J. Mol. Biol. (1996) **263**, 800–815

JMB

Structural Families in Loops of Homologous Proteins: Automatic Classification, Modelling and Application to Antibodies

Andrew C. R. Martin[1] and Janet M. Thornton[1,2]*

[1]*Biomolecular Structure & Modelling Unit, Department of Biochemistry & Molecular Biology, University College London, Gower Street London, WC1E 6BT United Kingdom*

[2]*Crystallography Department Birkbeck College, Malet Street London, WC1E 7HX United Kingdom*

**Corresponding author*

Loop regions of polypeptide in homologous proteins may be classified into structural families. A method is described by which this classification may be performed automatically and ''key residue'' templates, which may be responsible for the loop adopting a given conformation, are defined. The technique has been applied to the hypervariable loops of antibodies and the results are compared with the previous definition of canonical classes. We have extended these definitions and provide complete sets of structurally determining residues (SDRs) for the observed clusters including the first set of key residues for seven-residue CDR-H3 loops.

Keywords: loops; complementarity determining regions; cluster analysis; canonicals; modelling

Introduction

The structure of a protein is divided into regions of regular secondary structure (mainly α-helices and β-sheets) and so-called ''random coil'' regions that connect the secondary structure elements. In the last ten years, it has been recognised that these coil regions are not, after all, random. In particular, Sibanda and Thornton (1985; Sibanda *et al*., 1989) recognised that short loops connecting strands of β-hairpins are constrained to a limited number of conformations. Together with the work of Milner-White and co-workers (Milner-White, 1986; Milner-White & Poet, 1986, 1987) a classification scheme for β-hairpins has been derived. Leszczynski & Rose (1986) defined a sub-set of loops that they termed Ω-loops with a characteristic shape while, more recently, Martin *et al*. (1995) have performed an analysis of long loops in proteins and have defined two subsets of loops, which they term long-open and long-closed. The two classes are based on the distance between the ends of the loop and of adjoining secondary structure defining whether the loops are able to span intermediary secondary structures.

Abbreviations used: CDR, complementarity determining region; SDR, structurally determining residue; PDB, Protein Data Bank; MDS, multi-dimensional scaling; RMS, root-mean-square.

Within homologous families, the environment of any given loop may further restrict its conformation to a limited set of structures making additional analysis of loop conformation possible. Immunoglobulins are a special example in which nature has evolved a system designed to bind virtually any antigen by variation in the sequence and length of just six loops. Chothia *et al*. (1986, 1989, 1992) have performed a detailed visual analysis of five of the six hypervariable loops or ''complementarity determining regions'' (CDRs) of antibodies. From this analysis, they have defined a number of ''canonical classes'' for these loops. Given the conserved nature of the framework, these classes depend primarily on the length of the loop, but also on the presence of certain structurally determining residues (SDRs) which are involved in the packing of the loop. More generally, Rooman *et al*. (1989) have examined the predictive power of amino acid sequence templates for turn motifs in proteins.

It is now clear that whilst the immunoglobulin family is exceptional, many families of proteins have evolved to perform multiple related functions with variation in loop regions on a relatively conserved framework. For example, the chymotrypsin-like serine proteases, the aspartic proteases or the globins. Indeed, super-families such as the globin fold and the β-trefoil fold show a relatively conserved framework (in structure, if not in sequence) with variation of loop regions.

0022–2836/96/450800–16 $25.00/0

The current Brookhaven Protein Data Bank (Bernstein *et al*., 1977) has been shown to saturate the conformational space available to short peptide fragments (Fidelis *et al*., 1994). Jones & Thirup (1986), Claessens *et al*. (1989) and Unger *et al*. (1989) have all suggested the use of ''spare parts'' modelling to build structures using regions from unrelated proteins in the Protein Data Bank (PDB) and an automated spare-part modelling procedure has been developed by Sutcliffe *et al*. (1987). Recognising that loops within homologous protein families may adopt a limited range of conformations and identifying SDRs, which define these conformations, allows us to take an additional step forward with spare-parts modelling. Thus, within a homologous family of proteins, it is sometimes possible with relative ease to obtain a reasonably accurate model from the amino acid sequence alone.

The alternative to knowledge-based approaches to modelling is some form of *ab initio* conformational search. This approach is exemplified by the work of Fine *et al*. (1986) and of Bruccoleri & Karplus (1987). Martin *et al*. (1989) developed a ''combined'' algorithm using both knowledge-based and *ab initio* techniques. Both approaches have advantages and disadvantages: knowledge-based approaches and, in particular, the template methods exemplified by the canonical approach of Chothia *et al*. (1989), have the advantage of simplicity and a high degree of accuracy when the correct conformation is available in the database. The *ab initio* approaches remove the restriction imposed by requiring conformations to have been observed previously, but are more complex and impose a greater problem in selecting the correct conformation.

Our aim is to identify automatically the different conformations for a given loop in homologous proteins and the key residues that determine each conformer. This extends and automates the canonical approach used by Chothia and co-workers to identify template conformations for the CDR loops of antibodies. We have developed a method, which uses cluster analysis in internal coordinate space followed by a post-cluster merging in Cartesian space, to define conformational classes for a set of loops. A simple analysis of the residues within the loops and those that make conserved contacts with the loops is used to define ''sequence environments''. Key residue templates are then identified by searching these sequence environments for buried hydrophobics, hydrogen-bonding residues and conserved glycine and proline residues.

Results

In order to test the procedure, we applied it to the complementarity determining regions of antibodies. This system has been studied extensively before and our results are compared with the visual analysis performed by Chothia and co-workers (Chothia & Lesk, 1987; Tramontano *et al*., 1989; Chothia *et al*., 1989).

Table 1. Results of cluster analysis of conformations for CDR-L1

Length	Cluster	Canonical class	Members	Representative
10	10A	1	4	2fbj
11	11A	2	22	1ikf
17	17A	3	4	1hil
16	16A	4	9	1rmf
15	15A	5	1	1ggi
12	12A	6	1	1fig
13	13A	5λ	2	2fb4
14	14A	6λ	1	7fab
14	14B	7λ	3	1gig
11	11B	?	1	8fab
14	14C	?	1	1mcwB
14	14D	?	1	2mcgA
14	14E	?	1	1mcwA
14	14F	?	1	2bjlA
15	15B	?	2	1acy
16	16B	?	2	1nbv
16	16C	?	1	1jel

Clusters are defined by the automatic structural clustering protocol described herein. Clusters are named *N*A, *N*B. . ., where *N* is the number of residues in the loops and the letter indicates the specific cluster. Canonical classes are assigned on the basis of sequence using the allowed amino acid templates at key SDR positions defined by Chothia *et al*. Some canonical classes are described by these authors without assigning ''official'' numbers to them (classes 5λ, 6λ and 7λ). Representative structures are selected as the conformation closest to the centroid of the cluster in torsional space. The representatives are shown using their PDB codes; for references, see the entries in the Protein Data Bank (Bernstein *et al*., 1977). '?' under canonical class represents sequences that cannot be assigned to a class using the allowed amino acids at the SDR positions that Chothia *et al*. define.

Structural clustering

Five of the six CDRs have been analysed previously (Chothia *et al*., 1989) and the loop clustering method was applied to these loops in 49 unique antibody Fv regions and eight light chain dimers available in the Protein Data Bank.

