J. Mol. Biol. (1998) **278**, 457–479

JMB

Structural Determinants in the Sequences of Immunoglobulin Variable Domain

Cyrus Chothia[1], Israel Gelfand[2] and Alexander Kister[2]

[1]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England*

[2]*Department of Mathematics Rutgers University New Brunswick, NJ 08903 USA*

To determine the general relation between the sequence and structure of variable domains of immunoglobulins we have carried out an analysis of their atomic structures, some 5300 different expressed sequences and the human germline gene segments. Variable domains are formed by two β-sheets, packed face to face, and the inter-strand turns. Comparison of the different known structures shows that they have a core of 76 residues which has the same main-chain conformation in all structures. This common core contains almost all of the β-sheet structure and three inter-strand turns. The regions that differ in conformation are the three hypervariable regions, three other inter-strand turns and a few adjacent residues.

The 5300 expressed sequences currently known for variable domains were examined to determine the residues that occur at the 76 sites. Ignoring site conservations that occur for functional reasons, there are eight sites that have the same residue in almost all sequences; 12 that have one of a small group of very similar residues, and 52 where the chemical character of the residues is strongly conserved but not their volume.

The role of residues at each site in the core was determined from the examination of their accessible surface areas, contacts, packing and buried side-chain hydrogen bonds. The most strongly conserved sites form the ''deep'' structure of the domain at the centre of the interface between the β-sheets. It includes eight invariant sites and 11 sites that have one of a set of very similar residues. Around the deep structure there are buried hydrophobic residues that, in different variable domains, can differ greatly in volume. These differences in volume are accommodated by conformational changes in turn regions that are outside the common core. On the surface nearly all residues not involved in function or turn conformations strongly conserve hydrophilic or neutral residues.

The implications of these results for the general relations between the sequence and structure of proteins are discussed.

Keywords: conserved core; residue classification; deep structure

Introduction

Answers to the question: how does the amino acid sequence of a protein determine its structure? can cover different aspects and different levels of understanding. In recent years, considerable progress has been made in understanding the role of residues in the folding pathway(s) and their contributions to the protein stability and to secondary structure formation. Here we are concerned with an aspect where less progress has been made: the general relation between the sequence of a protein and its three-dimensional fold.

The major requirement for the creation of a stable three-dimensional structure for a protein is a hydrophobic interior, that is sufficiently large, close packed and free of strain, and a surface that is sufficiently hydrophilic. All residues make some contribution to meeting these requirements but the importance of their contribution can be very different. At some sites, substitution of the native residue by a wide range of other residues produces only very small changes in structure and stability. At other sites substitutions of any but the most conservative kind make the structure unstable. The

Abbreviations used: V, variable; C, constant; L, light; H, heavy; IR, invariant residue; SR, similar residue; RC, residue class.

0022–2836/98/170457–23 $25.00/0/mb981653

ability or inability of a site to accept substitutions of a native residue reflects its role in the structure and the constraints placed on the site by this role. This means that if a wide range of divergent sequences are known for a particular fold we can determine, to a first approximation, the nature of the residues at that site that are consistent with the native structure. If atomic structures are known for the protein, we can also determine, at least in outline, the actual role of the residue at each site in the formation of the hydrophobic interior and hydrophilic surface. Here we try to show how putting these two sets of data together gives a description of the major sequence determinants of a protein fold.

The use of homologous sequences to determine the different ranges of residues that can occur at the different sites in a protein is not new. Taylor (1986) aligned the sequences of 100 variable domains and 85 constant domains and described the sequence patterns associated with the two folds. He showed that these patterns are different to those found in certain other proteins built of β-sheets. (Smith & Xue (1997) recently brought this work up-to-date and extended it to other members of the immunoglobulin superfamily.) Bashford *et al.* (1987) carried out an analysis of 221 globin sequences and showed that the sequence pattern common to members of the globin family is essentially unique: it was not found to a significant extent in any non-globin sequence. Residue substitutions consistent with a native fold were also found by carrying out systematic mutagenesis experiments in which a large range of residues were substituted at successive single sites in the arc repressor (Bowie & Sauer, 1989), the N-terminal domain of the λ repressor (Lim & Sauer, 1989) and T4 lysozyme (Rennell *et al.*, 1991). However, none of these pieces of work, nor to our knowledge of later work, investigated the general structural basis of the residue substitution patterns that they found.

To describe the structural determinants in the sequences for a protein fold, we extend here the previous work on the sequences and structures of immunoglobulin variable domains by Gelfand & Kister (1995) and by Chothia *et al.* (1988). The variable (V) domains are an excellent subject for an investigation of this kind because of the considerable amount of data that is now available. The database of Kabat *et al.* (1991) now contains the sequences of some 5300 different variable domains. The atomic structures of some 140 variable domains have been determined and are available from the Protein Data Bank (Bernstein *et al.*, 1977). Another reason to focus on these molecules is that different variable domains can have very divergent sequences. The identities in the sequence of V_L and V_H domains are ~35% on average; but in extreme cases they can drop to ~20%. This divergence, together with the large number of different known sequences, should give us a good picture of the sequence variations that are allowed within the constraints of (i) the functional and structural requirements of variable domains and (ii) the sequence changes that are accessible by normal evolutionary processes (Lesk & Chothia, 1980).

In the first part of this paper we define the common core of variable domains: the region whose conformation is conserved in all, or almost all, variable domains. In the second part, we use a statistical analysis of the aligned sequences of 5300 V domains to determine the range of residues that occur at each site in the common core. In the third part, the general structural role of each residue is found from an examination of the atomic structures of immunoglobulins. Putting this information together we give an account of the principle relations between sequence and structure in variable domains.

Immunoglobulins and their Variable Domains

The immunoglobulin molecule is built from two heavy chains and two light chains. The light chain folds into two domains: the variable, V_L, and the constant, C_L, and the heavy chains folds into four domains: one variable, V_H, and three constant, C_{H1}, C_{H2} and C_{H3}. The chains are held together by disulphide bridges and by the packing together of domains: V_L against V_H; C_L against C_{H1}, and C_{H3} of one heavy chain against C_{H3} of the other heavy chain (Figure 1(a)). The domains are homologous. Different variable domains usually have sequence identities of about 35% or more and two-thirds or more of their structures have the same conformation. The same is found for different constant domains. However, if variable domains are compared with constant domains, the differences between are greater: they have residue identities that are typically 10 to 20% and the regions with the same fold usually comprise only one third to one half of each domain (Lesk & Chothia, 1982).

The variable domains of heavy (V_H) and light chains (V_L) are built from two β-sheets: one with four strands, A, B, E and D, and the other with six strands, A′, G, F, C, C′ and C″ (Figure 1(b)). In the immunoglobulin molecule, V_L and V_H associate to form a dimer with contacts that involve the GFCC′ β-sheet strands and adjacent residues of one V domain packing against the same region in the other domain. The binding site for antigens is formed by the inter-strand links BC, C′C″ and FG from each domain. Variations in sequence and conformation of these six regions are the major determinants of specificity and affinity for antigens.

On the basis of similarities in sequence, variable domains are placed in one of three classes: two classes of light chain domains, V_κ and V_λ, and one class of heavy chain domains, V_H. Inspection of the atomic structures known for these domains shows, of course, that the inter-strand links that form the antigen binding site, BC, C′C″ and FG, differ in conformation both within and between different classes. It also shows that three other inter-strand

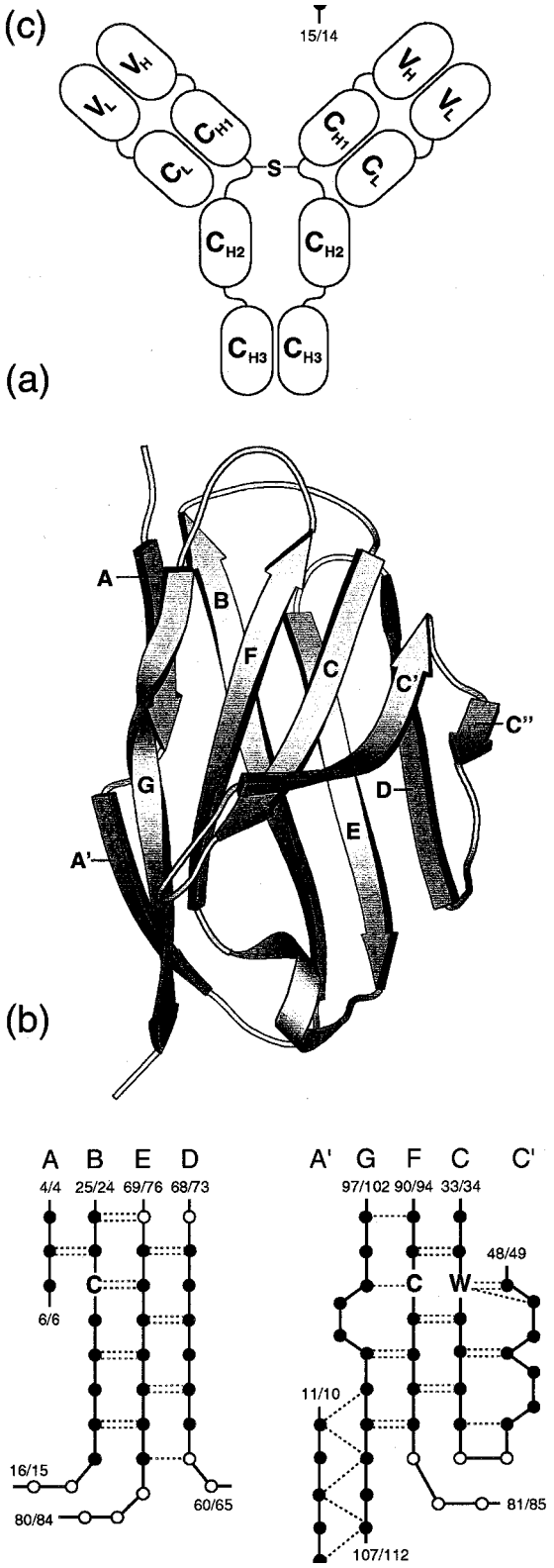

Figure 1. (a) The domain structure of the immunoglobulin molecule (see the text). Each ellipsoid represents a domain of about 100 residues. (b) The structure of a typical variable domain. β-Sheet strands are shown as ribbons and labelled A, A′, B, C, C′, C″, D, E, F and G. Note how a β-bulge produces a rotation in the G strand. The EF turn contains a short helix. (c) A plan of the common core of variable domains. Filled circles represent β-sheet residues and open circles turn residues. This structure is believed to be common to all, or almost all, immunoglobulin variable domains. Different families or classes of variable domains have different conformations for the regions outside this core.

regions, AA′, C″D and DE, have different lengths and conformations in the different classes.

Here we are concerned with the sequence determinants of not the whole variable domain but the large part that has the same fold in all, or almost all, variable domains. This means that we do not discuss the hypervariable regions, or the few other regions, whose conformations are not conserved. We also do not discuss the V_L-V_H interface. Some of the relations between the sequence and structure of these parts have been discussed in our previous papers (Chothia *et al.*, 1985, 1989; Gelfand & Kister, 1995).

The Common Core of Variable Domains

A conservative framework can be characterised as a set of atoms that occupy the same relative positions in space. To identify these positions different analytical or visual superposition methods are usually applied (Lesk & Chothia, 1982; Chothia & Lesk, 1987; Taylor & Orengo, 1989; Diamond, 1992; Shapiro *et al.*, 1992; Yee & Dill, 1993; Gerstein & Altman, 1995). All of them use multiple structural alignments to characterise structural similarity. Recently, two different approaches for comparison of members of the same structural family were proposed (Gelfand *et al.*, 1996, 1998). The advantage of these methods is that they allow one to compare protein structures and to find a common geometrical core without requiring the superposition procedure.

In the first of these approaches an invariant system of co-ordinates was derived for variable domains that is based on the inherent geometrical properties common to all members of the family. One of these common properties is the pseudo 2-fold symmetry which is observed to relate the V_L and V_H domains in all known immunoglobulin structures. Thus, the Y axis of this invariant co-ordinate system was chosen to coincide with the 2-fold axis. The line connecting two centres of mass for V_L and V_H domains becomes the X axis of the invariant system of co-ordinates. The origin of the co-ordinates system is the midpoint of the two centres of mass. One of the advantages of this system of co-ordinates is it provides information about possible structural or functional roles of residues. For example, residues which have a small value for their X-co-ordinate are mostly involved in domain-domain interactions, whilst residues with high Y values are usually located in antigen-binding regions.

Comparison of the co-ordinates of the C^{α} atoms in different V_L and V_H domains shows which resi-

dues have approximately the same position. At about 76 of the C^{α} positions the dispersion of the positions is very small: usually less than 1 Å. (These comprise about 70% of each variable domain.) We used this dispersion as a criterion for the selection of the geometrically conserved core of the variable domains. This subset of positions is an invariant feature of the fold of variable domains and the averaged co-ordinates of the C^{α} atoms of the residues that this selection procedure puts in the common core of the V_L and V_H domains are given by Gelfand *et al.* (1996).

