J. Mol. Biol. (1992) **227**, 799–817

Structural Repertoire of the Human V_H Segments

Cyrus Chothia[1,2]†, Arthur M. Lesk[3], Ermanno Gherardi[2]
Ian M. Tomlinson[1,2], Gerald Walter[1,2], James D. Marks[1]
Meirion B. Llewelyn[2] and Greg Winter[1,2]

[1]*Cambridge Centre for Protein Engineering*
[2]*MRC Laboratory of Molecular Biology*
[3]*Department of Haematology, University of Cambridge*
Hills Road, Cambridge CB2 2QH, England

(Received 6 February 1992; accepted 2 June 1992)

The V_H gene segments produce the part of the V_H domains of antibodies that contains the first two hypervariable regions. The sequences of 83 human V_H segments with open reading frames, from several individuals, are currently known. It has been shown that these sequences are likely to form a high proportion of the total human repertoire and that an individual's gene repertoire produces about 50 V_H segments with different protein sequences. In this paper we present a structural analysis of the amino acid sequences produced by the 83 segments.

Particular residue patterns in the sequences of V domains imply particular main-chain conformations, canonical structures, for the hypervariable regions. We show that, in almost all cases, the residue patterns in the V_H segments imply that the first hypervariable regions have one of three different canonical structures and that the second hypervariable regions have one of five different canonical structures. The different observed combinations of the canonical structures in the first and second regions means that almost all sequences have one of seven main-chain folds.

We describe, in outline, structures of the antigen binding site loops produced by nearly all the V_H segments. The exact specificity of the loops is produced by (1) sequence differences in their surface residues, particularly at sites near the centre of the combining site, and (2) sequence differences in the hypervariable and framework regions that modulate the relative positions of the loops.

Keywords: antibodies; hypervariable regions; canonical structures

1. Introduction

The antigen binding site of an antibody is formed by six loops of polypeptide: three from the light chain variable domain (V_L) and three from the heavy chain variable domain (V_H) (see Fig. 1). Great variation in the sequences that form the binding sites is achieved by a combinatorial process in which the complete gene for the protein is produced by the recombination of a number of gene segments, each of which is drawn from a pool of moderate size. The V_H domain is produced by the recombination of three gene segments: V_H, D and J_H. The V_H gene segment codes for residues 1 to 94 or 95 of the domain. This region includes the first and second binding site loop (Fig. 1). The third loop is formed by all three segments: the end of the V_H, D, and the beginning of J_H. The V_L domains are formed by a combination of two gene segments V_L and J_L. As in the V_H domain, the first segment codes for the first two binding site loops and the third loop is formed around the join of both gene segments.

The primary antibodies produced by the gene recombination are believed to be capable of recognizing all antigens with at least a moderate affinity. Subsequent somatic mutations of the rearranged gene increase the affinity and specificity.

Analysis of antibodies of known atomic structure has elucidated relationships between the sequence and three-dimensional structure of antibody combining sites (Chothia & Lesk, 1987; Tramontano *et al.*, 1990). These relationships imply that, except for the third region in V_H domains, binding site loops have one of a small number of main-chain conformations: canonical structures. The canonical

† Correspondence may be addressed to any author.

0022-2836/92/190799-19 $08.00/0

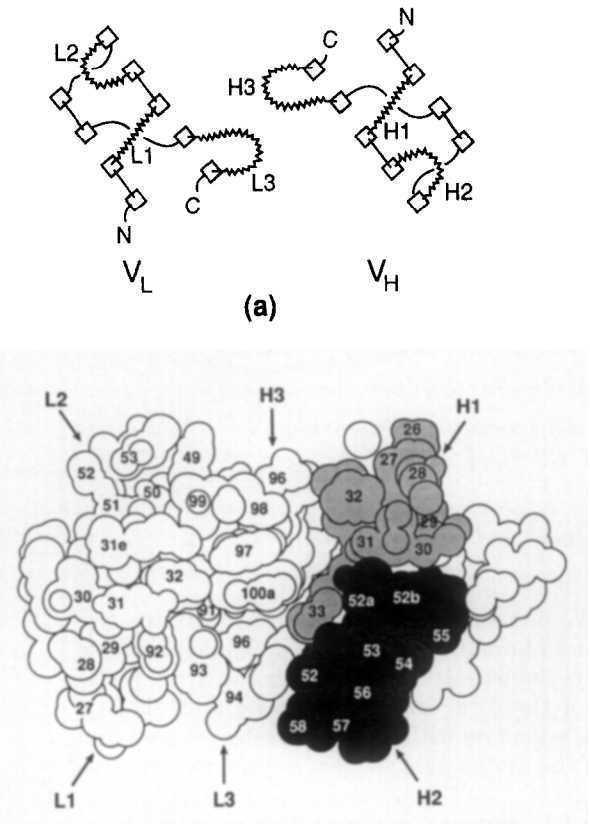


Figure 1. The antigen binding site. (a) Schematic diagram showing how the site is formed by 6 loops; 3 from the V_L domain, L1, L2 and L3 and 3 from the V_H domain, H1, H2 and H3. These loops are attached to strands of β-sheets that are conserved and form the framework structure. (b) A space filling drawing of the antigen binding site of an antibody. The regions formed by the V_H segment are shaded in this drawing. The region that includes H1 and CDR1 has light stippling and that which includes H2 and part of CDR2 has dark stippling.

structure formed in a particular loop is determined by its size and the presence of certain residues at key sites in both the loop and in framework regions. The general validity of the relationships has been demonstrated by their reasonably successful prediction of the structure of hypervariable regions in various antibodies prior to the experimental determination by X-ray crystallography (Chothia *et al.*, 1986, 1989).

Recently, Tomlinson *et al.* (1992) determined the sequences of the large majority of the functional germline V_H segments in a single individual (DP) and compiled a directory, which includes the previously determined V_H segments. As a result, we now know the sequences of 83 human V_H segments with open reading frames. The amino acid sequences implicit in these segments probably form a very high proportion of the human repertoire (Tomlinson *et al.*, 1992).

In this paper we discuss the structures implicit in the 83 known V_H segments. We note the extent to which sequences have a high proportion of identical residues and, therefore, similar framework structures. We then determine the canonical structures present in the sequence and hence the outline of the structural repertoire of the V_H segments.

Table 1
Structurally defective V_H segments

Sequence	Residue†	Normal role of residue
$V_{35}/V_{1\text{-}2b}$	69S (IM)	Part of strongly conserved hydrophobic core
	88V (A)	Small residue allows hydrogen bonding by buried polar groups
65-4/DP-39	23P (ATVK)	Main-chain hydrogen bonded in β-sheet
	78P (ALF)	Main-chain hydrogen bonded in β-sheet
V_H19/DP-59	38H (R)	Buried side-chain hydrogen bonds
	82T (LIM)	Part of hydrophobic core
65-2/DP-44	6H (QE)	Side-chain forms buried hydrogen bonds
DP-52	20P (VLI)	Main-chain hydrogen bonded in β-sheet
	39R (Q)	Side-chain H-bonded in V_L–V_H interface

† The residues that are normally found in V_H domains are given in parentheses.

Additional remarks:

(1) The 65-2 and 65-4 gene segments are orphons (Matsuda *et al.*, 1990).

(2) 65-4/DP-39, V_H19/DP-59 and 65-2/DP-44 have unusual heptamer sequences (Baer *et al.*, 1988; Matsuda *et al.*, 1990; Tomlinson *et al.*, 1992).

2. The Classification of the Sequences by Residue Identities

(a) *Defective protein sequences*

Tomlinson *et al.* (1992) list 83 V_H segments with open reading frames (see Table 5 below). For a product of the germline genes to be functional it must form a stable three-dimensional structure. The residues mainly responsible for the fold of variable domains have been determined by the analysis of the antibodies of known atomic structure (Padlan, 1977; Saul *et al.*, 1978; Lesk & Chothia, 1982; Davies & Metzger, 1983; Chothia *et al.*, 1988; Beale & Coadwell, 1989*a*,*b*). Using this information on the sequence requirements for stable V_H folds, we examined the amino acid sequences of the 83 segments to see if any contained residues that might hinder the formation of a stable three-dimensional structure.

Proteins have the ability to adapt, at least in part, to effects of mutations. This means that simple inspection of the sequences cannot give a totally unambiguous answer to the question of whether or not they will form a stable V_H fold. Inspection of the sequences of V_H segments does suggest, however, that five are unlikely to be functional (Table 1). Four have two residues that would be expected to hinder seriously the formation of the standard fold of V_H domains. Two of these four also have genetic defects. 65-4/DP-39 is an orphon, a gene transposed

Table 2

The ranges of residue identities for sequences within and between the families of human V_H segments

Family	Number of segments in the family†	Family 1	2	3	4	5	6
1	21	(61-98)	29-35	44-59	38-49	50-67	36-44
2	5	29-35	(76-95)	39-47	48-59	33-42	48-51
3	24	44-59	39-47	(69-98)	45-60	45-56	45-54
4	20	38-49	48-59	45-60	(76-98)	41-52	62-68
5	6	50-67	33-42	45-56	41-52	(82-97)	36-43
6	1	36-44	48-51	45-54	62-68	36-43	(—)

† The amino acid sequences of the segments are listed in Table 5.

to a region where it may not be able to take part in productive recombination, and also has a heptamer whose unusual sequence may prevent recombination. V_H19/DP-59 has an unusual heptamer sequence. The fifth sequence, 62-2/DP-44 has one defective residue, is an orphon and has an unusual heptamer sequence (Table 1).

On a different level, only a fragment of the sequence of DP-61 is known at present (see Table 5), and, therefore, it is not included in the structural analysis described here.

(b) *The extent of the sequence identities within and between the six V_H families*

The V_H gene segments are classified into six families on the basis of nucleotide homology (Kodaira *et al.*, 1986; Lee *et al.*, 1987; Shen *et al.*, 1987; Berman *et al.*, 1988; Humphries *et al.*, 1988; Buluwela & Rabbits, 1988). We determined the extent of the variations in their amino acid sequences. For all pairs of V_H segments, we calculated the number of homologous sites that contain identical residues. The results of these calculations are summarized in Table 2. Each sequence has more identities with all the other members of its own family than with any other sequence.

In most cases, pairs of sequences within the same family have 80 or more identical residues. The sequences in family 1 are more divergent than those in other families, in that two members, $V_{I\text{-}4.1b}$ and DP-21, have residue identities of between 61 and 65 with six members of the family. However, they have higher residue identities with the other 13 members of family 1, and lower residue identities with all sequences in the other families.