Tables 1 to 5 show the results of the cluster analysis of the three-dimensional conformations of the five CDRs together with the canonical class assignments made by using the key residue templates suggested by Chothia and co-workers (Chothia & Lesk, 1987; Tramontano *et al*., 1989; Chothia *et al*., 1989, 1992; Tomlinson *et al*., 1995). As an example, Figure 1 shows C^{α} traces of two

Table 2. Results of cluster analysis of conformations for CDR-L2

Length	Cluster	Canonical class	Members	Representative
7	7A	1	55	1lmk
7	7B	1	1	6fab

Note that 7fab (antibody NEWM) is excluded from the analysis of CDR-L2, since there is a deletion in this region of the structure (Saul & Poljak, 1992).

Table 3. Results of cluster analysis of conformations for CDR-L3

Length	Cluster	Canonical class	Members	Representative
9	9A	1	40	1tet
9	9B	2	1	2fbj
8	8A	3	1	2hfl
7	7A	4	1	1dfb
10	10A	5	1	1baf
9	9C	4λ + ?	1 + 1	7fab[a]
11	11A	5λ	2	2fb4
8	8B	?	1	1eap
9	9D	?	2	1gig
9	9E	1	1	1fig
9	9F	?	1	8fab
10	10B	?	1	1mcwB
10	10C	?	1	2mcgA
10	10D	?	1	1mcwA
11	11B	?	1	2bjlA

As with CDR-L1, two of the classes (4λ and 5λ) have been discussed by Chothia *et al*., but not officially named.

[a] The clustering indicates two members of cluster 9C (class 4λ). Only one of these, 7fab, is assigned to class 4λ by the key residue templates of Chothia *et al*. (the other, 1mfa, is unassigned).

members of the main 16-residue CDR-L1 cluster (cluster 16A) with one of the outlying conformations (cluster 16C). This example shows a clear difference in conformation. Multi-dimensional scaling (MDS) allows visualisation in two dimensions of a dataset in multi-dimensional space and was performed using the interactive program XGobi (Swayne *et al*., 1991) to view the clusters. Figures 2 to 4 show MDS results for selected examples, with Figure 2 showing the same example of 16-residue CDR-L1 loops.

Throughout this paper, we use the term cluster to refer to the three-dimensional conformational clusters identified by the methods described herein, while the term class refers to the canonical classes based on the sequence templates used by Chothia and co-workers.

In our clustering of conformations, post-cluster merging in Cartesian space (see Materials and Methods) is a common event. On average, approximately 44% of clusters identified in torsion space are merged in Cartesian space. The smallest percentage of merges occurs in CDR-Ll, where 19 clusters are reduced to 17, whilst the largest percentage occurs in CDR-L2, where 14 clusters identified in torsional space are reduced to just two clusters with the second cluster containing only one member.

Table 4. Results of cluster analysis of conformations for CDR-H1

Length	Cluster	Canonical class	Members	Representative
10	10A	1	43	2fbj
11	11A	2	1	1baf
12	12A	3	2	1ggi
10	10B	1	1	1igi
10	10C	1	1	1nbv
10	10D	1	1	1fig

Table 5. Results of cluster analysis of conformations for CDR-H2

Length	Cluster	Canonical class	Members	Representative
9	9A	1	8	1gig
10	10A	2	21	1bbd
10	10B	3	11	1igc
12	12A	4	2	1mcp
10	10C	3 + ?	1 + 1	1ind[a]
10	10D	2	1	1rmf
10	10E	2	1	6fab
10	10F	2	1	1fig
12	12B	4	2	4fab

Here, the CDR is defined as running from residue H50 to its strand partner at H58, a definition introduced by Martin *et al*. (1989).

[a] The clustering indicates two members of cluster 10C. One of these, 1ind, is assigned to class 3 by the key residue templates of Chothia *et al*. (the other, 1bbj is unassigned).

Prediction of conformational cluster by key residues

In our analysis of antibodies, we have adopted the numbering scheme used by Chothia & Lesk (1987). A more commonly used scheme is the Kabat scheme (Kabat *et al*., 1991). However, the positions of insertions in the Kabat scheme do not match the structural sites of insertions observed in the crystal structures. These differences affect only CDR-L1 and CDR-H1. Chothia's more recent papers (e.g. Chothia *et al*., 1989) describe the insertion in CDR-L1 as being at L31 rather than at L30 as used by Chothia & Lesk (1987), but our examination of the structures suggests that L30 is the more correct insertion site.

Table 6 shows the canonical class assignments made using the sequence templates used by Chothia *et al*. for each of the 49 antibodies and eight light chain dimers used in this study together with

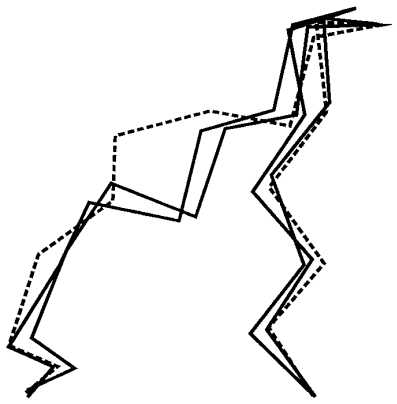

Figure 1. Examples of two members of the main 16-residue CDR-L1 cluster (cluster 16A, PDB codes: 1cgs (Guddat *et al*., 1994) and 1rmf (Jedrzejas *et al*., 1995)) with one of the outlying conformations (cluster 16C, PDB code: 1jel (Prasad *et al*., 1993)) shown with broken lines.

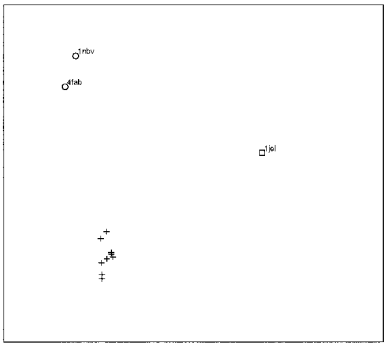

Figure 2. Projection of the clustering for 16-residue CDR-L1 loops into two dimensions by multi-dimensional scaling. The main cluster (cluster 16A) is shown as + symbols while the outliers are marked with their corresponding PDB codes.

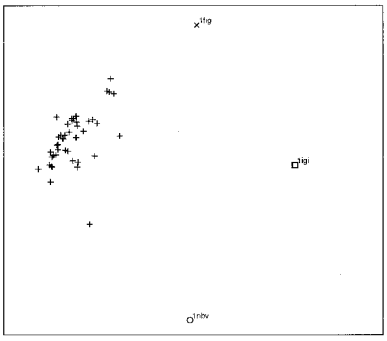

Figure 4. Projection of the clustering for ten-residue CDR-H1 loops into two dimensions by multi-dimensional scaling. The main cluster (cluster 10A) is shown as + symbols while the outliers are marked with their corresponding PDB codes.

cluster numbers based on structural clustering as described herein.

In general, the results support the canonical hypothesis of Chothia *et al*. extremely well. Using the allowed amino acids at the key positions they describe, it is possible to assign 86.2% of the CDRs to canonical classes. Where the canonical class can be assigned, the predictions match the conformational cluster resulting from the cluster analysis in 93.9% of cases; only 14 out of 231 predictions from sequence should be placed in new conformational clusters (Table 6). Of the 37 CDRs for which the SDRs are unable to make a classification, 19 (51.4%) adopt canonical conformations already observed. These are listed in Table 7 and can be used to extend the allowed residues in the sequence templates used by Chothia *et al*. The remaining 18 unclassified CDRs adopt new conformations.