The second core-finding method is independent of any particular system of co-ordinates (Gelfand *et al.*, 1998). This algorithm is based on the distances between C^{α} atoms, in the conserved core of the members of a protein family, being invariant. Thus, core is defined as the set of residues for which the distances between the $C^{\alpha}(i)$ of the residue at the ith position and the $C^{\alpha}(j)$ of the residue at the jth position are the same or very similar in all members of a protein family. (Holm & Sander (1993) also use the comparison of residue-residue distances to align pairs of protein structures and so measure the extent of their similarities.) We compared C^{α}–C^{α} distances between the residues at the equivalent positions in 100 immunoglobulin structures to define the geometrical core common to V_L and V_H domains.

The two methods for determining the residues that form the common core of variable domains give results that have no significant differences. The common core consists of 76 residues and includes all, or almost all, of strands A, A′, B, C, C′, D, E, F and G and inter-strand links, A′B, CC′ and EF. These regions are illustrated in Figure 1(c) and the sites are listed in Table 1. The regions that differ in conformation or position are the C″ strand, the other six inter-strand regions and a small number of adjacent residues.

The number of different structures that we have compared is small in comparison with the number of different sequences that are known. The sequences of these structures, however, do cover a large part of the observed range of residues found in the known variable domain sequences. We would expect, therefore, that all, or almost all sequences used here give structures with a core that has a conformation very close to that considered here.

Residue Frequency at Sites in the Common Core

The frequency of residues at common core sites of variable domains was calculated from sequences in the Kabat database (Kabat *et al.*, 1991). This contains a complete collection of the known immunoglobulin sequences. It includes the sequences of heavy and light chain variable domains from human, mouse, rat, rabbit, shark and other species; some 5300 in all. These sequences were aligned. The alignment was based on both sequence and structural information. In the common core this gives results which are the same as that of Kabat *et al.* (1991) except in the region of the A′ strand where the alignment is based on structural data. From this alignment of ~5300 sequences, we determined the frequencies with which different residues are found at each of the 76 core sites. These frequencies are listed in Table 1 where we give the Kabat numbering for each site and its position in the secondary structure: A1, A2, ... etc., for residues in strands and A′B1, A′B2, ... etc., for residues in turns.

The sequences in the Kabat database might be thought to be a biased selection of those produced by the immune system because of the various special reasons for which they were determined. We, therefore, checked our results using a second database that contained the sequences of the complete human repertoire of V_H, V_κ, V_λ, J_H, J_κ and J_λ germline gene segments (Tomlinson *et al.*, 1992, 1995; Matsuda *et al.*, 1993; Cook *et al.*, 1994; Schäble & Zachau, 1993; Williams *et al.*, 1996; Ignatovich *et al.*, 1997; Ravetch *et al.*, 1981; Hieter *et al.*, 1982; Udey & Blomberg, 1987). The second database is smaller and lacks the variations that arise from somatic mutations and species differences but, in being the complete germline repertoire of one species, will not have the bias that might be present in the sequences in the Kabat database. In spite of these differences the picture of residue conservation and variations at the different sites given by the two databases is remarkably similar. The extent of the residue variations at each site is somewhat greater in the Kabat database than that amongst the human gene segments but the type of conservation that occurs at each site is the same or very similar (see below and Table 1).

The Classification of Residue Conservation at Core Sites

Residue frequencies at the different sites that form the common core of variable domains is given in Table 1. Inspection of this Table shows that the nature and extent of residue conservation at different sites varies greatly. It can, however, be described as one of three kinds. At the first kind of site, only one residue is found in almost all sequences. At the second, one of a few closely related residues is found. At the third, a wider range of residues is found and the conservation is best described in terms of the chemical character of the residues that occur at the site.

Our analysis of residues at different sites (Table 1) shows that there are nine positions which are occupied by a single particular residue in almost all sequences. For example, W at the position 35/35 and C at the position 88/92 is found in more than 99% of the sequences. We refer to such sites as invariant residue (IR) sites (Table 2). (Note that residue sites are numbered here m/n where m

is the Kabat number of the V_L site and n is the Kabat number of the structurally equivalent V_H site.)

Further inspection of the residue frequencies in Table 1 shows that 17 sites have, in almost all sequences, one of a small number of very similar residues. An example of such sites is 21/20 where V, L, I or M are found in 99% of sequences: the side-chains of these four residues differ in shape but all are hydrophobic and occupy similar volumes. We refer to this class of sites as similar residue (SR) sites. Here similar residues are defined as those that have the same chemical character and whose volumes differ by no more than the equivalent of one methylene group. We make one exception to this rule: site 16/15. The main residues found here are G (87%) and S (10%), which have differences in volume a little greater than a methylene group. But this site is in a turn and both residues are characteristic of such regions.

Two thirds of the sites are occupied by residues that have a wider range in chemical character and/or volume than that found at SR sites. In nearly all cases, these sites do not contain all types of residues but are limited to certain classes of residues. We will refer to these sites as residue class (RC) sites.

Different types of RC sites can be defined by the particular class of residues that they conserve. To classify residues, we use a scheme based on two correlated properties: (i) their tendency to be on the surface or in the interior of a protein; and (ii) their hydrophobic character, as measured by the energy involved in their transfer between water and organic solvent. The extent to which different types of residues are found on the surface or in the interior of proteins of known structures has been determined (Miller *et al.*, 1987, and references therein). This work shows that residues can be put into one of three groups (Figure 2). The hydrophobic group contains residues that have a high probability of being buried in the protein interior and we will refer to them as b (buried) group residues. The s (surface) group contains hydrophilic residues that have a high probability of being on the surface and the n (neutral) group contains residues that have a roughly equal chance of being on the surface or in the interior. The residues in each group are:

s :	R,	K,	E,	D,	Q	and	N :	
n :	P,	H,	Y,	G,	A,	S	and	T;
b :	C,	V,	L,	I,	M,	F	and	W.

The mean hydrophobicity of residues in each of the three groups, as measured by the energy involved in their transfer between water and organic solvent, correlates well with the extent to which they are buried (Figure 2). RC sites in the variable domains, i.e. those sites whose residues are not confined to a single type of residue or to a few closely related types, are classified in terms of s, n and b. For example, a site labelled b would contain several, but in most cases not all, of the residues from the C, V, L, I, M, F, W group. A site that has residues from both the s and n groups, or from both the b and n groups, is labelled sn, or bn.

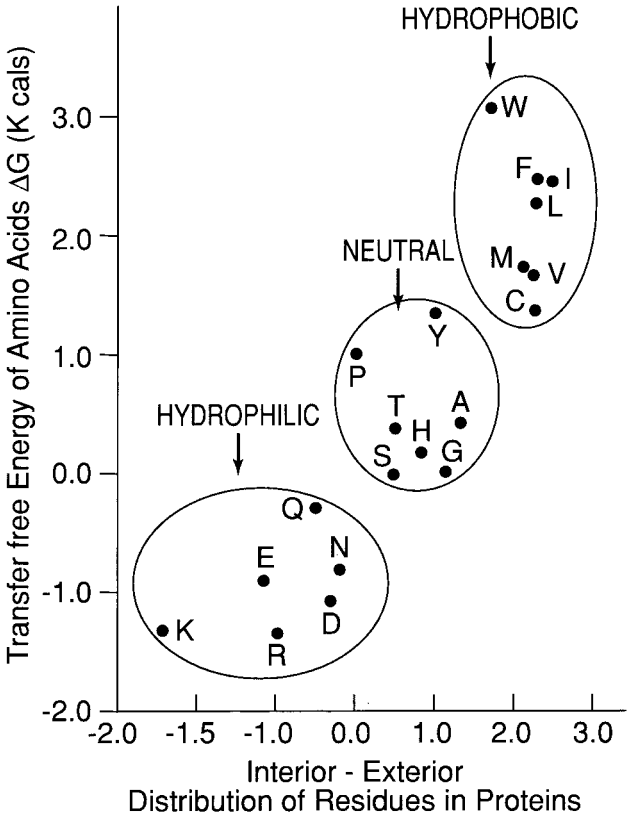

Figure 2. The classification of residues on the basis of their hydrophobicity and the extent to which they are distributed between the surface and interior of proteins. Free energies of the transfer of residues from water to organic solvent are taken from the work of Fauchere & Pliska (1983). The distribution of residues between the interior and the surface of proteins is given in terms of a partition coefficient, ln f, where f is:

$$\frac{N_S/\Sigma N_S}{N_B/\Sigma N_B}$$

N_S is the number of surface residues of a given type out of a total of ΣN_S surface residues of all types and N_B is number of surface residues of a given type out of a total of ΣN_B surface residues of all types (see Miller *et al.*, 1987).

The Extent of Conservation at the Sites in the Common Core

To put in context the extent of residue conservation at sites in variable domains, it is worth noting what would be found if, in the different V domain sequences, residues were to be distributed equally at each site. The amino acid composition of variable domains is described by Kabat *et al.* (1991). In the average variable domain, the frequency of individual residues varies: Met is the least common (2% of all residues) and Ser is the most common

Table 1. Residue frequencies at the sites that form the conserved common core of variable domains

	Seq	Hydrophilic residues (s)						Neutral residues (n)							Hydrophobic residues (b)							Total residues				Conservation
Site	type	R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W	s	n	b	all	type (%)
4/4	E				1	4	1	9	1		7		1	2		186	3642	47	1231	3		6	20	5109	5135	LM 95
A1	GL															3	84		24					111	111	LM 100
5/5	E	12	178	182	54	1032	12	7	2	1	1	20	59	2513		884	141	26	11			1470	2603	1062	5135	sn 79
A2	GL		3			12								61		33	1		1			15	61	35	111	bn 86
6/6	E	8	3	1308	10	3767	1	2	3		1	17	2	1		5	5				2	5097	26	12	5135	QE 99
A3	GL			34		77																111			111	QE 100
11/10	E	43	25	1108	73	19	60	8	4		1258	88	50	44		505	1532	12	515	9		1328	1452	2573	5353	bn 75
A'1	GL			13		1			1		32	8	2	1		20	29		2	2		14	44	53	111	bn 87
12/11	E	3		28	1	12	2	369	1	44	11	369	1718	176	3	369	2203	17	9	18		46	2688	2619	5353	bn 99
A'2	GL							7				1	52	2		17	32						62	49	111	bn 99
13/12	E	14	317	33	2	37	7				191	1168	7	164	3	3058	222	42	61	26	1	410	1530	3413	5353	bn 92
A'3	GL		14	2							8	24	0			55	7	1				16	32	63	111	bn 86
14/13	E	286	1465	17	1	688	100	41	8	2	9	238	2071	383	3	11	8	3	8	11		2557	2752	44	5353	sn 99
A'4	GL	1	30			18					1	5	43	12			1					49	61	1	111	sn 99
15/14	E	1	2	1				3580	2		8	274	85	29		458	885	49	3	14		4	3978	1409	5391	bn 99
A'5	GL							82					1	3		13	12						86	25	111	bn 100
16/15	E	27	22	29	3	1	1	6			4684	3	545	56	1	3	1		1	2	6	83	5294	14	5391	GS 97
A'B1	GL	2	1								93		12	3								3	108		111	GS 95
17/16	E	127	66	1388	776	884	5	6	5	2	1036	735	219	116	1	13	7	3	1	1		3246	2119	26	5391	sn 99
A'B2	GL	5	2	21	19	27					20	13	3	1								74	37		111	sn 100
18/17	E	970	617	4		216		104			10	37	2562	784		11	8	13	1	3		1807	3497	36	5340	sn 99
B1	GL	26	4					7					45	28					1			30	80	1	111	sn 99
19/18	E	36	1	2	5	4		2	17		3	737	2	4	7	2778	1573	98	67	4		48	765	4527	5340	bn 99
B2	GL											24				46	38	3					24	87	111	bn 100
20/19	E	599	1652	52		20	11	2		1	14	20	1092	1812	1	4	3	42	5	10		2334	2941	65	5340	sn 99
B3	GL	29	16				1						21	44								45	65		111	sn 100
21/20	E	3		1				2	1	1	2	8	4	7	1	206	2266	2031	773	34		4	25	5311	5340	VLIM 99
B4	GL															10	53	48						111	111	VLI 100
22/21	E	10	23	1	1	3	79	7		3	3	12	3390	1717	40	36	2	4		9		117	5132	91	5340	ST 96
B5	GL						1						60	50								1	110		111	ST 99
23/22	E	8	1							7	7	1	4		5310						2	9	19	5312	5340	C 99
B6	GL														111									111	111	C 100
24/23	E	1117	1689	46	4	142	7	4	9	6	27	863	726	615	1	61	6	13	2	1	1	3005	2250	85	5340	sn 98
B7	GL	28	15			2					5	27	9	24					1			45	65	1	111	sn 99
25/24	E	7		3	4			6	1	1	468	3223	659	306	1	517	22	87	1	34		14	4664	662	5340	bn 99
B8	GL	1									21	53	10			14	8	1	1	2		1	84	26	111	bn 99
33/34	E	4	4		1		10	7	3	115	2	118	5	7	2	564	1390	740	2052	21	310	19	257	5079	5355	bn 99
C1	GL							3				7				18	32	11	28		12		10	101	111	bn 100
		R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W					