3. The Canonical Structures in the First Hypervariable Regions

The residues at sites 31 to 35 in V_H domains are hypervariable and the region was designated the first complementarity determining region (CDR1†) (Kabat & Wu, 1971; Kabat *et al.*, 1991). Inspection of the antibodies of known structure shows that residues 33 to 35 are part of the framework β-sheet and show very little variation in main-chain conformation (Chothia & Lesk, 1987). The side-chain of residue 33 is on the surface; residue 34 is hydrophobic and packed in the interior of the domain and residue 35 is on the edge of the V_L–V_H interface. Residues 31, 32 and, in those sequences that have insertions, 31a and 31b, are at the end of a loop, formed by residues 26 to 32, that connects two of the strands of the framework β-sheet (Fig. 1).

Although the variations in residues at position 27 to 30 are less extensive than at positions 31 to 35, significant differences do occur and can influence the structure of CDR1. Thus, in structural terms, CDR1 can be treated as part of a single region covering residues 26 to 35. The part outside the framework β-sheet, 26 to 32, we refer to as H1.

Most V_H segments have neither insertions nor deletions in the H1 region and we will refer to these as being of standard size. Some sequences have one or two residues inserted.

(a) *Observed structures for the H1 regions*

There is now considerable evidence from the analysis of the observed three-dimensional structure of antibodies that standard size H1 regions have conformations close to that illustrated in Figure 2(a). We refer to this as structure 1 for the H1 regions. The early work on the determinants of this structure is described by Chothia & Lesk (1987) and this has been extended by the analysis of the antibody structures determined since then (our unpublished results). The main determinants are (1) a Gly at position 26 that produces a sharp turn by means of a conformation that would produce steric strain in other residues, and (2) a large hydrophobic residue at position 29 that packs deep in the interior of the domain between hydrophobic residues at positions 24 and 34. The conformation is also influenced by the residue at position 27, which packs into a surface cavity and the residue at position 94, which packs against H1 residues.

Although the residues most commonly found in the known structures at positions 24, 29 and 34 are Ala, Phe and Met, respectively, some variations in the volumes of the residues at these sites are seen; for example, the antibody D1.3 has Val, Leu and Val at these sites and the antibody HyHEL-10 has

† Abbreviations used: CDR, complementarity determining region; r.m.s., root-mean-square.

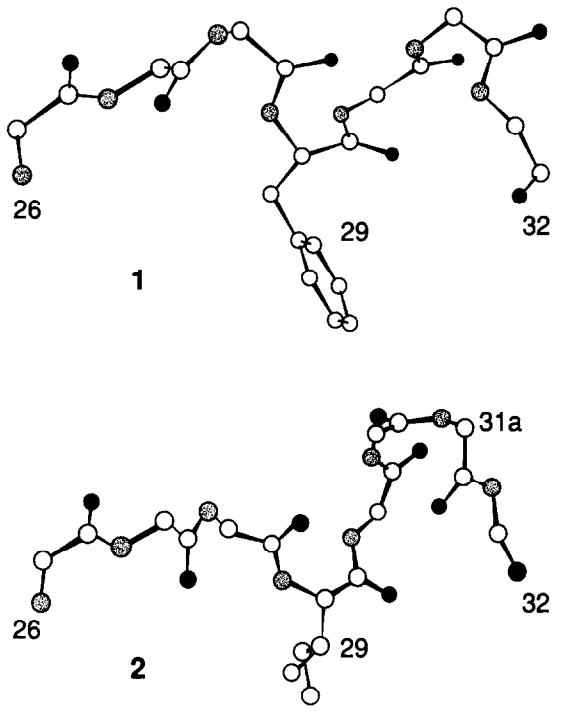


Figure 2. Canonical structures 1 and 2 for the H1 region. Structure 1: this drawing is taken from Chothia & Lesk (1987). Structure 2: this drawing is of the H1 region in AN02 (Brunger *et al.*, 1991; see the text). This structure was determined at medium resolution only. This means that, though the fold of the main-chain and the disposition of the side-chains is clear, the exact orientation of the carbonyl groups may be revised. Relative to the view of H1 in Fig. 1, the main-chain conformations are viewed from the "side": the antigen will make contacts with their top part and the bottom part, particularly residue 29 packs against the framework. Canonical structure 1 is found in standard size H1 regions and structure 2 is found in those with a one residue insertion (see the text).

Val, Ile and Trp. These variations produce only small differences in the conformation of H1 (Fischmann *et al.*, 1991; Padlan *et al.*, 1989; Chothia *et al.*, 1989). The residues Tyr and Arg are found frequently at positions 27 and 94. Exceptions do occur and produce changes that are structurally small but which can significantly affect affinity (Foote & Winter, 1992).

Less is known about the structure of those H1 regions that have insertions of one or two residues. Some information is available from the recent determination of the atomic structure of the Fab fragment of AN02 (Brunger *et al.*, 1991). Its H1 region has one inserted residue and the examination of its structure shows that (1) residues 26 to 29 and 32 to 35 have a conformation very similar to that found for H1 regions without insertions and (2) the insertion occurs in the surface loop formed by residues 30 and 31 (Brunger *et al.*, 1991). Inspection of the AN02 sequence shows that this might have been expected because it has residues at positions 24, 26, 29 and 34 very similar to those in the sequences of standard size H1 structures.

The same residue pattern is found in the other sequences that have one or two insertions in H1. This means that AN02 gives us a general structure for those H1 regions that have a one-residue insertion: canonical structure 2 for H1 regions (Fig. 2(b)). It also implies that, for H1 regions with a two-residue insertion, residues 26 to 29 and 32 to 35 have a conformation close to that in structure 1, and that the two insertions will form, with residues 30 and 31, a surface loop whose exact conformation is unknown at present. We can refer to this canonical structure 3 for H1 regions.

Note that the position in which the structural data and sequence patterns place the insertion(s), in the region of residue 31, is different to that given by Kabat *et al.* (1991). These authors place the insertion(s) after residue 35.

(b) *H1 structures implicit in the sequences of the V_H segments*

Of the 77 potentially functional V_H segments, 58 have standard-sized H1 regions; 9 have one insertion and 10 have two insertions. Inspection of sequences (Table 5 below) shows that all have residues at key sites that fit one of the defined canonical structures. In Table 3 we list the residues

Table 3

Residues at the key sites for the H1 canonical structures

Canonical structure	Family	Number of sequences	Sites†					
			24	26	27	29	34	94
1	1	21	A:19 V:2	G	Y:18 F:1 G:2	F:20 L:1	M:13 I:5 L:2 V:1	R:17 T:2 A:1
	3	24	A:23 G:1	G	F	F:22 V:2	M:23 T:1	R:18 K:5 T:1
	4	7	V	G	G	F:2 I:3 V:2	W	R
	5	6	G:5 T:1	G	Y	F	I	R
2	2	1	F	G	F	L	C	H
	4	8	V	G	Y:4 G:4	I	W	R
3	2	4	F:3 V:1	G	F	L	V	R:3 H:1
	4	5	G	G	G	I:4 V:1	W	R
	6	1	G	G	D	V	W	R

The amino acid sequences of the segments are given in Table 5.

† At sites that have more than one kind of residue we give the frequencies with which they occur.

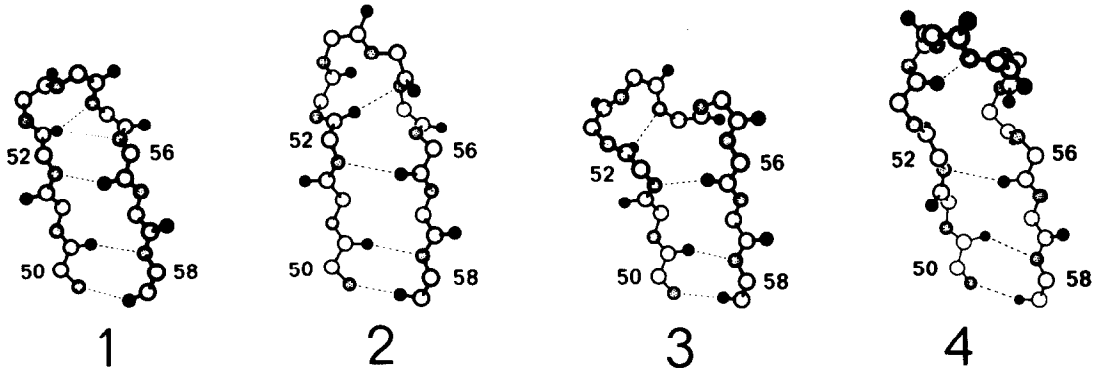


Figure 3. Canonical structures 1, 2, 3 and 4 of the H2 region. This drawing is taken from Tramontano *et al.* (1990). The face of the hairpins that is towards the viewer is that seen by the antigen; see Figs 1 and 4 to 9.

that occur in the sequences at positions 24, 26, 27, 29, 34 and 94.

The Gly residue at position 26 is absolutely conserved (Table 3). Half of the sequences have Ala, Phe and Met/Ile at positions 24, 29 and 34. The other half have somewhat different combinations of hydrophobic residues but in most, the range of residues is the same as that in the known H1 structures. In a few, it is just a little outside the range so far observed. This implies that the H1 region of standard size have conformations close to canonical structure 1, those with one insertion have conformations close to canonical structure 2 and those with two insertions have conformations close to canonical structure 3.

4. The Canonical Structures in the Second Hypervariable Region

Residues in the region 50 to 65 in V_H domains show considerable variations in sequence and this region was designated the second complementarity determining region (CDR2) (Kabat & Wu, 1971; Kabat *et al.*, 1991).

In the known V_H structures residues 61 to 65 are part of a surface loop distant from the antigen binding site. Residues 50 to 52 and 56 to 60 have very similar conformations (Padlan, 1977; Chothia & Lesk, 1987; Tramontano *et al.*, 1990): they form two strands of β-sheet that hydrogen bond to each other (Fig. 3). The residue at site 51 is buried and shows little variation. Residues 52 and 56 to 60 are on the surface and though sequence variations change the shape of this surface they have little effect on its main-chain structure.

The remaining residues, 52 to 56, form a hairpin loop (Fig. 3). In the known structures this loop varies in size, having five, six or eight residues, and conformation. We refer to this region as H2.

(a) *Observed structures for the H2 regions*

The conformations observed for the H2 regions, and the residues mainly responsible for these conformations have been discussed in some detail by Tramontano *et al.* (1990). Four different conformations are found and constitute canonical structures 1 to 4 for H2 regions.

