Reprinted from *J. Mol. Biol.* (1987) **196**, 901–917

Canonical Structures for the Hypervariable Regions of Immunoglobulins

Cyrus Chothia and Arthur M. Lesk

Canonical Structures for the Hypervariable Regions of Immunoglobulins

Cyrus Chothia[1,2] **and Arthur M. Lesk**[1,3]†

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

[2] *Christopher Ingold Laboratory*
University College London
20 Gordon Street
London WC1H 0AJ, England

[3] *EMBL Biocomputing Programme*
Meyerhofstr. 1, Postfach 1022.09
D-6900 Heidelberg
Federal Republic of Germany

(Received 13 November 1986, and in revised form 23 April 1987)

We have analysed the atomic structures of Fab and V_L fragments of immunoglobulins to determine the relationship between their amino acid sequences and the three-dimensional structures of their antigen binding sites. We identify the relatively few residues that, through their packing, hydrogen bonding or the ability to assume unusual ϕ, ψ or ω conformations, are primarily responsible for the main-chain conformations of the hypervariable regions. These residues are found to occur at sites within the hypervariable regions and in the conserved β-sheet framework.

Examination of the sequences of immunoglobulins of unknown structure shows that many have hypervariable regions that are similar in size to one of the known structures and contain identical residues at the sites responsible for the observed conformation. This implies that these hypervariable regions have conformations close to those in the known structures. For five of the hypervariable regions, the repertoire of conformations appears to be limited to a relatively small number of discrete structural classes. We call the commonly occurring main-chain conformations of the hypervariable regions "canonical structures".

The accuracy of the analysis is being tested and refined by the prediction of immunoglobulin structures prior to their experimental determination.

1. Introduction

The specificity of immunoglobulins is determined by the sequence and size of the hypervariable regions in the variable domains. These regions produce a surface complementary to that of the antigen. The subject of this paper is the relation between the amino acid sequences of antibodies and the structure of their binding sites. The results we report are related to two previous sets of observations.

The first set concerns the sequences of the hypervariable regions. Kabat and his colleagues (Kabat *et al.*, 1977; Kabat, 1978) compared the sequences of the hypervariable regions then known and found that, at 13 sites in the light chains and at seven positions in the heavy chains, the residues are conserved. They argued that the residues at these sites are involved in the structure, rather than the specificity, of the hypervariable regions. They suggested that these residues have a fixed position in antibodies and that this could be used in the model building of combining sites to limit the conformations and positions of the sites whose residues varied. Padlan (1979) also examined the sequences of the hypervariable region of light

† Also associated with Fairleigh Dickinson University, Teaneck-Hackensack Campus, Teaneck, NJ 07666, U.S.A.

0022-2836/87/160901-17 $03.00/0

chains. He found that residues that are part of the hypervariable regions, and that are buried within the domains in the known structures, are conserved. The residues he found conserved in V_λ sequences were different to those conserved in V_κ sequences.

The second set of observations concerns the conformation of the hypervariable regions. The results of the structure analysis of Fab and Bence-Jones proteins (Saul *et al.*, 1978; Segal *et al.*, 1974; Marquart *et al.*, 1980; Suh *et al.*, 1986; Schiffer *et al.*, 1973; Epp *et al.*, 1975; Fehlhammer *et al.*, 1975; Colman *et al.*, 1977; Furey *et al.*, 1983) show that in several cases hypervariable regions of the same size, but with different sequences, have the same main-chain conformation (Padlan & Davies, 1975; Fehlhammer *et al.*, 1975; Padlan *et al.*, 1977; Padlan, 1977*b*; Colman *et al.*, 1977; de la Paz *et al.*, 1986). Details of these observations are given below.

In this paper, from an analysis of the immunoglobulins of known atomic structure we determine the limits of the β-sheet framework common to the known structures (see section 3 below). We then identify the relatively few residues that, through packing, hydrogen bonding or the ability to assume unusual ϕ, ψ or ω conformations, are primarily responsible for the main-chain conformations observed in the hypervariable regions (see sections 4 to 9, below). These residues are found to occur at sites within the hypervariable regions and in the conserved β-sheet framework. Some correspond to residues identified by Kabat *et al.* (1977) and by Padlan (Padlan *et al.*, 1977; Padlan, 1979) as being important for determining the conformation of hypervariable regions.

Examination of the sequences of immunoglobulins of unknown structure shows that in many cases the set of residues responsible for one of the observed hypervariable conformations is present. This suggests that most of the hypervariable regions in immunoglobulins have one of a small discrete set of main-chain conformations that we call "canonical structures". Sequence variations at the sites not responsible for the conformation of a particular canonical structure will modulate the surface that it presents to an antigen.

Prior to this analysis, attempts to model the combining sites of antibodies of unknown structure have been based on the assumption that hypervariable regions of the same size have similar backbone structures (see section 12, below). As we show below, and as has been realized in part before, this is true only in certain instances. Modelling based on the sets of residues identified here as responsible for the observed conformations of hypervariable regions would be expected to give more accurate results.

2. Immunoglobulin Sequences and Structures

Kabat *et al.* (1983) have published a collection of the known immunoglobulin sequences. For the variable domain of the light chain (V_L)† they list some 200 complete and 400 partial sequences; for the variable domain of the heavy chain (V_H) they list about 130 complete and 200 partial sequences. In this paper we use the residue numbering of Kabat *et al.* (1983), except in the few instances where the structural superposition of certain hypervariable regions gives an alignment different from that suggested by the sequence comparisons.

In Table 1 we list the immunoglobulins of known structure for which atomic co-ordinates are available from the Protein Data Bank (Bernstein *et al.*, 1977), and give the references to the crystallographic analyses. Amzel & Poljak (1979), Marquart & Deisenhofer (1982) and Davies & Metzger (1983) have written reviews of the molecular structure of immunoglobulins.

The V_L and V_H domains have homologous structures (for references, see Table 1). Each contains two large β-pleated sheets that pack face to face with their main chains about 10 Å apart (1 Å = 0·1 nm) and inclined at an angle of −30° (Fig. 1). The β-sheets of each domain are linked by a conserved disulphide bridge. The antibody binding site is formed by the six hypervariable regions; three in V_L and three in V_H. These regions link strands of the β-sheets. Two link strands that are in different β-sheets. The other four are hair-pin turns: peptides that link two adjacent strands in the same β-sheet (Fig. 2). Sibanda & Thornton (1985) and Efimov (1986) have described how the conformations of small and medium-sized hair-pin turns depend primarily on the length and sequence of the turn. Thornton *et al.* (1985) pointed out that the sequence-conformation rules for hair-pin turns can be used for modelling antibody combining sites. The results of these authors and our own unpublished work on the conformations of hair-pin turns, are summarized in Table 2.

3. The Conserved β-Sheet Framework

Comparisons of the first immunoglobulin structures determined showed that the framework regions of different molecules are very similar

Table 1

Immunoglobulin variable domains of known atomic structure

Protein	Chain L	Type H	Reference
Fab'NEWM	λI	II	Saul *et al.* (1978)
Fab MCPC603	κ	I	Segal *et al.* (1974)
Fab KOL	λI	III	Marquart *et al.* (1980)
Fab J539	κ	III	Suh *et al.* (1986)
V_L REI	κ		Epp *et al.* (1975)
V_L RHE	λI		Furey *et al.* (1983)

† Abbreviations used: V_L and V_H, variable regions of the immunoglobulin light and heavy chains, respectively; r.m.s., root-mean-square; CDR, complementarity-determining region.

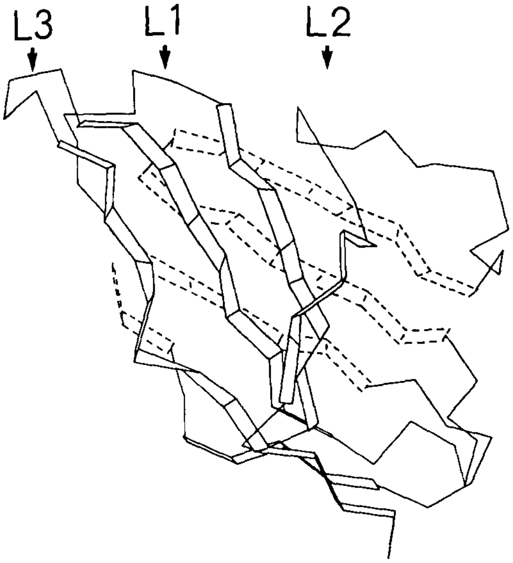


Figure 1. The structure of an immunoglobulin V domain. The drawing is of KOL V_L. Strands of β-sheet are represented by ribbons. The three hypervariable regions are labelled L1, L2 and L3. L2 and L3 are hairpin loops that link adjacent β-sheet strands. L1 links two strands that are part of different β-sheets. The V_H domains and their hypervariable regions, H1, H2 and H3, have homologous structures. The domain is viewed from the β-sheet that forms the V_L–V_H interface. The arrangement of the 6 hypervariable regions that form the antibody binding site is shown in Figure 2.

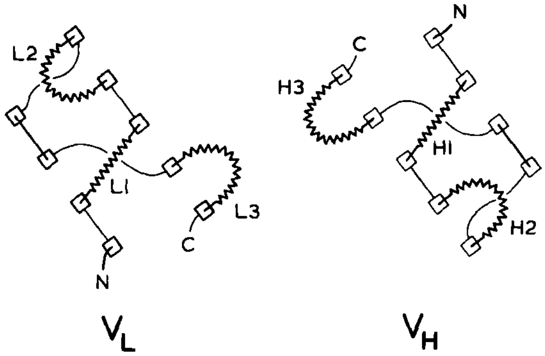


Figure 2. A drawing of the arrangement of the hypervariable regions in immunoglobulin binding sites. The squares indicate the position of residues at the ends of the β-sheet strands in the framework regions.

(Padlan & Davies, 1975). The structural similarities of the frameworks of the variable domains were seen as arising from the tendency of residues that form the interiors of the domains to be conserved, and from the conservation of the total volume of the interior residues (Padlan, 1977*a*, 1979). In addition, the residues that form the central region of the interface between V_L and V_H domains were observed to be strongly conserved (Poljak *et al.*, 1975; Padlan, 1977*b*) and to pack with very similar geometries (Chothia *et al.*, 1985).