Site	Seq type	Hydrophilic residues (s)						Neutral residues (n)							Hydrophobic residues (b)							Total residues				Conservation type (%)
		R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W	s	n	b	all	
34/35	E	8	32	268	67	62	1181	3	1357	216	357	671	887	101	61	9	13	36	5	10	11	1618	3592	145	5355	sn 97
C2	GL	1	1		5	2	13		19	9	7	20	32		1					1		22	87	2	111	sn 98
35/36	E	11	1			2				12	1	2	1		6	4	9		1	1	5304	14	16	5325	5355	W 99
C3	GL																				111			111	111	W 100
36/37	E	2	1	1	2	14	6	3	34	1938	6	22	11	4	6	2515	83	404	20	281	2	26	2018	3311	5355	bn 99
C4	GL								2	54						32	2	16		5			56	55	111	bn 100
37/38	E	1837	1015	7	3	2055	4	5	10	1	9	4	1	5	6	4	368	1	9		11	4921	35	399	5355	s 92
C5	GL	48				59											4					107		4	111	s 96
38/39	E	25	109	109	1	4996		4	64	3	1	2	3	3		5	17	10			3	5240	80	35	5355	Q 93
C6	GL					111																111			111	Q 100
39/49	E	932	1890	18	2	22	15	480	52	8	21	774	383	199	2	86	76	9	119	102		2879	1917	394	5190	sn 92
C7	GL	5	42			2		14	7			32	1	1			5		2			49	55	7	111	sn 94
40/41	E	14	3	2	2	51	11	4414	158		19	42	391	52		4	23	1	2	1		83	5076	31	5190	n 98
CC'1	GL	1				1		104				1	2	2								2	109		111	P 4
41/42	E	36	9	364	278	13	13	11	30	1	4244	32	98	9	2	7	3	6			34	713	4425	52	5190	sn 99
CC'2	GL			2	2						105	1	1									4	107		111	sn 100
42/43	E	326	1998	63	3	1729	127	13	171	1	167	35	159	358		1	7	3	2	22	5	4246	904	40	5190	sn 99
C'1	GL	2	62	1		35	1	1					3	5			1					101	9	1	111	sn 99
43/44	E	299	91	12	5	1	7	301			2263	620	1263	192	1	21	98	11		1	4	415	4639	136	5190	sn 97
C'2	GL	2						7			45	45	9			2				1		2	106	3	111	n 95
44/45	E	1	27	3		7	13	1956			10	4	2	3		125	2893	29		112	5	51	1975	3164	5190	bn 99
C'3	GL							61				1					49						62	49	111	bn 100
45/46	E	382	1600	2791	17	193	26	1	1	2	13	8	8	115		136	25	6	4	1	5	5009	148	177	5334	s 94
C'4	GL	18	28	50		5										8		2				101		10	111	s 91
46/47	E	303	1	5	1	1	3	74	5	83	114	21	19	203	4	46	1646	22	2	44	2732	314	524	4496	5334	bn 94
C'5	GL	3							1	7		1	2		3		44			2	48	3	11	97	111	bn 97
47/48	E	4		10	4	3		2		2	6	1	12	1		790	2131	1532	447	4	385	21	24	5289	5334	VLIM 92
C'6	GL															31	53	14	12	1				111	111	VLIM 100
48/49	E	5		5	3	1		1		18	2112	606	146	15	1	104	17	2264	23	5	8	14	2898	2422	5334	bn 99
C'7	GL										29	7	14				3	54	4				50	61	111	bn 100
60/65	E	28	39	54	1163	5	89	7	5	6	1791	649	1399	54	1	76	7	6		6		1378	3911	96	5385	sn 98
C'D5	GL			9	26		2	1			32	6	34	1								37	74		111	sn 100
61/66	E	3993	1091	5	1	116	3	10	19		23	4	10	18		4	7	7		13	9	5209	84	40	5333	RK 95
D1	GL	109				1			1													110	1		111	R 99
62/67	E	16	2		4	1				1	16	911	62	187	3	432	234	188	2	3274		23	1177	4133	5333	bn 99
D2	GL															23	4	1		83				111	111	b 100
63/68	E	22	50		3	2	9	1	2	4	21	77	2248	2617	1	19	3	237	2	15		86	4970	277	5333	ST 91
D3	GL												61	48		1			1				109	2	111	ST 98
		R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W					

continued overleaf

	Seq	Hydrophilic residues (s)						Neutral residues (n)							Hydrophobic residues (b)							Total residues				Conservation
Site	type	R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W	s	n	b	all	type (%)
64/69	E	4	5	2	14			5	1	1	2252	24	46	21	1	165	987	1583	104	118		25	2350	2958	5333	bn 99
D4	GL										60	1				1		41	7	1			61	50	111	bn 100
65/70	E	13	5	1	10	5	91	5		8	25	12	3698	1395	2	3	20	17		23		125	5143	65	5333	ST 95
D5	GL						1						98	11			1					1	109	1	111	ST 98
66/71	E	1179	439	7	8	2	30	79		4	1827	459	138	65		859	210	22	3	1	1	1665	2572	1096	5333	sn 79
D6	GL	28	15	1			4	1			33	4	7	2		11	3	2				48	47	16	111	sn 86
67/72	E	9	16	94	2739	1	75	2	16	19	16	12	2156	37	2	13	17	100	1	5	3	2934	2258	141	5333	sn 97
D7	GL				53		1			1		1	52				3					54	54	3	111	sn 97
68/73	E	79	926	52	163	8	804	7	1	4	2219	42	53	961		9	5	57	6	3	1	2032	3287	81	5400	sn 99
DE1	GL	3	3	1	3		17				59	3	1	20					1			27	83	1	111	sn 99
69/76	E	30	15	1	150	50	241	199	3	5	67	352	2390	1842	2	9	25	4	4	11		487	4858	55	5400	sn 99
DE2	GL				2		50				2	3	13	40			1					52	58	1	111	sb 99
70/77	E	8	136	118	1296	566	28	4	17	14	9	63	616	2212		56	20	265	39	17		2152	2935	397	5484	sn 93
E1	GL		3	6	29	15	1					2	15	37				2	1			54	54	3	111	sn 97
71/78	E	17	3	2	3	4	23	5	12	728	56	1716	24	25	6	428	822	14	19	1577		52	2566	2866	5484	bn 99
E2	GL	4								1	3	34		1		6	19			43		4	39	68	111	bn 96
72/79	E	38	22	10	37	14	6	6	33	2245		131	1041	1415	36	38	80	20	27	285		127	4871	486	5484	n 89
E3	GL									38		3	24	40		3	3						105	6	111	n 95
73/80	E	7	56	8	5	36	7	2		2	2	5	7	10		19	4029	29	1115	144	1	119	28	5337	5484	LIMF 97
E4	GL																99		10	2				111	111	LM 98
74/81	E	83	653	364	34	2122	109	5	28	1	32	44	102	1750		6	47	43	59	2		3365	1962	157	5484	sn 97
E5	GL	1	18	10	25						1	4		50		1	1					54	55	2	111	sn 98
75/82	E	15	8	3	2		64	2	5	2	17	13	69	9	2	67	1617	2401	1061	12	115	92	117	5275	5484	LIM 93
E6	GL															1	24	59	24	1	2			111	111	LIM 96
76/82a	E	331	40	12	51	6	1172	13	104	5	47	21	3102	471	1	7	65	31	1	2	2	1612	3763	109	5484	sn 98
E7	GL	1	1				23				1		76	8	1							25	85	1	111	sn 99
77/82b	E	522	50	8	32	3	445	120		5	427	155	3222	102	8	11	8	5		3		1060	4031	35	5126	sn 99
EF1	GL	13					8				23		65		2							21	88	2	111	sn 98
78/82c	E	1	1	16	3	4		22		3	2	114	4	25	3	1260	3208	11	447	1	1	25	170	4931	5126	VLM 96
EF2	GL											5		1		24	76	2	3				6	105	111	VL 90
79/83	E	840	430	1265	50	1011	4	5	5	3	7	27	14	1431	2	4	11	11	2	2	2	3600	1492	34	5126	sn 99
EF3	GL	28	8	14	3	44								12			2					97	12	2	111	sn 98
80/84	E	6		94	18	74	73	584	4	28	17	1996	1591	447	40	96	8	29	1	20		265	4667	194	5126	n 91
EF4	GL			1	1			26				52	22	8		1						2	108	1	111	n 97
81/85	E	3	18	4080	444	11	15	2	2	3	55	302	123	9	1	38	3	1	12		4	4571	496	59	5126	sn 99
EF5	GL			85	5						4	10	2			4			1			90	16	5	111	sn 95
82/86	E	2	1	27	5022	5	8		1	7	9	10	3	6	1	10	5	3	1	4	1	5065	36	25	5126	D 98
EF6	GL				111																	111			111	D 100
		R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W					

Site	Seq type	Hydrophilic residues (s)						Neutral residues (n)							Hydrophobic residues (b)							Total residues				Conservation type (%)
		R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W	s	n	b	all	
83/87	E	3	2	388	23	5	16	6	4	11	19	493	990	1853	4	260	435	171	54	387	2	437	3376	1313	5126	bn 91
EF7	GL			30								3		48		9		1	1	19		30	51	30	111	bn 73
84/88	E	5		1	2	1		6		8	523	4771	18	39	5	45	4	1	3	3		9	5365	61	5435	GA 97
F1	GL										7	102	2										111		111	GA 98
85/89	E	20	1	28	308	4	11		31	23	8	11	45	1458	10	2372	174	434	489	8		372	1576	3487	5435	bn 93
F2	GL			3	27					1		1		20		53	3		3			30	22	59	111	bn 73
86/90	E		1		3	6	3	1	6	5356	1	5	2	1	7		4	3		35	1	13	5372	50	5435	Y 99
F3	GL									111													111		111	Y 100
87/91	E	5			1	2	3	1	35	4292	3	3	8	4	27		20	6		1025		11	4346	1078	5435	YF 99
F4	GL								3	107										1			110	1	111	y 96
88/92	E	6	1			8	2	1		4	7	5	11	1	5378	2	1		1	2	5	17	29	5389	5435	C 99
F5	GL														111									111	111	C 100
89/93	E	9	36	7	4	1293	22	14	66	6	183	2890	162	228	15	124	167	3	86	91	29	1371	3549	515	5435	sn 91
F6	GL					27	1		2	1	4	49	4	3	2	1	8		9			28	63	20	111	sn 82
90/94	E	2538	184	8	15	1591	102	26	107	20	97	81	286	152	19	44	125	28	4	3	5	4438	769	228	5435	sn 96
F7	GL	40	5	1		31				1		7	11	7		2	4	2				77	26	8	111	sn 93
97/102	E	2	1	1	11	1	9	62	34	1018	6	15	50	1157	5	528	63	122	8	31	3	25	2342	760	3127	bn 99
G1	GL							1	1	1				5		5	1	1					8	7	15	bn 100
98/103	E	4		1		1				1	3					1	5			1444	1667	6	4	3117	3127	FW 99
G2	GL																			9	6			15	15	FW 100
99/104	E	1			1	1		1		1	3109	7	4			1					1	3	3122	2	3127	G 99
G3	GL										15												15		15	G 100
100/105	E	32	35	8	1	1499	1	103	11		802	422	117	90		3	3					1576	1545	6	3127	sn 99
G4	GL	1				8		1			4			1								9	6		15	sn 100
101/106	E	2		1	1	1		1			3109	6	3	1		1	1					5	3120	2	3127	G 99
G5	GL										15												15		15	G 100
102/107	E					1	3	4			3	20	16	3057		7	3	13				4	3100	23	3127	T 98
G6	GL													15									15		15	T 100
103/108	E	48	1315	35	1	17	14	8		3	2	4	260	680		11	629	2	95	1	2	1430	957	740	3127	sn 76
G7	GL	1	7			1								1			4		1			9	1	5	15	sn 67
104/109	E				1	1		1			2	3		3	1	1668	1439	4		3	1	2	9	3116	3127	VL 99
G8	GL															10	5							15	15	VL 100
105/110	E	1		1123	45	3	2	2	2		4	7	17	1846		27	6	40		2		1174	1878	75	3127	sn 98
G9	GL			4	1									10								5	10		15	sn 100
106/111	E	4	2				2				2	2	7	1	1	1956	241	888	12	9		8	12	3107	3127	VLI 99
G10	GL															10		5						15	15	VI 100
107/112	E	16	1113	2		2	12	2			3	4	1564	48	2	3	218	4		2	1	1145	1621	230	2996	sn 92
G11	GL		5										6				4					5	6	4	sn	sn 73
		R	K	E	D	Q	N	P	H	Y	G	A	S	T	C	V	L	I	M	F	W					

For each site we give: (i) the Kabat residue numbers in the form m/n where m is the residue number in V_L and n the number of the structurally homologous residue in V_H; (ii) the position of the residue in a strand, e.g. A2 or in an interstrand loop, e.g. A′B 2; (iii) the frequencies of residues found in expressed sequences in the row marked E and the frequencies in the products of the human germline genes in the row marked GL; (iv) the total number of residues found at this site (note that the overall totals for the expressed sequences vary because some are only known in part); and (v) the major type of conservation found at the site (see the text for the conventions used here) and the percentage of sequences that have this type of conservation.