Canonical structure 1 is found in the five-residue H2 regions (Fig. 3). It has the 3:5 conformation commonly found for loops of this size (Sibanda *et al.*, 1989). These loops usually have a Gly, Asn or Asp residue at the fourth position (residue 55 here) as the residue at this site has positive values for the main-chain torsion angles ϕ,ψ. The H2 loop packs against the residue at site 71 and the position of the

Table 4

Residues at the key sites for the H2 canonical structures

Canonical structure	Family	Number of sequences	Sites†			
			52a	54	55	71
1	2	5	—	—	D	K
	3	4	—	—	G	R
	4	20	—	—	G	V:19 I:1
2	1	7	P:3 T:2 A:2	—	G	A:3 T:2 L:2
	5	6	P	—	S	A
3	1	13	—	G:2 S:6 N:4 D:1	—	R
	3	16	—	G:13 S:3		R
4	3	3	—	S	Y	R

The amino acid sequences of the segments are given in Table 5.

† At sites that have more than one kind of residue we give the frequencies with which they occur.

—, Indicates that the identity of the residues at the site are not an important determinant for its conformation.

Table 5

Structural classification of the protein sequences of the human V_H segments

	1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9
	1|||5||||0||||5||||0||||5||||0|ab|||5||||0||||5||||0||abc||5||||0||||5||||0||||5||||0||abc||5||||0||||
Canonical structure class 1-1	
Family 3	
DP-42	EVQLVETGGGLIQPGGSLRLSCAASGFTVSS NYMSWVRQAPGKGLEWVSVIY SGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
8-1B[3]	EVQLVESGGGLVQPGGSLRLSCAASGFTVSS NYMSWVRQAPGKGLEWVSVIY SGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
DP-45	EVQLVQSGGGLVQPGGSLRLSCAGSGFTFSS YAMHWVRQAPGKGLEWVSAIG TGGGTYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDMAVYYCAR
13-2[3]/DP-48	EVQLVESGGGLVQPGGSLRLSCAASGFTFSS YDMHWVRQATGKGLEWVSAIG TAGDTYYPGSVKGRFTISRENAKNSLYLQMNSLRAGDTAVYYCAR
Family 4	
Tou-VH4.21[16]	QVQLQQWGAGLLKPSETLSLTCAVYGGSFSG YYWSWIRQPPGKGLEWIGEII HSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V_H5[10]/*$V_H4.21$*[17]/DP-63	QVQLQQWGAGLLKPSETLSLTCAVYGGSFSG YYWSWIRQPPGKGLEWIGEIN HSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V58[18]	QVQLQQWGAGLLKPSETLSLTCAVYGGSVSG YYWSWIRQPPGKGLEWIGYIY YSGSTNNNPSLKSRATISVDTSKNQFSLNLSSVTAADTAVYCCAR
V_{IV}-4[2]	QVQLQESGPGLVKPSETLSLTCTVSGGSISS YYWSWIRQPAGKGLEWIGRIY TSGSTNYNPSLKSRVTMSVDTSKNQFSLKLSSVTAADTAVYYCAR
$V_H4.11$[17]/DP-71	QVQLQESGPGLVKPSETLSLTCTVSGGSISS YYWSWIRQPPGKGLEWIGYIY YSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V_{71}-4[4]	QVQLQESGPGLVKPSETLSLTCTVSGGSVSS YYWSWIRQPPGKGLEWIGYIY YSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
$V_H4.16$[17]	QVQLQESGPGLVKPSDTLSLTCTVSGGSISS YYWSWIRQPPGKGLEWIGYIY YSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
Canonical structure class 1-2	
Family 1	
DP-3	EVQLVQSGAEVKKPGATVKISCKVSGYTFTD YYMHWVQQAPGKGLEWMGLVDP EDGETIYAEKFQGRVTITADTSTDTAYMELSSLRSEDTAVYYCAT
DP-10	QVQLVQSGAEVKKPGSSVKVSCKASGGTFSS YAISWVRQAPGQGLEWMGGIIP IFGTANYAQKFQGRVTITADESTSTAYMELSSLRSEDTAVYYCAR
hv1263[5]	QVQLVQSGAEVKKPGSSVKVSCKASGGTFSS YAISWVRQAPGQGLEWMGRIIP ILGIANYAQKFQGRVTITADKSTSTAYMELSSLRSEDTAVYYCAR
DP-21	QVQLVQSGSELKKPGASVKVSCKASGYTFTS YAMNWVRQAPGQGLEWMGWINT NTGNPTYAQGFTGRFVFSLDTSVSTAYLQICSLKAEDTAVYYCAR
V_{I}-4.1b[2]	QVQLVQSGSELKKPGASVKVSCKASGYTFTS YAMNWVRQAPGQGLEWMGWINT NTGNPTYAQGFTGRFVFSLDTSVSTAYLQISSLKAEDTAVYYCAR
DP-14	QVQLVQSGAEVKKPGASVKVSCKASGYTFTS YGISWVRQAPGQGLEWMGWISA YNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTAVYYCAR
VH1GRR[6]	QVQLVQSGAEVKKPGASVKVSCKASGYTFTS YGISWVRQAPGQGLEWMGWISA YNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTA
Family 5	
VH251[22]/DP-73	EVQLVQSGAEVKKPGESLKISCKGSGYSFTS YWIGWVRQMPGKGLEWMGIIYP GDSDTRYSPSFQGQVTISADKSISTAYLQWSSLKASDTAMYYCAR
V_HVJB[17]	EVQLVQSGAEVKKPGESLKISCKGSGYSFTS YWIGWVRQMPGKGLEWMGIIYP GDSDTRYSPSFQGQVTISADKPISTAYLQWSSLKASDTAMYYCAR
V_HVCW[17]	EVQLVQSGAEVKKPGESLKISCKGSGYSFTS YWIGWVRQMPGKGLEWMGIIYP GDSDTRYSPSFQGQVTISADKSISTAYLQWSSLKASDTAMYYCAR
1-v[3]	EVQLLQSAAEVKRPGESLRISCKTSGYSFTS YWIHWVRQMPGKELEWMGSIYP GNSDTRYSPSFQGHVTISADSSSSTAYLQWSSLKASDAAMYYCVR
VH32[23]	EVQLVQSGAEVKKPGESLRISCKGSGYSFTS YWISWVRQMPGKGLEWMGRIDP SDSYTNYSPSFQGHVTISADKSISTAYLQWSSLKASDTAMYYCAR
V_HVRG[17]/*V_HVMW*[17]	EVQLVQSGAEVKKPGESLRISCKGSGYSFTS YWISWVRQMPGKGLEWMGRIDP SDSYTNYSPSFQGHVTISADKSISTAYLQWSSLKASDTAMYYCAR

Canonical structure class 1-3

Family 1			
DP-1	QVQLVQSGAEVKKPGASVKVSCKASGYIFTD	YYMHWVRQAPGQELGWMGRINP	NSGGTNYAQKFQGRVTMTRDTSISTAYTELSSLRSEDTATYYCAR
V_{I-2}[2]	QVQLVQSGAEVKKPGASVKVSCKASGYTFTG	YYMHWVRQAPGQGLEWMGWINP	NSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
DP-8	QVQLVQSGAEVKKPGASVKVSCKASGYTFTG	YYMHWVRQAPGQGLEWMGWINP	NSGGTNYAQKFQGWVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
1-1[3]	QVQLVQSGAEVKKLGASVKVSCKASGYTFTG	YYMHWVXQAPGQGLEWMGWINP	NSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
DP-12	QVQLVQSGAEVKKPGASVKVSCKASGYTFTN	YCMHWVRQVHAQGLEWMGLVCP	SDGSTSYAQKFQARVTITRDTSMSTAYMELSSLRSEDTAMYYCVR
V_{71-5}[4]/DP-2	QMQLVQSGPEVKKPGTSVKVSCKASGFTFTS	SAVQWVRQARGQRLEWIGWIVV	GSGNTNYAQKFQERVTITRDMSTSTAYMELSSLRSEDTAVYYCAA
V_{I-3b}[2]/DP-25	QVQLVQSGAEVKKPGASVKVSCKASGYTFTS	YAMHWVRQAPGQRLEWMGWINA	GNGNTKYSQKFQGRVTITRDTSASTAYMELSSLRSEDTAVYYCAR
V_{I-3}[2]	QVQLVQSGAEVKKPGASVKVSCKASGYTFTS	YAMHWVRQAPGQRLEWMGWSNA	GNGNTKYSQEFQGRVTITRDTSASTAYMELSSLRSEDMAVYYCAR
DP-15	QVQLVQSGAEVKKPGASVKVSCKASGYTFTS	YDINWVRQATGQGLEWMGWMNP	NSGNTGYAQKFQGRVTMTRNTSISTAYMELSSLRSEDTAVYYCAR
21-2[3]/3-1[3]/DP-7	QVQLVQSGAEVKKPGASVKVSCKASGYTFTS	YYMHWVRQAPGQGLEWMGIINP	SGGSTSYAQKFQGRVTMTRDTSTSTVYMELSSLRSEDTAVYYCAR
HG3[7]	QVQLVQSGAEVKKPGASVKVSCKASGYTFNS	YYMHWVRQAPGQGLEWMGIINP	SGGSTSYAQKFQGRVTMTRDTSTSTVYMELSSLRSEDTAVYYCAR
7-2[3]	QMQLVQSGAEVKKTGSSVKVSCKASGYTFTY	RYLHWVRQAPGQALEWMGWITP	FNGNTNYAQKFQDRVTITRDRSMSTAYMELSSLRSEDTAMYYCAR
DP-4	QMQLVQSGAEVKKTGSSVKVSCKASGYTFTY	RYLHWVRQAPGQALEWMGWITP	FNGNTNYAQKFQDRVTITRDRSMSTAYMELSSLRSEDTAMYYCAR
Family 3			
DP-31	EVQLVESGGGLVQPGRSLRLSCAASGFTFDD	YAMHWVRQAPGKGLEWVSGISW	NSGSIGYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTALYYCAK
DP-32	EVQLVESGGGVVRPGGSLRLSCAASGFTFDD	YGMSWVRQAPGKGLEWVSGINW	NGGSTGYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTALYHCAR
DP-33	EVQLVESGGVVVQPGGSLRLSCAASGFTFDD	YTMHWVRQAPGKGLEWVSLISW	DGGSTYYADSVKGRFTISRDNSKNSLYLQMNSLRTEDTALYYCAK
22-2B[3]/DP-35	QVQLVESGGGLVKPGGSLRLSCAASGFTFSD	YYMSWIRQAPGKGLEWVSYISS	SGSTIYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
15-2B[3]/DP-40	EVQLVESGGGLVQPGGSLRLSCAASGFTFSN	HYTSWVRQAPGKGLEWVSYSSG	NSGYTNYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCVK
f1-p1[11]	EVQLVESGGGLVQPGGSLRLSCSASGFTFSS	YAMHWVRQAPGKGLEYVSAISS	NGGSTYYADSVKGRFTISRDNSKNTLYVQMSSLRAEDTAVYYCVR
hv3005[12]	QVQLVESGGGVVQPGRSLRLSCAASGFTFSS	YAMHWVRQAPGKGLEWVAVISY	DGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
hv3005f3[11]	QVQLVESGGGVVQPGRSLRLSCAASGFTFSS	YAMHWVRQAPGKGLEWVAVISY	DGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
GL-SJ2[13]/DP-46	QVQLVESGGGVVQPGRSLRLSCAASGFTFSS	YAMHWVRQAPGKGLEWVAVISY	DGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
VH26[14]/DP-47	EVQLLESGGGLVQPGGSLRLSCAASGFTFSS	YAMSWVRQAPGKGLEWVSAISG	SGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAK
DP-58	EVQLVESGGGLVQPGGSLRLSCAASGFTFSS	YEMNWVRQAPGKGLEWVSYISS	SGSTIYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
1.9III[3]/DP-49	QVQLVESGGGVVQPGRSLRLSCAASGFTFSS	YGMHWVRQAPGKGLEWVAVISY	DGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAK
3019b9[11]/DP-50	QVQLVESGGGVVQPGRSLRLSCAASGFTFSS	YGMHWVRQAPGKGLEWVAVIWY	DGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
DP-51	EVQLVESGGGLVQPGGSLRLSCAASGFTFSS	YSMNWVRQAPGKGLEWVSYISS	SSSTIYYADSVKGRFTISRDNAKNSLYLQMNSLRDEDTAVYYCAR
H11[15]/DP-53	EVQLVESGGGLVQPGGSLRLSCAASGFTFSS	YWMHWVRQAPGKGLVWVSRINS	DGSSTTYADSVKGRFTISRDNAKNTLYLQMNSLRAEDTAVYYCAR
DP-54	EVQLVESGGGLVQPGGSLRLSCAASGFTFSS	YWMSWVRQAPGKGLEWVANIKQ	DGSEKYYVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR

Canonical structure class 1-4

Family 3			
12-2[3]/DP-29	EVQLVESGGGLVQPGGSLRLSCAASGFTFSD	HYMDWVRQAPGKGLEWVGRTRNKANSYTTEYAASVKGRFTISRDDSKNSLYLQMNSLKTEDTAVYYCAR
VHD26[8]	EVQLLESGGGLVQPGGSLRLSCAASGFTFSD	HYMSWVRQAQGKGLELVGLIRNKANSYTTEYAASVKGRLTISREDSKNTLYLQMSSLKTEDLAVYYCAR
DP-30	EVQLVESGGGLVQPGGSLRLSCAASGFTFSD	HYMSWVRQAQGKGLELVGLIRNKANSYTTEYAASVKGRLTISREDSKNTLYLQMSSLKTEDLAVYYCAR

Table 5 *(continued)*

	1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9 1|||5||||0||||5||||0||||5||||0|ab|||5||||0||||5||||0||abc||5||||0||||5||||0||||5||||0||abc||5||||0|||||
Canonical structure class 2-1	
Family 2	
V_{II}-5b[2]	QITLKESGPTLVKPTQTLTLTCTFSGFSLSTS EWCGWIRQPPGKALEWLALIY WNDDKRYSPSLKSRLTITKDTSKNQVVLTMTNMDPVDTATYYCAHR
Family 4	
DP-67	QVQLQESGPGLVKPSETLSLTCAVSGYSISSG YYWGWIRQPPGKGLEWIGSIY HSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V_H^{SP19}/VH-JA[20]/V_H4.22[17]	QVQLQESGPGLVKPSETLSLTCTVSGYSISSG YYWGWIRQPPGKGLEWIGSIY HSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
hv4005[21]	QVQLQESGPGLVKPSQTLSLTCAVSGYSISSS NWWGWIRQPPGKGLEWIGYIY YSGSIYYNPSLKSRVTMSVDTSKNQFSLKLSSVTAVDTAVYYCAR
V_{12G}-1[18]/1.9II[3]/V_H4.13[17]/DP-68	QVQLQESGPGLVKPSDTLSLTCAVSGYSISSS NWWGWIRQPPGKGLEWIGYIY YSGSTYYNPSLKSRVTMSVDTSKNQFSLKLSSVTAVDTAVYYCAR
V_{11}[17]	QVQLQESGPGLVKPSETLSLTCVVSGGSISSS NWWSWVRQPPGKGLEWIGEIY HSGNPNYNPSLKSRVTISIDKSKNQFSLKLSSVTAADTAVYYCAR
V_H4.17[17]/V_H4.23[17]/DP-69	QVQLQESGPGLVKPSETLSLTCVVSGGSISSS NWWSWVRQPPGKGLEWIGEIY HSGSPNYNPSLKSRVTISVDKSKNQFSLKLSSVTAADTAVYYCAR
V_{79}[18]/V_H4.19[17]/V_{IV}-4b[2]	QVQLQESGPGLVKPPGTLSLTCAVSGGSISSS NWWSWVRQPPGKGLEWIGEIY HSGSTNYNPSLKSRVTISVDKSKNQFSLKLSSVTAADTAVYCCAR
DP-70	QVQLQESGPGLVKPSGTLSLTCAVSGGSISSS NWWSWVRQPPGKGLEWIGEIY HSGSTNYNPSLKSRVTISVDKSKNQFSLKLSSVTAADTAVYYCAR
Canonical structure class 3-1	
Family 2	
DP-26	QVTLKESGPVLVKPTETLTLTCTVSGFSLSNARMGVSWIRQPPGKALEWLAHIF SNDEKSYSTSLKSRLTISKDTSKSQVVLTMTNMDPVDTATYY
DP-27	QVTLRESGPALVKPTQTLTLTCTFSGFSLSTSGMCVSWIRQPPGKALEWLALID WDDDKYYSTSLKTRLTISKDTSKNQVVLTMTNMDPVDTATYY
DP-28	QVTLKESGPALVKPTQTLTLTCTFSGFSLSTSGMRVSWIRQPPGKALEWLARID WDDDKFYSTSLKTRLTISKDTSKNQVVLTMTNMDPVDTATYY
V_{II}-5[2]	QITLKESGPTLVKPTQTLTLTCTFSGFSLSTSGVGVGWIRQPPGKALEWLALIY WNDDKRYSPSLKSRLTITKDTSKNQVVLTMTNMDPVDTATYYCAHR
Family 4	
DP-64	QLQLQESGSGLVKPSQTLSLTCAVSGGSISSGGYSWSWIRQPPGKGLEWIGYIY HSGSTYYNPSLKSRVTISVDRSKNQFSLKLSSVTAADTAVYYCAR
DP-65	QVQLQESGPGLVKPSQTLSLTCTVSGGSISSGGYYWSWIRQHPGKGLEWIGYIY YSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V_{71}-2[4]/DP-66	QVQLQESGPGLVKPSETLSLTCTVSGGSVSSGSYYWSWIRQPPGKGLEWIGYIY YSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V_H4.18[17]	QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIY YSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
V_2-1[18]	QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIY YSGSTYYNPSLKSRVTISVDTSKNHFSLKLSSVTAADTAVYYCAR
Canonical structure class 3-5	
Family 6	
V_H-VI[24]/6-1G1[3]/DP-74	QVQLQQSGPGLVKPSQTLSLTCAISGDSVSSNSAAWNWIRQSPSRGLEWLGRTYYR SKWYNDYAVSVKSRITINPDTSKNQFSLQLNSVTPEDTAVYYCAR

Segments with uncertain canonical structures

Family 1			
DP-5	QVQLVQSGAEVKKPGASVKVSCKVSGYTLTE	LSMHWVRQAPGKGLEWMGGFDP	EDGETIYAQKFQGRVTMTEDTSTDTAYMELSSLRSEDTAVYYCAT
Family 3			
9-1[3]/DP-38	EVQLVESGGGLVKPGGSLRLSCAASGFTFSN	AWMSWVRQAPGKGLEWVGRIKSKTDGGTTDYAAPVKGRFTISRDDSKNTLYLQMNSLKTEDTAVYYCTT	
DP-61	TFSS	YAMHWVRQAPGKGLEYVSAISS	NGGSTYYAD

Segments with defective protein sequences

Family 1			
V35[1]/*VI-2b*[2]	QVQLVQSGAEVKKPGASVKVSCKASGYTFTG	YYMHWVRQAPGQGLEWMGRINP	NSGGTNYAQKFQGRVTSTRDTSISTAYMELSRLRSDDTVVYYCAR
Family 3			
65-4[9]/DP-39	EVQLVESGGGLVQPGGSLRLSCPASGFTFSN	HYMSWVRQAPGKGLEWVSYISG	DSGYTNYADSVKGRFTISRDNANNSPYLQMNSLRAEDTAVYYCVK
VH19[10]/DP-59	EVQLVESGGGLVQPGGSLRLSCAASGFTFSN	SDMNWVHQAPGKGLEWVSGVSW	NGSRTHYADSVKGRFIISRDNSRNTLYLQTNSLRAEDTAVYYCVR
65-2[9]/DP-44	EVQLVHSGGGLVQPGGSLRLSCAGSGFTFSS	YAMHWVRQAPGKGLEWVSAIG	TGGGTYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDMAVYYCAR
DP-52	EDQLVESGGGLVQPGGSLRPSCAASGFAFSS	YVLHWVRRAPGKGPEWVSAIG	TGGDTYYADSVMGRFTISRDNAKKSLYLQMNSLIAEDMAVYYCAR

The structural classification of the protein sequences is described in the text and illustrated in Figs 4 to 8. DP sequences are taken from Tomlinson *et al.* (1992). Previously published genes are shown in italics and suffixed according to source: [1]Matsuda *et al.* (1988); [2]Shin *et al.* (1991); [3]Berman *et al.* (1988); [4]Kodaira *et al.* (1986); [5]Chen *et al.* (1989); [6]Friedman *et al.* (1991); [7]Rechavi *et al.* (1983); [8]Buluwella *et al.* (1988); [9]Matsuda *et al.* (1990); [10]Baer *et al.* (1988); [11]Olee *et al.* (1991); [12]Chen (1990); [13]Pascual *et al.* (1990); [14]Mathyssens & Rabbitts (1980), corrected by Chen *et al.* (1988); [15]Rechavi *et al.* (1982); [16]van Es *et al.* (1991); [17]Sanz *et al.* (1989); [18]Lee *et al.* (1987); [19]Baer *et al.* (1985); [20]Denny *et al.* (1986); [21]Chen & Yang (1990); [22]Shen *et al.* (1987), corrected by Sanz *et al.*, 1989); [23]Humphries *et al.* (1988), corrected by Sanz *et al.* (1989); [24]Buluwela & Rabbitts (1988).

loop relative to the framework is mainly determined by the size of the residue at this site.