In this section we define and describe the exact extent of the structurally similar framework regions in the known Fab and V_L structures. This was determined by optimally superposing the main-chain atoms of the known structures (Table 1) and calculating the differences in position of atoms in homologous residues[†].

In Figure 3(a) we give a plan of the β-sheet framework that, on the basis of the superpositions, is common to all six V_L structures. It contains 69 residues. The r.m.s. difference in the position of the main-chain atoms of these residues is small for all pairs of V_L domains; the values vary between 0·50 and 1·61 Å (Table 3A). The four V_H domains share a common β-sheet framework of 79 residues (Fig. 3(b)). For different pairs of V_H domains the r.m.s. difference in the position of the main-chain atoms is between 0·64 and 1·42 Å.

The combined β-sheet framework consists of V_L residues 4 to 6, 9 to 13, 19 to 25, 33 to 49, 53 to 55, 61 to 76, 84 to 90, 97 to 107 and V_H residues 3 to 12, 17 to 25, 33 to 52, 56 to 60, 68 to 82, 88 to 95 and 102 to 112. A fit of the main-chain atoms of these 156 residues in the four known Fab structures gives r.m.s. differences in atomic positions of main-chain atoms of:

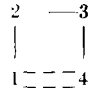
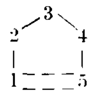
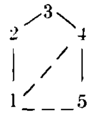
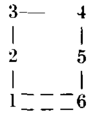
	NEWM	McPC603	J539
KOL	1·39 Å	1·15 Å	1·14 Å
NEWM	—	1·47 Å	1·37 Å
McPC603	—	—	1·03 Å

The major determinants of the tertiary structure of the framework are the residues buried within and between the domains. We calculated the accessible surface area (Lee & Richards, 1971) of each residue in the Fab and V_L structures. In Table 4 we list the residues commonly buried within the V_L and V_H domains and in the interface between them. These are essentially the same as those identified by Padlan (1977*a*) as buried within the then known structures and conserved in the then known sequences. Examination of the 200 to 700 V_L sequences and 130 to 300 V_H sequences in the Tables of Kabat *et al.* (1983) shows that in nearly all the sequences listed there the residues at these positions are identical with, or very similar to, those in the known structures.

There are two positions in the V_L sequences at which the nature of the conserved residues depends on the chain class. In V_λ sequences, the residues at positions 71 and 90 are usually Ala and Ser/Ala, respectively; in V_κ sequences the corresponding residues are usually Tyr/Phe and Gln/Asn. These residues make contact with the hypervariable loops and play a role in determining the conformation of

[†] For these and other calculations we used a program system written by one of us (see Lesk, 1986).

Table 2
Conformation of hair-pin turns

Structure	Sequence[a]			Conformation[b] (°)						Frequency[c]
	1 2 3 4			ϕ2,	ψ2	ϕ2,	ψ3			
	X- G- G- X			+55	+35	+85	−5[d]			
					or					6/6
				+65	−125	−105	+10[c]			
	X- G- X- X			+70	−115	−90	0[e]			6/7
	X- X- G- X			+50	+45	+85	−20[d]			7/8
	X- X- X- X			+60	+20	+85	+25[f]			4/4
		ϕ1	ψ1	ϕ2	ψ2	ϕ3	ψ3	ϕ4	ψ4	
	X- X- X- G	−135	+175	−50	−35	−95	−10	+145	+155	4/4
	1 2 3 4 5[g]	ϕ2	ψ2	ϕ3	ψ3	ϕ4	ψ4	ϕ5	ψ5	
	X X X X G	−75	−10	−95	−50	−105	0	+85	−160	3/3
	X X X X X	+50	+55	+65	−50	−130	−5	−90	+130	1/1(3/3)
	1 2 3 4 5[h]	ϕ2	ψ2	ϕ3	ψ3	ϕ4	ψ4			
	G									
	X- X- X- N- X	−60	−25	−90	0	+85	+10			13/15
	D									
	1 2 3 4 5 6[i]	ϕ2	ψ2	ϕ3	ψ3	ϕ4	ψ4	ϕ5	ψ5	
	G									3/3
	X- X- X- X- N- X	−65	−30	−65	−45	−95	−5	+70	+35	2/2
	X									1/1

The data in this Table are from an unpublished analysis of proteins whose atomic structure has been determined at a resolution of 2 Å or higher. The conformations described here for the 2-residue X-X-X-G turn and the 3-residue turns are new. The other conformations have been described by Sibanda & Thornton (1985) and by Efimov (1986). We list only conformations found more than once.

[a] X indicates no residue restriction except that certain sites cannot have Pro, as this residue requires a ϕ value of ~ −60° and cannot form a hydrogen bond to its main-chain nitrogen.

[b] Residues whose ϕ,ψ values are not given have a β conformation.

[c] Frequencies are given as n_1/n_2, where n_2 is the number of cases where we found the structure in column 1 with the sequence in column 2 and n_1 the number of these cases that have the conformation in column 3. Except for the frequencies in brackets, data is given only for non-homologous proteins.

[d,e,f] These are type I', II' and III' turns.

[g] Different conformations are found for the single cases of X-D-G-X-X and X-G-X-G-X.

[h] Different conformations are found for the single cases of X-X-X-X-X, X-G-G-X-X and X-G-X-X-G. The 2 cases of X-X-X-X-X- have different conformations.

[i] Different conformations are found for the 2 cases of X-G-X-X-X-X.

these loops. This is discussed in sections 5 and 7, below.

The conservation of the framework structure extends to the residues immediately adjacent to the hypervariable regions. If the conserved frameworks of a pair of molecules are superposed, the differences in the positions of these residues is in most cases less than 1 Å and in all but one case less than 1·8 Å (Table 5). In contrast, residues in the hypervariable region adjacent to the conserved framework can differ in position by 3 Å or more.

The six loops, whose main-chain conformations vary and which are part of the antibody combining site, are formed by residues 26 to 32, 50 to 52 and 91 to 96 in V_L domains, and 26 to 32, 53 to 55 and 96 to 101 in the V_H domains L1, L2, L3, H1, H2 and H3, respectively. Their limits are somewhat different from those of the complementarity-determining regions defined by Kabat *et al.* (1983) on the basis of sequence variability: residues 24 to 34, 50 to 56 and 89 to 97 in V_L and 31 to 35, 50 to 65 and 95 to 102 in V_H. This point is discussed in section 11, below.

4. Conformation of the L1 Hypervariable Regions

In the known V_L structures, the conformations of the L1 regions, residues 26 to 32, are characteristic of the class of the light chain. In V_λ domains their conformation is helical and in the V_κ domains it is extended (Padlan *et al.*, 1977; Padlan, 1977*b*; de la Paz *et al.*, 1986). These conformational differences are the result of sequence differences in both the L1 region and the framework (Lesk & Chothia, 1982).

(a) V_λ *domains*

Figure 4 shows the conformation of the L1 regions of the V_λ domains. The L1 regions in RHE

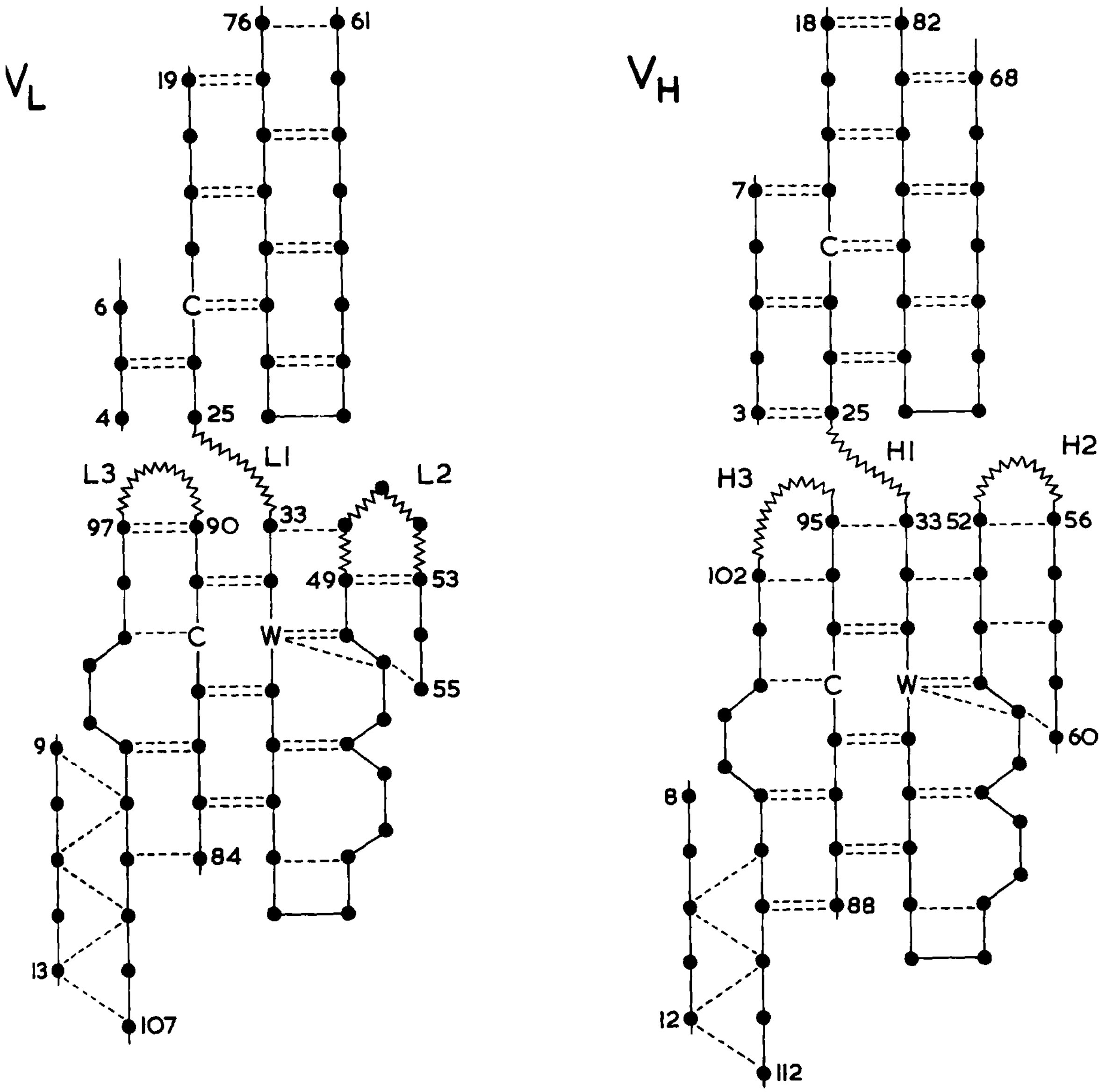


Figure 3. Plane of the β-sheet framework that is conserved in the V_L and V_H domains of the immunoglobulins of known atomic structure.

