	0.58	0.60	0.00	0.61
LYS	0.11	0.37	0.58	0.36	0.04	0.53
ARG	0.24	0.11	0.54	0.55	0.05	0.67
PRO	0.18	0.21	0.55	0.64	0.01	0.68
CYS	0.08	0.12	0.29	0.65	0.00	0.67

in full agreement with another estimation of goodness-it by the method of small probe contact dots[11], in that ı complementarity has been found for buried residues, ıgh in the other algorithm inclusion of explicit hydros has been considered essential for calculation. In the ;ent case, calculation of average S_m for the 20 amino ls did not yield any additional insight with the incluı of explicit hydrogen atoms. However, in order to judge alternative conformations within the core of a singl protein, hydrogen atoms should be included for greate sensitivity to torsional rotations of the side-chain.

Conclusion

Although information regarding the three-dimensiona structure of a protein is contained within its sequence o amino acid residues, not all the regions of the molecule have the same informational content. The information ('the difference which makes a difference') could lie either in the pattern of hydrophobic/hydrophylic residues down the chain or in the shapes of amino acids which enable them to pack efficiently to form compact molecular interiors. There have been attempts to find conformational specificity within the core[12], perhaps in the hope of finding two amino acids with complementary surfaces rather like adjacent pieces of a jigsaw puzzle which would then occur with increased pairwise frequency within the cores. To date this hope has however been belied. Simple rules have yet to emerge, elucidating the complex arrangement of atoms distributed over the whole polypeptide chain by which a protein achieves the high inner complementarity of surfaces.

The jigsaw puzzle model of protein folding states that the acquisition of the overall fold by the protein is governed predominantly by internal packing considerations primarily due to the exquisite surface complementarity between the side-chains. That is, the packing directs the fold. In view of the present calculations, the very high overall 'goodness of fit' of buried residues is indeed borne out. It is also true that this fit is primarily mediated by the amino acid side chains. Global average in S_m for the 20 highly resolved structures in Table 5 is 0.48 for Case 1 (calculated without hydrogen). For the same set of structures S_m calculated between the side-chain atoms of the buried residues against the main-chain atoms alone of the surrounding cavity averages to only 0.09. The global average for Case 2 is 0.59. Although the steric fit is dominated by side-chain atoms, main-chain atoms do make a non-trivial contribution. This is dramatically demonstrated where the fit between side-chain atoms shows serious packing defects which are then adequately compensated by main-chain atoms (Table 6). On the other hand, where the side-chains show excellent 'goodness of fit', the S_m for side-chains (of buried residues) against surrounding main-chain atoms collapses almost to zero (Table 6). Thus, it is the combination of main and side-chain atoms in protein interiors that makes packing remarkably tight and uniform, with side-chain atoms playing the pre-eminent role. So the empirical foundations of the jigsaw puzzle model of protein folding are by and large valid.

It is however still controversial whether the tight packing found in proteins determines its overall fold. Most experimental approaches[14,15] in deciding this issue make a distinction between acquisition of the gross overall native fold on one hand, and subsequent optimization of the structure in terms of thermal stability and catalytic efficiency on the other. There is absolutely no doubt that the stability of protein molecules is critically dependent on detailed packing interactions in the core of the molecule. However, alternative core sequences can support a unique fold though with different thermal stabilities[14]. The present calculation indicates that no amino acid seems to be preferred by virtue of its shape given the fact that all residues pack with the same measure of complementarity upon burial (Case 3). Although the global average for the hydrophobic residues (ala, val, ile, leu, phe) in Case 1 is 0.49 and thus relatively higher than polar and charged residues (0.43), the burial of apolar residues in the interior has more to do with their hydrophobicities rather than their intrinsic ability to pack better than other amino acids. The difference in hydrophobicities between ala, val, leu, ile, phe and other polar or charged amino acids appears to be more significant in accounting for their increased tendency to bury in the protein interiors, than their packing ability. The partitioning of hydrophobic amino acids in the interior of the protein molecule is a step which occurs early in the protein folding pathway. It is in this context that perhaps the primary initiating step in protein folding may be due to hydrophobicities rather than specific packing interactions. However, given that all residues exhibit high steric fit within a narrow range (Case 3) complementarity should act as a powerful filter in rejecting all structures which do not satisfy the rather stringent packing recquirements within the core of a protein.