For the 14 predictions that, according to the clustering of conformations, are mis-classified, one must ask whether they really are true outliers, or if they have fallen outside the clustering limits by an insignificant amount. In all cases, MDS illustrates the validity of the clustering. Results are shown for 16 residue CDR-L1 loops (Figure 2), CDR-L2 (Figure 3) and ten-residue CDR-H1s (Figure 4).

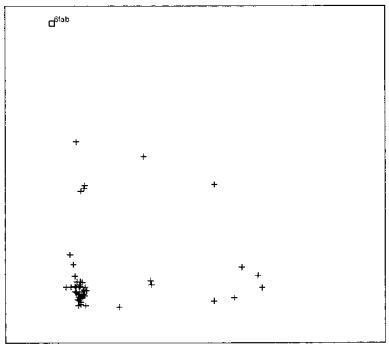

Figure 3. Projection of the clustering for CDR-L2 loops into two dimensions by multi-dimensional scaling. The main cluster (cluster 7A) is shown as + symbols while the outlying conformation of 6fab (antibody 36–71 (Strong *et al*., 1991)) is shown as a square marked with its PDB code.

Updating the sequence templates

Having defined conformational clusters in an automated fashion, we proceeded to define key residue positions responsible for dictating loop conformation. The high level of accuracy obtained using the positions defined by Chothia *et al*. suggests that a limited number of positions are indeed critical in defining conformation.

Our automated procedure for defining the SDRs identifies virtually all the positions previously identified by Chothia *et al*., but additionally defines a number of other positions. In the main, these have been excluded from Chothia's analysis as they are conserved throughout the antibodies for other reasons (for example the conserved Cys at L23 or the Trp at L35). Alternatively, they are additional residues necessary to define the conformation of longer loops and loops of lengths not considered in Chothia's analysis.

Tables 8 to 11 show the key residues that we define. The positions defined by Chothia *et al*. are indicated with an asterisk and their allowed residues at each position are indicated in capital letters. These are derived by merging the descriptions of SDRs given in seven papers by these authors (Chothia *et al*., 1986, 1989, 1992; Chothia & Lesk, 1987; Tramontano *et al*., 1989; Barré *et al*., 1994; Tomlinson *et al*., 1995). The Tables show the extensions to these definitions resulting from our analysis. A number of residues suggested as allowed by Chothia *et al*. (1992) have no crystal structure to confirm that this is indeed the case. Note that we identify no key residues for CDR-L2 (see Discussion).

Our analysis has shown that there are more conformational clusters than defined by Chothia *et al*. Clearly, it would be possible to adjust the clustering thresholds such that some of these additional conformational clusters merge into other clusters, but this would result in merging of some

Table 6. Canonical class and cluster assignments

	Canonical Class					Cluster				
PDB code	L1	L2	L3	H1	H2	L1	L2	L3	H1	H2
1acy	?	1	1	3	1	15B	7A	9A	12A	9A
1baf	1	1	5	2	1	10A	7A	10A	11A	9A
1bbd	3	1	1	1	2	17A	7A	9A	10A	10A
1bbj	2	1	1	1	?	11A	7A	9A	10A	10C
1cgs	?	1	1	1	2	16A	7A	9A	10A	10A
1dbb	4	?	1	1	?	16A	7A	9A	10A	10A
1dfb	2	1	4	1	3	11A	7A	7A	10A	10B
1eap	?	1	?	?	2	11A	7A	8B	10A	10A
1fai	2	1	1	1	2	11A	7A	9A	10A	10A
1fbi	2	1	1	?	2	11A	7A	9A	10A	10A
1fgv	2	1	1	1	2	11A	7A	9A	10A	10A
1fig	6	1	[1]	[1]	[2]	12A	7A	9E	10D	10F
1for	1	1	1	1	2	10A	7A	9A	10A	10A
1fpt	4	1	1	1	2	16A	7A	9A	10A	10A
1frg	3	1	1	1	3	17A	7A	9A	10A	10B
1fvc	?	1	1	1	2	11A	7A	9A	10A	10A
1fvd	?	1	1	1	2	11A	7A	9A	10A	10A
1ggi	5	1	1	?	1	15A	7A	9A	10A	9A
1gig	7λ	1	?	1	1	14B	7A	9D	10A	9A
1hil	3	1	1	1	3	17A	7A	9A	10A	10B
1ibg	?	1	1	1	1	15B	7A	9A	10A	9A
1igc	?	1	1	1	3	11A	7A	9A	10A	10B
1igf	?	1	1	1	3	16A	7A	9A	10A	10B
1igi	4	1	1	[1]	2	16A	7A	9A	10B	10A
1igm	2	1	1	1	3	11A	7A	9A	10A	10B
1ikf	2	1	1	1	3	11A	7A	9A	10A	10B
1ind	7λ	1	?	?	[3]	14B	7A	9D	10A	10C
1jel	?	1	1	1	2	16C	7A	9A	10A	10A
1jhl	2	1	1	1	2	11A	7A	9A	10A	10A
1lmk	?	1	1	1	2	16A	7A	9A	10A	10A
1mam	2	1	1	1	4	11A	7A	9A	10A	12A
1mcp	3	1	1	1	4	17A	7A	9A	10A	12A
1mfa	7λ	1	?	1	2	14B	7A	9C	10A	10A
1mlb	2	1	1	1	2	11A	7A	9A	10A	10A
1nbv	[4]	1	1	[1]	[4]	16B	7A	9A	10C	12B
1ncb	?	1	1	1	4	11A	7A	9A	10A	10A
1rmf	4	1	1	1	[2]	16A	7A	9A	10A	10D
1tet	?	1	1	1	2	16A	7A	9A	10A	10A
1vfa	2	1	1	1	1	11A	7A	9A	10A	9A
2cgr	?	1	1	1	2	16A	7A	9A	10A	10A
2fb4	5λ	1	5λ	1	3	13A	7A	11A	10A	10B
2fbj	1	1	2	1	3	10A	7A	9B	10A	10B
2gfb	2	1	1	1	3	11A	7A	9A	10A	10B
2hfl	1	1	3	1	2	10A	7A	8A	10A	10A
3hfm	2	1	1	1'	1	11A	7A	9A	10A	9A
4fab	[4]	1	1	1	[4]	16B	7A	9A	10A	12B
6fab	2	[1]	1	1	[2]	11A	7B	9A	10A	10E
7fab	6λ	—	4λ	1'	1	14A	—	9C	10A	9A
8fab	?	?	?	1	3	11B	7A	9F	10A	10B
1ivlA	2	1	1	—	—	11A	7A	9A	—	—
1mcwA	?	1	?	—	—	14E	7A	10D	—	—
1mcwB	?	1	?	—	—	14C	7A	10B	—	—
1reiA	2	1	1	—	—	11A	7A	9A	—	—
1wtlA	?	1	1	—	—	11A	7A	9A	—	—
2bjlA	?	?	[5λ]	—	—	14F	7A	11B	—	—
2mcgA	?	1	?	—	—	14D	7A	10C	—	—
2rheA	5λ	?	5λ	—	—	13A	7A	11A	—	—

Canonical class assignments are made on the basis of the sequence templates of Chothia *et al*., while cluster numbers come from the 49 antibodies and eight light chain dimers used in this study. Note that no canonical assignment or clustering is attempted for CDR-L2 of 7fab (antibody NEWM (Saul & Poljak, 1992)), which has a deletion in this region. Boxed canonical class numbers represent structures that have been mis-classified by sequence template prediction according to the structural cluster analysis. Loops that are unassigned using the templates of Chothia *et al*. are indicated with a ?.