and KOL contain nine residues designated 26 to 30, 30a, 30b, 31 to 32; NEWM has one additional residue. The L1 regions in RHE and KOL have the same conformation: their main-chain atoms have a r.m.s. difference in position of 0·28 Å. Superposition of the L1 region of NEWM with those of KOL and RHE shows that the additional residue is inserted between residues 30b and 31 and has little effect on the conformation of the rest of the region: superpositions of the main-chain atoms of 26 to 30b and 31 to 32 in NEWM to 26 to 32 in KOL and RHE give r.m.s. differences in position of 0·96 Å and 1·25 Å. Thus, the sequence alignment for the V_λ L1 regions of KOL, RHE and NEWM implied by the structural superposition is:

Position	26	27	28	29	30	30a	30b	30c	31	32
RHE	Ser	Ala	Thr	Asp	Ile	Gly	Ser		Asn	Ser
KOL	Thr	Ser	Ser	Asn	Ile	Gly	Ser		Ile	Thr
NEWM	Ser	Ser	Ser	Asn	Ile	Gly	Ala	Gly	Asn	His

In all three structures, residues 26 to 29 form a type I turn with a hydrogen bond between the carbonyl of 26 and the amide of 29. Residues 27 to 30b form an irregular helix (Fig. 4). This helix sits across the top of the β-sheet core. The side-chain of residue 30 penetrates deep into the core occupying a cavity between residues 25, 33 and 71. The major determinant of the conformation of L1 in the observed structures is the packing of residues 25, 30, 33 and 71. V_λ RHE, KOL and NEWM have the

Table 3

Differences in immunoglobulin framework structures (Å)

For pairs of V domains we give the r.m.s. difference in the atomic positions of framework main chain atoms after optimal superposition.

A. *V_L domains*

Framework residues are 4 to 6, 9 to 13, 19 to 25, 33 to 49, 53 to 55, 61 to 76, 84 to 90 and 97 to 107.

	KOL	NEWM	REI	MCPC603	J539
RHE	0·74	1·47	1·46	1·61	1·41
KOL		1·13	1·23	1·36	1·15
NEWM			1·24	1·28	1·53
REI				0·50	0·77
MCPC603					0·76

B. *V_H domains*

Framework residues are 3 to 12, 17 to 25, 33 to 52, 56 to 60, 68 to 82, 88 to 95 and 102 to 112.

	NEWM	MCPC603	J539
KOL	1·42	0·64	0·89
NEWM	—	1·27	1·29
MCPC603	—		0·89

same residues at these sites: Gly25, Ile30, Val33 and Ala71. (Another L1 residue, Asp29 or Asn29, is buried by the contacts it makes with L3.)

Kabat *et al.* (1983) listed 33 human V_λ domains for which the sequences of the L1 regions are known. The 21 sequences in subgroups I, II, V and VI have L1 regions that are the same length as those found in RHE, KOL or NEWM. Of these, 18 conserve the residues responsible for the observed conformations:

Residue position	Residue in KOL/RHE/NEWM	Residues in 18 V_λ sequences
25	Gly	18 Gly
30	Ile	17 Val, 1 Ile
33	Val	17 Val, 1 Ile
71	Ala	18 Ala
29	Asp/Asn	11 Asp, 6 Asn, 1 Ser

The conservation of these residues implies that these 18 L1 regions have a conformation that is the same as that in RHE, KOL or NEWM.

Subgroups III and IV have 13 sequences for which the L1 regions are known (Kabat *et al.*, 1983). These regions are shorter than those in RHE and KOL and in the other V_λ subgroups. They also have a quite different pattern of conserved residues.

Kabat *et al.* (1983) listed 29 mouse V_λ domains for which the sequence of the L1 region is known. These L1 regions are the same size as that in NEWM. They also have a pattern of residue conservation similar to, but not identical with, that in KOL/NEWM: Ser at position 25, Val at 30, Ala at 33 and Ala at 71. This suggests that the fold of

Table 4

Residues commonly buried within V_L and V_H domains

V_L domains			V_H domains		
Position	Residues in known structures	A.S.A.[a] (Å^2)	Position	Residues in known structures	A.S.A.[a] (Å^2)
4	L,M	6	4	L	14
6	Q	12	6	Q,E	16
19	V	11	18	L	21
21	I,M	1	20	L	0
23	C	0	22	C	0
25	G,A,S	13	24	S,V,T,A	8
33	V,L	3	34	M,Y	4
35	W	0	36	W	0
37	Q	30	38	R	13
47	L,I,W	8	48	I,V	1
48	I	24	49	A,G	0
62	F	11	51	I,V,S	4
64	G,A	13	69	I,V,M	13
71	A,F,Y	2	78	L,F	0
73	L,F	0	80	L	0
75	I,V	0	82	M,L	0
82	D	4	86	D	2
84	A,S	11	88	A,G	3
86	Y	0	90	Y	0
88	C	0	92	C	0
90	A,S,Q,N	7	104	G	11
97	V,T,G	18	106	G	19
99	G	3	107	T,S	17
101	G	11	109	V	2
102	T	1			
104	L,V	2			

[a] Mean accessible surface area (A.S.A.) of the residues in the Fab structures NEWM, MCPC603, KOL and J539 and in the V_L structures REI and RHE.

the mouse V_λ L1 regions is a distorted version of that found in the known human structures.

(b) V_κ *domains*

In Figure 5 we illustrate the conformation of the L1 regions in the three known V_κ structures: J539, REI and MCPC603. In J539 L1 has six residues, in REI it has seven and in MCPC603 13. The L1 region of J539 has an extended conformation. In REI, residues 26 to 28 have an extended conformation and 29 to 32 form a distorted type II turn. The six additional residues in MCPC603 all occur in the region of this turn (Fig. 5). In the three structures the main chain of residues 26 to 29 and 32 have the same conformation. A fit of the main-chain atoms of these residues in J539, REI and MCPC603 gives r.m.s. differences in position of 0·47 to 1·03 Å. The sequence alignment implied by the structural superposition is:

Residue	26	27	28	29	30	31	31a	31b	31c	31d	31e	31f	32
J539	Ser	Ser	Ser	Val	Ser		—	—	-	—	--	—	Ser
REI	Ser	Glu	Asp	Ile	Ile	Lys				—	—	—	Tyr
MCPC603	Ser	Glu	Ser	Leu	Leu	Asn	Ser	Gly	Asn	Glu	Lys	Asn	Phe

In J539, REI and MCPC603, residues 26 to 29 extend across the top of β-sheet framework with one, 29, buried within it. The main contacts of 29 are with residues 2, 25, 33 and 71. The penetration of residue 29 into the interior of the framework is not as great as that of residue 30 in the V_λ domains, and the deep cavity that exists in V_λ domains is filled in V_κ domains by the large side-chain of the residue at position 71. In J539, REI and MCPC603, the residues involved in the packing of L1 (2, 25, 29, 33 and 71) are very similar: Ile, Ala/Ser, Val/Ile/Leu, Leu and Tyr/Phe, respectively.

The six residues 30 to 30f in MCPC603 form a hair-pin loop that extends away from the domain (Fig. 5) and does not have a well-ordered conformation (Segal *et al.*, 1974).

Kabat *et al.* (1977) noted that residues at certain positions in the L1 regions of the V_κ sequences then known were conserved, and suggested that they have a structural role. The structural role of residues at positions 25, 29 and 33 is confirmed by the above analysis of the V_κ structures and the pattern of residue conservation in the much larger number of sequences known now. Kabat *et al.* (1983) listed 65 human and 164 mouse V_κ sequences for which the residues between positions 2 and 33 are known. For about half of these, the residue at position 71 is also known. These data show that there are 59 human and 148 mouse sequences that have residues very similar to those in the known structures at the sites involved in the packing of L1:

Position	J539/REI/MCPC603	Human V_κ	Mouse V_κ
2	Ile	57 Ile, 1 Met, 1 Val	134 Ile, 14 Val
25	Ala Ser	52 Ala, 7 Ser	104 Ala, 4 Ser
29	Val Ile Leu	30 Ile, 21 Val, 8 Leu	59 Leu, 51 Val, 38 Ile
33	Leu	57 Leu, 2 Val	94 Leu, 44 Met, 7 Val, 3 Ile
71	Tyr Phe	28 Phe, 1 Tyr	54 Phe, 26 Tyr

The number of residues in the L1 region in these sequences varies:

Residue size of L1	6	7	8	9	10	11	12	13
Number of human V_κ		38	14	—	–	1	4	2
Number of mouse V_κ	17	40	—	—		32	35	30

The conservation of residues at the positions buried between L1 and the framework implies that in the large majority of V_κ domains residues 26 to 29 have a conformation close to that found in the known structures and that the remaining residues, if small in number, form a turn or, if large, a hair-pin loop.

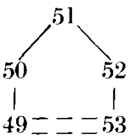
5. Conformation of the L2 Hypervariable Regions

The L2 regions have the same conformation in the known structures (Padlan *et al.*, 1977; Padlan, 1977*b*; de la Paz *et al.*, 1986) expect for NEWM, where it is deleted. We find that the similarities in the L2 structures arise from the conformational requirements of a three-residue turn and the conservation of the framework residues against which L2 packs.

The know structures L2 consists of three residues, 50 to 52:

Residue	RHE	KOL	REI	MCPC603	J539
50	Tyr	Arg	Glu	Gly	Glu
51	Asn	Asp	Ala	Ala	Ile
52	Asp	Ala	Ser	Ser	Ser

These three residues link two adjacent strands in the framework β-sheet. Residues 49 and 53 are hydrogen bonded to each other so that the L2 region is a three-residue hair-pin turn (Fig. 6).