Canonical structures 2 and 3 are found in six-residue H2 regions. Structure 2 occurs when residues 52a and 71 are small or medium sized hydrophobic residues. Structure 3 is found when residue 71 is Arg or Lys. Both structures have residues with positive ϕ,ψ values: residue 55 in the case of structure 2 and 54 in the case of structure 3. Positive ϕ,ψ values are allowed for Gly, Asn and Asp and partially allowed for other residues. Inspection of the known structures shows that though the two conformations often have Gly at positions 55 or 54, other residues can occur, e.g. Ser (Tramontano *et al.*, 1990).

Canonical structure 4 is found in eight-residue H2 regions. The main determinants of this structure are Tyr at positions 55 and Arg at position 71. Again 54 has positive ϕ,ψ values and Gly does occur at this position but is not a requirement.

(b) *H2 structures implicit in the sequences of the V_H segments*

Inspection of the H2 regions in the 77 potentially functional V_H segments shows that 74 have a size, and the residues at the key sites, that corresponds to one of the four known canonical structures. In Table 4 we list the residues found at the key sites in 74 sequences.

The three sequences that did not fit the size and/or the sequence requirements of the known canonical structures are DP-5, 9-1, and V_H-VI (see Table 5 below). V_H-VI is the single member of family 6 and is found in expressed sequences. Neither the size nor the sequence of its H2 region correspond to the requirements of the known H2 canonical structures. It clearly has its own distinct conformation and we will call this canonical structure 5.

The H2 regions in DP-5 and 9-1 correspond in size and sequence to canonical structures 2 or 3 and 4, respectively, except for the residue at one key site. DP-5 has Glu at position 71 and 9-1 has Gly at position 55. It is difficult to predict the effect of these residues in the loop conformations.

5. Structural Classification of the V_H Segments

In the previous sections of the paper we have described (1) the sequences that contain residues that are likely to result in defective three-dimensional structures; (2) the number of the residue identities shared by the different sequences, and (3) the canonical structures that are expected to be present in the first and second hypervariable regions. In Table 5 these features of the sequences are put together to give a structural classification of the V_H segments.

Of the currently known V_H segments with open reading frames, 74 have sequences that correspond to the requirements of known canonical structures in both the H1 and H2 regions, three have sequences that do not correspond to a known H2 structure and five have sequences that are unlikely to give a stable structure. For one segment, only half the sequence is known (Table 5).

Sequences that have the same canonical structure for both H1 and H2 can be grouped together into "canonical structure classes" (Table 5). These classes are numbered in the form N-M where N is the number of the H1 canonical structure and M the number of the H2 structure. The 74 sequences fall into six canonical structure classes 1-1, 1-2, 1-3, 1-4, 2-1 and 3-1. The total number of sequences in each class is 11, 13, 29, 3, 9 and 9, respectively. V_H-VI, the sole segment in family 6, will give the seventh structure, class 3-5, though the conformation of its H2 region is unknown at present.

6. Outline Structures of the Binding Site Loops Formed by the V_H Segments

The knowledge of the canonical structures produced by the V_H segments means that we can describe the structures of their antigen binding site loops at least in outline. In Figures 4 to 8 we show, for different canonical structure classes, schematic diagrams of the arrangement of the V_H segment residues that form the antigen binding sites. The small variations in conformation that occur within canonical structure classes are discussed below.

7. Discussion: the Structural Basis of Antibody Specificity

The analysis presented in the previous sections implies that almost all of the known V_H segments produce structures with one of seven main-chain folds. These results have implications for the mechanisms by which sequence variations at hypervariable sites, key sites and in the framework determine antibody specificity.

(a) *Sequence variations at the hypervariable sites*

For each canonical structure class we determined the extent of the sequence variability in the regions of the binding site. The results of these calculations are given in Table 6 and clearly show that the sites in these regions differ greatly in the extent of their variability.

The H2 regions have more variability than the H1. (This is also true if the calculations are made for sequences within V_H gene families rather than within canonical structure classes, see Tomlinson *et al.* (1992).) The sites in H2 that generally have the greatest variability are 50, 52 and 53 (Table 6). Within H1, the most variable site is 33.

Inspection of the structure of the antigen binding sites shows that residues 50, 52 and 53 are adjacent to 33 which, in turn, is adjacent to H3 and L3. This means that, for sequences that belong to the same canonical structure class, the residues with the greatest variability are those that form the centre of the antigen binding site (Fig. 9).

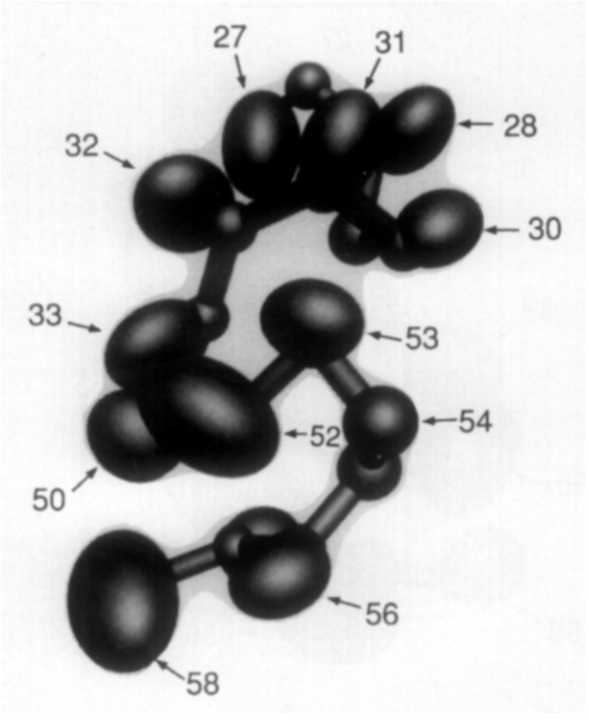


Figure 4. Schematic drawing of the structure of the H1 and H2 regions formed by V_H gene segments in canonical structure class 1-1. A tracing of the main-chain is given as rods joining C^{α} atoms. Side-chains are shown as ellipsoids. This schematic Figure should be compared with the space filling atomic structure shown in Fig. 1(b). The ellipsoids shown here have been fitted to the side-chain of the residue that is either the most common to occur at the site or, where the range is wide, it is one of average size. Their conformations have been taken from known structures in almost all cases. Alternative side-chains and conformations would modify somewhat the shape and orientation of the ellipsoids. For sequences in canonical structure class 1-1, the residues in the binding sites and hypervariable regions and, marked by *, the residues at the key sites are:

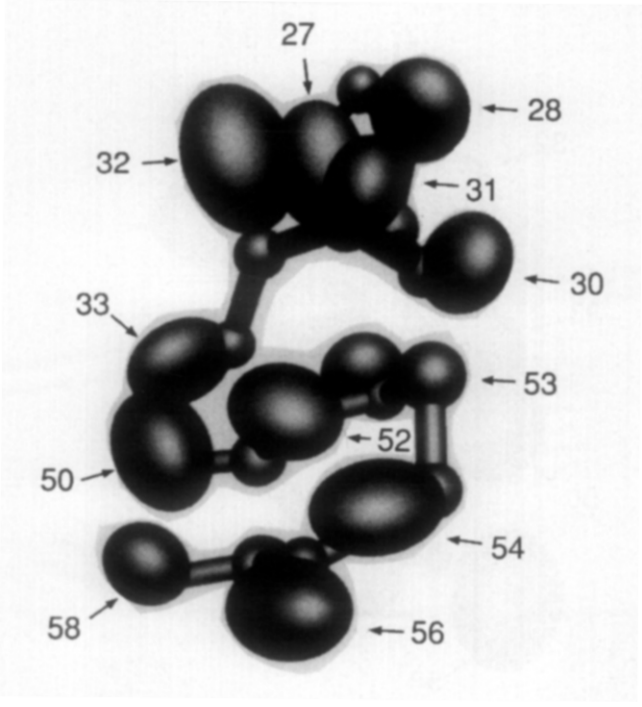
	H1			H2	
	2 3	2	9	5 5 5	7
	6----------5	4	4	0-2abc-----8	1
Family 3					
	** * *	*	*	*	*
DP-42	GFTVSS--NYMS	A	R	VIY---SGGSTY	R
8-1B	GFTVSS--NYMS	A	R	VIY---SGGSTY	R
DP-45	GFTFSS--YAMH	G	R	AIG---TGGGTY	R
13-2	GFTFSS--YDMH	A	R	AIG---TAGDTY	R
Family 4					
	** * *	*	*	*	*
Tou-VH4.21	GGSFSG--YYWS	V	R	EII---HSGSTN	V
VH5	GGSFSG--YYWS	V	R	EIN---HSGSTN	V
V58	GGSVSG--YYWS	V	R	YIY---YSGSTN	V
VIV-4	GGSISS--YYWS	V	R	RIY---TSGSTN	V
VH4.11	GGSISS--YYWS	V	R	YIY---YSGSTN	V
V71-4	GGSISS--YYWS	V	R	YIY---YSGSTN	V
VH4.16	GGSVSS--YYWS	V	R	YIY---YSGSTN	V

Full sequences are given in Table 5.

(b) *Sequence variations at the key sites and within the framework*

Sequence differences also occur at the key sites and in the framework regions. The extent of these variations within the different canonical structure classes are given in Table 7.