(13%). The n group of residues form 49% of the all residues in variable domains; the b group form 26% and the s group form 25%. This means that residues in the sn group are 74% of all residues and those in the bn group are 75%.

Examination of the residue frequencies in the ~5300 Kabat sequences shows that there is very significant residue conservation at 72 of the 76 core sites: details are given in Table 1 and summarised below in Table 2. These 72 sites show conservation in at least 89% of the sequences and, at 49 sites, the conservation occurs in at least 97% of sequences. There are nine IR sites where a single type of residue is found in 93 to 99% of sequences and 17 SR sites where one of a set of closely related residues is found in 91 to 99% of sequences. At 46 sites the conservation involves the chemical character of the residues. Of the 46, two sites conserve neutral (n) residues (89 and 98% of all sequences); and two conserve hydrophilic residues (92 and 94%). On a broader, but still very significant, level there are 16 sites with bn residues in 91 to 99% of sequences and 26 sites with sn residues in 91 to 99% of sequences.

There are four sites where there is a little or no conservation of residue type. At two, 5/5 and 66/71 the proportion of s, n and b residues found in different sequences is close to what would be given by a random distribution of residue types. At 11/10 there is some bias towards s type residues and at 103/108 there is a bias towards b type residues; in both cases the bias is at the expense of n type residues (Table 1).

Examination of the residue frequencies in the human germline sequences shows that they give a classification of core sites that is the same as that given by the expressed sequences at 72 of 76 sites. At three of the other four sites the human germline sequences give somewhat more conservative classifications; at 40/41, Pro 94% rather than sn 98%; at 43/44, n 95% rather than sn 97%; and at 62/67, b 100% rather than bn 99%. (The fourth site is 5/5 which is not strongly conserved.) As might be expected, the extent of the residue variations within the classes tends to be narrower in the human germline sequences than in the expressed sequences (Table 1). Nevertheless, given the absence of somatic mutations and species differences in the human germline sequences, the agreement of the descriptions of sequence conservation in variable domains given by two sets of data is remarkable.

Buried and Surface Sites in Variable Domains

In discussing the role of residues in the structure of a protein it is useful to know whether their sites are buried in the protein interior, highly exposed on the surface or in an intermediate position. A good quantitative measure of this property is the area that it has accessible to the solvent (Lee & Richards, 1971; Shake & Rupley, 1973; Chothia, 1976; Miller *et al*., 1987). We calculated the accessible surface area (*ASA*) of residues in six V_κ, two V_λ and eight V_H domains using the procedure described by Miller *et al*. (1987). The *ASA* values of structurally equivalent sites in the different domains are similar and, using the data from the 16 structures, we determined the mean accessible surface area of the residues that occupy the 76 common core sites. The mean accessible surface area of each core site is given in Table 2.

To give an overview of the positions of residues with different types of conservation, we put each site in one of three groups: interior, exposed or highly exposed. Interior residues are those with accessible surface areas in the range 0 to 20 Å^2; partly exposed residues are those with accessible surface areas in the range 20 to 50 Å^2, and highly exposed residues are those with accessible surface areas of 50 Å^2 or greater (see Miller *et al*. (1987) for a discussion of the data on which the ranges are based and for references to related publications). In Table 3 we give the distribution of the different types of sites in these three groups.

The sites that are strongly conserved, the IR and SR, sites are strongly biased in their distribution: 21 are found in the interior, two in the partly exposed surface and three in the highly exposed surface (Table 3). The distribution of 45 sites that conserve just the class of residue (the RC group) is less biased: 14 sites are found in the interior, nine on the exposed surface and 22 on the highly exposed surface.

The *ASA* values for the b, bn, n and sn sites correspond to what would be generally expected. All b sites are in the interior as are two-thirds of the bn sites. The n sites are divided between the interior and surface. The sn sites are largely on the surface: 22 out of 25. The conserved hydrophilic(s) sites, however, are the opposite of what might be expected generally: three are in the interior, two in the exposed surface and only one in the highly exposed surface.

Conservation of the Residue-Residue Contacts

The structural role of a residue is determined by the nature and extent of the interactions it makes with other residues. In this section we describe a systematic analysis of residue-residue interactions in the variable domain structures and list the contacts that are conserved. These conservative contacts characterise the specific folding motif of the immunoglobulin family.

We are interested here in the intra-domain contacts that are involved in forming the three-dimensional structure of the variable domains, as opposed to those involved in inter-domain or local structure. This means that we do not consider contacts between domains or between residues within the same strand or loop. In our analysis we dis-

Table 2. Residue conservation and accessible surface areas at the sites that form the common core of the variable domains

Site	Res	%	*ASA*	Site	Res	%	*ASA*	Site	Res	%	*ASA*
A. *Sites that have an invarient residue or one of closely related residues*											
Sites with invariant residues (IR)											
23/22	C	99	0	35/36	W	99	1	38/39	Q	93	16
82/86	D	98	5	86/90	Y	99	0	88/92	C	99	0
99/104	G	99	3	101/106	G	99	13	102/107	T	98	5
Sites that conserve closely related residues (SR)											
4/4	LM	95	8	6/6	QE	99	14	16/15	GS	97	58
21/20	VLIM	99	1	22/21	ST	96	42	47/48	VLIM	92	7
61/66	RK	95	41	63/68	ST	91	58	65/70	ST	95	64
73/80	LIMF	97	0	75/82	LIM	93	0	78/82c	VLM	96	3
84/88	GA	97	6	87/91	YF	99	6	98/103	FW	99	18
104/109	VL	99	0	106/111	VLI	99	7				

Site	%	*ASA*	Site	%	*ASA*	Site	%	*ASA*	Site	%	*ASA*
B. *Sites that conserve residue class(es) (RC sites)*											
Sites that conserve hydrophobic and neutral residues (bn)											
11/10	75	36	12/11	99	73	13/12	92	12	15/14	99	74
19/18	99	12	25/24	99	11	33/35	99	2	36/37	99	1
44/45	99	7	46/47	94	17	48/49	99	0	62/67	99	10
64/69	99	11	71/78	99	1	83/87	91	58	85/89	93	31
97/102	99	34									
Site that conserves neutral residues (n)											
40/41	98	105	72/79	89	36						
Sites that conserve hydrophilic and neutral residues (sn)											
5/5	79	85	14/13	99	74	17/16	99	74	18/17	99	116
20/19	99	77	24/23	98	73	34/35	97	8	39/40	92	62
41/42	99	82	42/43	99	95	43/44	97	33	60/65	98	110
66/71	79	73	67/72	97	74	68/71	99	56	69/74	99	46
70/77	93	47	74/81	97	29	76/82a	98	57	77/82b	99	49
79/83	99	82	80/84	97	77	81/85	99	113	89/93	91	1
90/94	96	7	100/105	99	76	103/108	76	96	105/110	98	58
107/112	92	60									
Site that conserves hydrophilic residues (s)											
37/38	92	21	45/46	94	89						

Residue frequencies are calculated from data in Table 1.
Mean accessible surface area of each site (*ASA*) is given in Å^2.

tinguish between two types of contacts: P-contacts, which are between residues in neighbouring strands (or loops) within a single β-sheet, and F-contacts which are all other contacts between residues in a variable domain. Most F-contacts are between residues in different β-sheets.

Conserved residue-residue contacts were calculated in the variable domains of 27 different structures† and are listed in Table 4. We assume that two residues are in contact if any two heavy atoms of these residues are closer than 5.0 Å. We consider here that the residues at the given position have conservative contacts if they are observed in more than 25 of the analysed structures. Conserved P contacts are made by residues at 67 sites in both V_L and V_H domains; at three sites (42, 79 and 81) in V_L domains alone, and at one site (41) in V_H domains alone (Table 4). Conserved F contacts are made by residues at 32 sites in both V_L and V_H domains; at four sites (11, 22, 36 and 81) in V_L domains alone, and at four sites (17, 83, 84, and 111) in V_H domains alone.

Residues that make P-contacts are important for the formation and mutual orientation of strands in a β-sheet (Gelfand & Kister, 1995). Residues at half the sites have intersheet F-contacts in addition to the intrasheet contacts. They are mainly responsible for interactions between the two β-sheets. This list of contacts helps to characterise the structural role of residues (see below). Note that many of the residues that make F-contacts belong to the IR or SR groups.

The Structural Role of Residues

In the preceding sections, we have described the nature of the residue conservation at the 76 sites in the common core; the extent to which these sites are buried or on the surface of the domain and the residue-residue contacts. In this section, we

† The structures are identified by the Protein Data Bank names: 1BAF, 1BBD, 1BBJ, 1CBV, 1DBA, 1DFB, 1F19, 1FDL, 1GGC, 1HIL, 1IGF, 1IGI, 1MAM, 1MCO, 1MCP, 1NBV, 1TET, BL,2FB4, 2FBJ, 2HFL, 2IG2, 2MCP, 4FAB, 6FAB, 7FAB, 8FAB and 3HFM.

describe the role of each site in the three-dimensional structure of variable domains.

The wealth of structural detail involved in such a discussion can be confusing, therefore, to try and make it easier to follow we put our discussion in the context of a simple hydrophilic/hydrophobic model that incorporates two assumptions. The first assumption is a general relation between sequence and structure: a sequence consisting of hydrophilic/neutral (s/sn) and hydrophobic/neutral (b/bn) sites produces the fold of the protein by placing all s/sn sites on the surface and all b/bn sites in the interior. The second assumption is that the structure of the common core of the variable domains can be seen as two β-sheets packed face to face: the inside faces of both β-sheets forming the interior of the domain; the outside face of the ABED β-sheet exposed to solvent, and the outside face of the A′GFCC′ β-sheet having one part in the interface of the V_L-V_H dimer, a second part in contact with the C domains and a third part exposed to solvent.

The outside of the ABED β-sheet

Sites: 5/5, 18/17, 20/19, 22/21, 24/23, 63/68, 65/70, 67/72, 70/77, 72/79, 74/81, 76/82a

On a plan of the ABED β-sheet in Figure 3(a), we show the nature of residue conservation and the accessible surface area at the 12 sites whose side-chains are on the outside surface. Residue conservation at eight of the 12 sites fits the simple model: they have hydrophilic/neutral (sn) residues in 79 to 99% (on average 95%) of the sequences.

At three sites, 22/21, 63/68 and 65/70, the conservation and chemical character of the residues does not fit the simple model: at all three Ser or Thr occur in 91 to 95% of sequences. At a fourth site, 72/79, n group residues are found in 89% of sequences. There is no apparent structural reason for the conservation at these sites. If the sequences of V_L and V_H domains are examined separately, we find more specific conservation of specific residues. In V_L sequences we find:

		67 Ser : 91%
22 Ser/Thr : 95%	72 Ser/Thr/Ala : 97%	65 Ser/Thr : 95%
		63 Ser/The : 91%.

and in V_H sequences we find:

		72 Asp : 92%
22 Ser/Thr : 95%	79 Tyr/Phe : 89%	70 Ser/Thr : 95%
	81 Ser/Thr/Ala : 97%	68 Ser/The : 91%.

Apart from the hydrophilic/hydrophobic character of the sites, this model has no residue specificity. In what follows we look at each of the 76 sites and ask does it have the non-specific character expected from this simple model and, if it does not, what are its structural/functional role(s) that require that the residue has some specific characteristics:

Table 3. Nature of the conservation at sites in the interior and on surface of variable domains

Residue class	Interior sites *ASA*: 0 to 20 Å² IR and SR sites	Interior sites *ASA*: 0 to 20 Å² RC sites	Partly exposed sites *ASA*: 20 to 50 Å² IR and SR sites	Partly exposed sites *ASA*: 20 to 50 Å² RC sites	Highly exposed sites *ASA*: 50 + Å² IR and SR sites	Highly exposed sites *ASA*: 50 + Å² RC sites	Total
b	12	–	–	–	–	–	12
bn	1	11	–	3	–	3	18
n	5	–	1	1	3	1	11
sn	–	3	–	5	–	21	29
s	3	–	1	1	–	1	6
Total	21	14	2	10	3	26	76

The strong conservation of particular types of residues at these surface sites suggests that they are involved in a recognition process whose identity is not known at present.

The outside of the A′GFCC′ β-sheet

V_L-V_H Interface sites: 34/35, 36/37, 38/39, 43/44, 44/45, 46/47, 87/91, 89/93, 98/103, 100/105; V-C interface sites: 12/11, 83/87, 105/110, 107/112; surface sites: 14/13, 85/89, 103/108

On a plan of the A′GFCC′ β-sheet in Figure 3(a), we show the nature of residue conservation and the accessible surface area at the 17 sites on the outside surface. This surface is largely involved in the packing of the V_L and V_H domains and in the contacts made between V and C domains.

The packing of the V_L and V_H domains buries, or partly buries, ten residues at the top and right-hand side of the β-sheet (as viewed in Figure 3(a)). The nature of these residues and their relation to the geometry of the V_L-V_H interface have been described in some detail previously (Chothia *et al.*, 1985) and, as was mentioned at the beginning of the paper, it is outside the scope of the present work.