The conformations of L2 in the five structures are very similar: r.m.s. differences in position of their main-chain atoms are between 0·1 and 0·97 Å. The only difference among the conformations is in the orientation of the peptide between residues 50 and 51. In MCPC603 this difference is associated with the Gly residue at position 50. The side-chains of L2 all point towards the surface. The main-chain packs

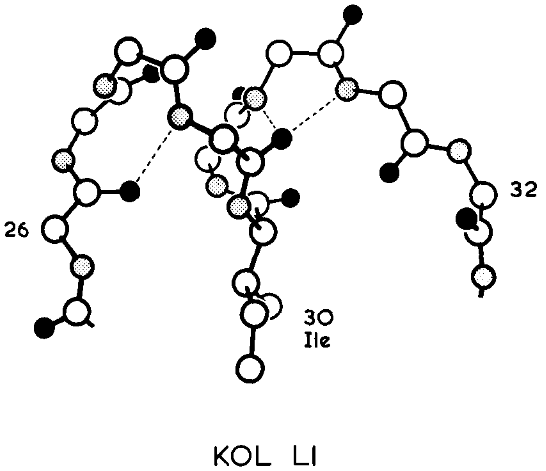


Figure 4. The conformation of the L1 region of V_λ KOL. The side-chain of Ile30 is buried within the framework structure; see section 4.

against the conserved framework residues Ile47 and Gly64/Ala64 (Fig. 6).

Kabat *et al.* (1983) give the sequences of the L2 regions of 174 V_L domains. In all cases they are three residues in length. Of the 174, 122 do not contain Gly and 49 have, like MCPC603, a Gly residue at position 50. The residues at position 48 and 64 are almost absolutely conserved as Ile and Gly. These size and sequence identities imply that almost all L2 regions have a conformation close to that found in the known structures.

Table 5

Differences in the positions of the framework residues adjacent to the hypervariable regions in immunoglobulin structures

Hypervariable region	Adjacent framework residues		Differences in position (Å)	
L1	25	33	0·2–1·1	0·5–0·8
L2	49	53	0·3–0·5	0·5–1·4
L3	90	97	0·8–1·0	0·8–1·2
H1	25	33	0·5–1·2	0·3–1·2
H2	52	56	0·8–2·1	1·2–1·7
H3	95	102	0·5–1·2	0·4–1·7

6. Conformation of the L3 Hypervariable Regions

The L3 region, residues 91 to 96, forms the link between two adjacent strands of β-sheet. Our analysis of the structures and sequences known for this region suggests that the large majority of κ chains have a common conformation that is quite different from the conformations found in λ chains.

(a) V_λ domains

The L3 region of V_λ NEWM has six residues and those of KOL and RHE have eight. Superposition of the three regions gives the following alignment:

	91	92	93	93a	93b	94	95	96
NEWM	Tyr	Asp	Arg	—	—	Ser	Leu	Arg
KOL	Trp	Asn	Ser	Ser	Asp	Asn	Ser	Tyr
RHE	Trp	Asn	Asp	Ser	Leu	Asp	Glu	Pro

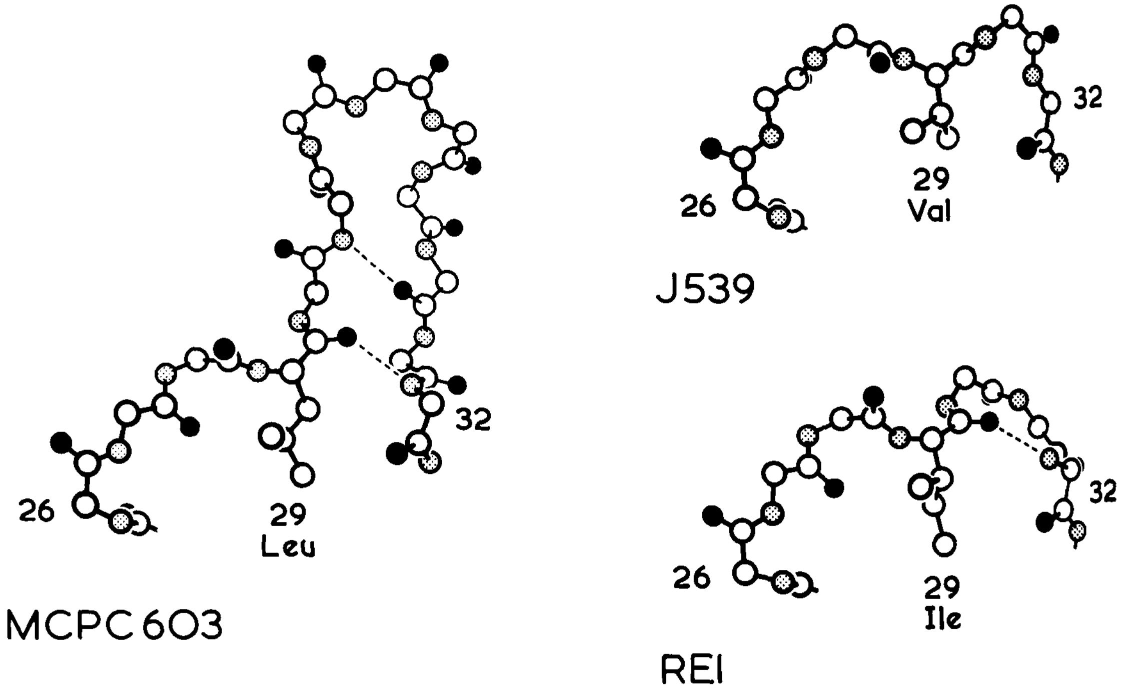


Figure 5. The conformation of the L1 regions of V_κ MCPC603, V_κ REI and V_κ J539. Residues 26 to 29 and 32 have the same conformation in the 3 structures. The side-chain of residue 29 is buried within the framework structure; see section 4.

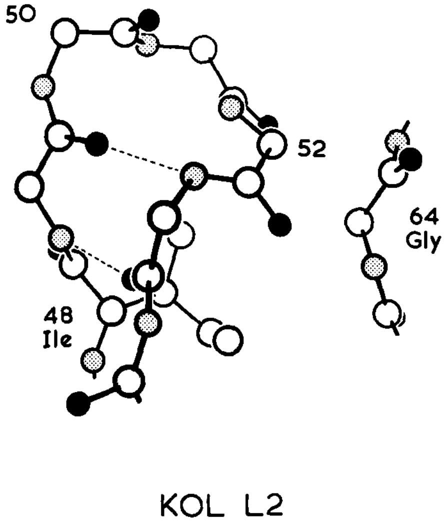


Figure 6. The conformation of the L2 region of V_λ KOL. This region packs against framework residues Ile47 and Gly64.

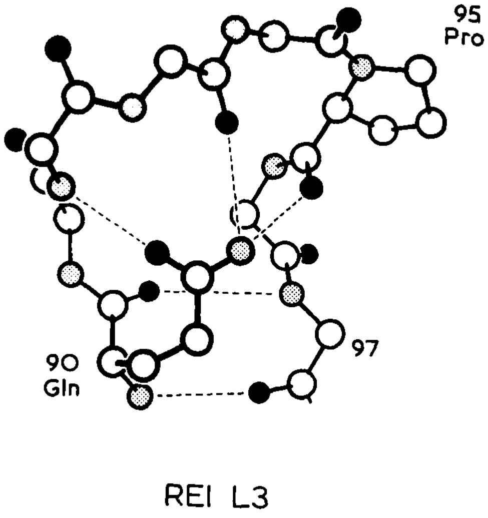


Figure 7. The conformation of the L3 region of V_κ REI. The conformation is stabilized by the hydrogen bonds made by the framework residue Gln90 and by the *cis* conformation of the peptide of Pro95.

In all three V_λ structures, residues 91 to 92 and 95 to 96 form an extension of the β-sheet framework with main-chain hydrogen bonds between residues 92 and 95:

```
                      93a——93b
                       |     |
93_____94             93    94
 |      |              |     |
92= = =95             92= = =95
 |      |              |     |
91     96             91    96
 |      |              |     |
90= = =97             90= = =97
```

Residues 93 and 94 in NEWM form a two-residue type II′ turn (see Table 2). Residues 93, 93a, 93b and 94 in RHE and KOL form a four-residue turn with the same conformation: the r.m.s. difference in the position of their main-chain atoms is 0·19 Å. This conformation is found in almost all four residue turns that, like KOL and RHE, have Gly or Asn in the fourth position of the turn, position 94 here (Sibanda & Thornton, 1985; Efimov, 1986; and see Table 2).

Kabat *et al.* (1983) listed 27 human and 25 mouse V_λ domains for which the sequence of the whole of the third hypervariable region is known. The distribution of sizes of the L3 region in these sequences is:

Residue size	5	6	7	8
Number of human V_λ	1	7	12	7
Number of mouse V_λ	—	25		—

In the L3 regions with six residues we would expect, as in NEWM, 91 to 92 and 95 to 96 to continue the β-sheet of the framework and 93 to 94 to form a two-residue hair-pin turn. Rules relating the sequence and conformation of two-residue turns (Sibanda & Thornton, 1985; Efimov, 1986) are given in Table 2. Similarly, in L3 regions with eight residues we would expect 91 to 92 and 95 to 96 to continue the β-sheet framework and 93, 93a, 93b and 94 to form a four-residue turn.

(b) V_κ *domains*

The L3 regions in REI, MCPC603 and J539 are the same size:

	91	92	93	94	95	96
REI	Tyr	Gln	Ser	Leu	Pro	Tyr
MCPC603	Asp	His	Ser	Tyr	Pro	Leu
J539	Trp	Thr	Tyr	Pro	Leu	Ile

In REI and MCPC603, the L3 regions have the same conformation: the r.m.s. difference in the positions of the main-chain atoms of residues 91 to 96 is 0·43 Å. L3 in J539 has a conformation different from that in REI and MCPC603.