Table 7. Unassigned CDRs that adopt previously observed conformations

	Canonical class			
PDB code	L1	L2	H1	H2
1cgs	4			
1dbb		1		2
1eap	2		1	
1fbi			1	
1fvc	2			
1fvd	2			
1ggi			3	
1igc	2			
1igf	4			
1lmk	4			
1tet	4			
2cgr	4			
8fab		1		
1wtlA	2			
2bjlA		1		
2rheA		1		

These CDRs are not assigned by Chothia's structurally determining residues, but do adopt canonical conformations. The class numbers provided are the canonical classes into which these CDRs fall based on conformation rather than sequence matching at the SDR positions.

of the canonical classes that they define. Their analysis of CDR-L1 and CDR-L3 has largely been limited to κ-light chains; the λ-light chains appear to show more conformational variability, particularly in CDR-L1. For example, the eight 14-residue CDR-L1s (all from λ-light chains) adopt six different conformations, whereas the 22 11-residue CDR-L1s (from κ-light chains) all adopt the same conformation.

In no case do the SDRs defined by Chothia *et al*. result in mis-assignment between the conformations corresponding to the canonical classes that they define. For example, in CDR-H2, no conformation matching the conformational cluster of canonical class 2 gets placed into class 3 on the basis of the key residue template. In a number of cases, however, the SDRs of Chothia *et al*. place loops into previously defined canonical classes whereas our analysis of the conformations indicates that they should be placed into new structural classes. In general, where a loop is assigned to a canonical class by Chothia's SDRs and the loop should be in a new conformational cluster, the cluster selected by the SDRs is, in fact, the closest of the recognised clusters. For these mis-assigned loops, Table 12 shows the cluster predicted by Chothia's SDRs, the actual cluster and the nearest cluster on C^{α} RMS deviation between representative members.

Only in the case of CDR-L3 of antibody 1F7 (Haynes *et al*., 1994, PDB code 1fig) have the Chothia SDRs failed to select the most similar recognised canonical class. In this case, not only have they failed to identify a new conformational class, but they have also assigned the loop to the less similar of the two recognised canonical classes. The loop is predicted as cluster 9A (class 1), whereas the conformation falls into cluster 9E (no canonical class); cluster 9C (class 4λ) has been observed by Chothia *et al*. and would have been a better prediction.

Rogue clusters

In some cases, key residues, including our extended templates (and the complete sequence environments, i.e. all positions making conserved contacts), seem to be unable to predict the conformation correctly. Such clusters, where all the key residues match those for another cluster, we term ''rogue'' clusters and indicate them in Tables 8 to 11 with a dagger (†). Figures 2 to 4 illustrate that these are true conformational outliers. For example, CDR-L1 of 1nbv (antibody BV04-01; Herron *et al*., 1991) has allowed residues for cluster 16A at every SDR position, but falls into Cluster 16B. We performed an extended analysis of all residue positions that interact with the different conformational clusters in an attempt to identify possible additional SDR positions, but still fail to find any residues in 1nbv that are forbidden in cluster 16A (data not shown).

For lower-resolution structures, it is possible that the loops have been incorrectly built (CDR-H2-1bbj (Beady *et al.*, 1992), CDR-L3-2bjlA (Schiffer *et al.*, 1989) and CDR-H2-1rmf (Jedrzejas *et al*., 1995)). In higher-quality structures, some differences can be explained by crystal packing contacts (CDR-L1-4fab (Herron *et al*., 1989) and CDR-L2-6fab (Strong *et al*., 1991)), or possible induced fit in complexes (CDR-H2-1ind; Love *et al.*, 1993).

In CDR-H1-1igi (antibody 26-10; Jeffrey *et al*., 1993), side-chains from other CDRs block the region normally occupied by the CDR. However, all these residues are allowed by cluster 10A and presumably move to fill the gap left by CDR-H1 rather than causing its distorted conformation. The temperature factors in the distorted region of CDR-H1 are high, so it is possible that the conformation here is poorly determined. However, there is strong evidence that the conformation adopted does not match the usual canonical form (P.D. Jeffrey, personal communication) and it is likely that the loop is flexible. For this and CDR-L1 1nbv (antibody BV04-01; Herron *et al*., 1991), there is currently no explanation as to why they do not adopt standard canonical conformations.

Application to CDR-H3

As an example of application of the technique to a novel loop, we have applied the analysis to CDR-H3 of the same 49 antibody Fv regions. Earlier analysis of canonical class has not considered CDR-H3, since it is far more variable than any of the other five CDRs. However, it has been widely believed that it should be possible to derive similar canonical conformations and structurally determining residues once a large enough number of examples becomes available.

Table 8. CDR-L1 structurally determining residues

Class	1	2	3	4	5	6	5λ	6λ	7λ								
Cluster	10A	11A	17A	16A	15A	12A	13A	14A	14B	11B	14C	14D‡	14E	14F	15B	16B†	16C
Length	10	11	17	16	15	12	13	14	14	11	14	14	14	14	15	16	16
Members	4	22	4	9	1	1	2	1	3	1	1	1	1	1	2	2	1
L2*	I	I	I	lVi	I	N	s	s	aq	—	s	s	s	s	i	v	v
L4	l	ml	m	ml	l	l	l	l	v	l	l	l	l	l	ml	m	m
L23	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
L24					r										r		
L25*	A	A	[S]	pS	A	A	G	[G]	[S]	a					a	s	s
L26		s		sn	s			s	s	n	t	t	h	s	s	s	s
L27				q												q	q
L28		nsde			s			s	g	l	s	s	s	s	s		
L29*	V[IL]	IV[L]	L[IV]	Li	V	V	nd	n	at	p	d	d	d	n	v	l	i
L30*	—				d		I	I	V		v	v	v	i	ds		
L30A	—	—		hl			g			—						h	h
L30B	—	—		s		—				—						s	g
L30C	—	—		nds	g	—	—			—					g	nq	n
L30D	—	—		g		—	—	—	—	—	—	—	—	—		g	g
L31								h	n		n	n	n	n			
L32				ys				n	yh		y	y	s	s		y	y
L33*	ML	LVi	L	Lf	L	L	V	V	A	a	v	v	i	v	mi	l	l
L34		agnshvf		hen	h					y					h	hr	e
L35	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w
L36		ylf								y							
L46		lrv								m							
L48								i	i		i	i	i	i			
L49		yhfk								y							
L51		atgv		v	s		dn	n	t	d	v	v	v	d	a	v	i
L66								k	l		k	k	k	k			
L71*	Y[F]	YF	F[Y]	F	F	Y	A	A	A	v	a	a	a	a	f	f	f
L88	c			c	c	c	c	c	c		c	c	c	c	c	c	c
L90	q	hq	n	q	q	q	a	s	l	a	s	s	s	a	qh	q	q
L91						y											
L92				ts	n										nr	t	s
L93	syr	gsntrea	ns	h	e	g	vd	r	sn	n	g	g	s	d	e	h	h
H100				wdfal												a	y
H100A		sydrlm								a							

Key positions identified by Chothia *et al*. are indicated with an asterisk. Horizontal rules delineate the CDR itself. Rogue clusters (see the text) that cannot be distinguished on the basis of sequence templates from main clusters are indicated with a dagger (†) or with a double dagger (‡) if they cannot be distinguished from another outlying conformation rather than from a main cluster. Amino acids shown in upper case are those that Chothia *et al*. identify as allowed residues. When these are shown in square brackets they are not identified by our procedure as allowed residues, or, in the cases where a residue in square brackets is the only residue at a position (e.g. Ser for L25 in cluster 17A), then this position is not identified as critical by our procedure. Note that the most recent paper from Chothia's group (Tomlinson *et al*., 1995) does suggest that position L30D is a key residue in 16-residue loops.

