J. Mol. Biol. (1992) **224**, 629–638

Analysis of Non-polar Regions in Proteins

M. Ya. Karpeisky

Engelhardt Institute of Molecular Biology
U.S.S.R. Academy of Sciences, Vavilov St 32, Moscow, U.S.S.R.

and V. A. Ilyin

Shubnikov Institute of Crystallography
U.S.S.R. Academy of Sciences, Leninsky St 59, Moscow, U.S.S.R.

(Received 10 June 1991; accepted 23 October 1991)

A new method for finding hydrophobic nuclei and microclusters in protein structure is proposed. The method uses simple and clear-cut criteria based on an analysis of distances between the hydrocarbon groups of all residues.

A detailed analysis of the composition and properties of hydrophobic nucleic and microclusters for proteins of different types has been carried out. This approach reveals that a hydrophobic nucleus can be composed not merely of classical hydrophobic amino acids, but also of dicarboxylic acids, their amides, arginine, lysine, histidine and tyrosine. The hydrophobic nucleus defined by this method should be considered as an individual structural unit along with such elements of the secondary structure as alpha-helices, beta-turns and beta-sheets.

Keywords: hydrophobic nuclei; protein stability

1. Introduction

Non-covalent interactions are known to be of importance in the stabilization of the native conformations of proteins (Kauzmann, 1959; Singh & Thornton, 1990; Privalov & Gill, 1988; Wood & Thompson, 1990; Spolar *et al.*, 1989; Sali & Blundell, 1990; Dill, 1990). The role of main-chain hydrogen bonds in maintaining the structure of the regular elements of protein globules is well documented (for a review, see Chothia, 1984). In addition, the role of non-covalent side-chain interactions in the stabilization of the mutual orientation of regular elements of protein structure, as well as their role in protein folding, has been discussed (Chou *et al.*, 1990; Nemethy & Scheraga, 1979; Creighton & Chothia, 1989). To determine the contributions of non-covalent interactions between the side-chains of amino acid residues to hydrophobic nuclei and clusters in a protein the non-polar contacts must be analyzed according to the data on three-dimensional structures. A commonly used approach to assign a residue to a hydrophobic cluster or nucleus is based on calculations of the accessibility of a residue to water molecules (Richards, 1977; Lesser & Rose, 1990; Plochocka *et al.*, 1988). This information is necessary but insufficient to study the topology, structure and properties of the non-polar regions of a protein. Also, there is no agreed definition of either a hydrophobic cluster or a nucleus. Thus, a protein residue is assigned to a hydrophobic nucleus mostly using intuition.

The contribution of interactions between non-polar groups of side-chains to the total interaction energy of the side-chains in the protein can be estimated by specifying the number of pairs of these groups (atoms) that are in contact. This necessitates calculating and analyzing distances between the non-polar side-chain groups (CH, CH_2, CH_3, S) of all the residues of a protein molecule.

The aim of this study was to work out a clear-cut approach for the identification of the hydrophobic nuclei and clusters of a protein globule and to define certain quantitative criteria for their characterization by analyzing the non-polar contacts between the side-chain atoms. To that end we used "contact maps" and computerized graphics programs for personal computers (Nemethy & Scheraga, 1979; Karpeisky *et al.*, 1990*a*,*b*).

The use of contact maps is a comparatively new strategy in structural studies, and is based on analyzing the distance between non-covalently bonded protein atoms (Levitt & Warshell, 1975; Rossmann & Liljas, 1974). The amino acid sequence

0022-2836/92/070629-10 $03.00/0

of a protein is plotted on the axes of the contact map and the points indicate the number of contacts that this residue has with the environment (Tanaka & Sheraga, 1975; Godzik & Sander, 1989), calculated from the set of atomic co-ordinates of the protein on the basis of a selected cut-off distance. The analysis should include certain definitions related to the term "hydrophobic nucleus", and it should use a computer program aimed at detection and analysis of the non-polar contact regions. In addition, one has to justify the selection of the cutoff distance value for interacting groups (atoms), and to formulate clear-cut criteria for the assignment of an amino acid residue to that region.

2. Method

(a) *Cutoff distance for interacting non-polar groups*

Non-polar group (united atom) i and non-polar group (united atom) j of a protein side-chain (CH, CH_2, CH_3, S) are considered to be in contact if the distance R between centers of their non-hydrogen atoms is $\leq D$ (Chou *et al.*, 1983), i.e.:

$$R \leqslant D = (2-2^{-1/6})(R_i^0 + R_j^0),$$

where R_i^0 and R_j^0 are the van der Waals' radii. The use of the equation as well as united-atom van der Waals' radii values have been justified (Chou *et al.*, 1985). Thus, according to the equation, groups (atoms) are considered to be in contact whenever they are at a distance D corresponding to the sum of their van der Waals' radii.

Consequently, to select, a maximum cutoff distance for interacting non-polar groups of a protein, one should take into account the data in Table 1 as well as the estimated accuracy of the atomic co-ordinates. The accuracy at 2 Å (1 Å = 0·1 nm) resolution was found to be not more than 0·2 Å (Thomas *et al.*, 1982). The maximum difference between R-values is less than the van der Waals' radius for a hydrogen atom. For these reasons we used 5·2 Å as a common value for the maximum cutoff distance.