For a given canonical structure, some key sites require a particular residue. Other sites allow a range of residues, for example a large or medium-



1-2

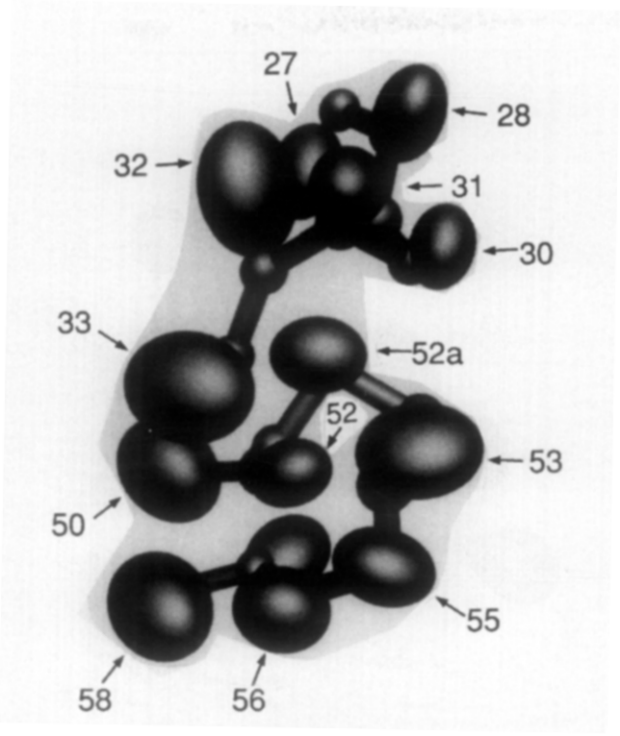
Figure 5. Schematic drawing of the structure of the H1 and H2 regions formed by V_H gene segments in canonical structure class 1-2, see the legend to Fig. 4 for details of the drawing. For sequences in canonical structure class 1-2, the residues in the binding sites and hypervariable regions and, marked by *, the residues at the key sites are:

	H1			H2	
	2 3	2	9	5 5 5	7
	6----------5	4	4	0-2abc-----8	1
Family 1					
	** * *	*	*	* *	*
DP-3	GYTFTD--YYMH	V	T	LVDP--EDGETI	A
DP-10	GGTFSS--YAIS	A	R	GIIP--IFGTAN	A
hv1263	GGTFSS--YAIS	A	R	RIIP--ILGIAN	A
DP-21	GYTFTS--YAMN	A	R	WINT--NTGNPT	L
VI-4.1b	GYTFTS--YAMN	A	R	WINT--NTGNPT	L
DP-14	GYTFTS--YGIS	A	R	WISA--YNGNTN	T
VH1GRR	GYTFTS--YGIS	A	-	WISA--YNGNTN	T
Family 5					
	** * *	*	*	* *	*
VH251	GYSFTS--YWIG	G	R	IIYP--GDSDTR	A
VHVJB	GYSFTS--YWIG	G	R	IIYP--GDSDTR	A
VHVCW	GYSFTS--YWIG	G	R	IIYP--GDSDTR	A
1-v	GYSFTS--YWIH	T	R	SIYP--GNSDTR	A
VH32	GYSFTS--YWIS	G	R	RIDP--SDSYTN	A
VHVRG	GYSFTS--YWIS	G	R	RIDP--SDSYTN	A

Full sequences are given in Table 5.

sized hydrophobic residue. The variation at these latter sites can be systematic in some cases. For example, sequences in canonical class 1-1 come from families 3 and 4. At sites 24, 27, 34 and 71 the sequences from family 3 have Ala/Gly, Phe, Met and Arg; those from family 4 have Val, Gly, Trp and Val. A converse example is canonical structure class 1-3. Although its sequences are found in two families, 1 and 3, they have no systematic differences.

In different antibodies canonical structures with identical key residues have very similar local conformations. If atomic structures determined at high resolution are compared, the r.m.s. difference in the local position of their main-chain atoms are 0·2 to 0·5 Å (1 Å = 0·1 nm). Variations in residues at key sites produce small changes and give r.m.s. differences closer to 1 Å (Chothia & Lesk, 1987; Chothia *et al.*, 1989; and our unpublished results).



1-3

Figure 6. Schematic drawing of the structure of the H1 and H2 regions formed by V_H gene segments in canonical structure class 1-3, see the legend to Fig. 4 for details of the drawing. For sequences in canonical structure class 1-3, the residues in the binding sites and hypervariable regions and, marked by *, the residues at the key sites are:

	H1			H2	
	2 3	2	9	5 5 5	7
	6----------5	4	4	0-2abc-----8	1
Family 1					
	** * *	*	*	*	*
V71-5	GFTFTS--SAVQ	A	A	WIVV--GSGNTN	R
DP-1	GYIFTD--YYMH	A	R	RINP--NSGGTN	R
VI-2	GYTFTG--YYMH	A	R	WINP--NSGGTN	R
DP-8	GYTFTG--YYMH	A	R	WINP--NSGGTN	R
1-1	GYTFTG--YYMH	A	R	WINP--NSGGTN	R
DP-12	GYTFTN--YCMH	A	R	LVCP--SDGSTS	R
VI-3b	GYTFTS--YAMH	A	R	WINA--GNGNTK	R
VI-3	GYTFTS--YAMH	A	R	WSNA--GNGNTK	R
DP-15	GYTFTS--YDIN	A	R	WMNP--NSGNTG	R
21-2	GYTFTS--YYMH	A	R	IINP--SGGSTS	R
HG3	GYTFNS--YYMH	A	R	IINP--SGGSTS	R
7-2	GYTFTY--RYLH	A	R	WITP--FNGNTN	R
DP-4	GYTFTY--RYLH	A	R	WITP--FNGNTN	R
Family 3					
	** * *	*	*	*	*
DP-31	GFTFDD--YAMH	A	K	GISW--NSGSIG	R
DP-32	GFTFDD--YGMS	A	R	GINW--NGGSTG	R
DP-33	GFTFDD--YTMH	A	K	LISW--DGGSTY	R
22-2B	GFTFSD--YYMS	A	R	YISS--SGSTIY	R
15-2B	GFTFSN--HYTS	A	K	YSSG--NSGYTN	R
f1-p1	GFTFSS--YAMH	A	R	AISS--NGGSTY	R
hv3005	GFTFSS--YAMH	A	R	VISY--DGSNKY	R
hv3005f3	GFTFSS--YAMH	A	R	VISY--DGSNKY	R
GL-SJ2	GFTFSS--YAMH	A	R	VISY--DGSNKY	R
VH26	GFTFSS--YAMS	A	K	AISG--SGGSTY	R
DP-58	GFTFSS--YEMN	A	R	YISS--SGSTIY	R
1.9III	GFTFSS--YGMH	A	K	VISY--DGSNKY	R
3019b9	GFTFSS--YGMH	A	R	VIWY--DGSNKY	R
DP-51	GFTFSS--YSMN	A	R	YISS--SSSTIY	R
H11	GFTFSS--YWMH	A	R	RINS--DGSSTT	R
DP-54	GFTFSS--YWMS	A	R	NIKQ--DGSEKY	R

Full sequences are given in Table 5.

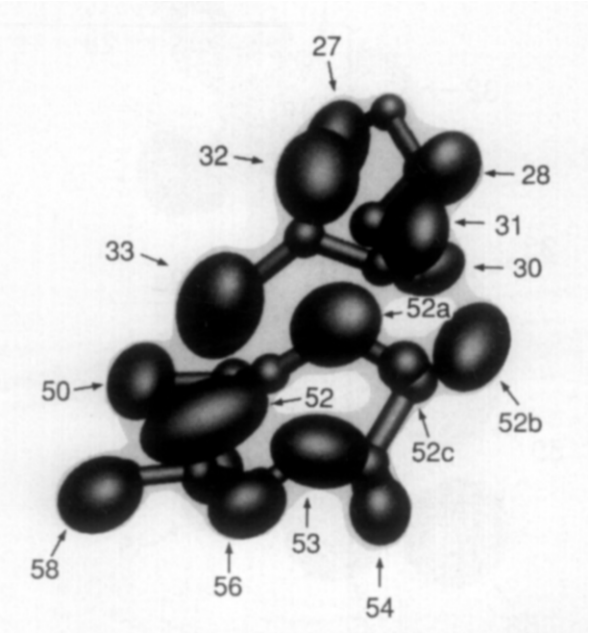


Figure 7. Schematic drawing of the structure of the H1 and H2 regions formed by V_H gene segments in canonical structure class 1-4, see the legend to Fig. 4 for details of this drawing. For sequences in canonical structure class 1-4, the residues in the binding sites and hypervariable regions and, marked by *, the residues at the key sites are:

	H1			H2	
	2 3	2	9	5 5 5	7
	6----------5	4	4	0-2abc-----8	1
Family 3	** * *	*	*	**	*
12-2	GFTFSD--HYMD	A	R	RTRNKANSYTTE	R
VHD26	GFTFSD--HYMS	A	R	LIRNKANSYTTE	R
DP-30	GFTFSD--HYMS	A	R	LIRNKANSYTTE	R

Full sequences are given in Table 5.

All but two of the canonical structure classes contain sequences from two different families (Table 5). This means that, because of the low residue identities between sequences in different families (Table 2), V_H segments in the same canonical structure class can have sequences that differ by up to about 50% (Table 7).

Within the larger canonical structure classes, 1-1, 1-2, 1-3 and 2-1, about half the framework residues can change their identity (Table 7). The direct relation between sequence and structure divergence (Chothia & Lesk, 1986) means that, although the same canonical structures in different antibodies have very similar local conformations, their posi-

Table 6

Variability of binding site residue in the canonical structure classes

Canonical structure class	Number of segments	Sites																			
		30	31	31a	31b	32	33	34	35	50	51	52	52a	52b	52c	53	54	55	56	57	58
1-1	11	1·0	2·8			2·4	3·7	3·1	2·4	13·8	1·0	6·3				11·0	4·7	1·0	3·7	1·0	3·1
1-2	13	2·4	2·2			1·0	8·7	2·6	8·7	19·5	2·2	16·3	4·3			19·5	10·8	3·7	19·5	4·3	8·7
1-3	29	8·9	8·5			4·6	26·1	6·0	5·8	29·0	4·6	16·9	23·2			16·1	7·7	3·1	15·8	4·6	14·5
1-4	3	1·0	1·0			1·0	1·0	1·0	3·0	3·0	3·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0
2-1	9	1·0	2·3	2·6		4·5	2·6	2·3	3·6	9·0	1·0	1·0				4·5	2·3	2·3	3·9	7·2	6·8
3-1	9	1·0	5·4	5·4	5·4	5·4	11·3	3·6	3·0	15·0	1·0	4·5				9·0	5·4	3·6	5·4	3·6	9·0
3-5	1	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0	1·0		1·0	1·0	1·0	1·0	1·0	1·0

The variability at a site is the number of different amino acids that occur at that site divided by the frequency of the most common amino acid at that site (Kabat *et al.*, 1991). The amino acid sequences of segments in the canonical structure classes are given in Table 5.