Among the 49 crystal structures, there are CDR-H3 loops with all lengths from 5 to 15 residues and 17 residues being represented. Consequently, the maximum number of loops of any one length is still limited (ten loops of length 11).

Figure 5 summarises the results of applying the cluster analysis to CDR-H3. For each loop length, the graph shows the number of loops, the maximum number of loops in any one cluster and the ratio of (number of clusters)/(number of loops). These results show that the number of clusters available to a loop (represented by this ratio) is weakly correlated with the length of the loop (correlation coefficient = 0.62), i.e. the longer loops have more conformations accessible to them and many of the observed conformers are singlets. Although 49 antibodies are considered, there is still a low-counts problem with some of the loop lengths. Lengths of 6, 8, 14 and 15 residues have only one or two examples. Removing these and recalculating the correlation coefficient raises the value to 0.91. By definition, the conformational space available to short loops is limited compared with longer loops (which have more conformations available since they have more rotatable torsion angles). The correlation between loop length and number of clusters indicates that this is still true for the CDR-H3 loops even with the constraints applied by the need to connect to the framework. While this is what might be expected for any given loop, it is not observed in the other CDRs. For example, CDR-L1 (Table 1), has four examples of 17-residue loops that all adopt a single conformation. Shorter, 14-residue loops show much more conformational variability.

The length and sequence variability for CDR-H3 means that none of the clusters is very densely populated; the best is one of the clusters for seven-residue loops, which has four members (Figure 5). There are two other clusters seen for

Table 9. CDR-L3 structurally determining residues

Class	1	2	3	4	5	4λ	5λ								
Cluster	9A	9B	8A	7A	10A	9C	11A	8B	9D	9E	9F	10B	10C‡	10D	11B†
Length	9	9	8	7	10	9	11	8	9	9	9	10	10	10	11
Members	40	1	1	1	1	2	2	1	2	1	1	1	1	1	1
L2	ilv	i				qs			a	n	—				
L3	vqle	v				iv			v	v	e				
L4	ml	l			l	vl	l		v	l	l	l	l	l	l
L28	sndte	s				gs			g	s	l				
L30	dlyvisnfhgt	—				vi			v	s	n				
L31	sntkg	s				nh			n	t	q				
L32	fynahsr	s		w	y	hn			y	y	y	y	y	s	
L33	mlvif	l				av	v		a	l	a				v
L34				a											
L36			y	y	y			y				y	y	f	
L88	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
L89	qsgfl	q	q		q	aq	a	l	a	q	q	s	s	m	a
L90*	QNH	Q	Q	Q	Q	ls	a	q	l	q	a	s	s	s	a
L91	nfgsrdhtyv	w	w	y	w	wy		y	w	y	w	y	y	y	
L92	nywtsrqhad	t			s	sd	dn		y	s	d	e	e	l	d
L93	enghtsra	y				nr			s	g	n				
L94*	dytvlhniwps	P				ns	[DNG]		n	y	s				
L95*	P	L	[P]			hl			hl	p	a				
L95A	—	—	—	—	P	—		—	—	—	—	n	n	s	
L96	plyriwf	i	—	—	i	wr		—	w	l	s	f	f	f	
L97	t	t		—	t	iv	vg		v	t	i	v	v	v	v
L98	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f
H47	wy	w			w	w			w	w	w	—	—	—	
H100	edwstvyfagl	y				yg			d	y	t				
H100B	yflmgadtw	—				—			f	f	f				
H100I				f											

See Table 8 for details. Note that the most recent paper from Chothia's group (Tomlinson *et al*., 1995) does suggest that position L97 is a key residue.

seven-residue CDR-H3s, each of which has a single member.

The automated key residue analysis identifies key residues for seven-residue CDR-H3 loops at H99 and H101 in the loop and at L36, L46 and H103 outside the loop. Of these only L46 in LFR2, which is Arg in cluster 7C (the other clusters have a hydrophobic at this position), appears to have any power to discriminate between the conformations.

Visual examination of the seven-residue CDR-H3 loops, the extended sequence templates and the sequence environments (all residues making conserved contacts) shows that the prime difference between the three conformations is the separation of the termini of the loop resulting from changes in the framework. For the other CDR loops. the analysis of key residues may be restricted to the loop and its contacting residues because the framework is highly conserved. In CDR-H3, the conformation of the framework is less conserved and analysis of key residues must be extended into the neighbouring framework regions. Visual analysis shows that the conformation of the CDR-H3 takeoff region of the framework is determined primarily by the nature of the residue at H94 (immediately preceding the first residue of the CDR). In >80% of the heavy chains in the Kabat sequence database (April 1996 release) where this residue is known, it is Arg and this residue is observed in the four-member cluster (cluster 7A). The Arg projects up from the bottom of the loop such that the distal charged atoms are exposed on the protein surface.

The other two conformations observed are outliers and have different residues at this position (Table 13). In cluster 7B, H94 is Gly. This residue is in the normal allowed (β) area of the Ramachandran plot, but the absence of a side-chain causes distortion of the backbone such that the α-carbon packs against Met at H34. In cluster 7C, there is a His at position H94 and this residue packs against Tyr-H27, Leu-H4, Ile-H34. The bulk of this residue packed inside the protein also distorts the takeoff position of this loop compared with clusters 7A and 7B. Chothia & Lesk (1987) have also suggested that H94 may influence the conformation of CDR-H3.

Further visual analysis of the extended sequence templates and sequence environments for these three clusters suggests that residue H98 within the CDR may be important, since residue types not observed in cluster 7A are seen in clusters 7B and 7C. Cluster 7B has Gly at H98. This residue is in an allowed region of the Ramachandran plot (α_R), but the preceding Tyr is in a disallowed region (bottom right of the plot, positive ϕ / negative ψ). This is not currently detected as a critical residue by the automated technique as there must be at least two members in a cluster for glycine residues to be defined as critical.

For longer CDR-H3 loops, Pedersen (1993)

Table 10. CDR-H1 structurally determining residues

Class	1	2	3			
Cluster	10A	11A	12A	10B†	10C	10D
Length	10	11	12	10	10	10
Members	43	1	2	1	1	1
H2	vig	v		v	v	v
H4	lv			l	p	l
H20	limv	l	l	m	l	i
H22	c	c	c	c	c	c
H24*	TAVGS	V[F]	VF[G]	s	a	a
H26*	G	[G]	G	g	g	g
H28			s			
H29*	IFLs[V]	I[L]	IL[V]	f	f	l
H31A	—	d		—	—	—
H32	ihyftnced			f	n	h
H33	yawgtlv	a		y	a	n
H34*	IVMw[TL]	W[C]	WV	m	m	i
H35	henqsyt			n	n	n
H36	w	w	w	w	w	w
H48	imvl	m	ml	i	v	i
H50		y				
H51	livtsn			i	i	i
H53		y	yw			
H69	ilfmv	i		l	i	l
H76		n				
H78	alvyf	f	fv	a	l	l
H80	lm	l	il	m	l	m
H90	yf			y	y	y
H92	c	c	c	c	c	c
H94*	RKgshn[TA]	[HR]	[HR]	g	r	r
H96		w				
H102	yhvisdg			y	y	y

See Table 8 for details. The definition used here for the CDR encompasses both the Chothia (H26-H32) and Kabat (H31-H35) definitions. Chothia *et al*. suggest that residue H27 is also a key residue, but we do not find that this residue influences the conformation. Note that we do not identify Gly at H26 as key for cluster 11A as we require two members of a cluster before a conserved Gly is defined as key.

identified the residue at position H102, which is normally Tyr or Val, as determining the conformation of the takeoff region and Searle *et al*. (1994) went on to define seven general classes of CDR-H3 loops based on length and the conservation of residues at positions H94 and H101. It thus appears that the mechanisms underlying the determination of the conformation of CDR-H3 are a little more complex than for the other loops. The takeoff region is more variable and the residues that determine the geometry of the framework region leading into the loop must be considered in addition to key residues that determine the conformation of the loop itself.