(b) *Assignment of hydrophobic nuclei*

One can look for a hydrophobic nucleus in a protein by finding and analyzing all the contacts for all non-polar groups of each amino acid residue. The analysis has been carried out using the following definitions:

(1) *The hydrophobic family of a residue.* The "family" is defined by the number of amino acid residues having at least 2 non-polar groups (hydrocarbon groups and/or sulfur atoms) separated from any analogous groups and/or atoms of the residue under consideration by a distance $\leq 5{\cdot}2$ Å (the criterion for a non-polar neighbor). To decrease bias in the selection criterion for "contacting residues", resulting from the limited accuracy of the X-ray co-ordinates of protein atoms, we used a minimum value of 2 rather than 1.

It should be kept in mind that the number of hydrophobic families determined by this method is far greater than the actual number of clusters because, in terms of the above criterion, different families include the same amino acids.

(2) *Library.* This is a list of hydrophobic families for every residue of a protein under consideration arranged according to the protein sequence.

(3) *The non-polar region of a protein globule* is a part of a protein globule comprising amino acid residues selected out of the library on the basis that each of them has at least one non-polar neighbor from the library.

(4) The *hydrophobic cluster* is the part of the non-polar region of the protein in which every residue has 2 or more non-polar neighbors at a cutoff distance of 5·2 Å.

(5) The *hydrophobic nucleus* (γ-nuclei) is the most compact part of the non-polar region of the protein globule, comprising amino acid residues having no fewer than 2 non-polar neighbors at the cutoff distance of 4·5 Å, provided at least 3 of them contact each other to form an "interacting triangle". If a residue has only a single non-polar neighbor at a cutoff distance of 4·5 Å then it is considered as a residue "associated with" the nucleus.

The co-ordinates of side-chain hydrocarbon atoms were used as the co-ordinates of united atoms (CH, CH_2, CH_3). The number of contacts was computed on the basis of atomic van der Waals' radii. The set of atomic co-ordinates of a protein was analyzed using these definitions in order to construct a hydrophobic library. For a non-polar region(s) of a protein to be identified, there must be present a network of non-polar contacts of amino acid residues of maximally "populated" (basic) families with F, Y, I, L, V or M as the central residue. To that end, a scheme of interacting residues was built up on the basis of the library of data on hydrophobic groups. A visual inspection of the scheme allows one to make definite conclusions regarding the composition and borders of the non-polar clusters of a protein molecule.

The average density of the contacts in a nucleus is calculated by:

$$D = \frac{\Sigma(C_i)}{n}, \quad i = 1, 2, \ldots, n,$$

where C_i is the percentage of non-polar contacts between the ith residue and other residues in a particular nucleus:

$$C_i = \frac{\Sigma(K_{i,j})}{\text{NORM (TYPE } (R_i))}, \quad j! = i, \quad j = 1, 2, \ldots, n;$$

where: NORM (TYPE (R_i)) is the average maximum value of the number of surrounding side-chain atoms (for a given amino acid type) that are within 4·5 Å (Heringa & Argos, 1991); n is number of residues in a non-polar nucleus; $K_{i,j}$ is the number of non-polar contacts between 2 residues i and j.

The aforementioned criteria were used to develop an algorithm and to create a program for the determination of hydrophobic families. On the basis of the scheme of non-polar regions and in accordance with definition (5) above, the composition and pattern of interactions among residues of hydrophobic nuclei were defined.

Table 1

Parameters for united-atom Lennard–Jones interaction energy (Lawrence et al., *1978; Nemethy* et al., *1983)*

Atom groups	E_{min} (kcal/mol)	D (Å)
Aliphatic		
CH . . . CH	−0·13	4·75
CH_2 . . . CH_2	−0·14	4·45
CH_3 . . . CH_3	−0·18	4·25
Aromatic		
CH . . . CH	−0·12	4·20
S . . . S	−0·223	4·15

$R_i = D_i/2$.
E_{min} is interaction energy.