Table 4. Conserved strand-strand and loop-strand contacts in variable domains

Site		Conserved contacts that are made to 1. Residues in the same β-sheet (P)	2. Residues in the other β-sheet (F)
4/4	A1	B6 B7 B8	F5 F6 F7 G1 G2 G3 G4
5/5	A2	B5 B6 B7	
6/6	A3	B5 B6	C3L F3 F4 F5 G4 G5 G6
11/10	A′1	G6L G7 G8 G9	B2L
12/11	A′2	G9	
13/12	A′3	EF2 G9 G10 G11	A′B2 B1H B2
14/13	A′4	G11	
15/14	A′5	G10H G11	
16/13	A′B1	EF1 EF2	
17/16	A′B2	E6 E7 EF1 EF2	A′3
18/17	B1	E6 E7 EF1 EF2	A′3H
19/18	B2	E4 E5 E6 EF2	A′1L A′3 G8
20/19	B3	E3L E4 E5	
21/20	B4	E2 E3 E4	C3 F3 G6
22/21	B5	A2 A3 E2 E3	C3L
23/22	B6	A1 A2 A3 E1 E2	C1L C3 F5
24/23	B7	A1 A2 E1	
25/24	B8	A1	
33/34	C1	C″7L F5L F6 F7	B6L E2L
34/35	C2	C″7 F4L F5 F6	
35/36	C3	C′4L C′5 C′6 C′7 F3L F4 F5	A3 B4 B5L B6 E2 E3 E4
36/37	C4	C′4L C′5 C′6 C′7 F3L F4 F5	G2L
37/38	C5	C′2L C′3 C′4 C′6 EF6 F1L F2 F3	
38/39	C6	C′1L C′2 C′3 F1 F2 F4	
39/40	C7	EF5L EF7 F1	
40/41	CC′1	EF7H	
41/42	CC′2	–	
42/43	C′1	C6L	
43/44	C′2	C5L C6	
44/45	C′3	C4L C5 C6	
45/45	C′4	C3L C4 C5	
46/46	C′5	C3 C4	
47/48	C′6	C3 C4 C5	D2 E4 F3
48/47	C′7	C1L C2 C3	D4 E4L
60/65	C″D5	D2	
61/66	D1	E5 E6 E7 EF1 EF3L	EF5L EF6
62/67	D2	E4 E5 E6	C′6 EF6L F3L
63/68	D3	E3 E4 E5	
64/69	D4	E3 E4	C′4
65/70	D5	E2 E3	
66/71	D6	E1 E2 E3H	
67/72	D7	E1 E2L	
68/73	DE1		
69/76	DE2		
70/77	E1	B6 B7 D6 D7	
71/78	E2	B4 B5 B6 D5 D6 D7L	C1L C3
72/79	E3	B3L B4 B5 D3 D4 D5 D6H	C3
73/80	E4	B2 B3 B4 D2 D3 D4	C3 C′6 C″7L F3
74/81	E5	B2 B3 D1 D2 D3	
75/82	E6	A′B2 B1 B2 D1 D2	EF6L F3
76/82a	E7	A′B2 B1 D1	
77/82b	EF1	A′B1 A′B2 B1 D1	
78/82c	EF2	A′3 A′5 A′B1 A′B2 B1 B2	G8L G10H
79/83	EF3	D1L	G10H
80/84	EF4		G10H
81/85	EF5	C7L	D1L
82/86	EF6	C5 E6L F3 G8	D1 D2L
83/87	EF7	C7 CC′1H G8 G9 G10	
84/88	F1	C5L C6 C7 G6 G7 G8	
85/89	F2	C4L C5 C6 G6 G7	
86/90	F3	C3L C4 C5 EF6 G5 G6 G8	A3 B4 C′6 D2L E4 E6
87/91	F4	C2L C3 C4 C6 G2 G3 G4 G5	A3
88/92	F5	C1L C2 C3 G1 G2 G3	A1 A3 B6
89/93	F6	C1 C2 C4L G1 G2	A1
90/94	F7	C1 G1	A1
91/95	F8		
97/102	G1	F5 F6 F7	A1
98/103	G2	F4 F5 F6	A1 C4L
99/104	G3	F4 F5	A1 A3 C4L
100/105	G4	F4	A1 A3
101/106	G5	F3 F4	A3
102/107	G6	A′1L F1 F2 F3	A3.B4
103/108	G7	A′1 F1 F2	
104/109	G8	A′1 EF6 EF7 F1 F3	B2 EF2L
105/110	G9	A′1 A′2 A′3 EF7	
106/111	G10	A′3 A′5H EF7	EF2H EF3H EF4H
107/112	G11	A′3 A′4 A′5	

L or H is added to the contact site if these contacts are found generally only in V_L or V_H domains.

Here residues are listed as being in contact if they have atomic groups closer than 5 Å. Loop-residue contacts are not listed if they are separated in the sequence by three residues or less.

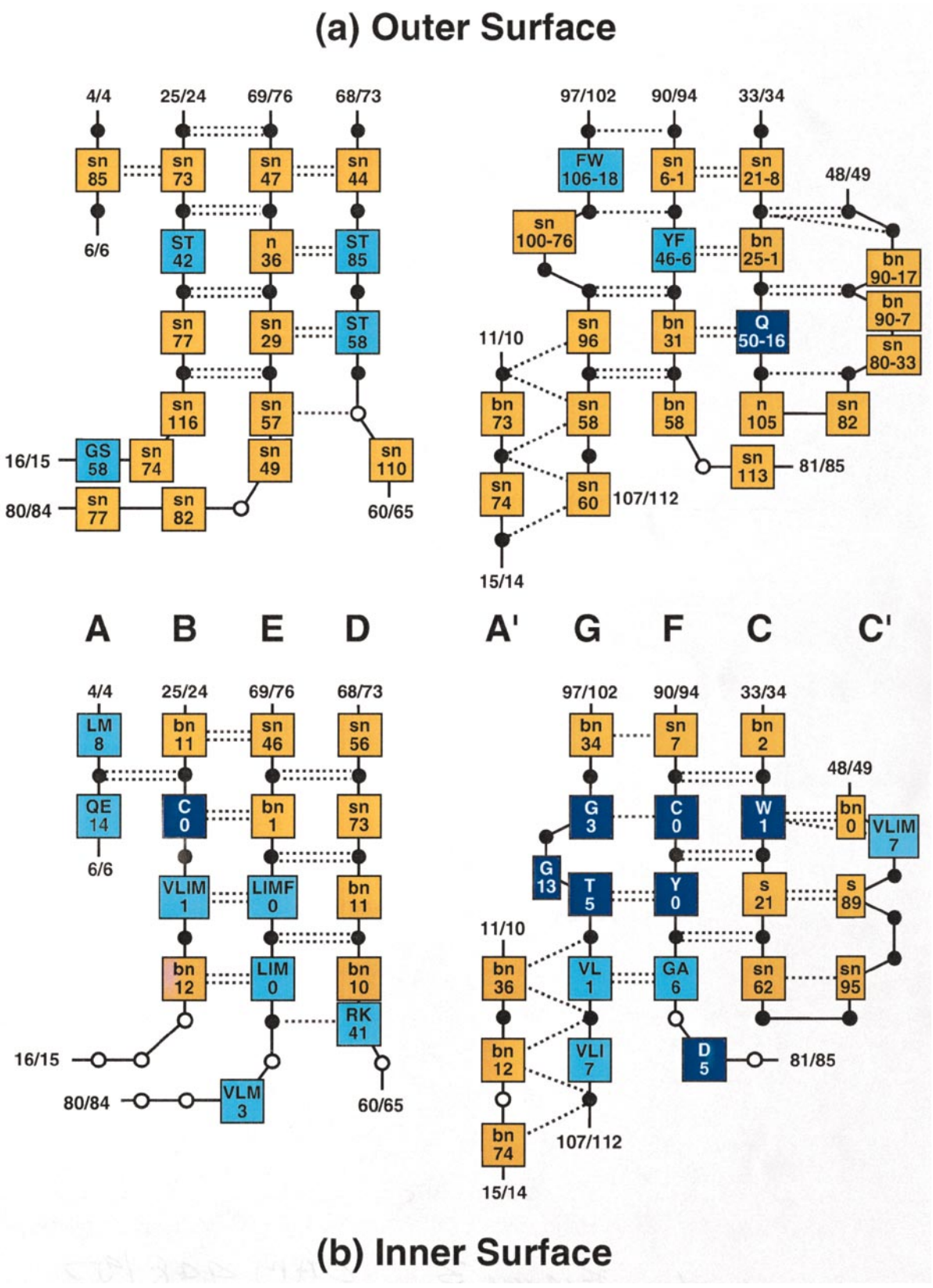

Figure 3. Residue conservation and accessible surface area at sites that form the common core: (a) sites on the outer-surface of the ABED β-sheet (left) and of the A′GFCC′ β-sheet (right), and (b) sites on the inner-surface of the ABED β-sheet (left) and of the A′GFCC′ β-sheet (right). For each site we indicate the nature of the conservation in upper case letters for residues at IR (in dark blue) and SR sites (in light blue) and in lower case letters for residue classes at RC sites (in ochre) (see Table 1). The mean accessible surface area of residues at each site is also given; the units are $Å^2$. For the residues that form the V_L-V_H interface (on the outside of the A′GFCC′ β-sheet) we give mean accessible areas of residues in both the absence and presence of the other domain. Residues 47/48 and 101/106 are shown here for the sake of clarity bulging ''away'' from the β-sheet. They do in fact fold in between the β-sheets and are buried: see Figure 4 which shows the relative position of 47/48.

V_L-V_H dimers can rotate some 50° relative to the C_L-C_{H1} dimers. This movement is facilitated by torsion angle changes in the peptide links of the V and C domains and by a molecular ball-and-socket joint between V_H and C_{H1} (Lesk & Chothia, 1988). The ball-and-socket joint is formed by residues 11, 110 and 112 in V_H and 149 and 150 in C_{H1}. The hydrophobic residue at 11 and the neutral residues at 110 and 112 are absolutely or strongly conserved in V_H. Depending upon the relative orientation of domains, addition V-C contacts can be made particularly by the V_L residues 40 and 83, which are usually hydrophobic or neutral (Table 1 and Figure 3(a)).

The three remaining residues are 14/13, 85/89 and 103/108. Residues at 85/89 belong to the bn group. This is because they are partly covered by the conserved Phe/Tyr at 87/91 that forms part of the V_L-V_H interface. Residues at 14/13 and 103/108 are members of the sn group.

The inner faces of the ABED and A′GFCC′ β-sheets

The inner face of the ABED β-sheet has 15 sites and that of the A′GFCC′ β-sheet has 20 sites: 35 in all (Figure 3(b)). Of these, 17 are IR or SR sites. The other 18 are RC sites: ten bn sites and eight sn or s sites. Thus, two-thirds of the sites are not what would be expected from the simple model and in the following paragraphs we describe the nature of the interactions that account for the specific features of these sites. In these descriptions it is useful to include two sites in the EF turn, 78/82c and 82/86, that are buried and in contact with the interior residues.

The interface between the two β-sheets can be divided into four regions. Three regions, a central hydrophobic region and the two adjacent polar regions, form what we call the ''deep'' structure of the variable domain fold. It involves contacts made by a set of 21 residues almost all of which are strongly conserved; see below. Adjacent to the deep structure there is a ''peripheral'' region formed by residues that are buried but not well conserved.

The central hydrophobic region: 4/4, 19/18, 21/20, 23/22, 35/36, 47/48, 73/80, 75/82, 78/82c, 84/88, 86/90, 88/92, 102/107, 104/109 and 106/111

Residues at 15 sites form a hydrophobic region that fills the centre of the interface between the β-sheets. Nearly all these sites are strongly conserved, i.e. IR or SR sites. Going from top bottom of the molecule to the bottom, as viewed in Figure 4, the residues in this region are 4/4 LM, 23/22 C, 88/92 C, 35/36 W, 102/107 T, 21/20 LIM, 73/80 LIMF, 47/48 VLIM, 86/90 Y, 104/109 VL, 19/18 bn, 84/88 GA, 75/82 LIM, 106/111 VLI and a residue from the EF turn 78/82c VLM. Note that the 19/18 RC site is close to being a SR site in that 85% of the residues at that site are V, L, I or M (Table 1). On two sides of this central hydrophobic region there are sets of buried polar residues.