In an extreme case, antibody CNJ206 (Golinelli-Pimpaneua *et al*., 1994; PDB code: 2gfb) has a hugely distorted framework region to the C terminus of CDR-H3 giving the ten-residue loop an extended conformation rather than being a distorted hairpin. The terminal C^{α} atoms are separated by 22.7 Å rather than around 8 Å, which is typical for the other antibody crystal structures. While the resolution of the CNJ206 crystal structure is quite low (3.0 Å, $R = 21.3\%$), given that the crystal structure was solved by molecular replacement, it seems unlikely that such a large deviation could be explained solely by an error in the structure.

Conformation of seven-residue loops in the protein databank

We have also applied the cluster analysis technique to all seven-residue loops extracted from a non-homologous set of proteins from the Protein Data Bank. The set of proteins used was as described by Martin *et al*., (1995). From this dataset, a total of 253 loops was identified, which fell into 170 clusters. These loops vary widely in conformation, some being quite extended while others are more classically ''loop-shaped''.

To examine whether the conformations seen for the seven-residue loops in antibodies represent all the conformations available to a loop of this length given the end-point separation restrictions enforced by the antibody framework, we analysed the distances between the terminal residues of CDR-H3 loops (of all lengths) in the known antibody structures by calculating the mean ($\bar{x}$) and standard deviation (σ_{n-1}). A range was then calculated as $\bar{x} \pm 3\sigma_{n-1}$. Equivalent ranges were calculated for the distance between the N-terminal residue of a loop and the residue before the C terminus and the distance between the C-terminal residue and that following the N terminus. Similar distance ranges have been used previously to extract loops from the Protein Data Bank for antibody loop modelling (Martin *et al*., 1989).

These distance constraints were applied to the selection of seven-residue loops from the non-homologous set of proteins to identify the loops that satisfy the distance restraints required to attach the loops to the framework. This resulted in 31 loops being extracted, which fell into 29 conformational clusters. This conformational diversity contrasts with the seven-residue CDR-L2 loop, where the sequence essentially has no effect on the conformation. Visual comparison and MDS of the four clusters seen for seven-residue CDR-H3s show that the conformations are fairly well spread amongst the 29 conformations available to seven-residue loops selected from the non-homologous set (data not shown). For each of the four conformations seen for CDR-H3, the most similar of the 29 clusters from the non-homologous set was identified on the basis of C^{α} RMS deviation. The conformations seen in the antibodies were not found to match the two better-saturated (two-member) clusters from the non-homologous set.

Discussion

Conformational clusters compared with canonicals

Our analysis clearly shows the power of the canonical method when applied to antibodies. Analytical clustering has supported the earlier manual analysis reported by Chothia *et al*. and has shown that in the available crystal structures, the majority of antibody CDRs L1, L2, L3, H1 and H2 (88.1%) fall into conformational clusters with the

Table 11. CDR-H2 structurally determining residues

Class	1	2	3	4					
Cluster	9A	10A	10B	12A	10C†	10D†	10E	10F	12B
Length	9	10	10	12	10	10	10	10	12
Members	8	21	11	2	2	1	1	1	2
L94				ly					v
H33		ywgatl	agvyw		at	a	g	n	
H47	wy	wy	w	w	w	w	w	w	w
H50		rewygqvlnka	gtyfiev	fa	yt	v	y	n	rq
H51	imv	li	iv	is	it	i	n	i	i
H52		dlnsy	sfwh		sl	s	n	d	
H53		agysktn	dgsn		g	y	g	y	
H54*		nstkdg	SG[ND]	[KS]	ng	s	n	y	
H55*	G[D]	[GS]	[GS]	[Y]					
H56		yredgvsa	sytndr		df	d	y	g	
H58		kntsdrgfy	gyhfdn		kf	n	a	n	
H59	yl	y	y	y	y	y	y	f	y
H69	im	iflm	i	iv	li	m	l	l	i
H71*	RKV[I]	vAL[T]	R	R	ar	v	v	v	r
H78		alv	l	l	al	a	a	l	lv

See Table 8 for details. Chothia *et al*. also identify residue H52A as a key residue in determining the conformation of class 2 and 3; we do not find this residue to be important. Here we define CDR-H2 as running from residue H50 to its strand-partner, H58.

canonical conformations Chothia *et al*. described previously, although some of these (8.1%) were not assigned by the key residues that they have defined. Our automated analysis has identified additional conformational clusters for lengths of CDRs previously unclassified and for sequences that do not match earlier sequence specifications.

In addition, we have shown that a small percentage of the sequences assigned to canonical classes on the basis of previously presented SDRs are actually in new clusters or are outlying conformations. The additional clusters are necessary in order to apply the same clustering criteria to all loops while ensuring that all canonicals defined by Chothia *et al*. are assigned to separate conformational clusters. The majority of the outliers can be explained by additional residues that we consider to be structurally determining; judging the predictive value of these will have to await the availability of new crystal structures.

However, for nine of the loops there appears to be no explanation in terms of key residues. For six of these ''rogue cluster'' loops there are possible explanations in terms of crystal packing, potential induced fit or low resolution of the structure determination. Alternatively, and for the remaining rogue loops, individual SDRs may be insufficient to define the conformation. It is possible, therefore, that pairs of residues need to be considered together. In the most obvious example, a sequence template might suggest that Arg and Glu are both allowed residues at contacting positions X and Y.

Table 12. Loops incorrectly predicted by Chothia key residues

CDR	Antibody	Chothia predicted cluster	Actual cluster	Nearest cluster	C^{α} RMS (Å)
L1	1nbv/4fab	16A	16B	16A	1.320
L2	6fab	7A	7B	7A	2.043
L3	1fig	9A	9E	9C	1.674
L3	2bjlA	11A	11B	11A	1.234
H1	1fig	10A	10D	10A	1.142
H1	1igi	10A	10B	10D	2.373
H1	1nbv	10A	10C	10A	0.999
H2	1ind	10B	10C	10B	1.188
H2	1rmf	10A	10D	10E	2.967
H2	1fig	10A	10F	10A	0.914
H2	6fab	10A	10E	10A	1.235
H2	1nbv/4fab	12A	12B	12A	0.873

The predicted, actual and closest clusters for the loops incorrectly predicted by the SDRs of Chothia *et al*. are shown. The RMS deviation supplied is between the representative members of the SDR-predicted cluster and the actual cluster to which the conformation belongs.