Normally, for six-residue loops, we might expect the main-chain atoms of residues 92 and 95 to form hydrogen bonds, and residues 93 and 94 to form a turn (see the discussion of L3 in the V_λ chains, section 6(a), above). This conformation is prevented in the two V_κ structures REI and MCPC603 by a Pro residue at position 95. In these two V_κ structures, residue 92 has an α_L conformation and Pro95 has a *cis* peptide. This puts residues 93 to 96 in an extended conformation (Fig. 7). Important determinants of this particular L3 conformation are the hydrogen bonds formed to its main-chain atoms by the side-chain of framework residue 90. Though the side-chains at position 90 are not identical (REI has Gln and MCPC603 has Asn), the amides are in the same position and play the same role: the NH group forms hydrogen bonds to the carbonyls of 93 and 95 and the O atom forms a hydrogen bond to the amide of 92 (Fig. 7).

Although L3 in J539 is six residues in length, it has Leu, not Pro, at position 95 and forms a two-residue hair-pin turn:

```
Tyr93______Pro94
  |          |
Thr92 = = = Leu95
  |          |
Trp91       Ile96
  |          |
Gln90 = = = Thr97
```

Because of the Pro residue at position 94, this turn has a conformation different from those in V_λ chains and those commonly found (see Table 2): Tyr93 has ϕ,ψ values $-51°$, $+131°$; Pro94 has a *cis*-peptide and ϕ,ψ values of $-46°$, $-54°$.

Kabat *et al.* (1977) found that residues at positions 90 and 95 in L3 regions of V_κ sequences are conserved and suggested that they have a structural role. Kabat *et al.* (1983) listed 121 human and mouse V_κ domains for which the sequence of the whole of the third hypervariable region is known. The size distribution of the L3 regions is:

Residue size	5	6	7	12
Number of human V_κ	1	36	—	—
Number of mouse V_κ	1	81	2	1

Of the 117 L3 regions that contain six residues, 93 have Pro at position 95 and Gln or Asn at position 90. Their size and sequence identities imply that these 93 have an L3 conformation that is the same as that found in REI and MCPC603. A further 16 of the 117 have Pro at position 94 but not at 95 and are likely to have the L3 conformation found in J539.

7. Conformation of the H1 Hypervariable Region

The H1 regions are the same size in four known structures:

	26	27	28	29	30	31	32
KOL	Gly	Phe	Ile	Phe	Ser	Ser	Tyr
MCPC603	Gly	Phe	Thr	Phe	Ser	Asp	Phe
J539	Gly	Phe	Asp	Phe	Ser	Lys	Tyr
NEWM	Gly	Thr	Ser	Phe	Asp	Asp	Tyr

They pack across the top of the V domain (Fig. 2). Padlan (1977*b*) noted that the folds of H1 in NEWM and MCPC603 are very similar. They are also similar to those found in KOL and J539 (Fig. 8). For these four structures the r.m.s. differences in the position of the main-chain atoms 26 to 32 are between 0·4 and 1·4 Å. Small differences occur in residues 30 to 32 (Fig. 8), which [illegible] H[illegible] (Fig. 2). They appear to be the result of changes in conformation and residue identity in H2.

In the observed H1 structures the Gly at position 26 produces a sharp turn through a ϕ,ψ value $(+75,0)$ outside the range allowed for non-glycine residues. The Phe at position 29 is deeply buried within the framework structure, packing against the side-chain of residue 34 and the main chain of residues 72 and 77. The residues at position 27, Phe or Thr, are partially buried in a surface cavity next to residue 94. In the four structures the residues at positions 26, 34 and 94 are identical or similar; Gly, Phe, Met/Tyr and Arg.

Kabat *et al.* (1983) listed 185 human and mouse V_H domains for which the sequence of the first hypervariable region is known. Of 178, 170 are the same length as those found in the known structures, one mouse sequence is one residue longer, and six human sequences are two residues longer.

Of the 170 with seven residues, there are 115 for which the residue at position 94 is also known. Of these, three-quarters have residues at positions 26, 27, 29, 34 and 94 that are the same as or very close to those found in the known structures:

Residues					Number of sequences
26	27	29	34	94	
Gly	Phe	Phe	Met	Arg	50
Gly	Tyr	Phe	Met	Arg	29
Gly	Phe	Phe	Met	Lys	4
Gly	Tyr/Phe	Phe	Ile	Arg/Lys	5

The conservation of the length of the loop and of the residues at the sites involved in the packing of H1 against the framework implies that in at least these V_H domains the conformation of H1 is close to that found in the known structures.

8. Conformation of the H2 Hypervariable Region

The H2 region forms the link between the framework residues 52 and 56, which are in adjacent strands of β-sheet. The H2 loops differ in length in the known V_H structures: in NEWM it contains three residues, in KOL and J539 four, and in MCPC603 six. Kabat *et al.* (1983) list 127 human and mouse V_H sequences for which the sequence of the whole H2 region is known. In all but one, H2 has a length that is the same as one of the known structures:

Residue size of H2	3	4	5	6
Number of sequences	13	71	1	42

The 42 H2 regions with six residues are all mouse sequences in subgroup III.

The three residues in the H2 region of NEWM (Tyr53, His54, Gly55) form the apex of a seven-

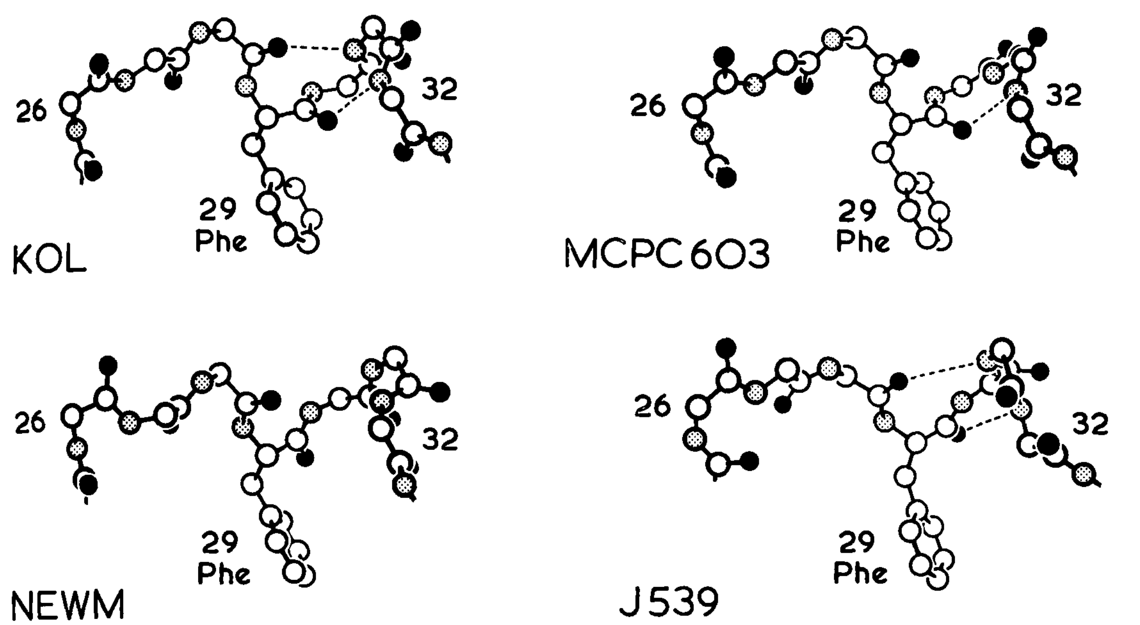
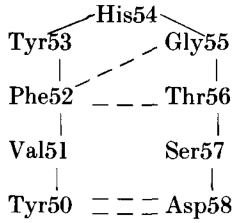


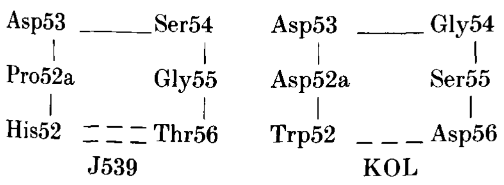
Figure 8. The conformation of the H1 regions of V_H KOL, V_H NEWM, V_H MCPC603 and V_H J539. The side-chain of Phe29 is buried within the framework structure.

residue turn. The other four residues in the turn are part of the framework structure:



The conformation of seven-residue turns is described in Table 2. The conformation found in NEWM is the conformation found for nearly all seven-residue turns that have Gly, Asn or Asp at the fifth position, position 55 in NEWM (Sibanda & Thornton, 1985; Efimov, 1986). Of the 13 three-residue H2 regions listed by Kabat *et al.* (1983), nine have a Gly residue at position 55 and four have Asp. We would expect these H2 regions to have the conformation found in NEWM.

The H2 regions in J539 and KOL, 52a to 55, form four-residue turns:



The conformation of these turns is determined by the position of the Gly residue. H2 in J539 has the conformation most commonly found for four-residue turns (Sibanda & Thornton, 1985; Efimov, 1986; see Table 2): the first three residues are in an approximately α_R conformation and the fourth (Gly55) in an α_L conformation. H2 in KOL is different; Gly54 is in an α_L conformation and the other three residues are in an α_R conformation.

Of the 71 H2 regions with four residues, Gly, Asn or Asp residues occur at position 54 in ten cases, at position 55 in 12 cases and at both positions 54 and 55 in 32 cases. In those with a Gly, Asn or Asp residue at position 54 only, we should expect an H2 conformation like that in KOL. In those cases

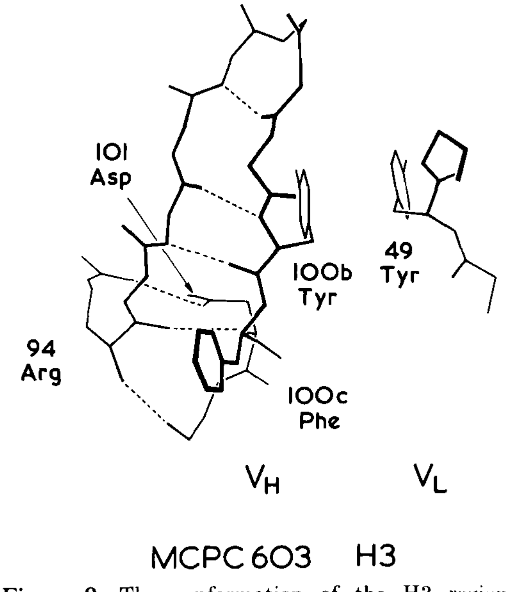


Figure 9. The conformation of the H3 region of MCPC603. Residue 100b Tyr100b packs against Tyr49 of the V_L domain. Phe100c also packs in the V_L V_H interface. Residues Arg94 and Asp101 form a salt bridge.

where a Gly residue occurs at position 55 we should expect an H2 conformation like that in J539.