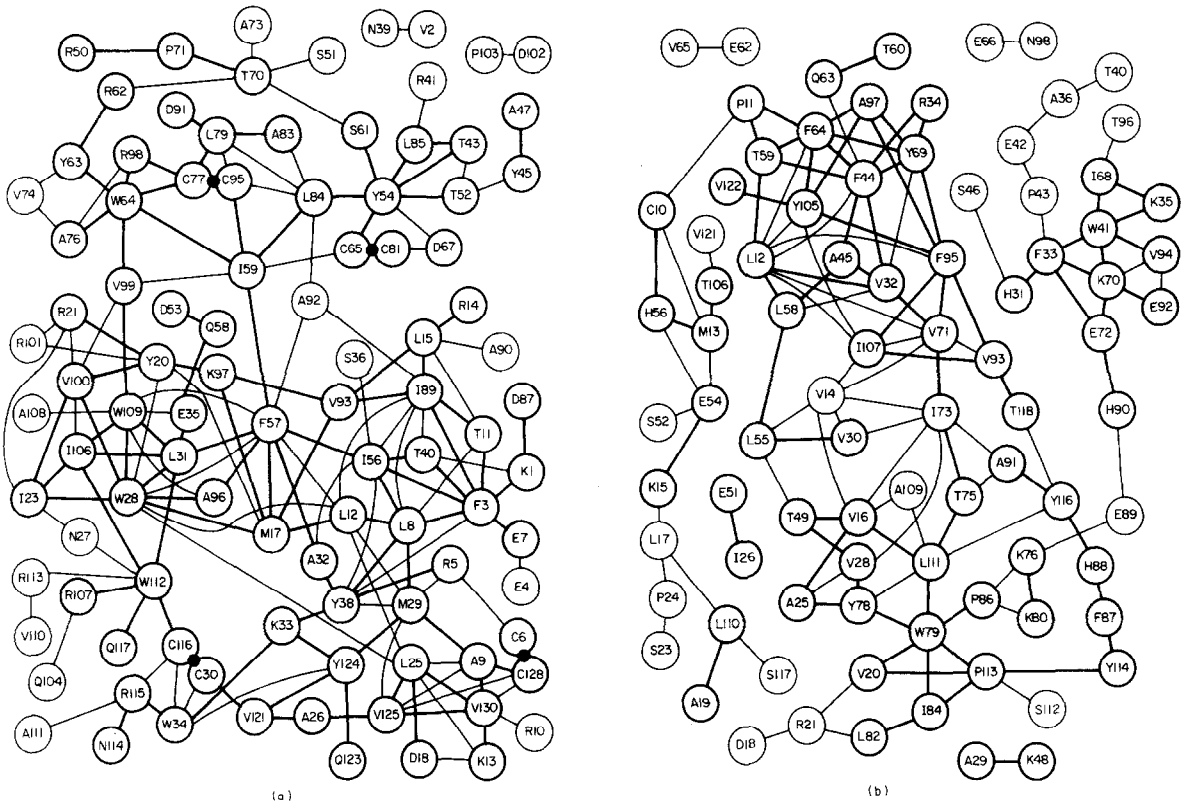

Figure 1. Graf-schemes of non-polar regions in the proteins. Contacts between residues at 4·5 Å are represented by thick lines and those at 5·2 Å by thin lines: (a) lysozyme and (b) prealbumin.

3. Results

(a) *Proteins*

Calculations were performed for four different proteins (Levitt & Chothia, 1976) whose three-dimensional structures have been determined at high resolution in crystalline state by X-ray analysis: all-alpha-myoglobin from sperm whale, all-beta-human prealbumin (transphyretin), alpha/beta-ribonuclease from *Aspergillus oryzae* (RNase T_1), alpha and beta-human lysozyme. The atomic co-ordinates of all proteins were taken from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977; 1MBD, 2PAB, 3RNT, 1LZ1, respectively).

(b) *Non-polar regions in three-dimensional protein structures*

The proteins selected were analyzed using the above-mentioned criteria and definitions. The hydrophobic family libraries were compiled and schemes for non-polar contacts of side-chains were drawn up for each protein at a cutoff distance of 5·2 Å. The schemes of contacts for lysozyme and prealbumin are given as examples in Figure 1(a) and (b). Subdivision of a non-polar region into individual clusters has been carried out by visual inspection. In most cases, clusters can be defined unambiguously. There are only a few locations where divisions are to some extent uncertain. However, this uncertainty could not affect the conclusions.

The analysis of the given schemes in accordance with the definition (5) was carried out to discern the most compact parts of these clusters, hydrophobic nuclei. The results are summarized in Table 2.

It should be noted that these proteins belong to a single-domain protein family and that the presence of several separate non-polar nuclei might be of interest. These nuclei consist of classical hydrophobic amino acids side-chains, but they also include hydrocarbon groups of polar and charged residues.

The quantitative characteristics of hydrophobic nuclei for the proteins under consideration were calculated at a cutoff distance of 4·5 Å and are listed in Table 3.

4. Discussion

(a) *Characteristics of hydrophobic nuclei*

The data listed in Table 3 allow one to characterize the composition and properties of each nucleus. They show that a residue possesses on average three to four non-polar neighbors and more than ten contacts with the residues of the nucleus. These findings from the interaction energies of the contacts (Table 1) allow the assessment of the mean energy of binding of a residue in the hydrophobic

Table 2

Composition of non-polar nuclei in proteins and contact density of residues in the nuclei