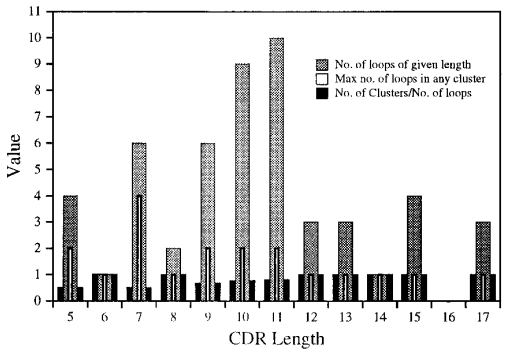

Figure 5. Results of cluster analysis of CDR-H3 from the 49 antibody crystal structures. For each loop length, the graph shows the number of loops (grey bars), the size of the largest cluster (white bars) and the ratio of the number of clusters to the number of loops (black bars); i.e. if this value is 1.0, every loop adopts a different conformation; lower values indicate more conformational restriction.

Table 13. Key residues for seven-residue CDR-H3s

	Cluster		
	7A	7B	7C
L46	L	V	R
H94	R	G	H
H98	NQSTY	G	D

Residues observed at positions H94 and H98 that define the conformation of seven-residue CDR-H3s.

However, this does not convey the fact that Arg at position X is always paired with Glu at Y and *vice versa*. A new sequence having Arg at both positions would match the template, but clearly would have a different conformation as the two Arg residues would repel one another. Much more subtle pairings may also be important.

Chothia *et al*.'s early description of canonical classes (Chothia *et al*., 1989) defined class 1' for CDR-H1. This class was characterised by differences in the residue at position H27. However, their more recent paper (Chothia *et al*., 1992), while stating that the residue at position H27 influences the conformation, no longer makes this distinction. Our analysis reveals no significant differences in the conformations of structures 3hfm (antibody HyHEL-10 (Padlan *et al*., 1989)) and 7fab (antibody NEWM (Saul & Poljak, 1992)), which would be placed in class 1' suggesting that the definition is, indeed, redundant. Using the criteria that we describe in Materials and Methods (Derivation of sequence templates), we do not identify H27 as a key residue.

Extending the SDRs

We have been able to expand and refine the sets of allowed residues at each key position that Chothia *et al*. define. These templates have been extended to include additional conformational clusters and this has required the consideration of additional key positions. We have derived complex expanded sequence templates and, from these, have identified a smaller number of positions that appear to be important in defining outlying conformations. These are generally a superset of the key residues defined previously. Our extended residue sets are more complex than the very limited sets of key positions defined by Chothia *et al*., but all the positions we define appear to be important, given that they all satisfy a simple set of rules that identify positions involved in conserved hydrogen-bonding or hydrophobic packing interactions.

The identification of conformational clusters and extended sequence templates is completely automated and can be repeated easily as new crystal structures become available. This allows us to maintain an up-to-date key residue list, which has been made available on the World Wide Web together with software that identifies the appropriate conformational cluster from a supplied antibody sequence. These can be accessed using the URL http://www.biochem.ucl.ac.uk/~martin/antibodies.html.

Sutcliffe (1988) attempted to define key structurally determining residues in an automated fashion, but took a very different approach. For a group of loops of the same length, each residue is assigned as being structurally determining if it causes the next C^{α} to be moved relative to the position of that C^{α} in any of the other loops. When modelling a novel sequence, a template loop is selected by scoring these structurally determining residues using the Dayhoff mutation matrix (Dayhoff *et al*., 1987). This method considers only SDRs within the loops and does not identify many of the key residues defined by Chothia *et al*., or ourselves.

The possibility of inter-loop SDRs has been suggested by Gregory (1992) and Chothia & Lesk (1987) suggest that the conformation of residues H30 to H32 in CDR-H1 is influenced by CDR-H2. Amongst the additional SDR positions that we have identified are some positions in other CDRs. For example, 15-residue CDR-L1s shown in Table 8 adopt two conformational clusters. It appears that residue L51 (from CDR-L2) may be responsible for the difference in conformation. Examination of the structures using interactive molecular graphics shows that there is a difference is packing around the area of contact of CDR-L1 with this residue.

In CDR-L1, the most important residue in defining the conformation is the hydrophobic Leu, Ile or Val at position L29 (κ-light chains) or L30 (λ-light chains). In one case (the λ-light chain antibody Hil, PDB code 8fab: F. A. Saul & R. J. Poljak, unpublished results) there is an Asn at L30 with the expected hydrophobic appearing at L28 and a Pro at L29. Clearly this leads to a very different conformation for CDR-L1. These characteristics are successfully identified by our automatic procedure (cluster 11B).

For CDR-L2, we find no key residues. While Chothia *et al*. (1989) suggest that L48 and L64 are key residues, their earlier paper (Chothia & Lesk, 1987) simply observes that these residues are conserved amongst the small number of structures they analysed. CDR-L2 encompasses a three-residue hairpin turn linking adjacent β-strands and the conformational requirements for such a turn appear to provide sufficient restrictions on the conformation as they do in other proteins (Sibanda *et al*., 1989). Of the currently available structures, all but one of which adopt canonical class 1, residue L48 is Ile, Met or Val, while L64 is Ala, Gly, Ser or Val, rather than the absolutely conserved Gly suggested by Chothia *et al*., suggesting that this residue at least is not critical in determining the conformation of CDR-L2. The one exception we have found has been antibody 36-71 (Strong *et al*., 1991, PDB code 6fab). This structure is at relatively high resolution (1.9 Å, R = 20.9%) and the temperature factors in the region of CDR-L2 are not unusually high, suggesting that the conformation is correct. The differences are restricted to the loop

and to the framework residue (L57) immediately to the C-terminal side of the loop. There is no obvious explanation for the difference in conformation in the residues of the loop or its surroundings, but the loop does make crystal contacts including a salt-bridge between Arg-L53 in the loop and Glu-H85 in the adjacent molecule and this is likely to explain the difference in conformation.

Sequence template analysis

While our automated identification of key residues appears to perform well, analysis of conserved contacting residues presents a dilemma: other residues that do not make a contact in any of the available structures may be conserved and mutation of these to larger more bulky side-chains may result in conformational change in the loop. This was illustrated by the work of Tramontano *et al*. (1990), who only observed the importance of framework residue H71 in determining the conformation of ten-residue (H50-H58) CDR-H2s when modelling the antibody HyHEL-5 (Sheriff *et al*., 1987). All the structures available at that time had a short side-chain at H71 and when this side-chain is short, the CDR adopts the conformation of canonical class 2 (cluster 10A). However, HyHEL-5 has Arg at this position and, when such a bulky residue is present, the conformation of CDR-H2 flips to canonical class 3 (cluster 10B).

Conversely, conservation of amino acid types may not be a requirement of canonical class definition, but may result from lack of selective pressure. For example, the N-terminal half of CDR-L1 (residues L24 to L27) is generally quite conserved with a sequence of the form (SR)(ASG)S(SRQT). However, this part of the loop is away from the combining site and makes no contact with antigen in any of the antibody-antigen complex crystal structures available in the Protein Data Bank (MacCallum *et al*., 1996). The side-chains are also exposed to solvent and, therefore, unlikely to be critical to loop conformation providing their hydrophilic nature is conserved. Indeed, this region of CDR-L1 has been mutated to introduce a metal binding site (Gregory, 1992; Gregory *et al*., 1993) and preliminary binding assays (Staunton *et al*., 1993) indicate that the antibody still binds its antigen as well as metal ions, suggesting that no major change in structure has occurred.