Buried polar region I: 37/38, 61/65, 84/88 and 86/90

On one side of the hydrophobic core there are three buried polar residues 82/86 D in the EF turn, 37/38 s (usually R, K or Q) in the C strand and the

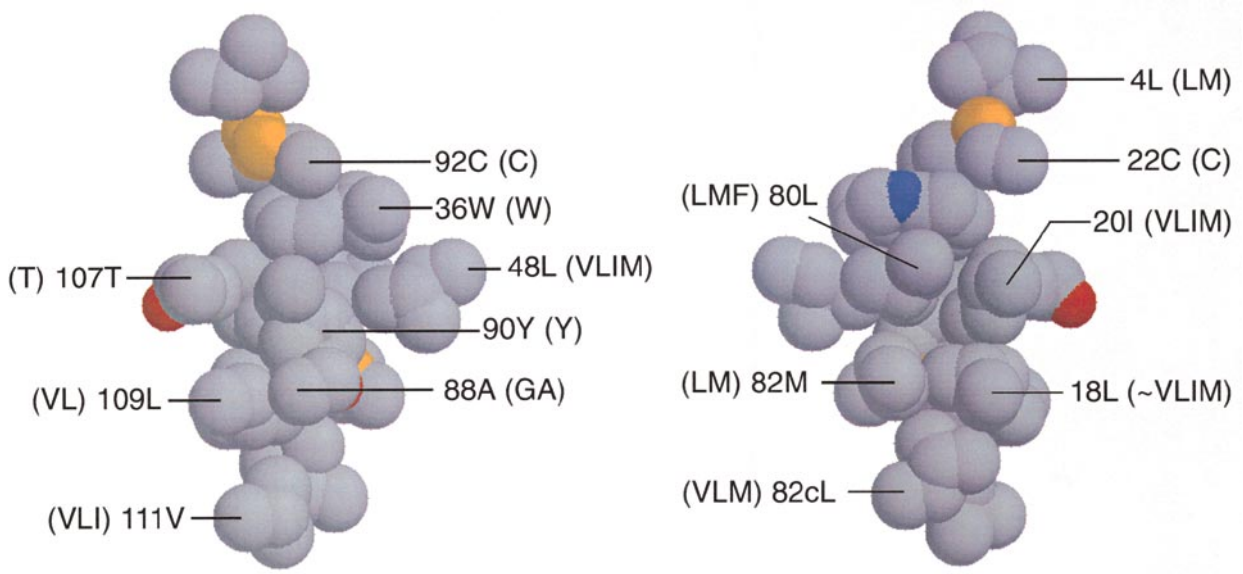

Figure 4. The packing in the central buried hydrophobic region. The side-chains of the 15 residues that form the central hydrophobic region are shown in space-filling drawings. On the left we show a ''front'' view of the residues, i.e. from the point of view used in Figures 1 and 3. On the right we show a ''back'' view of the residues. In brackets we list the residues commonly found at these sites in other variable domains. At 14 of the sites the residues listed are almost absolutely conserved. At site 19/18 residues VLIM are found in 85% of sequences (Table 1).

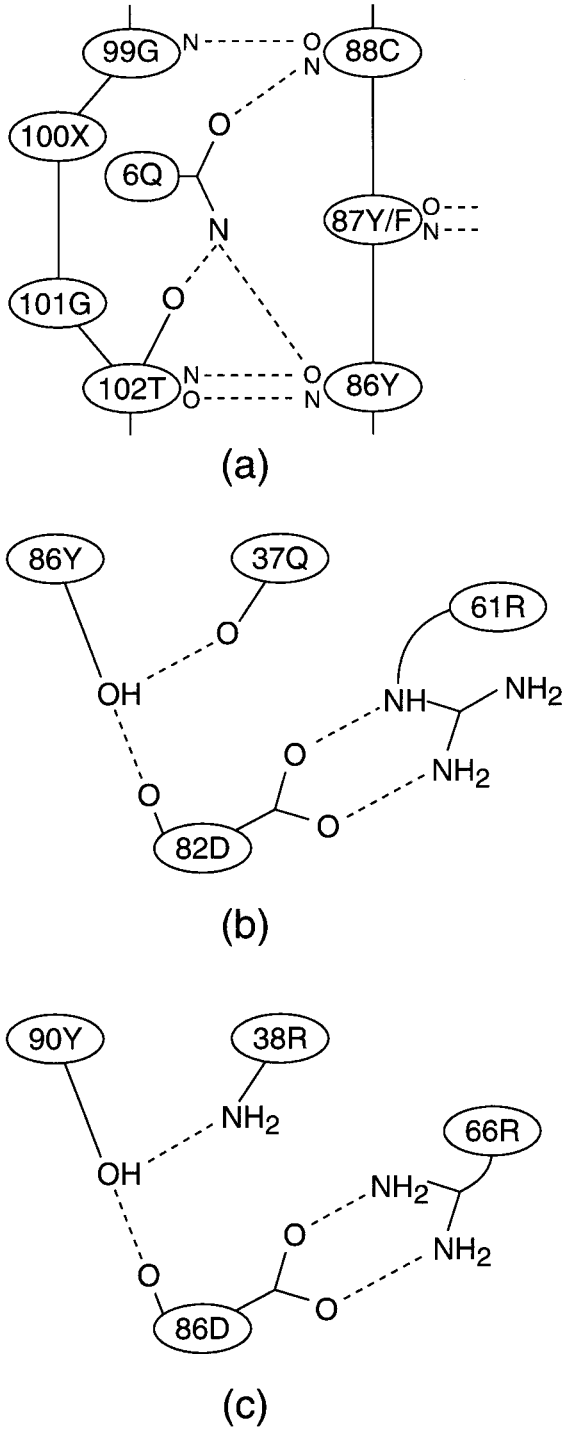

Figure 5. Schematic diagrams of the hydrogen bonds formed in the buried polar regions. These regions are adjacent to the central hydrophobic region (see the text and Figure 3). (a) The interactions in the second polar region. In cases where Gln occurs at 100/105 the hydrogen bonding can be modified. (b) and (c) The interactions in the first polar region. These are a little different in V_L (b) and V_H (c) domains (see the text).

partly accessible residue 61/65 RK in the D strand (Figure 3). These hydrogen bond to each other and to the hydroxyl of 86/90 Y, a residue that is part of the hydrophobic core. These hydrogen bonds stabilise the buried polar side-chains and the conformation of the EF turn. At site 37/38 the presence of Q in place of R/K, modifies the pattern of the hydrogen bonds (Figure 5(b) and (c)).

The conformation of residues 82/86 to 86/90 and the hydrogen bonding made by the side-chain of the β-sheet residue 86/90 Tyr to the main-chain polar atom of a residue four residues back in a turn, 82/86, make this region an example of what has been called a ''tyrosine corner''. They occur commonly in β-sheet sandwich structures (Hemmingsen *et al.*, 1994).

Buried polar region II:: 6/6, 99/104, 101/106 and 102/107

On the other side of the hydrophobic region 6/6 QE, 99/104 G and 101/106 G pack together with the hydrophobic core site 102/107 T (Figure 3). The buried polar side-chain atoms of 6/6 QE and 102/107 T and main-chain polar groups form hydrogen bonds that stabilise the β-bulge in the G strand (Figure 5(a)). The two Gly in the G strand, 99/104 and 101/106, are buried and the second has a conformation that is only allowed for Gly residues. The high conservation of residues in this region is functionally important because the conformation of the bulged G strand is crucial for the formation of the interface between V_L and V_H (Chothia *et al.*, 1985).

Sites around the central hydrophobic region: bn sites: 11/10, 13/12, 25/24, 33/34, 48/49, 62/67, 64/69, 71/78 and 97/102, and the sn site 90/94

The preceding paragraphs have described how a set of sites, nearly all of which have invariant residues or closely related residues, pack together to form the deep structure of the variable domains. It is surrounded by nine bn sites and one sn site. Eight of these are buried and two are partly exposed (Figure 3(b)). All have, within the constraints of their residue classes, a very wide range of residues (Table 1). We will refer to this set of ten residues as the peripheral buried structure of the variable domain fold.

How are the variations in the size of the residues in the peripheral structure accommodated? Inspection of the atomic structures of immunoglobulins shows that most of the variations are accommodated by the differences in the conformations and size of loop regions that are not part of the common core. Indeed differences in residues at the common core sites 25/24, 33/34, 48/29, 71/78, 90/94, and 97/102 play an important role in the different conformations of loops that form the hypervariable regions (see Lesk & Chothia, 1982; Chothia & Lesk, 1987; Chothia *et al.*, 1992; Tomlinson *et al.*, 1995). To give one example: at site 71/78 V_λ domains usually have Ala but V_κ domains usually have Phe or Tyr. This large difference in volume is accommodated by the BC loops

having quite different conformations that move a hydrophobic side-chain in the loop close to the C^{α} atom of the residue at 71/78 in V_{λ} domains, or further from it in V_{κ} domains.

Sites 11/10 and 13/12 have access to the surface and increases in volume can be accommodated by putting the additional atoms on the surface. Indeed, residues such as R, K and E are found occasionally at these sites. The hydrophobic part of these side-chains is packed between the β-sheets and the charged head group is on the surface.

Saul & Poljak (1993) have described how the effects of sequence differences at position 67 in V_H spread right across the β-sheet. If F replaces L, I, V, T or A at site 67 the conformation of the neighbouring side-chain, 82, and its neighbour, 18, are switched. The switch in conformation at 18 is allowed by a change in conformation of the AA′ link.

To summarise: the variations in the size of these buried residues are accommodated by the changes in conformation of those parts of the protein not in the common core, small changes in these regions or by the residues having access to the solvent.

sn sites on the corners of the β-sheets: 39/40, 42/43, 45/46, 66/71, 68/73 and 69/76

At two corners of the inner faces of the β-sheets there are six sn sites (Figure 3(b)). The twist and coil of the β-sheet at these corners exposes these ''inner'' residues to solvent (Figure 1(b)).

Inter-strand loops

The conserved inter-strand loops have the sequence patterns:
A′B: bn(15/14)-G/S(16/15)-sn(17/16)-sn(18/17);
CC′: n(40/41)-sn(41/42), and
EF: sn(77/82b)-VLM(78/82c)-sn(79/83)-sn(80/84)-sn(81/85)-D(82/86)-bn(83/87).

The A′B link has the conformation of a type II′ turn with the G/S residue at 16/15 having a $+\phi, +\psi$ conformation. The side-chain of the hydrophobic/neutral residues at the 15/14 site does have one side exposed to the solvent but the other side packs against the hydrophobic buried core. The CC′ turn has the typical turn residues P-G at sites 40/41 - 41/42 in over 80% of sequences and alternative residues at the two sites are typical of turns in nearly all other cases (Table 1). The EF link forms a short helix. The residues at sites 78/82c and 82/86 are buried between the β-sheets and form part of the deep structure: see above.

Residues in the Variable Domains of P_0, CD8, CD4 and CD2

So far in this paper we have been concerned with the identity and structural role of residues in the variable domains of immunoglobulins. Homologues of variable domains are found in other members of the immunoglobulin superfamily and structures are known for five of these: domain 1 of CD2 (Jones *et al.*, 1992), domains 1 and 3 of CD4 (Wang *et al.*, 1990; Ryu *et al.*, 1990; Garrett *et al.*, 1993; Brady *et al.*, 1993), CD8 (Leahy *et al.*, 1992) and P_0 (Shapiro *et al.*, 1996). The first three of these proteins are adhesion molecules involved in cell-cell recognition by T cells. The forth, P_0, mediates membrane adhesion in the formation of the myelin sheath. Here we discuss three features of the structures and sequences of these proteins: (i) the extent to which they have a structural core that is the same as that in immunoglobulins; (ii) the identity of the residues in their central buried regions; and (iii) the identity in P_0 of residues at the surface sites homologous to those that are conserved in the immunoglobulin variable domains.

Superposition calculations show that P_0 and CD8 have a core structure that is largely co-extensive in conformation with the 76 residues in the common core immunoglobulin variable domains. P_0 has four fewer residues because it does not have the β-bulge in the G strand and there is a different conformation at the end of the D strand. CD8 has five fewer residues because of small differences in the regions of AA′, CC′ and DE turns. The variable domains in CD4 and CD2 are less similar and have only 59 to 66 residues in the same conformation. This is because of larger differences in the CC′ turn that affect adjacent C′ strand residues; the absence of the β-bulge in the G strand; some strands being shorter, and the absence of the A strand.

We determined the residues in these homologues at sites equivalent to those that form the deep structure in the variable domains. We collected sequences from different species: four sequences for CD2; 11 for CD4; three for CD8 and seven for P_0. These sequences were aligned with those of the variable domains and the residues at the sites homologous to those that form the deep structure in variable domains thus found are given in Table 5.

As discussed above, the deep structure of variable domains has a central hydrophobic part and two adjacent polar parts. The central hydrophobic region in variable domains is formed by five IR sites, nine SR sites and one RC site that is close to being an SR site. In P_0 and CD8, 12 of these 15 sites have the same residues (Table 5). The three sites that have differences are, in P_0, sites equivalent to 4/4, 86/90 and 102/107 and, in CD8, sites equivalent to 4/4, 78/82c and 102/107. These sites are on the periphery of the region and the differences are all small: at 4/4 LM to V or LF; at 78/82c VLIM to F; at 86/90 Y to F, and at 102/107 T to V (Table 5).