The six residues in the H2 region of MCPC603 are part of a ten-residue hair-pin turn. At present there is too little experimental and theoretical evidence to formulate rules governing the conformations of such large turns. Therefore we do not know if the other six-residue H2 regions in the mouse subgroup III have a conformation close to that found in MCPC603.

9. Conformation of the H3 Hypervariable Region

The H3 region consists of residues 96 to 101. The V_H structure is formed by the recombination of three genes: V_H, which codes for residues 1 to 94 or 95; *D*, which codes for between one and 13 residues, and J_H (for a review, see Tonegawa, 1983). There are six human J_H germline genes that code for the following amino acid sequences:

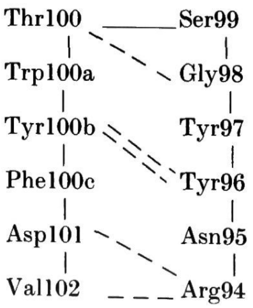
J_{H1}:	Ala- Glu- Tyr- Phe- Gln- His- Trp- Gly- Gln- Gly- Thr- Leu- Val- Thr- Val- Ser- Ser
J_{H2}:	Tyr- Trp- Tyr- Phe- Asp- Leu- Trp- Gly- Arg- Gly- Thr- Leu- Val- Thr- Val- Ser- Ser
J_{H3}:	Ala- Phe- Asp- Val- Trp- Gly- Gln- Gly- Thr- Met- Val- Thr- Val- Ser- Ser
J_{H4}:	Tyr- Phe- Asp- Tyr- Trp- Gly- Gln- Gly- Thr- Leu- Val- Thr- Val- Ser- Ser
J_{H5}:	Asn- Trp- Phe- Asp- Ser- Trp- Gly- Gln- Gly- Thr- Leu- Val- Thr- Val- Ser- Ser
J_{H6}:	Tyr- Gly- Met- Asp- Val- Trp- Gly- Gln- Gly- Thr- Thr- Val- Thr- Val- Ser- Ser

(Ravetch *et al.*, 1981) and four mouse genes that code for the following amino acid sequences:

J_{H1}:	Trp- Tyr- Phe- Asp- Val- Trp- Gly- Ala- Gly- Thr- Thr- Val- Thr- Val- Ser- Ser-
H_{H2}:	Tyr- Phe- Asp- Val- Trp- Gly- Gln- Gly- Thr- Thr- Val- Thr- Val- Ser- Ser
J_{H3}:	Trp- Phe- Ala- Tyr- Trp- Gly- Trp- Gly- Thr- Leu- Val- Thr- Val- Ser- Ala
J_{H4}:	Asp- Tyr- Trp- Gly- Trp- Gly- Thr- Ser- Val- Thr- Val- Ser- Ser

(Sakano *et al.*, 1980). Because the joining ends of the *D* and J_H genes can be varied, the residues coded at the beginning of J_H genes may not be present in the final structure. Further sequence diversity in this region is produced by somatic mutations. So it is not surprising that the H3 regions of J539, NEWM, MCPC603 and KOL differ greatly in size (6, 7, 9 and 15 residues, respectively), sequence and conformation. Here we shall confine our discussion to the H3 region of MCPC603, as our analysis suggests that its conformation is found at least in part in several other immunoglobulins.

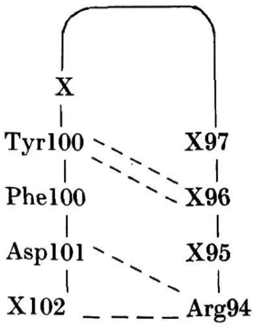
The H3 region in MCPC603 forms a large hair-pin loop:



For such large loops the range of allowed ϕ,ψ values will permit several conformations and the one actually found will depend upon the packing against the rest of the protein. In MCPC603 the conformation of H3 is determined mainly by the interactions of residues Arg94, Tyr100b, Phe100c and Asp101 within the V_H domain and at the V_L–V_H interface. (The importance of Tyr100b–Phe100c in making the conformation of H3 in MCPC603 different from that in NEWM was noted by Padlan *et al.* (1977).) The side-chains of residues 96 to 100a are on the surface of the protein.

Arg at position 94 packs across the H3 hair-pin and forms a surface salt bridge with Asp101. Tyr100b and Phe100c pack in the V_L–V_H interface (Fig. 9). Residues at positions equivalent to 100c are usually part of the conserved core of the V_L–V_H interface (Chothia *et al.*, 1985). Residues at this position are Phe or Leu in 83% of known sequences. In MCPC603, Tyr100b packs into a large cavity adjacent to Tyr49 of V_L (Fig. 9). The hydroxyl groups of both Tyr residues are on the surface. Tyr or Phe occurs at position 49 in 82% of V_L domains. Different residues at the position equivalent to 100b can produce different H3 conformations. For example, in KOL it is Gly and the cavity adjacent to Tyr49 in V_L is filled by Phe at a position equivalent to 100a. This contributes to making the conformation of H3 in KOL very different from that found in MCPC603.

The sequence Tyr-Phe-Asp at positions 100-100-101 is found in the human genes J_{H2} and J_{H4} and the mouse genes J_{H1} and J_{H2} (see above). The human J_{H5} gene has Trp in place of Tyr. If these residues are not removed during gene recombination or by somatic mutation, we should normally expect these J_4 genes to produce an H3 conformation close to that found in MCPC603. We inspected the sequence tables of Kabat *et al.* (1983) for H3 regions that are at least six residues in length, have an Arg residue at position 94, Asp at position 101 and Tyr-Phe in the two positions preceding 101, i.e. those that have the form:



Kabat *et al.* (1983) listed the entire H3 region for 28 human and 77 mouse sequences. Of these, one human and 48 mouse sequences fulfil exactly the size and sequence conditions. Another five human sequences are close in that they differ only by having a Lys residue at position 94, Phe or Trp in place of Tyr, or Met in place of Phe. The size distribution of the H3 regions in these 54 sequences is:

H3 residue length	6	7	8	9	10	11	12	13	14
Number of sequences	18	4	1	5	20	3	2	0	1

For these sequences we would expect the stem of H3 to have the same conformation as that found in MCPC603. The conformation of the remaining (distal) part of the small and medium-sized H3 loops may be given by the turn rules described in Table 2.

10. The Effects of Environment on the Structure of the Hypervariable Regions

The descriptions of the hypervariable regions given above suggest that their main-chain conformations are determined solely by particular residues within each region. In reality we should expect the conformations to be affected by their environment. The effects on a particular region can be divided into two parts; local changes in conformation and changes in the relative position of the region in the binding site.

A measure of the difference in conformation of two peptides is the r.m.s. difference in position of their atoms after they have been optimally superposed. In the sections above we report the r.m.s. differences for the main-chain regions of hypervariable regions with the same fold in different immunoglobulin structures. The r.m.s. differences in position are small. For the structures determined at high resolution they are less than 0·5 Å. For those determined at medium resolution they are usually 1·0 Å or less and are due mainly to differences in the orientation of peptides. It is only in the H1 region that we find significant, though small, differences in conformation (see section 7, above).

To determine differences in the relative positions of hypervariable loops in the immunoglobulin structures we did the following calculation. The Fab proteins NEWM, MCPC603, KOL and J539 were superposed by a fit of the V_L–V_H framework residues listed in section 3, above. The V_L structures REI and RHE were superposed on the Fabs by a fit of the V_L framework residues. After the superposition of the framework, we calculated the additional shift required to superpose hypervariable regions of the same fold; for example, the common residues in the L1 regions of J539, REI and MCPC603.

The results of these calculations are given in Figure 10. In the Fab proteins, hypervariable regions of the same fold differed in position by 0·2 to 1·5 Å. In part these differences occur because, although the V_L–V_H dimers have the same pattern of residue contacts and very similar packing geometries (Poljak *et al.*, 1975; Padlan, 1979; Chothia *et al.*, 1985), there are small differences in the orientation of V_H relative to V_L (Davies & Metzger, 1983).

The REI V_L structure was determined from a Bence-Jones protein. This contains a V_L–V_L dimer and their packing in REI is very similar to the V_L–V_H packing in the Fabs (Epp *et al.*, 1975). The positions of the REI hypervariable regions relative to the framework are the same as those that occur in the Fabs (Fig. 10).

In these structures, therefore, differences in the environment of the hypervariable regions produce only small differences in main-chain conformation and differences of no more than 1·5 Å in their position relative to the framework. In the Bence-Jones proteins RHE and MCG, the hypervariable regions have environments very different from those that would normally be found in the immunoglobulins.

The packing of V_L–V_L dimer in RHE is quite different from that found for V_L–V_H dimers (Furey *et al.*, 1983) and so the environments of its hypervariable regions are very different from those found in the other immunoglobulins discussed here. These differences in environment have little effect on the conformations of L1, L2 and L3 in RHE: they fit the homologous regions in Fab KOL, with r.m.s. differences in their co-ordinates of less than 0·3 Å (see above). They do have some effect on the

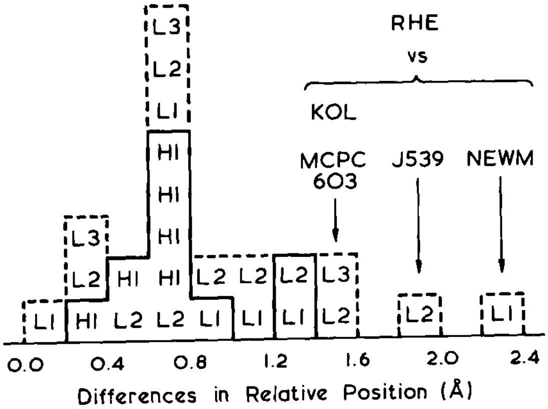


Figure 10. Differences in the position, relative to the β-sheet framework, of homologous hypervariable regions. The method used to determine the differences is described in the text. Continuous lines enclose the differences found between hypervariable regions in Fab structures. Broken lines enclose the differences found between hypervariable regions in Bence-Jones proteins and Fabs. The large differences found between the regions in RHE and the Fabs are labelled. Differences were determined for the relative position of the main-chain atoms of the residues of:

L1 in RHE, KOL and NEWM, 26 to 32;
L1 in J539, REI and MCPC603 26 to 29 and 32;
L2 in RHE, KOL, J539, REI and MCPC603, 50 to 52;
L3 in RHE and KOL, 91 to 96;
L3 in REI and MCPC603, 91 to 96, and
H1 in KOL, NEWM, MCPC603 and J539.

position of L1, L2 and L3 relative to the framework. The positions in RHE can differ by up to 2·2 Å from the positions found in the Fabs (Fig. 10).