N	Res.	*K*	*C* (%)	Ass. res.	N	Res.	*K*	*C* (%)	Ass. res.
A. *Lysozyme*									
1	T43	4	11·1		2b	A26	4	16·7	
	Y54	20	42·6	T52, S61		W28	39	78·0	
	C65	11	39·3			M29	15	40·5	
	C81	16	57·1	D67		C30	24	85·7	
	L84	4	11·1			L31	27	75·0	
	L85	7	19·4			A32	6	25·0	
						K33	16	43·2	
2a	I59	6	16·2			W34	20	40·0	
	Y63	5	10·6	R62		E35	7	18·9	
	W64	28	56·0	A76		Y38	16	34·0	R5
	C77	14	50·0			T40	5	13·9	
	L79	4	11·1	A83, D91		I56	13	35·1	
	C95	13	46·4			F57	19	43·2	
	R98	5	11·4			Q58	4	9·8	D53
	V99	15	45·5			I89	14	37·8	
						V93	12	36·4	
2b	K1	8	21·6	D87		A96	7	29·2	
	F3	29	65·9	E7		K97	17	45·9	
	L8	29	80·6			V100	14	42·4	
	A9	6	25·0	C128		I106	24	64·9	
	T11	4	11·1			W109	25	50·0	
	L12	14	38·9			W112	22	44·0	R107, Q117
	L15	6	16·7	R14		R115	14	31·8	N114
	M17	26	70·3			C116	16	57·1	
	Y20	17	36·2	R21		V121	11	33·3	
	I23	17	45·9			Y124	23	48·9	Q123
	L25	9	25·0	D18		V125	8	24·2	
						V130	7	21·2	K13
	mc3: R50, T70, P71					mc4: Y45, A47			
B. *Myoglobin*									
1a	L2	22	61·1	K133	2	F43	13	29·5	
	E6	5	13·5			F46	23	52·3	R45
	W7	32	64·0	L137		L49	6	16·7	H48, S58
	V10	7	21·2			E52	2	5·4	K34
	K79	10	27·0	L11		M55	13	35·1	
	A130	3	12·5	L9		K56	3	8·1	
	A134	5	20·8			L61	19	52·8	
						H64	5	12·2	
1b	W14	26	52·0			V68	2	6·1	
	V17	18	54·5			I107	6	16·2	
	V21	2	6·1	V66					
	H24	4	9·8		3	H36	9	22·0	S35
	I28	8	21·6			K42	5	13·5	
	L69	9	25·0			H93	3	7·3	
	L72	10	27·8			H97	2	4·9	
	L76	9	25·0			I99	13	35·1	
	I111	8	21·6			P100	15	39·5	
	I112	5	13·5			Y103	21	44·7	
	V114	6	18·2	D27		F106	12	27·3	K102
	L115	14	38·9						
	R118	3	6·8	H119	4a	I75	6	16·2	
	F123	15	34·1	**		K82	3	7·3	H82
	M131	12	32·4			L86	6	16·7	D141
	L135	11	30·6	S108		L89	6	16·7	
						A90	2	8·3	K145
** → V13, H116, A127, Q128						L104	2	5·6	
						F138	6	13·6	
2	L29	6	16·7			I142	13	35·1	
	I30	8	21·6						
	L32	10	27·8		4b	A94	7	29·2	
	F33	17	38·6	L40		I101	7	18·9	A143
	T39	3	8·3			Y146	21	44·7	
						Y151	13	27·7	K98, Q152

Table 2 *(continued)*

N	Res.	*K*	*C* (%)	Ass. res.	N	Res.	*K*	*C* (%)	Ass. res.
mc5: A22, Q26, K62. mc6: R31, A110. mc7: K87, P88.									
C. *Prealbumin*									
1	F33	10	22·7	K31	3	P11	7	18·4	
	K35	17	45·9			L12	12	33·3	
	W41	21	42·0	V94		V32	11	33·3	
	I68	9	24·3			R34	11	25·0	
	K70	9	24·3	E92		F44	27	61·4	
	E72	2	5·4	H90		A45	6	25·0	
						L55	5	13·9	V30
2	V16	8	24·2			L58	11	30·6	
	V20	6	18·2			T59	11	30·6	
	A25	8	33·3			F64	29	65·9	
	V28	7	21·2			Y69	35	74·5	
	T49	4	11·1			V71	6	18·2	I73
	I73	3	8·1			V93	5	15·2	T118
	T75	5	13·9			F95	13	29·5	
	Y78	19	40·4			A97	9	37·5	
	W79	30	60·0	P86		Y105	10	21·3	V122
	I84	5	13·5	L82		I107	6	16·2	
	L111	10	27·8						
	P113	13	34·2						
mc4: F87, H88, A91, Y114, Y116. mc5: C10, M13, K15, E54, H56, T106. mc6: A19, L110. mc7: I26, E51. mc8: A29, K48. mc9: K76, K80. mc10: T60, Q63.									
D *RNase* T_1									
1	Y42	23	48·9	N44	3	Y4	15	31·9	D3
	E46	8	21·6	Y45		C6	13	46·4	D76
	F48	22	50·0	E102		Y11	14	29·8	D15
	F50	25	56·8			V16	10	30·3	L86
	Y56	15	31·9	S54, N81		I61	15	40·5	A19
	E58	2	5·4	R77		V78	4	12·1	
	V79	14	42·4			V89	16	48·5	
	A87	8	33·3			T91	3	8·3	T93
	I90	22	59·5			C103	12	42·9	
	F100	17	38·6	H92					
					mc4: Q20, Y24, H27, E28, K41, P55, Y57, F80, E82, N84.				
2	A22	2	8·3	V67	mc5: C2, T5, C10.				
	L26	6	16·7	K25, E31	mc6: T32, S37.				
	V33	4	12·1		mc7: V52, Q85				
	Y38	22	46·8	N36, H40					
	P39	11	28·9						
	W59	28	56·0						
	P60	35	92·1						
	Y68	27	57·4	L62					
	P73	13	34·2						

N, non-polar nucleus; Res., residues involved in a hydrophobic nucleus; K, number of contacts between particular residue and other residues in the nucleus; *C*, percentage of non-polar contacts between a particular residue and other residues in the nucleus (see the text); Ass. res., residues associated with the nucleus; mc, composition of microclusters.

nucleus (about 2·5 to 3·5 kcal/mol (1 cal = 4·184 J)), which is approximately equivalent to the energy of one hydrogen bond. The assessment of mean energy per atom in these nuclei is about 0·5 kcal/mol, which is also close to the analogous value in regular structure elements (in the alpha-helix this is 0·5 to 0·7 kcal/mol per atom of the main chain).