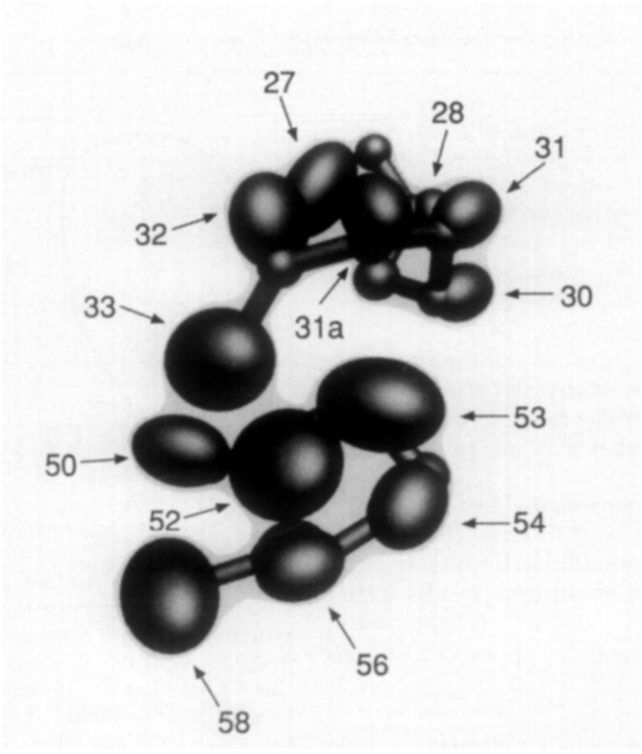


Figure 8. Schematic drawing of the structure of the H1 and H2 regions formed by V_H gene segments in canonical structure class 2-1, see the legend to Fig. 4 for details of this drawing. Sequences in canonical structure class 3-1 will have a similar fold except for the insertion of residue 31b in the region of 30-31-31a; see the text. For sequences in canonical structure classes 2-1 and 3-1, the residues in the binding sites and hypervariable regions and, marked by *, the residues at the key sites are:

	H1			H2	
	2 3 3	2	9	5 5 5	7
	6----1ab---5	4	4	0-2abc-----8	1
Canonical structure class 2-1					
Family 2					
	** * *	*	*	*	*
VII-5b	GFSLSTS-EWCG	F	H	LIY---WDDDKR	K
Family 4					
	** * *	*	*	*	*
DP-67	GYSISSG-YYWG	V	R	SIY---HSGSTY	V
VHSP	GYSISSG-YYWG	V	R	SIY---HSGSTY	V
hv4005	GYSISSS-NWWG	V	R	YIY---YSGSIY	V
V12G-1	GYSISSS-NWWG	V	R	YIY---YSGSTY	V
VH4.17	GGSISSS-NWWS	V	R	EIY---HSGSPN	V
VH79	GGSISSS-NWWS	V	R	EIY---HSGSTN	V
DP-70	GGSISSS-NWWS	V	R	EIY---HSGSTN	V
V11	GGSISSS-NWWS	V	R	EIY---HSGNPN	V
Canonical structure 3-1					
Family 2					
	** * *	*	*	*	*
DP-26	GFSLSNARMGVS	V	R	HIF---SNDEKS	K
DP-27	GFSLSTSGMCVS	F	R	LID---WDDDKY	K
DP-28	GFSLSTSGMRVS	F	R	RID---WDDDKF	K
VII-5	GFSLSTSGVGVG	F	H	LIY---WNDDKR	K
Family 4					
	** * *	*	*	*	*
DP-64	GGSISSGGYSWS	V	R	YIY---HSGSTY	V
DP-65	GGSISSGGYYWS	V	R	YIY---YSGSTY	V
V71-2	GGSVSSGSYYWS	V	R	YIY---YSGSTN	V
V2-1	GGSISSSSYYWG	V	R	SIY---YSGSTY	V
VH4.18	GGSISSSSYYWG	V	R	SIY---YSGSTY	V

Full sequences are given in Table 5.

tions relative to the framework and to each other can vary. These differences in position are in the range 2 to 4 Å. They are produced by the net effect of the differences in the identities of the residues at the key sites, at neighbouring positions in the framework and at the V_L–V_H interface (Chothia & Lesk, 1987; Lascombe *et al.*, 1989; Chothia *et al.*, 1989; and our unpublished results).

Shifts in the relative positions of the binding site loops affect more than the static structure of the binding site. They also alter the range of low energy conformational changes that it may use to facilitate antigen binding.

Protein engineering experiments by Foote & Winter (1992) showed that conservative changes in residues at key sites changed affinity by factors of 3 to 10. Similar though more qualitative results were obtained by Reichmann *et al.* (1988) and Kettleborough *et al.* (1991).

(c) *The extent of the human V_H germline segment repertoire*

The extent to which structures described in this paper form a significant proportion of the total structural repertoire depends, of course, upon the extent to which the currently known V_H segments describe the total human repertoire.

The detailed investigation of the V_H segments in DP by Tomlinson *et al.* (1992) produced 51 sequences with open reading frames. In this paper we showed that four DP sequences with open reading frames have residues that probably prevent the formation of a stable three-dimensional structure. For family 2 it was subsequently found that the primers used by Tomlinson *et al.* (1992) would not have amplified sequences related to $V_{II\text{-}5}$ and $V_{II\text{-}5b}$ in family 2. But the small size of this family would imply that there is only a small number of such sequences. Thus, the results obtained from the determination of the DP V_H gene segments probably give a close to complete picture of a human functional repertoire and suggest that it consists of about 50 V_H segments (Tomlinson *et al.*, 1992).

If this is the case we would expect that the genes in other individuals will differ only by polymorphisms due to mutation or deletion/insertion. Prior to the work on the DP genes, sequences were known for 55 V_H genes that had open reading frames. Of these, 23 are identical in nucleotide sequence to those found in DP (Tomlinson *et al.*, 1992). For the other 32, we find that 21 differ by between zero and four residues from the closest DP sequence and nine by between five and 14 residues. These sequence differences are small and are consistent with, though do not prove, that the bulk of the known V_H germline genes not found in DP are mutation polymorphs of those that are found in DP.

The view that the currently known V_H gene segments form a high proportion of the total human repertoire is strongly supported by the analysis of the 292 known rearranged genes made by Tomlinson *et al.* (1992). None of these comes from the DP source and very few come from the same source as the other known germline genes. Of the 292 rearranged genes, 215 are derived from V_H segments identical to those found in DP and a

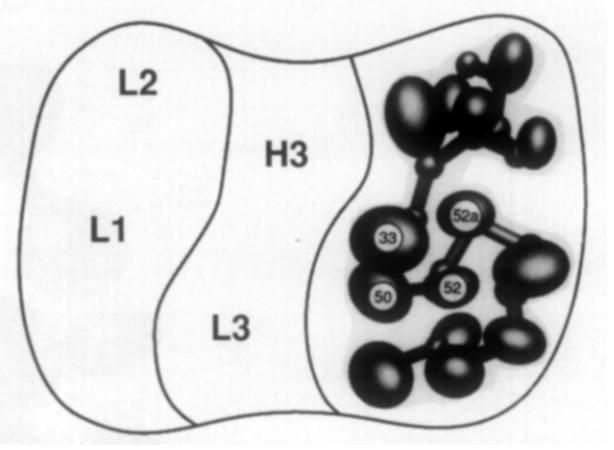


Figure 9. The relative position of the most variable sites in the V_H segments. For the V_H segments in canonical structure class 1-3 the most variable sites are 33, 50, 52 and 52a; see Table 6. These sites cluster together next to the regions that are the most variable of all: H3 and L3 (see Kabat *et al.*, 1991). A very similar picture is given by the other large canonical structure sets; see Table 6 and Figs 4 to 8.

Table 7

Sequence variations within the canonical structure sets and at key sites

Canonical structure class	Number of segments	Range of sequence identities	Residues at key sites					
			24	27	29	34	71	94
1-1	11	45-97	GAV	FG	VIF	MW	RV	R
1-2	13	54-97	GATV	YG	F	MI	ATL	RT
1-3	29	44-98	A	YF	F	VLIM	R	RKA
1-4	3	89-99	A	F	F	M	R	R
2-1	9	84-98 (54-59)	V(F)	GY(F)	I(L)	W(C)	VI(K)	R(H)
3-1	9	55-98	VF	FG	VLI	VW	KV	RH
3-5	1	—	I	D	V	W	P	R

Note: Canonical structure class 2-1 contains 7 very similar sequences from family 4 and one, VII-5b, from family 2. The data for VII-5b are given in parentheses.

further 53 are derived from other known V_H segments. Inspection of the remaining 24 sequences suggests that they are derived from just a small number of V_H segments (Tomlinson *et al.*, 1992). The size of their H1 and H2 regions and the residues at the key sites shows that these sequences belong to the canonical structure classes described here.

(d) *Canonical structures in human expressed V_H segments*

This paper is concerned with the structural repertoire of the human V_H germline segments. This work does raise, however, the question of whether the same repertoire is found in expressed genes whose sequences have been changed by somatic mutation.

We inspected the sequences of the rearranged genes to determine the extent and nature of the mutations that have occurred at the positions of the key residues: i.e. at positions 24, 26, 27, 29, 34, 52a, 54, 55, 71 and 94 (see Tables 3 and 4). Of the 268 sequences that are derived from the germline sequences discussed here, 184 (70%) had no mutations at these sites, 63 (25%) had one mutation, 19 had two mutations, one had three mutations and one had four.

The vast majority of these mutations are to residues consistent with the canonical structure given by the germline sequence. For example the mutations at site 34 are Met to Val, Ile, Leu or Phe and, with one exception, those at site 29 are Phe to Leu, Leu to Val, and Ile to Val or Leu. Such mutations will modify the position but not the conformation of the canonical structures (see above).

There are five cases where a somatic mutation is likely to have a disruptive effect on the germline canonical structure. These involve the mutations Ala24 to Pro, Phe29 to Gly, Leu29 to Ser, Phe29 to Thr and Arg71 to Met. There are also seven cases where Gly26 is mutated to a non-glycine residue. This is likely to induce a small amount of steric strain.

Thus, almost all the currently known human expressed V_H segments have canonical structures that are the same as those in the germline gene from which they are derived, though the relative positions of these canonical structures may have been modified by somatic mutations.

8. Conclusion

The analysis presented in the previous sections implies that almost all of an individual's repertoire of V_H segments produce structures with one of seven main-chain folds. The folds for canonical structure classes 1-1, 1-2, 1-3, 1-4 and 2-1 are known from the structural data currently available. For canonical structure classes 3-1 and 3-5 the fold of H1 is known only approximately and for 3-5, the H2 conformation is unknown.

Sequence variations in the binding site residues, particularly those that form its centre, modulate the surface that the canonical structures present to antigens. Sequence variations at the key sites and in the framework change the relative positions of the canonical structures and the range of low energy conformational changes that may be used for antigen binding.

A.M.L. thanks the Kay Kendall Foundation for support. I.M.T. and G. Walter were supported by the Medical Research Council Human Genome Mapping Project, J.D.M. by the Medical Research Council Aids Directed Programme and M.B.L. by a Medical Research Council/Celltech Ltd. training fellowship.

References

Baer, R., Chen, K. C., Smith, S. D. & Rabbitts, T. H. (1985). Fusion of an immunoglobulin variable gene and a T cell receptor constant gene in the chromosomal 14 inversion associated with T cell tumors. *Cell*, **43**, 705–713.

Baer, R., Forester, A., Lavenir, I. & Rabbitts, T. H. (1988). Immunoglobulin V_H genes are transcribed by T cells in association with a new 5′ exon. *J. Exp. Med.* **167**, 2011–2016.