At the sites that form the first polar region there is only one difference in P_0 and CD8: 37/38 RKQ to YF or F. In the second polar region of variable domains residues at 6/6 and 102/107 form buried hydrogen bonds that support a β-bulge which is an important part of the V_L-V_H interface (Figure 5(a)). CD8 also forms a dimer and a similar

Table 5. Residues in P_0, CD8, CD4 and CD2 at sites homologous to those that form the central buried region in variable domain of immunoglobulins

Protein:	V	P_0	CD8	CD4d1	CD2d1	CD4d3
core size:	76	72	71	66	59	58
Site	Residues					
	Central hydrophobic region					
4/4	LM	V	LF	[a]	[a]	[a]
19/18	(VLIM)	V	V	VA	I	VA
21/20	VLIM	L	L	L	L	F
23/22	C	C	C	C	I	F
35/36	W	W	W	W	W	L
47/48	VLIM	I	LI	I	VI	[a]
73/80	LIMF	I	L	LM	L	LF
75/82	LIM	I	L	I	I	LI
78/82c	VLM	L	F	L	LM	VA
84/88	GA	G	G	GDEQ	GD	G
86/90	Y	F	Y	Y	Y	G
88/92	C	C	C	C	V	L
102/107	T	V	V	V	LF	V
104/109	VL	L	V	L	LV	L
106/111	VLI	V	L	V	LI	VLI
	First polar region					
37/38	RKQ	YF	F	NHSDF	RK	W
61/66	R	R	RLQ	R	[a]	[a]
82/86	D	D	NE	D	DH	YF
86/90	Y	F	Y	Y	Y	G
	Second polar region					
6/6	QE	T	VLI	[a]	[a]	[a]
99/104	G	S	S	E	K	Q
101/106	G	QD	VLF	E	A	E
102/107	T	V	V	V	LF	V

[a] The structures of variable domains in CD2 and CD4 do not have an A strand and CD4d3 does not have a residue equivalent to 47/48. CD2 and CD4 domain 3, have at the beginning of the D strand conformations different to variable domains. Note that 19/18 has VLIM in only 85% of sequences.

β-bulge is involved but it is not supported by buried polar groups: the residues at the two sites in CD8 are hydrophobic (Table 5). P_0 also has quite different residues in this region: P_0 is not a dimer and, as was mentioned above, does not have a β-bulge in the G strand.

In contrast to P_0 and CD8, domain 1 of CD2 and domain 3 of CD4 have several differences in the residues that form the central buried region (Table 5). For example, the disulphide bridge is lost in both domains and the CD4 domain 3 has neither Trp at 35/36 nor Tyr at 86/90. For domain 1 of CD4 the number of differences are in between those of P_0/CD8 and CD2/CD4 domain 3 (Table 5).

The differences in the residues that form the central buried regions in CD2, and CD4 domain 3, are related to the differences in the conformation of regions equivalent to those that form the common core of variable domains. Here we give two examples of this. In CD2 the cysteine residues are replaced by Ile and Val. This increase in volume is accommodated in CD2 by residues, that would form the end of B strand in variable domains, moving away from the D strand into space let free by the absence of the A strand. Also in CD2, Asp is found at the position equivalent in variable domains to the buried 84/88 Gly in the F strand (Figure 5). The Asp is accommodated by the site being accessible to solvent because the C and C′ strands in CD2 are shorter than those in variable domains.

We described in the previous section how residues at sites of the surface of the ABED β-sheet in variable domains are conserved and suggested that they are part of a recognition site of unknown function. Also sites on the surface of the A″GFCC′ β-sheet have three conserved residues that form the centre of the V_L-V_H interface. P_0 has almost the same core structure as variable domains and their buried regions have very similar residues. However, its function and domain interactions are quite different (Shapiro *et al.*, 1996). What happens in P_0 at sites homologous to those in variable domains that are on the surface and believed to be conserved for functional reasons? In P_0 the residues at five of these sites are different:

Site	V_L/V_H	P_0	Site	V_L/V_H	P_0
22/21	ST	HYS	38/39	Q	Q
63/68	ST	QE	87/91	YF	T
65/70	ST	V	98/103	FW	TS

The site where the same residue is found, that homologous to 38/39, is involved in a P_0-P_0 adhesion contact but its geometry is quite different to the V_L-V_H contacts. The use of the same residue at this site is likely to be a either a coincidence or a fortuitous use and conservation of a residue found in the common ancestor of P_0 and immunoglobulin variable domains.

Structural Determinants in the Sequences of Immunoglobulin Variable Domains

Summary of the analysis of the structures and sequences of variable domains

The major requirement for the creation of a stable three-dimensional structure for a protein is a hydrophobic interior, that is sufficiently large, close packed and free of strain, and a surface that is sufficiently hydrophilic. These requirements place different constraints upon different sites. We have described a core structure of 76 residues that we expect to have the same conformation in all, or almost all, immunoglobulin variable domains. We have also described, for each site in this core, the nature and extent of the constraints placed on these residues that is implied by their frequency in the sequences of some 5300 different variable domains (Tables 1 and 2). The constraints are described in terms of sites conserving an invariant residue (IR), one of a small number of closely related residues (SR) or the chemical class of the residues at that site (RC).

The common core is formed by nine IR, 17 SR and 46 RC sites, and four sites that have little or no conservation. Analysis of the structures of immunoglobulin suggests the conservation at six surface sites, one IR site (38/39) and five RC (22/21,

63/68, 65/70, 87/91 and 98/103) is for functional reasons (dimer formation and other recognition roles) rather than structural reasons. This view is strongly supported by the analysis of homologues of immunoglobulin variable domains where these sites have sn residues if they are not involved in function.

This means that the common structure of the 76 residues in the core structure is produced by eight IR and 12 SR sites, 52 RC sites and four with little or no conservation. Inspection of the distribution and structural role of these sites shows that they can be assigned to one of four sets each with its distinct features and roles. Two sets form the interior of the core and two form the surface. The two interior sets are:

(1) The deep structure.

The 21 residues that form the central hydrophobic core and two adjacent buried polar regions form what we call the deep structure of the variable domains. The residues in this set are:

4/4	6/6	19/18	21/20	23/22	35/36	37/38
47/48	61/66	73/80	75/82	78/82c	82/86	84/88
86/90	88/92	99/104	101/106	102/107	104/109	106/111

They are highly conserved consisting of eight IR and 11 SR sites together with a bn site (19/18), that is close to being an SR site, and an s site (37/38) that usually has R, K or Q (see Figures 4 and 5 and Tables 2 and 5). Sites that have residues with very specific properties, such as Cys or Trp, are invariant. Sites with less specific residues, such as those with medium sized hydrophobic side-chains, have one of a closely related set of residues. For the less specific sites, the requirements of close packing and the limitations on strain mean that we would expect the nature and extent of substitutions permitted at different sites to vary and that only particular combinations of alternatives are allowed. This view is supported by protein engineering experiments on λ-repressor (Lim & Sauer, 1989), T4 lysozyme (Karpusas *et al.*, 1989; Gassner *et al.*, 1996) and Barnase (Buckle *et al.*, 1996; Axe *et al.*, 1996). Note that residues that form the deep structure lie along opposite diagonals of the β-sheets (Figure 3(b)). The twist of the β-sheets brings the residues at the corners of the β-sheets close to the centre of their interface (see Figure 4; Cohen *et al.*, 1981; Chothia & Janin, 1981).

(2) The peripheral buried structure.

The ten residues that form the peripheral buried structure are:

11/10	13/12	25/24	33/34	48/49	62/67	64/69	71/78	90/94	97/102

These sites are around the edge of the deep buried structure. Eight sites are buried. Two sites, 11/10 and 97/102, are on the corners of the β-sheet and have some exposure to the solvent (Figure 4). These sites consist of nine bn sites and one sn site and all have a wide range of different bn or sn residues (Table 1). The variations in the size of residues at these sites are accommodated mostly by conformational differences in the regions of the variable domains that are not part of the common core (see above).

The two surface sets are:

(1) Specific residues in turn regions.

The four sites in two turn regions have particular residues that tend to be conserved. The A′B turn produced by 15/14-16/15 has Pro in the first position in 66% of sequences and at the second position it has Gly in 87% of sequences and Ser in 10%. At the CC′ turn, 40/41-41/42, Pro occurs at the first site in 85% of sequences and Ser in 8%; at the second site Gly is found in 82% of sequences. Note that, though this conservation is very significant, it is not as high as that at IR sites in the interior.

(2) General surface sites.

At the remaining 41 sites, 31 have a residue of the s or n groups. If the functional constraints are removed from six sites involved in the recognition of other domains or proteins, we would expect them to also conserve sn residues.

The striking feature of these results is the contrast between (i) the high conservation of invariant residues and closely related residues at the sites that form the deep structure and (ii) the far less restrictive conservation of chemical class(es) at most other sites.

Experimental studies of the relative importance of residues at protein interfaces

Two protein engineering studies of the importance of residues at the centre and periphery of recognition sites are relevant to our results. The human growth hormone activates its receptor by producing a complex that consists of one hormone molecule and two receptor molecules (Vos *et al.*, 1992). Formation of the complex involves an initial interaction that brings 33 hormone residues into contact with, or close to, 31 residues on one receptor molecule. All but four of these side-chains were replaced by alanine and their effects on the affinity measured (Cunningham & Wells, 1993). For the hormone, mutations at eight sites reduced binding energies by between 1.1 and 2.4 kcal. At the other

24 sites the effects of mutation are smaller. For the receptor, mutations at 11 sites reduced binding energies by 1 to 4.5 kcal and those at 17 other sites give smaller reductions (Clackson & Wells, 1995). The size of the change in binding energy of a residue on mutation correlates neither with the number of intermolecular contacts it makes nor with the extent of surface it buries in the recognition site. It does correlate with the position of the residue in the recognition site. The eight sites in the hormone and the 11 in the receptor that produce large changes form contiguous patches at the centre of the contact regions. In the complex they pack together to form the central half of the recognition site with hydrophobic residues at the core flanked by polar and charged groups (Clackson & Wells, 1995).

The action of the extracellular ribonuclease barnase is inhibited within the bacterium by the inhibitor barstar. The two proteins have very high affinity, the K_a is 10^{14}. At the recognition site 15 residues on the enzyme make contacts with, and form 14 hydrogen bonds to, 15 residues on the inhibitor (Guillet *et al*., 1993; Buckle *et al*., 1994). The inhibitor binds to the active site of the enzyme. The function of the enzyme, to recognise and cleave RNA, means that the active site contains Arg, Lys and His residues and the inhibitor recognises these using Asp and Glu residues. The presence of charge-charge interactions at the centre of the contact is the opposite of what is found in the growth hormone-receptor complex just described. Mutations to Ala at the individual sites of a Lys, a His and two Arg at the centre of the barnase recognition site, and a Glu and two Asp at the centre of the barstar site, reduce the binding energy by 5 to 7.7 kcal (Schreiber & Fersht, 1993, 1995). These values are larger than those in the hormone-receptor complex because the interlocking set of charged hydrogen bonds formed by these residues means that the mutations leave buried charged groups without partners. Mutations at peripheral sites have only small effects on binding energies.

The experiments on these two very different complexes show that the conservation of identity of the residues that form the central part of the binding site, whatever their type, is of major importance for recognition. Changes to residues in the peripheral half can be accommodated and give only very small changes in affinity. This view of the relative importance of residues in protein-protein interfaces, obtained by experiments, is very close to the view of their importance in the β-sheet–β-sheet interface in variable domains, obtained by our analysis of sequences and structures.

Acceptance and non-acceptance of sequence changes

In certain variable domains, a Tyr residue is found at the sites 71/78 and at 86/90. At both sites the Tyr residue is buried and close packed. We would expect them to make similar contributions to the stability of the domain. However, whilst the Tyr at 86/90 is conserved in more than 99% of all variable sequences, the Tyr at site 71/78 is limited to certain V_κ domains: in V_λ domains it is replaced by Ala. Why does one site allow such large changes and not the other?

The change of Tyr to Ala at site 71 is accommodated by a large change in conformation of the BC loop that results in the cavity that would be formed by this change being filled by a residue from the BC loop (Lesk & Chothia, 1982). (It is, of course, very likely that the differences between the V_κ and V_λ structures are the result of a series of mutations.) Though 86/90 is conserved in the variable domains of immunoglobulins it does change to Gly in CD4 domain 3 (Table 5). There its accommodation includes changes in the conformation of several residues that, in variable domains, are part of the common core.

This suggests that large sequence changes at 86/90 (and by implication at other sites in the deep structure) are not consistent with the retention of the common core structure in variable domains. If large changes do occur at this site, within the context of normal free energy of the protein and without a change in the conformation of part of the core stricture, the lack of accommodation results in the loss of stability.

Recent experimental evidence supports this conclusion, and also explains how large changes do occasionally take place at sites in the deep structure. In the natural antibody ABPC48 the Cys at 92 in V_H, which is part of the conserved disulphide, is replaced by Tyr (Lieberman *et al*., 1975). Proba *et al*. (1997) found that the V_LV_H dimer of this antibody is significantly less stable than that of the average V_LV_H dimer. They also showed that a mutant, in which the Cys has been restored, has a stability significantly greater than that of the average V_LV_H dimer. Thus, in this case, the natural mutation Cys to Tyr at site 92 is made tolerable by the fortuitous evolution of a variable domain whose structure has an exceptionally high stability. A similar conclusion can be drawn from the protein engineering experiments of Frisch *et al*. (1996).

Whole variable domains

Here we have been concerned so far with the features of the variable domain sequences that are responsible for their common core structure. The advantage of this approach is that it allows us to use the information from all the 5300 sequences currently known. But this approach does raise the question how would the results differ if we considered smaller sets of variable domains that conserved not the 70% of their structure found in all domains but all, or almost all, of their structure?