In the structure of the Bence-Jones protein MCG (Schiffer *et al.*, 1973), a more complex situation is observed. The crystal of this protein has the dimer in the asymmetric unit, with the two V_L monomers in different environments. The L1 region of one monomer is in the helical conformation that we would expect from its sequence. The L1 region of the other monomer is prevented from having this conformation by the close approach of residues 31 and 32 to a neighbouring molecule and it is quite disordered (Schiffer, 1980). The observation that the close contact produces disorder, rather than an alternative conformation, suggests that the L1 region has only a limited flexibility.

11. The Residues that Form the Immunoglobulin Binding Sites and their Surface Area

(a) *Residues in the region of the binding sites*

In the preceding sections we have made a precise structural distinction between two parts of the variable domains: the conserved β-sheet framework and the regions of variable main-chain conformation. What are the contributions of these two parts to the antigen-binding site?

The hypervariable regions cluster at one end of the V_L–V_H dimer and present a surface, part of which interacts with the antigen. Figure 11 shows a space-filling drawing of this region in MCPC603. The residues accessible to the solvent in this part of the protein are 27 to 32, 49 to 53 and 92 to 94 in V_L and 28 to 33, 52 to 56 and 96 to 100a in V_H. The regions outside the β-sheet defined in section 3 are: 26 to 32, 50 to 52 and 91 to 96 in V_L and 26 to 32, 52 to 56 and 96 to 101 in V_H. The limits of some of the regions of accessible residues differ by up to two residues from the limits of these regions. In Table 6 we list the accessible residues that form the same region in J539, KOL and NEWM. The limits of the accessible loops in KOL, NEWM, J539 and MCPC603 are very similar but not identical. Taken together they show that the residues available for binding to antigens are largely those in the structurally variable regions defined in section 3, above. Variations of loop size and sequence may result in one or two residues at the loop ends becoming buried or one or two framework residues becoming exposed.

Except for H2, the regions listed in Table 6 are similar to the complementarity-determining regions (CDRs) determined from sequence variability

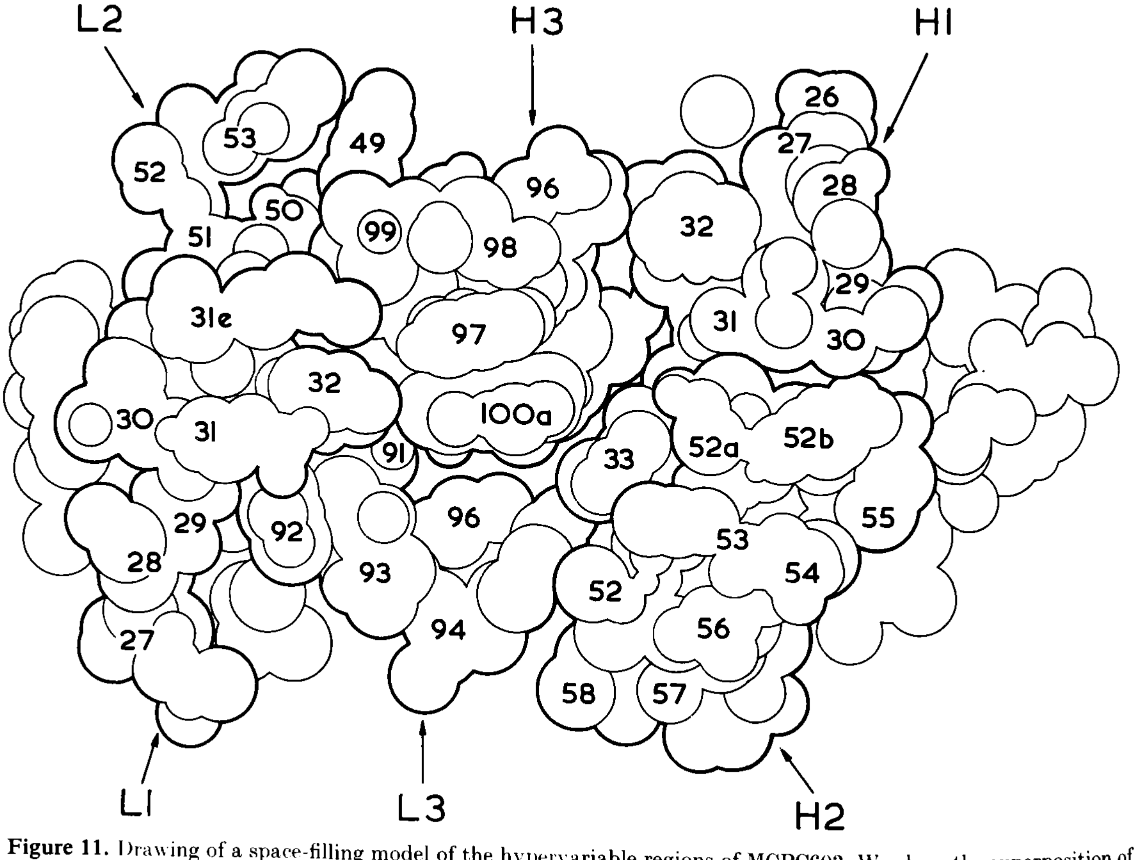


Figure 11. Drawing of a space-filling model of the hypervariable regions of MCPC603. We show the superposition of 5 sections cut through a model at 2 Å intervals. Just above the section shown here are residues 31a to 31d in V_L and 52c in V_H.

Table 6

Immunoglobulin binding sites: the residues and their accessible surface areas (Å^2)

Regions of variable main-chain structure	Residues accessible in the region of the binding sites							
	KOL		NEWM		MCPC603		J539	
	Residues	A.S.A.	Residues	A.S.A.	Residues	A.S.A.	Residues	A.S.A.
L1:26–33	27–32	270	27–33	450	27–32	770	27–30	240
L2:50–52	49–53	360	—		49–53	190	49–53	310
L3:91–96	91–94	240	91–96	254	92–94	170	91–96	330
H1:26–32	28–32	280	28–33	380	28–33	240	28–33	360
H2:53–55	52–58	440	52–58	430	52–56	620	50–58	430
H3:96–101	96–100g	600	96–100	250	96–100a	330	95–100	550
Total A.S.A.		2190		1764		2320		2220

A.S.A., mean accessible surface area.

(Kabat *et al.*, 1983). H2 in Table 6 covers residues 50 to 58; the corresponding CDR covers residues 50 to 65. Padlan (1977*a*) found that the first three and last six residues of this CDR had the same structure in NEWM and MCPC603. We find that this is also true for KOL amd J539. Residues 59 to 65 run down one side of the V_H domains and are fairly remote from the other hypervariable regions. The side-chains of 59 to 65 are accessible to the solvent and the variation in their sequence may reflect only a lack of structural and functional constraint.

(b) *Surface area of residues in the binding sites*

Table 6 also lists the accessible surface areas of the loops that make up the binding sites. H3 in KOL is unusually large, 15 residues, and it makes the largest contribution to the total surface. This is not the case in J539, NEWM or MCPC603, in which medium-sized H3 regions make contributions similar to these of the other loops. The important role of H3 in antibody specificity arises not from its size but from its central position in the binding site (Fig. 11).

The total accessible surface area of the region of the hypervariable loops in J539, KOL and MCPC603 is ~2250 Å^2; in NEWM, which is unique in having L2 deleted, it is 1760 Å^2 (Table 6). Analysis of oligomeric proteins shows that the cases where the structures in the isolated and associated states are very similar, stable associations are formed by surfaces with surface area that are smaller than the total found for these binding sites (Chothia & Janin, 1975, and our unpublished work). Typically, each monomer buries 500 to 1000 Å in the interface, a quarter to half of the total accessible surface area of the hypervariable loops. The number of hydrogen bonds and salt bridges in these interfaces varies. (In those cases in which association involves changes in structure or the stabilization of loops that do not have a fixed structure, larger surface areas are involved.)

The expectation that antibody–protein interactions would involve surfaces similar to those found in oligomeric proteins is supported by a description of the complex formed by immunoglobulin D1.3 and the antigen hen egg-white lysozyme (Amit *et al.*, 1986). The association does not involve significant changes in main-chain conformation. The antibody residues that make contact with the lysozyme are 30, 32, 49 to 50, 91 to 93 in V_L and 30 to 32, 52 to 54 and 96 to 99 in V_H. The interface consists of 690 Å^2 of the antibody surface and 750 Å^2 of the enzyme surface.

12. Conclusion

In this paper we have attempted to identify the residues that determine the conformations of the hypervariable regions. We have proposed that, if the residues we have identified are found in the sequences of other immunoglobulins, their hypervariable regions will have the same conformations as those found in the known structure that shares the same characteristic residues. The analysis of the immunoglobulin sequences implies that most of the hypervariable regions have one of a small set of main-chain conformations. We call these common conformations "canonical structures".

The atomic structures of the V_κ domains AU and ROY (Fehlhammer *et al.*, 1975; Colman *et al.*, 1977) give support to some of the conclusions of our analysis. The sequences of these two proteins differ from that of REI at 18 and 16 positions, respectively. Eight of the changed positions are in hypervariable regions:

Position	30	31	32	50	91	92	93	96
REI	Ile	Lys	Thr	Glu	Tyr	Gln	Ser	Tyr
AU	Ser	Asp	Tyr	Asp	Tyr	Asp	Tyr	Trp
ROY	Ser	Ile	Phe	Asp	Phe	Asp	Asn	Leu

The residue changes at positions 30, 31, 93 and 96 involve large differences in volume and chemical character. However, from the analysis given above we would not expect them to produce a main-chain conformation for L1 and L3 different from that found in REI and in fact no differences in main-chain conformation are seen (Fehlhammer *et al.*, 1975; Colman *et al.*, 1977).