The temperature factor values were used to characterize the atom mobility in gamma-nuclei (Table 3). According to the average *B*-values, $\langle B \rangle$, the nuclei should be considered as relatively rigid parts of the protein globule. In RNase T_1 it may be seen that the temperature factor values for atoms that belong to the first three gamma-nuclei have the same value as those in the alpha-helix ($\langle B \rangle = 9{\cdot}8$) or in beta-sheets ($\langle B \rangle = 9{\cdot}2$) and markedly lower than that in the other parts of the protein molecule ($\langle B \rangle = 13{\cdot}8$). The last row in Table 3 shows the packing density of the nuclei. To avoid any bias toward large residues (W, F, Y) we used a normalizing factor (Heringa & Argos, 1991).

In addition to gamma-nuclei one can also identify microclusters (mc†) in the non-polar regions of a

† Abbreviation used: mc, microcluster.

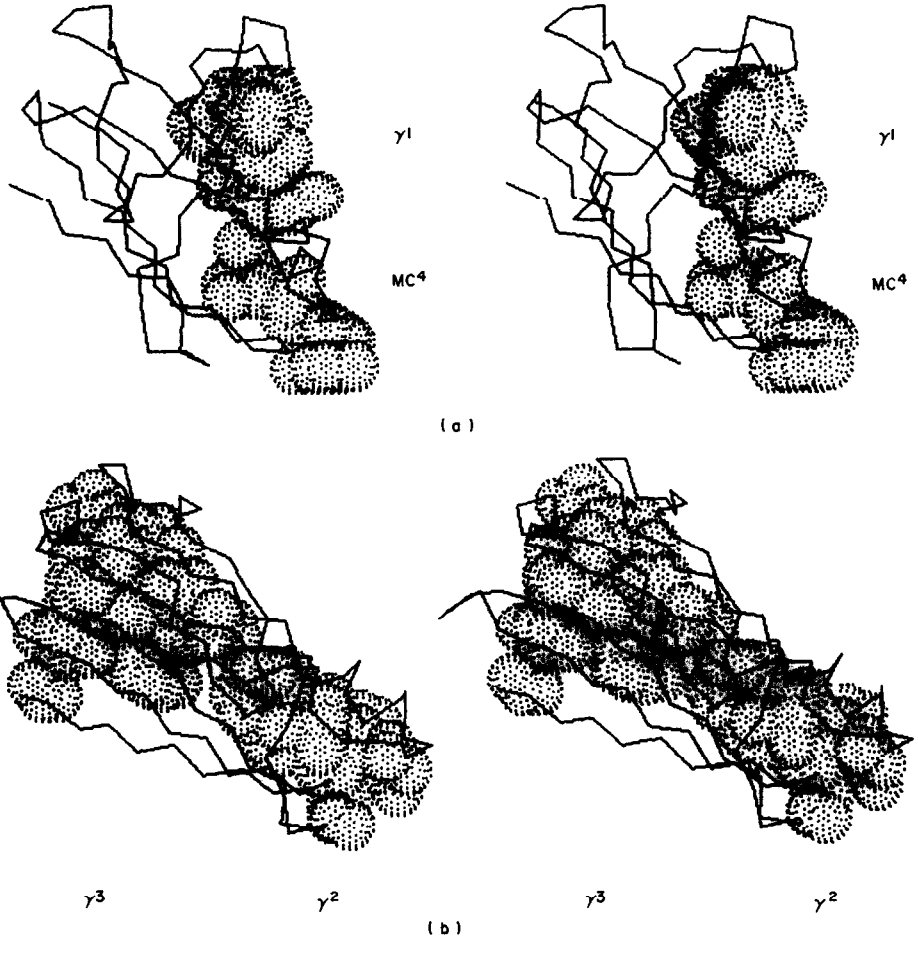
γ1
MC4
γ1
MC4
(a)
γ3
γ2
γ3
γ2
(b)
γ1
γ2
γ1
γ2
(c)
MC4
γ3
MC4
γ3
(d)