For the mouse and human variable domains of known structure, all but one of the regions outside the common core have one of a small repertoire of main-chain conformations. There is good evidence

that this is true for mouse and human immunoglobulins of unknown structure and for many of the regions in quite different species (Barré *et al.*, 1994). In the one region where there is considerable variation in conformation, the FG loop in V_H domains, most of it occurs in the part of the loop that protrudes into the solvent.

Inspection of variable domains that conserve all or almost all their conformation shows that residue conservation differs in two aspects from that found in the common core. First, the residues found at the ten buried sites adjacent to the deep structure are much more restricted. Some of these sites join the IR category; most are in the SR category or just outside it. For example, nearly all V_L domains have the same short C′C″ turn conformation that packs against IR sites 48 and 64 which conserve Ile and Gly, respectively. In V_H domains, on the other hand, the C′C″ turn is longer and the two homologous sites belong to the n and SR categories: Gly, Ala or Ser are usually found at 49 and Met or Ile at 69.

The second aspect involves the conservation at sites within the loops. Most loop sites conserve sn residues or, occasionally, Pro or Gly. However, several loops also bury residues in the regions adjacent to the deep structure; as was described above for the EF loop. Such residues are found at site 2 in V_κ, site 29 or 30 in V_L and V_H, site 52 in V_L and site 62 in V_H.

Thus, in sets of proteins that conserve the whole structure, rather than just a common core, the deep structure is more extensive with all, or nearly all, buried residues in the IR or SR groups. The changes at SR sites are usually accommodated by small shifts in adjacent regions (Gassner *et al.*, 1996; Chothia & Gerstein, 1997).

Conclusion

Here we have described in outline the roles of residues at different sites in determining the structure of the common core of variable domains. Of central importance are the sites that form the deep structure. Residues at these sites are largely invariant or limited to one of a group of closely related residues. The common core of variable domains comprises some 70% of individual variable domains and the residues in the deep structure are about a quarter of all residues. Proteins with larger common cores have a larger proportion of their interior residues forming their deep structures.

In future papers we will show that other families of proteins formed by large β-sheets packed face to face, β-sandwich structures, have sets of sites where conserved residues form a deep structure similar to that found in variable domains.

Acknowledgements

We thank Oleg Ptitsyn for discussions of the problems dealt with here and Jane Clarke for her comments on an earlier version of the paper. I.G. and A.K. thank Mrs M. Goldman for continuous encouragement. A.K. is supported by the Gabriela and Paul Rosenbaum Foundation.

References

Axe, D. D., Foster, N. W. & Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic cores. *Proc. Natl Acad. Sci. USA,* **93**, 5590–5594.

Barré, S., Greenberg, A. S., Flajnik, M. F. & Chothia, C. (1994). Structural conservation of hypervariable regions in the evolution of the immunoglobulins. *Nature Struct. Biol.* **1**, 915–920.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold: unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macro-molecular structures. *J. Mol. Biol.* **112**, 535–542.

Bowie, J. U. & Sauer, R. T. (1989). Identification determinants of folding and activity for a protein of unknown structure. *Proc. Natl Acad. Sci. USA,* **86**, 2152–2156.

Brady, R. L., Dodson, E. J., Dodson, G. G., Lange, G., Davis, S. J., Williams, A. F. & Barclay, A. N. (1993). Crystal structure of domains 3 and 4 of rat CD4: relation to the amino-terminal domains. *Science,* **260**, 979–983.

Buckle, A. M., Schreiber, G. & Fersht, A. R. (1994). Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0 Å resolution. *Biochemistry,* **33**, 8878–8889.

Buckle, A. M., Cramer, P. & Fersht, A. R. (1996). Structural and energetic responses to cavity creating mutations in hydrophobic cores. *Biochemistry,* **35**, 4298–4305.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–14.

Chothia, C. & Gerstein, M. (1997). How far can sequences diverge. *Nature,* **385**, 379–380.

Chothia, C. & Janin, J. (1981). Relative orientation of close-packed β-pleated sheets in proteins. *Proc. Natl Acad. Sci. USA,* **78**, 4146–4150.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–916.

Chothia, C., Novotny, J., Bruccoleri, R. & Karplus, M. (1985). Domain associations in immunoglobulin molecules. I. The packing of variable domains. *J. Mol. Biol.* **186**, 651–663.

Chothia, C., Boswell, D. R. & Lesk, A. M. (1988). The outline structure of the T cell αβ receptor. *EMBO J.* **7**, 3745–3755.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature,* **342**, 877–883.

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter,

G. (1992). Structural repertoire of the human V_H segments. *J. Mol. Biol.* **227**, 799–817.

Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science,* **267**, 383–386.

Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. (1981). Analysis of the tertiary structure of protein β-sheet sandwiches. *J. Mol. Biol.* **148**, 253–272.

Cook, G. P., Tomlinson, I. M., Walter, G., Riethman, H., Carter, N. P., Buluwela, L., Winter, G. & Rabbitts, (1994). A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome 14Q. *Nature Genet.* **7**, 162–168.

Cunningham, B. C. & Wells, J. A. (1993). Comparison of a structural and functional epitope. *J. Mol. Biol.* **234**, 554–563.

Diamond, R. D. (1992). On the multiple simultaneous superposition of molecular structures by rigid-body transformations. *Protein Sci.* **1**, 1279–1287.

Fauchere, J. L. & Pliska, V. (1983). Hydrophobic parameters, π, of amino acid side-chains from the partitioning of N-acetyl amino acid amides. *Eur. J. Med. Chem.* **18**, 369–375.

Frisch, C., Kolmar, H., Schmidt, A., Kleemann, G., Reinhardt, A., Pohl, I., Schneider, T. R. & Fritz, H.-J. (1996). Contribution of the intramolecular disulphide bridge to the folding stability of REI, the variable domain of a human immunoglobulin κ light chain. *Folding Design,* **1**, 431–440.

Garrett, T. P. J., Wang, J., Yan, Y., Liu, J. & Harrison, S. C. (1993). Refinement and analysis of the structure of the first two domains of human CD4. *J. Mol. Biol.* **234**, 763–778.

Gassner, N. C., Baase, W. A. & Matthews, B. W. (1996). A test of the jigsaw puzzle model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc. Natl Acad. Sci. USA,* **93**, 12155–12158.

Gelfand, I. M. & Kister, A. E. (1995). Analysis of the relation between the sequence and secondary and three dimensional structures of immunoglobulin molecules. *Proc. Natl Acad. Sci. USA,* **92**, 10884–10888.

Gelfand, I. M., Kister, A. E. & Leschiner, D. (1996). The invariant system of co-ordinates of antibody molecules: prediction of the standard Ca framework of V_L and V_H domains. *Proc. Natl Acad. Sci. USA,* **93**, 3675–3678.

Gelfand, I. M., Kister, A. E., Kulikovwski, C. & Stoyanov, O. (1998). Algorithmic determination of immunoglubulin core positions in the V_L and V_H domains. *J. Comput. Biol.* In the press.

Gerstein, M. & Altman, R. B. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175.

Guillet, V., Lapthorn, A., Hartley, R. W. & Mauguen, Y. (1993). Recognition between a bacterial ribonuclease, barnase, and its natural inhibitor, barstar. *Structure,* **1**, 165–177.

Hemmingsen, J. M., Gernert, K. M., Richardson, J. S. & Richardson, D. C. (1994). The tyrosine corner: a feature of most Greek key β-barrel proteins. *Protein Sci.* **3**, 1927–1937.

Hieter, P. A., Maizel, J. V. & Leder, P. (1982). Evolution of human immunoglobulin κ J region genes. *J. Biol. Chem.* **257**, 536–540.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–128.

Ignatovich, O., Tomlinson, I. M., Jones, P. T. & Winter, G. (1997). The creation of diversity in the human immunoglobulin V_γ repertoire. *J. Mol. Biol.* **268**, 69–77.

Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. & Stuart, D. I. (1992). Crystal structure at 2.8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature,* **360**, 232–239.

Kabat, E., Wu, T., Perry, H., Gottesman, K. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest.* National Institutes of Health Publ. No 91–3242, 5th edit., DHHS, PHS, Natl. Inst. of Health, Bethesda, MD.

Karpusas, M., Baase, W. A., Matsumura, M. & Matthews, B. W. (1989). Hydrophobic packing in T4 lysozyme probed by cavity filling mutants. *Proc. Natl Acad. Sci. USA,* **86**, 8237–8241.

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992). Crystal structure of a soluble form of the human T cell receptor CD8 at 2.6 Å. *Cell,* **68**, 1145–1162.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 223–268.

Lesk, A. M. & Chothia, C. (1982). The evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. *J. Mol. Biol.* **160**, 325–342.

Lesk, A. M. & Chothia, C. (1988). Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature,* **335**, 188–190.

Lieberman, R., Potter, M., Humphrey, W., Jr, Mushinski, E. B. & Vrana, M. (1975). Multiple individual and cross-specific idiotypes of 13 levan-binding myeloma proteins of BALB/c mice. *J. Exp. Med.* **142**, 106–119.

Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of λ-repressor. *Nature,* **339**, 31–36.

Matsuda, F., Shin, E. K., Nagaoka, H., Matsumura, R., Haino, M., Fukita, Y., Taka-ishi, S., Imai, T., Riley, J. H., Anand, R., Soeda, E. & Honjo, T. (1993). Structure and physical map of 64 variable segments in the 3′ 0. 8-megabase region of the human immunoglobulin heavy-chain locus. *Nature Genet.* **3**, 88–94.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656.

Proba, K., Honegger, A. & Plückthun, A. (1997). A natural antibody missing a cysteine in V_H: consequences for thermodynamic stability and folding. *J. Mol. Biol.* **265**, 161–172.

Ravetch, J. V., Siebenlist, U., Korsmeyer, S., Waldmann, T. & Leder, P. (1981). Structure of the human immunoglobulin μ locus: characterisation of embryonic and rearranged J and D genes. *Cell,* **27**, 583–591.

Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–87.

Ryu, S. E., Kwong, P. D., Truneh, A., Porter, T. G., Arthos, J., Rosenberg, M., Dai, X. P., Xuong, N. H., Axel, R., Sweet, R. W. & Hendrickson, W. A. (1990).

Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature,* **348**, 419–426.
Saul, F. A. & Poljak, R. J. (1993). Structural patterns at residue positions 9, 18, 67 and 82 in the V_H framework regions of human and murine immunoglobulins. *J. Mol. Biol.* **230**, 15–20.
Schäble, K. F. & Zachau, H. G. (1993). The variable genes of the human immunoglobulin κ locus. *Biol. Chem. Hoppe-Seyler,* **374**, 1001–1022.
Schreiber, G. & Fersht, A. R. (1993). Interaction of barnase with its polypeptide inhibitor barstat studied by protein engineering. *Biochemistry,* **32**, 5145–5150.
Schreiber, G. & Fersht, A. R. (1995). Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **248**, 478–486.
Shapiro, A., Botha, J. D., Pastore, A. & Lesk, A. M. (1992). A method for multiple superposition of structures. *Acta Crystallog, sect. A,* **48**, 11–14.
Shapiro, L., Doyle, J. P., Hensley, P., Colman, D. R. & Hendrickson, W. A. (1996). Crystal structure of the extracellular domain from P_0: the major structural protein of peripheral nerve myelin. *Neuron,* **17**, 435–449.
Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351–371.
Smith, D. K. & Xue, H. (1997). Sequence profiles of immunoglobulin and immunoglobulin-like domains. *J. Mol. Biol.* **274**, 530–545.
Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258.
Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.
Tomlinson, I. A., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). The repertoire of human germline V_H sequences reveals about 50 groups of V_H segments with different hypervariable loops. *J. Mol. Biol.* **227**, 776–798.
Tomlinson, I. A., Cox, J. P. L., Gherardi, E., Lesk, A. M. & Chothia, C. (1995). The structural repertoire of the human V_κ domain. *EMBO J.* **14**, 4628–4638.
Udey, J. A. & Blomberg, B. (1987). Human λ light chain locus: organisation and DNA sequences of three genomic J regions. *Immunogenet.* **25**, 63–70.
Vos, A. M. de. , Ultsch, M. & Kossiakoff, A. A. (1992). Human growth hormone and extracellular domain of its receptor: crystal structure of hr complex. *Science,* **255**, 306–312.
Wang, J. H., Yan, Y. W., Garrett, T. P., Liu, J. H., Rodgers, D. W., Garlick, R. L., Tarr, G. E., Husain, Y., Reinherz, E. L. & Harrison, S. C. (1990). Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature,* **348**, 411–418.
Williams, S. C., Frippiat, J.-P., Tomlinson, I. M., Ignatovitch, O., Lefranc, M.-P. & Winter, G. (1996). Sequence and evolution of the human germline V_λ repertoire. *J. Mol. Biol.* **264**, 220–232.
Yee, D. P. & Dill, K. A. (1993). Families and the structural relatedness among globular proteins. *Protein Sci.* **2**, 884–899.

Edited by A. R. Fersht

(Received 2 September 1997; received in revised form 20 January 1998; accepted 22 January 1998)