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

Canonical Structures for the Hypervariable Regions of Immunoglobulins

Cyrus Chothia and Arthur M. Lesk

J. Mol. Biol. (1987) 196, 901–917

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.

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

† 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.

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)

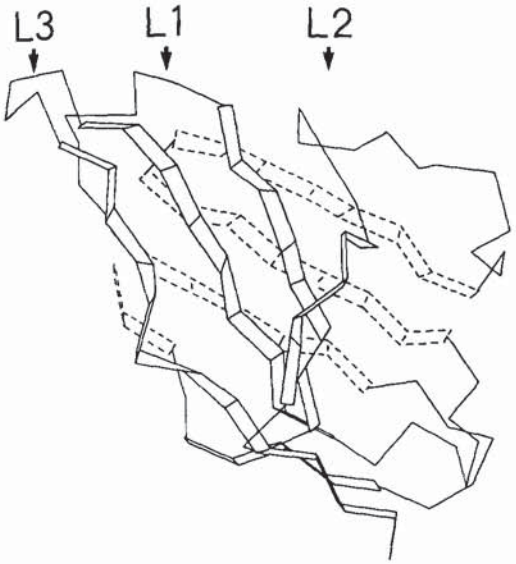

Figure 1. The structure of an immunoglobulin V domain. The drawing is of KOL V_L. Strands of β-sheet are represented by ribbons. The three hypervariable regions are labelled L1, L2 and L3. L2 and L3 are hairpin loops that link adjacent β-sheet strands. L1 links two strands that are part of different 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.

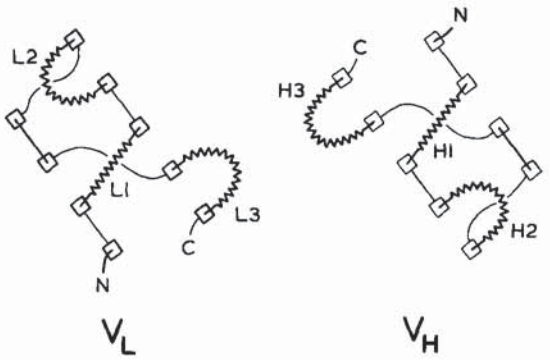

Figure 2. A drawing of the arrangement of the hypervariable regions in immunoglobulin binding sites. The squares indicate the position of residues at the ends of the β-sheet strands in the framework regions.

(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			$\phi 2$, $\psi 2$	$\phi 2$, $\psi 3$					
	X-G-G-X			+55 +35	+85 −5[d]					
				or						6/6
				+65 −125	−105 +10[e]					
2—3										
| |	X-G-X-X			+70 −115	−90 0[e]					6/7
1—4	X-X-G-X			+50 +45	+85 −20[d]					7/8
	X-X-X-X			+60 +20	+85 +25[f]					4/4
		$\phi 1$	$\psi 1$	$\phi 2$	$\psi 2$	$\phi 3$	$\psi 3$	$\phi 4$	$\psi 4$	
	X-X-X-G	−135	+175	−50	−35	−95	−10	+145	+155	4/4
	1 2 3 4 5[g]	$\phi 2$	$\psi 2$	$\phi 3$	$\psi 3$	$\phi 4$	$\psi 4$	$\phi 5$	$\psi 5$	
3 / \ 4	X X X X G	−75	−10	−95	−50	−105	0	+85	−160	3/3
| |										
1—5	X X X X X	+50	+55	+65	−50	−130	−5	−90	+130	1/1(3/3)
3 / \	1 2 3 4 5[h]	$\phi 2$	$\psi 2$	$\phi 3$	$\psi 3$	$\phi 4$	$\psi 4$			
2 4	G									
| / |	X-X-X-N-X	−60	−25	−90	0	+85	+10			13/15
1—5	D									
3— 4	1 2 3 4 5 6[i]	$\phi 2$	$\psi 2$	$\phi 3$	$\psi 3$	$\phi 4$	$\psi 4$	$\phi 5$	$\psi 5$	
| |	G									3/3
2 5	X-X-X-X-N-X	−65	−30	−65	−45	−95	−5	+70	+35	2/2
| |	X									1/1
1—6										

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 $\sim -60°$ 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-N-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

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.