Table 3

Comparative characteristics of nuclei at cutoff distance 4·5 Å

Nucleus	*R*	*C*	*N*	*K*	*L*	*S*	⟨*B*⟩	*T*	*T*/*R*	*D* (%)
A. *RNase* T_1							13·8*			
γ1	10	73	45	3·3	15	3·6	11·2	8	0·80	39
γ2	9	59	38	3·1	13	2·9	9·2	5	0·56	39
γ3	9	41	31	2·7	9	2·4	8·3	2	0·22	32
mc4	10	31	36	1·7	6	1·8	12·1	—	—	—
mc5	3	8	6	2·6	5	1·3	12·2	—	—	—
mc6	2	2	3	1·3	2	1·0	12·6	—	—	—
mc7	2	4	5	1·6	4	1·0	11·8	—	—	—
B. *Myoglobin*							14·9*			
γ1a	7	42	24	3·5	12	2·0	13·8	1	0·14	31
γ1b	17	84	70	2·4	10	2·8	12·7	3	0·18	26
γ2	15	70	63	2·2	9	2·5	13·7	3	0·20	23
γ3	8	47	33	2·8	12	2·5	10·3	3	0·38	24
γ4a	8	23	31	1·5	6	2·0	19·0	—	—	15
γ4b	4	20	17	2·4	10	2·0	17·3	1	0·25	30
mc5	3	6	7	1·7	4	1·3	23·6	—	—	—
mc6	2	2	4	1·0	2	1·0	11·8	—	—	—
mc7	2	2	7	0·6	2	1·0	18·9	—	—	—
C. *Prealbumin*							—			
γ1	6	37	30	2·5	12	2·7	—	3	0·50	27
γ2	12	51	44	2·3	9	2·5	—	2	0·17	26
γ3	17	92	68	2·7	11	3·4	—	10	0·59	32
mc4	5	20	23	1·7	8	1·6	—	—	—	—
mc5	6	13	17	1·5	4	1·7	—	—	—	—
mc6	2	2	5	0·8	2	1·0	—	—	—	—
mc7	2	2	6	0·7	2	1·0	—	—	—	—
mc8	2	2	5	0·8	3	1·0	—	—	—	—
mc9	2	3	8	0·7	3	1·0	—	—	—	—
mc10	2	3	4	1·5	3	1·0	—	—	—	—
D. *Lysozyme*							23·7*			
γ1	6	26	20	2·6	9	2·7	14·4	1	0·17	30
γ2a	9	42	37	2·3	9	2·2	22·4	2	0·22	31
γ2b	40	269	161	3·3	13	3·1	17·8	10	0·25	40
mc3	3	6	8	1·5	4	1·3	24·4	—	—	—
mc4	2	2	7	0·6	2	1·0	22·7	—	—	—

R, number of residues in a nucleus; *C*, number of contacts in a nucleus; *N*, number of CH and S atoms in a nucleus; *K*, average number of contacts for an atom of a nucleus; *L*, average number of contacts for a residue in a nucleus; *S*, average number of neighbors for a residue in a nucleus; ⟨*B*⟩, average temperature factor for atoms in a nucleus; *T*, number of triangles in a nucleus; γ, a hydrophobic nuclei; mc, micro cluster.

An asterisk indicates averaged temperature factor for all atoms in a protein.

protein; i.e. chains and/or associations of mutually linked residues, each of which has not more than two non-polar neighbors and fewer than ten contacts with the environment. Thus, the difference between microclusters and nuclei is equivalent to the number of residues involved and their interaction pattern: there are no "triangles" in microclusters and consequently even two residues can form a microcluster. It should be emphasized that the mean energy of linking of a single residue in this microcluster is about 1 to 1·5 kcal/mol, and that the energy of one atom in a microcluster is 0·2 to 0·3 kcal/mol. These values are significantly less than the analogous values in the gamma-nuclei. The microclusters are also characterized by a larger (by 5 to 20%) mobility of atoms in comparison with those in hydrophobic nuclei.

From a study of the hydrophobic nuclei in the molecules of lysozyme and prealbumin (transphyretin) (Fig. 1) one can specify central residues around which each nucleus should form. Thus, for prealbumin these are residues W41 (nucleus gamma 1), F44 and Y105 (gamma 3); Y78 and W79 (gamma 2). The microcluster mc4 is a chain of linked residues and it has no central residue. In lysozyme, the hydrophobic nuclei gamma 1 and gamma 2 are formed around Y54 and W64, respectively. Nucleus gamma 2b occupies a large non-polar region made up of 40 residues. The central part of the hydrophobic nucleus involves residues of alpha-helix frag-

Figure 2. Mutual orientation of gamma-nuclei and secondary structure elements in proteins. Prealbumin: (a) nucleus gamma 1 and mc4; (b) gamma 2 and gamma 3. RNase T_1: (c) nuclei gamma 1 and gamma 2; (d) gamma 3 and mc4. Each non-polar group of a residue included in a nuclei is shown as van der Waals' sphere.

Table 4

Conservation of residues in the non-polar nuclei on the basis of the alignment of the microbial ribonucleases family (Ilyin & Karpeisky, 1992)

G	*Sq.*	% Gm.	RNase	*N*	I+C	Replacements *R*
I	6	64 (60%)	T_1	28	26 (93%)	2 (Y4A; A22K)
			C2	26	24 (93%)	2 (Y4A; A22K)
			Pb1	25	24 (96%)	1 (A22K)
II	4	68 (64%)	Th1	31	29 (94%)	2 (R27Q; I89A)
All	12	51 (48%)	T_1	28	21 (75%)	7
			Pb1	25	19 (76%)	6
			C2	26	19 (73%)	7
			Th1	31	22 (71%)	9

G, group of proteins with identical location of disulfide bridges; *Sq.*, number of proteins studied; %Gm, percentage of amino acid homology among whole sequences; *N*, number of residues involved in the non-polar nuclei; I, C, identical and conserved residues of the non-polar nuclei; *R*, non-equivalent replacements in non-polar nuclei.

ment W28–E35. An inspection of Figure 1 suggests that the nucleus has three parts with central residues W28, I56 and V125, classical hydrophobic amino acids. However, there are no limitations on the nature of the amino acids comprising the microclusters (except, naturally, for glycine).