1. McDonald, I. K. and Thornton, J. M., *J. Mol. Biol.*, 1994, **238**, 773–793.
2. Dill, K., *Biochemistry*, 1990, **29**, 7133–7155.
3. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. and Sauer, R. T., *Science*, 1990, **247**, 1306–1310.
4. Lim, W. A., Farruggio, D. C. and Sauer, R. T., *Biochemistry*, 1992, **31**, 4324–4333.
5. Varadarajan, R. and Richards, F. M., *Biochemistry*, 1992, **31**, 12315–12327.
6. Crick, F. H. C., *Acta Crystallogr.*, 1953, **6**, 689–697.
7. Bashford, D., Chothia, C. and Lesk, A. M., *J. Mol. Biol.*, 1987, **196**, 199–216.
8. Richards, F. M., *J. Mol. Biol.*, 1974, **82**, 1–14.
9. Richards, F. M., *Annu. Rev. Biophys. Bioeng.*, 1977, **6**, 151–176.
10. Dahiyat, B. I. and Mayo, S. L., *Science*, 1997, **278**, 82–87.
11. 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., *J. Mol. Biol.*, 1999, **285**, 1711–1733.
12. Behe, M. J., Lattman, E. E. and Rose, G. D., *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 4195–4199.
13. Gassner, N. C., Baase, W. A. and Matthews, B. W., *Proc. Natl. Acad. Sci. USA.*, 1996, **93**, 12155–12158.
14. Lim, W. A. and Sauer, R. T., *J. Mol. Biol.*, 1991, **219**, 359–376.
15. Lazar, G. A., Desjarlais, J. R. and Handel, T. M., *Protein Sci.*, 1997, **6**, 1167–1178.
16. Sikorski, A. and Skolnick, J., *Proc. Natl. Acad. Sci. USA.*, 1989, **86**, 2668–2672.
17. Covell, D. G. and Jernigan, R. L., *Biochemistry*, 1990, **29**, 3287–3294.
18. Kametkar, S., Schiffer, J. M., Xiong, H., Babik, J. M. and Hecht, M. H., *Science*, 1993, **262**, 1680–1685.
19. Lawrence, M. C. and Colman, P. M., *J. Mol. Biol.*, 1993, **234**, 946–950.
20. Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C., *J. Mol. Biol.*, 1995, **247**, 536–540.
21. Abola, E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. and Weng, J., in *Crystallographic Databases–Information Content,*

Software Systems, Scientific Applications (eds Allen, F. H., Bergerhof, G. and Sievers, R.), Commision of the International Union of Crystallography, Bonn, 1987, pp. 107–132.

22. Ponder, J. W. and Richards, F. M., *J. Mol. Biol.*, 1987, **193**, 775–791.
23. Connolly, M. L., *J. Appl. Crystallogr.*, 1983, **16**, 548–558.
24. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P., *J. Am. Chem. Soc.*, 1984, **106**, 765–784.
25. Brunger, A. T., Kuriyan, J. and Karplus, M., *Science*, 1987, **235**, 458–460.
26. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. Jr., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. and Kollman, P. A., *J. Am. Chem. Soc.*, 1995, **117**, 5179–5197.
27. Moodie, S. L., Mitchell, J. B. and Thornton, J. M., *J. Mol. Biol.*, 1996, **263**, 486–500.
28. Gray, T. M. and Matthews, B. W., *J. Mol. Biol.*, 1984, **175**, 75–81.
29. Tsai, C. J., Lin, S. L., Wolfson, H. J. and Nussinov, R., *Protein Sci.*, 1997, **6**, 53–64.
30. Janin, J., Miller, S. and Chothia, C., *J. Mol. Biol.*, 1988, **204**, 155–164.
31. Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. and Bell, J. A., *Proteins*, 1997, **28**, 494–514.
32. Jones, S. and Thornton, J. M., *J. Mol. Biol.*, 1997, **272**, 121–132.
33. Gregoret, L. M. and Cohen, F. E., *J. Mol. Biol.*, 1990, **211**, 959–974.
34. Ke, H., Mayrose, D., Belshaw, P. J., Alberg, D. G., Schreiber, S. L., Chang, Z. Y., Etzkorn, F. A., Ho, S. and Walsh, C. T., *Structure*, 1994, **2**, 3.
35. Jones, T. A., *J. Appl. Crystallogr.*, 1978, **11**, 268–272

ACKNOWLEDGEMENTS. We thank Profs M. Lawrence and P. Colman for help regarding surface complementarity programs. We also thank Rakhi Sengupta and Ravi Chandran for stimulating discussions and Abhijit Bhattacharya for technical support.

Received 6 August 1999; revised accepted 4 November 1999