To test the accuracy of the analysis we are applying our results to predict the structures of the variable domains of new immunoglobulins. In all cases our predictions are being recorded prior to the determination of the structures by X-ray analysis.

There is a fundamental difference between the method of prediction based on the work described here and the methods used by previous workers (Padlan *et al.*. 1977; Davies & Padlan, 1977; Potter *et al.*, 1977; Stanford & Wu, 1981; Feldmann *et al.*, 1981; de la Paz *et al.*, 1986). Those authors compared the sequences of the hypervariable regions in their immunoglobulins with the sequences of the corresponding hypervariable regions in the known structures and then built a model of each loop from the region closest in size and overall sequence homology. In some cases *ad-hoc* adjustments were made to accommodate differences in sequence (Padlan *et al.*, 1977; Feldmann *et al.*, 1981).

In the prediction method based on the work described here, we are only concerned with the presence in the sequence whose structure is to be predicted of the few particular residues that are responsible for the canonical structures. For example, to determine the conformation of L1 we would examine the residues at positions 2, 25, 29, 30, 33 and 71. If the residues found at these positions matched one of the sets listed above in section 4, we would expect L1 to have the corresponding canonical structure whatever residues occurred at the other positions. If the residues at these positions did not match one of those sets, we would not expect one of the known canonical structures, however close the homology in the rest of the hypervariable region.

A prediction was made for the structure of immunoglobulin D1.3 and sent to the group carrying out the X-ray analysis prior to its structure determination (Chothia *et al.*, 1986). The conformation of the main chain was predicted using the analysis described here; for the conformation of the side-chains we used a procedure described previously (Lesk & Chothia, 1986). After the prediction was made, the atomic structure of D1.3 was determined from a 2·8 Å electron density map (Amit *et al.*, 1986). Predictions were made of the framework structure and of each of the six hypervariable regions. Of the 62 residues buried within or between the domains (Table 3), 56 in D1.3 are identical with those in the known structures and the other six differ by no more than a methyl group. The prediction that the structure of the β-sheet framework of D1.3 is the same as that in the known structures was confirmed by the crystal structure analysis.

Three of the hypervariable regions of D1.3 (L1, L2 and H2) are the same size as one of the known canonical structures and, at the sites we identified as important in determining their conformation, they contain identical residues. The prediction that the folds of these three regions would be close to those of the canonical structures was confirmed by crystallographic analysis (Chothia *et al.*, 1986).

The other three hypervariable regions have sequences that are the same as, or similar in size to, known canonical structures but, at the sites responsible for their conformation, they have similar but not identical residues. In making predictions of the structure of these regions, we had to judge whether the differences would produce a different main-chain conformation. Our prediction that they would not was correct in one case, H3, and partly incorrect in the other two, L3 and H1.

This one test carried out so far supports our assertions that we have identified the residues responsible for the conformation of the hypervariable regions in the known structures, and that if these particular residues occur in other immunoglobulins their hypervariable regions will have the same structure. It also suggests that when the residues are not identical it is difficult to predict the structure with confidence, even if the changes are small. Further predictions have been made for the structure of the variable domains of four immunoglobins whose X-ray analysis is in progress: DB3, NC41, H20 and NQ10/12.5 (Stura *et al.*, 1987; Laver *et al.*, 1987; Mariuzza *et al.*, 1985). Co-ordinates of the predicted structures have been sent to the groups carrying out the structure analyses.

Our analysis of the immunoglobulin sequences shows that many of their hypervariable regions form one of the canonical structures found in the six V_L domains or four V_H domains of known structure. The conclusion that most hypervariable regions have one of a small number of main-chain conformations may have only limited application to H3, where the variation in size and sequence is much greater than that found in the other regions. Our analysis does suggest, however, that half the H3 regions in the known mouse sequences have a conformation that, at least in part, is close to that in MCPC603; a point confirmed by the successful prediction of H3 in D1.3 (Chothia *et al.*, 1986).

The analysis of additional antibody crystal structures will extend the repertoire of canonical structures. Attempts to predict additional structures, and their tests after the structures have been determined crystallographically, will improve our ability to understand the effects of the changes that can occur in the residues responsible for their conformation.

The prediction of antibody structures is of use not only in testing the accuracy of our identification of the residues responsible for the conformation of canonical structures. It is of central importance in engineering antibodies of a prescribed specificity.

We thank John Cresswell for the drawings, Drs A. Feinstein, M. Levitt, R. Mariuzza and G. Winter for discussion and comments on the manuscript, and the Royal Society, the U.S. National Science Foundation (PCM83-20171), the National Institute of General Medical Sciences (GM25435), and the European Molecular Biology Organization for support.

References

Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986). *Science*, **233**, 747–753.

Amzel, L. M. & Poljak, R. J. (1979). *Annu. Rev. Biochem.* **48**, 961–997.

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). *J. Mol. Biol.* **112**, 535–542.

Chothia, C. & Janin, J. (1975). *Nature (London)*, **256**, 705–708.

Chothia, C., Novotny, J., Bruccoleri, R. & Karplus, M. (1985). *J. Mol. Biol.* **186**, 651–663.

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986). *Science*, **233**, 755–758.

Colman, P. M., Schramm, H. J. & Guss, J. M. (1977). *J. Mol. Biol.* **116**, 73–79.

Davies, D. R. & Metzger, H. A. (1983). *Annu. Rev. Immunol.* **1**, 87–117.

Davies, D. R. & Padlan, E. A. (1977). In *Antibodies in Human Diagnosis and Therapy* (Haber, E. & Krause, M., eds) p. 119, Raven, New York.

de la Paz, P., Sutton, B. J., Darsley, M. J. & Rees, A. R. (1986). *EMBO J.* **5**, 415–425.

Efimov, A. (1986). *Mol. Biol. (U.S.S.R.)*, **20**, 250–260.

Epp, O., Latham, E., Shiffer, M., Huber, R. & Palm, W. (1975). *Biochemistry*, **14**, 4943–4952.

Fehlhammer, H., Schiffer, M., Epp, O., Colman, P. M., Lattman, E. E., Schwager, P. & Steigemann, W. (1975). *Biophys. Struct. Mech.* **1**, 139–146.

Feldmann, R. J., Potter, M. & Glaudemans, C. P. J. (1981). *Molec. Immunol.* **18**, 683–698.

Furey, W., Wang, B. C., Yoo, C. S. & Sax, M. (1983). *J. Mol. Biol.* **167**, 661–692.

Kabat, E. A. (1978). *Advan. Protein Chem.* **32**, 1–75.

Kabat, E. A., Wu, T. T. & Bilofsky, H. (1977). *J. Biol. Chem.* **252**, 6609–6616.

Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Milner, M. & Perry, H. (1983). Sequences of Proteins of Immunological Interest, 3rd. edit., Public Health Service, NI. H. Washington, D.C.

Laver, W. G., Webster, R. G. & Colman, P. M. (1987). *Virology*, **156**, 181–184.

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

Lesk, A. M. (1986). In *Biosequences: Perspectives and User Services in Europe*, (Saccone, C., ed.), pp. 23–28, European Economic Commission, Strasbourg.

Lesk, A. M. & Chothia, C. (1982). *J. Mol. Biol.* **160**, 325–342.

Lesk, A. M. & Chothia, C. H. (1986). *Phil. Trans. R. Soc. Lond. ser. A*, **317**, 345–356.

Mariuzza, R. A., Boulot, G., Guillon, V., Poljak, R. J., Berek, C., Jarvis, J. M. & Milstein, C. (1985). *J. Biol. Chem.* **260**, 10268–10270.

Marquart, M. & Deisenhofer, J. (1982). *Immunology Today*, **3**, 160–166.

Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. (1980). *J. Mol. Biol.* **141**, 369–391.

Padlan, E. A. (1977*a*). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 2551–2555.

Padlan, E. A. (1977*b*). *Q. Rev. Biophys.* **10**, 35–65.

Padlan, E. A. (1979). *Mol. Immunol.* **16**, 287–296.

Padlan, E. A. & Davies, D. R. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 819–823.

Padlan, E. A., Davies, D. R., Pecht, I., Givol, D. & Wright, C. (1977). *Cold Spring Harbor Symp. Quant. Biol.* **41**, 627–637.

Poljak, R. J., Amzel, L. M., Chen, B. L., Phizackerley, R. P. & Sane, F. (1975). *Immunogenetics*, **2**, 393–394.

Potter, M., Rudikoff, S., Padlan, E. A. & Vrana, M. (1977). In *Antibodies in Human Diagnosis and Therapy* (Haber, E. & Krause, R. M., eds), pp. 19–28, Raven, New York.

Ravetch, J. V., Siebenlist, U., Korsmeyer, S., Waldmann, T. & Leder, P. (1981). *Cell*, **27**, 583–591.

Sakano, H., Maki, R., Kurosawa, Y., Roeder, W. & Tonegawa, S. (1980). *Nature (London)*, **286**, 676–683.

Saul, F., Amzel, L. & Poljak, R. J. (1978). *J. Biol. Chem.* **253**, 585–597.

Schiffer, M. (1980). *Biophys. J.* **32**, 230–231.

Schiffer, M., Girling, R. L., Ely, K. R. & Edmundson, A. B. (1973). *Biochemistry*, **12**, 4620–4631.

Segal, D., Padlan, E., Cohen, G., Rudikoff, S., Potter, M. & Davies, D. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 4298–4302.

Sibanda, B. L. & Thornton, J. M. (1985). *Nature (London)*, **316**, 170–174.

Stanford, J. M. & Wu, T. T. (1981). *J. Theoret. Biol.* **88**, 421–439.

Stura, E. A., Feinstein, A. & Wilson, I. A. (1987). *J. Mol. Biol.* **193**, 229–231.

Suh, S. W., Bhat, T. N., Navia, M. A., Cohen, G. H., Rao, D. N., Rudikoff, S. & Davies, D. R. (1986). *Proteins*, **1**, 74–80.

Thornton, J. M., Sibanda, B. L. & Taylor, W. R. (1985). In *Investigation and Exploitation of Antibody Combining Sites* (Reid, R., Cook, G. M. W. & Morse, D. J., eds), pp. 23–31, Plenum, New York.

Tonegawa, S. (1983). *Nature (London)*, **302**, 575–581.

Edited by R. Huber