One may note the influence of disulfide bridges on the composition and the compactness of a nucleus. In lysozyme (Fig. 1(a)) it is apparent that the nuclei that contain the disulfide bridges (gamma 1, gamma 2a and a part of gamma 2b) consist of a smaller number of residues, are less saturated with non-polar contacts between residues, and have on average fewer neighbors per residue than the more compact part of gamma 2b or nuclei without disulfide bridges in other proteins.

It is also easily seen from Figure 1 that the hydrophobic nuclei touch. Thus, in lysozyme, nuclei gamma 2a and gamma 2b are linked due to contacts between two tryptophan residues (W64 and W109) with V99, which is situated between them. The latter should be considered as a bridge residue. If the distance between these nuclei is limited strictly to 4·5 Å, a contact remains between I59 and F57. Another bridge residue is L84, which links nuclei gamma 1 and gamma 2a. Nucleus gamma 2b form a ring that surrounds the second alpha-helix (W28–E35) in the molecule. In prealbumin (Fig. 1(b)) at 4·5 Å there remains between nuclei gamma 2 and gamma 3 only a chain of contacts between residues V71, I73, T75 and L111, where the bridge residues are I73 and T75. The nucleus gamma 1 is completely separate from the other nuclei.

The results obtained suggest multiple functions for large hydrophobic residues such as W, F and Y. Some of them were found to be essential in a large number of families of proteins having non-polar nuclei. However, there are a number of cases when these residues are not integrated into non-polar nuclei. For example, Y45 in lysozyme, F87, Y114, Y116 in prealbumin, as well as Y24, Y55, F80 in ribonuclease T_1 (Fig. 1(a) and (b), Table 2). At present, our data do not allow us to draw definite conclusions on the role that these large hydrophobic residues play in protein folding. In addition, one can notice the importance of other hydrophobic residues such as L8, M17, I56, or (for lysozyme) a group of interacting residues L25, V125 and V130 in the structure of non-polar nuclei.

A detailed analysis of the composition and properties of non-polar nuclei and microclusters of microbial ribonucleases with known sequence has been carried out (Ilyin & Karpeisky, 1992) on the basis of the three-dimensional structures of RNases Pb1 and T_1. The number and position of hydrophobic residues located in the hydrophobic nuclei are conserved throughout the family (Table 4).

(b) *Gamma-nuclei and secondary structure*

The relative positions of the gamma-nuclei and elements of the secondary structure are shown in Figure 2(a) to (d) for prealbumin and ribonuclease T_1. Each hydrophobic nucleus contains side-chains of residues whose main chain forms part of some regular component. Thus, each hydrophobic nucleus stabilizes the mutual relative placement of several elements of secondary structure. In prealbumin, the nuclei gamma 2 and gamma 3 lie between two beta-sheets and give rise to the twisted conformation (in saddle form) of these sheets, it being the most energetically favorable form (Chou *et al.*, 1990). Nucleus gamma 2 also binds a small section of alpha-helix. Nucleus gamma 1 and mc4, which are on the surface of the molecule (Fig. 2(a)), have, apparently, not so great a stabilizing effect on the tertiary structure, but take part in the formation of the quarternary structure of the protein. From RNase T_1 (Fig. 2(d)) it is clear that nuclei gamma 3 and gamma 4 stabilize the binding of alpha-helix and beta-sheet and, in conjunction with nuclei gamma 1 and gamma 2, form a twisted (saddle

form) beta-sheet. It is noteworthy that, in both prealbumin and RNase T_1, the gamma-nuclei are in the marginal, less-stable parts of regular structures and also bind to irregular parts.

Thus, the hydrophobic nuclei stabilize the conformation of the protein globule to a significant extent by fixing the interrelations of elements of secondary structure, by maintenance of the energetically favorable form of the beta-sheets, and by limiting the mobility of functionally significant irregular components of polypeptide chain. As a consequence of this, the hydrophobic nuclei may be considered as separate elements of the structure of a protein, together with such structural elements as alpha-helix and beta-sheet.

5. Conclusions

A detailed analysis has been made of the composition and properties of the hydrophobic nuclei of four proteins that have different spatial organization and whose crystal structure is known to high resolution. It has been shown that in each protein there are a number of separate regions formed by non-polar groups, the atoms of which can be identified, in which hydrophobic nuclei and microclusters can be distinguished. The hydrophobic nuclei are characterized by dense packing of residues, a large number of mutual contacts, and, consequently, by high stability. The microclusters are characterized by a larger distance between residues and a lower number of hydrophobic contacts per residue, i.e. they are more "friable".