Beale, D. & Coadwell, J. (1989*a*). Some observations on conserved polar side chains in immunoglobulin V-domains. *Int. J. Biochem.* **21**, 227–232.

Beale, D. & Coadwell, J. (1989*b*). Some observations on the replacement of the conserved amino acid residues of immunoglobulin domains. *Int. J. Biochem.* **21**, 1033–1037.

Berman, J. E., Melliks, S. J., Pollock, R., Smith, C. L., Suh, H., Heinke, B., Kowal, C., Surti, U., Cantor, C. R. & Alt, F. W. (1988). Content and organization of the human Ig V_H locus: definition of three new V_H families and linkage to the Ig C_H locus. *EMBO J.* **7**, 727–738.

Brunger, A. T., Leahy, D. J., Hynes, T. R. & Fox, R. O. (1991). 2·9 Å resolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody Fab fragment with bound hapten. *J. Mol. Biol.* **221**, 239–256.

Buluwela, L. & Rabbitts, T. H. (1988). A V_H gene is located within 95 kb of the human immunoglobulin heavy chain constant region genes. *Eur. J. Immunol.* **18**, 1843–1845.

Buluwela, L., Albertson, D. G., Sherrington, P., Rabbitts, P. H., Spurr, N. & Rabbitts, T. H. (1988). The use of chromosomal translocations to study human immunoglobulin gene organization: mapping D_H segments within 35 kb of the C_γ gene and the identification of a new D_H locus. *EMBO J.* **7**, 2003–2010.

Chen, P. P. (1990). Structural analyses of human developmentally regulated V_H 3 genes. *Scand. J. Immunol.* **31**, 257–267.

Chen, P. P. & Yang, P. M. (1990). A segment of the human V_H gene locus is duplicated. *Scand. J. Immunol.* **31**, 593–599.

Chen, P. P., Liu, M. F., Sinha, S. & Carson, D. A. (1988). A16/6 idiotype-positive anti-DNA antibody is encoded by a conserved V_H gene with no somatic mutation. *Arthritis Rheum.* **31**, 1429–1431.

Chen, P. P., Leu, M. F., Glass, C. A., Sinha, S., Kipps, T. J. & Carson, D. A. (1989). Characterization of two immunoglobulin V_H genes that are homologous to human rheumatoid factors. *Arthritis Rheum.* **32**, 72–76.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Biol. Chem.* **196**, 901–917.

Chothia, C., Lesk, A. M. Levitt, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986). The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science*, **233**, 755–758.

Chothia, C., Boswell, R. & Lesk, A. M. (1988). The outline structure of the T cell $\alpha\beta$ 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 (London)*, **342**, 877–883.

Cleary, M. L., Meeker, T. C., Levy, S., Lee, E., Trela, M., Sklar, J. & Levy, R. (1986). Clustering of extensive somatic mutations in the variable region of an immunoglobulin heavy chain gene from a human B cell lymphoma. *Cell*, **44**, 97–106.

Davies, D. R. & Metzger, H. (1983). The structural basis of antibody function. *Annu. Rev. Immunol.* **1**, 87–117.

Denny, C. T., Yoshikai, Y., Mak, T. W., Smith, S. D., Hollis, G. F. & Kirsch, I. R. (1986). A chromosome 14 inversion in a T cell lymphoma is caused by site specific recombination between immunoglobulin and T cell receptor loci. *Nature (London)*, **320**, 549–551.

Fischmann, T. O., Bentley, G. A., Bhat, T. N., Boulot, G., Mariuzza, R. A., Phillips, S. E. V., Tello, D. & Poljak, R. J. (1991). Crystallographic refinement of the three-dimensional structure of the Fab D1.3-lysozyme complex at 2·5 Å resolution. *J. Biol. Chem.* **266**, 12915–12920.

Foote, J. & Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *J. Mol. Biol.* **224**, 487–499.

Friedman, D. F., Cho, E. A., Goldman, J., Carmack, C. E., Besa, E. C., Hardy, R. R. & Silberstein, L. E. (1991). The role of clonal selection in the pathogenesis of an autoreactive human B cell lymphoma. *J. Exp. Med.* **174**, 525–537.

Humphries, C. G., Shen, A., Kuziel, W. A., Capra, J. D., Blattner, F. R. & Tucker, P. W. (1988). A new human immunoglobulin V_H family preferentially rearranged in immature B cell tumours. *Nature (London)*, **331**, 446–449.

Kabat, E. A. & Wu, T. T. (1971). Attempts to locate complementarity determining residues in the variable positions of light and heavy chains. *Ann. N. Y. Acad. Sci.* **190**, 382–393.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*, 5th edit., Public Health Service, N.I.H. Washington, DC.

Kettleborough, C. A., Saldanha, J., Heath, V. J., Morrison, C. J. & Bendig, M. M. (1991). Humanization of a mouse monoclonal antibody by CDR-grafting: the importance of framework residues on loop conformation. *Protein Eng.* **4**, 773–783.

Kodaira, M., Kinashi, T., Umemura, I., Matsuda, F., Norma, T., Ono, Y. & Honjo, T. (1986). Organization and evolution of variable region genes of the human immunoglobulin heavy chain. *J. Mol. Biol.* **190**, 529–541.

Kon, S., Levy, S. & Levy, R. (1987). Retention of an idiotypic determinant in a human B cell lymphoma undergoing immunoglobulin variable region mutation. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 5053–5057.

Lascombe, M.-B., Alzari, P. M., Boulot, G., Saludjian, P., Tougard, P., Berek, C., Haba, S., Rosen, E. M., Nisonoff, A. & Poljak, R. J. (1989). Three-dimensional structure of Fab R19.9, a monoclonal murine antibody specific for the p-azobenzene-arsonate group. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 607–611.

Lee, K. H., Matsuda, F., Kinashi, T., Kodaira, M. & Honjo, T. (1987). A novel family of variable region genes of the human immunoglobulin heavy chain. *J. Mol. Biol.* **195**, 761–768.

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

Mathyssens, G. & Rabbitts, T. H. (1980). Structure and multiplicity of genes for the human immunoglobulin heavy chain variable region. *Proc. Nat. Acad. Sci., U.S.A.* **77**, 6561–6565.

Matsuda, F., Lee, K. H., Nakai, S., Sato, T., Kodaira, M., Zong, S. Q., Ohno, H., Fukuhara, S. & Honjo, T. (1988). Dispersed localization of D segments in the human immunoglobulin heavy chain locus. *EMBO J.* **7**, 1047–1051.

Matsuda, F., Shin, E. K., Hirabayashi, Y., Nagaoka, H., Yoshida, M. C., Zong, S. Q. & Honjo, T. (1990). Organization of variable region segments of the human immunoglobulin heavy chain: duplication of the D5 cluster within the locus and interchromosomal translocation of variable region segments. *EMBO J.* **9**, 2501–2506.

Mensink, E. J., Schuurman, R. K., Schot, J. D., Thompson, A. & Alt, F. W. (1986). Immunoglobulin heavy chain gene rearrangements in X-linked agammaglobulinemia. *Eur. J. Immunol.* **16**, 963–967.

Olee, T., Yang, P. M., Siminovitch, K. A., Olsen, N. J., Hillson, J., Wu, J., Kosin, F., Carson, D. A. & Chen, P. P. (1991). Molecular basis of an autoantibody associated restriction fragment length polymorphism that confers susceptibility to autoimmune diseases. *J. Clin. Invest.* **88**, 193–203.

Padlan, E. A. (1977). Structural implications of sequence variability in immunoglobulins. *Proc. Nat. Acad. Sci., U.S.A.* **74**, 2551–2555.

Padlan, E. A., Silverton, E. W., Sheriff, S., Cohen, G. H., Smith-Gill, S. J. & Davies, D. R. (1989). Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 5938–5942.

Pascual, V., Randen, I., Thompson, K., Sioud, M., Forre, O., Natvig, J. & Capra, J. D. (1990). The complete nucleotide sequence of the heavy chain variable regions of six monospecific rheumatoid factors derived from Epstein–Barr virus transformed B cells isolated from the synovial tissue of patients with rheumatoid arthritis. Further evidence that some autoantibodies are unmutated copies of germ line genes. *J. Clin. Invest.* **86**, 1320–1328.

Rechavi, G., Bienz, B., Ram, D., Ben, N. Y., Cohen, J. B., Zakut, R. & Givol, D. (1982). Organisation and evolution of immunoglobulin V_H gene subgroups. *Proc. Nat. Acad. Sci., U.S.A.* **79**, 4405–4409.

Rechavi, G., Ram, D., Glazer, L., Zakut, R. & Givol, D. (1983). Evolutionary aspects of immunoglobulin heavy chain variable region (V_H) gene subgroups. *Proc. Nat. Acad. Sci., U.S.A.* **80**, 855–859.

Riechmann, L., Clark, M., Waldmann, H. & Winter, G. (1988). Reshaping human antibodies for therapy. *Nature (London)*, **332**, 323–327.

Sanz, I., Kelly, P., Williams, C., Scholl, S., Tucker, P. & Capra, J. D. (1989). The smaller human V_H gene families display remarkably little polymorphism. *EMBO J.* **8**, 3741–3748.

Saul, F. A., Amzel, L. M. & Poljak, R. J. (1978). Preliminary refinement and structural analysis of the Fab fragment from human immunoglobulin New at 2·0 Å resolution. *J. Biol. Chem.* **253**, 585–597.

Shen, A., Humphries, C., Tucker, P. & Blattner, F. (1987). Human heavy chain variable region gene family nonrandomly rearranged in familial chronic lymphocytic leukemia. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 8563–8567.

Shin, E. K., Matsuda, F., Nagaoka, H., Fukita, Y., Imai, T., Yokoyama, K., Soeda, E. & Honjo, T. (1991). Physical map of the 3′ region of the human immunoglobulin heavy chain locus: clustering of autoantibody related variable segments in one halotype. *EMBO J.* **10**, 3641–3645.

Sibanda, B. L., Blundell, T. L. & Thorton, J. M. (1989). Conformation of β-hairpins in protein structures. *J. Mol. Biol.* **206**, 759–777.

Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). The repertoire of human germline V_H segments reveals fifty groups of V_H segments with different hypervariable loops. *J. Mol. Biol.* **227**, 776–798.

Tramontano, A., Chothia, C. & Lesk, A. M. (1990). Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. *J. Mol. Biol.* **215**, 175–182.

van Es, J. H., Meyling, F. H. J. G., van de Akker, W. R. M., Aanstoot, H., Derksen, R. H. W. M. & Logtenberg, T. (1991). Somatic mutations in the variable regions of a human IgG anti-double-strand DNA antibody suggests a role for antigen in the induction of systemic lupus erythematosus. *J. Exp. Med.* **173**, 461–470.

Edited by J. Karn