The stability of hydrophobic nuclei, the comparatively high binding energy for the residues they contain (of the order of 2 to 3 kcal/mol), the conservation of their composition, and their position in the protein globule suggest that they play an important functional part in protein structure, and possibly form the same kind of structural elements as alpha-helices and beta-sheets. An interesting peculiarity of the hydrophobic nuclei is the fact that the side-chains of residues that occur in irregular parts of the polypeptide chain can contribute to them. Considering the rigidity of the structure of the hydrophobic nuclei, marked limitations on the mobility of the corresponding sections of the polypeptide chain may be expected. The structural role of the hydrophobic microclusters is apparently that they serve as a connecting link between elements of secondary structure and/or of rigid hydrophobic nuclei.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Mayer, E. F., Jr, Bryce, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank, a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **122**, 535–542.

Chothia, C. (1984). Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **53**, 537–572.

Chou, K. C., Nemethy, G. & Scheraga, H. A. (1983). *J. Phys. Chem.* **87**, 2869–2881.

Chou, K. C., Nemethy, G., Rumsey, S., Tuttle, R. W. & Scheraga, H. A. (1985). Interactions between an alpha-helix and beta-sheets. Energetics of p/b packing in proteins. *J. Mol. Biol.* **186**, 591–609.

Chou, K. C., Nemethy, G. & Scheraga, H. A. (1990). Energetics of interactions of regular structural elements in proteins. *Accts Chem. Res.* **23**, 134–141.

Creighton, T. E. & Chothia, C. (1989). Selecting buried residues. *Nature (London)*, **339**, 14 15.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7157.

Godzik, A. & Sander, Ch. (1989). Conservation of residue interactions in family of Ca-binding proteins. *Prot. Eng.* **2**, 589–596.

Heringa, J. & Argos, P. (1991). Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **220**, 151–171.

Ilyin, V. A. & Karpeisky, M. Ya. (1992). Conservativeness of residues in nonpolar nuclei of microbial ribonucleases. *Biofisika (SSSR)*, in the press.

Karpeisky, M. Ya., Ilyin, V. A. & Sevcik, J. (1990*a*). Non-covalent interactions in ribonucleases Pb1, Sa and Bi. In *Metabolism and Enzymology of Nucleic Acids Including Gene Manipulations* (Zelinka, J. & Balan, J., eds), pp. 265–275, Plenum Press, Bratislava.

Karpeisky, M. Ya., Kolbanovskaya, E. Yu. & Borisov, V. V. (1990*b*). Intramolecular interactions in pancreatic ribonucleases. *DAN SSSR*, **315**, 495–500.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.* **14**, 1–64.

Lawrence, G., Dunfield, ., Antony, W., Burgess, . & Scheraga, H. (1978). United atoms potentials for polypeptide and proteins. *J. Phys. Chem.* **82**, 2609–2616.

Lesser, G. J. & Rose, G. D. (1990). Hydrophobicity of amino acid subgroups in protein. *Proteins, Struct. Funct. Genet.* **8**, 6–13.

Levitt, M. & Chothia, C. (1976). Classification of protein structure. *Nature (London)*, **261**, 552–567.

Levitt, V. & Warshell, A. (1975). Computer simulation of protein folding. *Nature (London)*, **253**, 694–698.

Nemethy, G. & Scheraga, H. A. (1979). A possible folding pathway of bovine pancreatic RNase. *Proc. Nat. Acad. Sci., U.S.A.* **76**, 6050–6054.

Nemethy, G., Pottle, M. S. & Scheraga, H. A. (1983). Energy parameters in polypeptide. 9. Updating of geometrical parameters nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.* **87**, 1883–1887.

Plochocka, D., Zielenkiewicz, P. & Rabczenko, A. (1988). Hydrophobic microdomains as structural invariant regions in proteins. *Protein Eng.* **2**, 115–119.

Privalov, P. L. & Gill, S. J. (1988). Stability of protein structure and hydrophobic interactions. *Advan. Protein Chem.* **39**, 193–234.

Richards, F. M. (1977). Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Rossmann, M. G. & Liljas, A. (1974). Recognition of structural domains in globular proteins. *J. Mol. Biol.* **85**, 177–181.

Sali, A. & Blundell, T. L. (1990). Definition of general

topological equivalence in protein structures. *J. Mol. Biol.* **212**, 403–428.

Singh, J. & Thornton, J. M. (1990). SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups. *J. Mol. Biol.* **211**, 595–615.

Spolar, R. S., Jeung-Hol, H. & Records, J. T. (1989). Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 8382–8385.

Tanaka, S. & Sheraga, H. A. (1975). Model of protein folding, inclusion of short-, medium-, and long-range interactions. *Proc. Nat. Acad. Sci., U.S.A.* **72**, 3802–3806.

Thomas, K. A., Smith, G. M., Thomas, T. M. & Feldman, R. J. (1982). Electronic distributions within protein phenylalanine aromatic rings are reflected by the three dimensional oxygen atom environments. *Proc. Nat. Acad. Sci., U.S.A.* **79**, 4843–4847.

Wood, R. H. & Thompson, P. T. (1990). Differences between pair and bulk hydrophobic interactions. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 946–949.

Edited by A. R. Fersht