In CDR-L3 of 1fig (antibody 1F7 (Haynes *et al*., 1994)) an apparently small change in residue type has led to a switch of conformational cluster. CDR-L3 of 1fig has all the conserved interacting residues for cluster 9A (class 1) with the exception of Asn at position L2; all members of cluster 9A have a hydrophobic (Ile, Leu or Val) at this position. Despite this small difference, CDR-L3 of 1fig adopts a different conformation.

CDR-H2 of 1ind (antibody CHA255 (Love *et al*., 1993)) matches the requirements of cluster 10A (class 3), but a hydrogen-bonding pair of Thr residues replaces a hydrophobic packing interaction and this appears to be sufficient to distort the loop conformation. As this is a complex, induced fit may also play a rôle. However, the distorted conformation that is adopted matches that of one of the other clusters. This suggests that the canonical conformations are all relatively stable structures, but that the energy barriers between them are small, such that it is possible to ''push'' a conformation from one canonical to another with relatively little effort.

It is interesting to note that in four of the structures (1fig, Haynes *et al*. (1994); 1nbv, Herron *et al*. (1991); 4fab, Herron *et al*. (1989); and 6fab, Strong *et al*. (1991)), pairs of CDRs (three in the case of 1fig and 1nbv) are assigned to the wrong structural cluster when using the SDR key residue templates used by Chothia *et al.* In three of the four cases, at least one of these adopts a unique conformation not observed in any of the other structures. One might, therefore, propose that these changes in conformation are coupled. However, with the exception of 1fig, the loops involved do not make significant contacts with one another.

Our sequence templates are rather more complex than the SDR templates used by Chothia *et al*.; one of the great attractions of the canonical analysis has been its simplicity. A number of the additional positions are absolutely conserved and other positions are important only in defining the conformation of loop lengths not previously analysed.

We have presented a number of sequence templates for outlying clusters having only a single member. In these cases, it is impossible to tell conclusively whether all the specified amino acids are absolutely required for a given conformation, or if they are simply the residues observed in this single structure. For example, in the ten-residue CDR-H1 loops (Table 10) position H48 in cluster 10A has allowed residues of I, M, V or L. Cluster 10C shows only V as an allowed residue (this is the residue observed in the one structure having this conformation). While it would be reasonable to assume that the other residues would be allowed at this position, without more examples one has no conclusive proof and must, therefore, err on the side of caution. In this way, one makes a trade-off between confidence (and accuracy) and the ability to predict the conformation of a wider range of new sequences.

Computer software has been implemented that allows sequences to be screened against these more complex templates and is accessible over the World Wide Web (http://www.biochem.ucl.ac.uk/~martin/antibodies.html). To use the software it is necessary only to supply the amino acid sequence. The software aligns this with a consensus antibody sequence in order to derive the standard numbering scheme and then applies the sequence templates to identify the predicted conformational cluster. Since our sequence templates are more restrictive than the standard Chothia key residue lists, they are likely to be more accurate, but are

only able to make a smaller number of predictions. Thus, where exact matches to the templates are not made, the nearest template is reported together with the locations of any differences. The user may also choose to apply only the Chothia key residue templates. In all cases, the representative member of the cluster is indicated and this information may be used to create a three-dimensional model of the antibody.

Application to CDR-H3 and seven-residue loops

We have applied the method to antibody CDR-H3 loops as a novel loop example. The fact that well-saturated clusters have not been observed for these loops (with the exception of short seven-residue loops) may suggest that the canonical hypothesis cannot be applied to this loop. Even in the case of the seven-residue loops, the primary factors responsible for the different conformations have arisen in the framework region to which the loop is attached. This has shown the importance of the conservation of framework in obtaining good results from the application of key residue templates. For the seven-residue loops, we find that relatively small distortions of the framework lead to conformational changes in the loop and the residues responsible are not currently identified using the criteria for defining key residues in the other CDRs.

By looking at seven-residue loops from a non-homologous set of proteins in the Protein Data Bank, we have shown that such segments are diverse in conformation. Even when the distance requirements for attaching to an antibody framework are imposed on these loops, no well-populated clusters are observed. This emphasises the importance of homology (both within the loops themselves and in the framework to which they are attached and against which they pack) in imposing conformation on the loops. Thus in the general case of non-homologous structures, there are many conformational clusters, even when distance constraints are applied, while for five of the six antibody loops (where there is a high level of homology and conserved key residues), a small number of well-saturated clusters is observed. In the case of CDR-H3, there is much more sequence and length variability in the loop itself and in the takeoff region of the framework.

Conclusions

We have developed a novel technique for the automatic clustering of loop conformations of mixed length and for the generation of templates of key residues responsible for the conformation. Recently, Wintjens *et al*. (1996) have used cluster analysis to classify αα-turn motifs, although their methodology was rather different.

We observe that antibody loops are more constrained in their conformation than loops of the same length in a set of non-homologous proteins from the Protein Data Bank. Our analysis of conformations, compared with the key residues that may be used to predict them, suggests that key residues do successfully predict the conformation of loops, but that they work only when the framework to which they are attached is sufficiently conserved. Thus for CDRs L1, L2, L3, H1 and H2 they perform extremely well. The example of seven-residue CDR-H3 loops shows that relatively small variation in the takeoff region of the framework will lead to changes in the conformation of the loop. This suggests that extension of the canonical hypothesis to other proteins may not prove successful unless the frameworks are equally well conserved. However, in these cases, it may be possible to define additional key residues within the framework that determine the conformation of the takeoff regions.

We intend to extend our analysis to loops in other families of homologous proteins. In order to do this we need to generate numbering schemes for these families where topologically equivalent residues have the same numbers. This can be achieved using a structural alignment program such as SSAP (Taylor & Orengo, 1989). Our initial experience with looking at the CDR-H3 region in antibodies suggests that the other CDRs may be rather special cases in having such a restricted repertoire of available conformations that are largely defined by such simple key residue templates. However, the genetics of antibody diversity are designed to maximise the variability of CDR-H3, so it may prove that CDR-H3 is unusual.

Materials and Methods

Structural loop cluster analysis

In order to define conformational classes for loops within a homologous protein family, we have used a cluster analysis technique. This gives a completely automatic assignment of clusters and having defined a clustering threshold, removes any possibility of arbitrary decisions in assignment of structural class. Cluster analysis has been used to classify αα-turn motifs by Wintjens *et al*. (1996), but their approach was rather different. Like us, they use a two-stage clustering, first on torsion angles, second in Cartesian space. However, their first stage is a simple classification on Ramachandran assignments while the more detailed cluster analysis in Cartesian space is performed second. In contrast, we perform the detailed cluster analysis first in torsional space and follow this by a reduction of the number of clusters by merging in Cartesian space.

Our method assumes that the loops emerge from a conserved framework. The conformation of the loops is defined in terms of ϕ, ψ and ω torsion angles. One residue before and after the loop is required in order to assign all torsion angles. To overcome the problem of the discontinuity between torsion angles of $\pm 180.0°$, each torsion is represented by two continuous values, the sine and the cosine of the angle. This technique has been used by Reczko *et al*. (1995) and was chosen as simple

preliminary tests using freely available clustering software indicated a successful clustering.

For each loop, we define a maximum length and an insertion scheme. This simply specifies, for loops shorter than the maximum length, at what positions the deletions should be considered to occur and in which order these positions should be filled. Initially every sine and cosine is set to a null value (10.0) selected to be outside the possible ± 1.0 range. Values are then assigned from the torsion angles of the loop in the order specified by the insertion scheme. Any remaining positions (representing deletions compared with the maximum loop length) will retain their null values. This allows us to represent all lengths of loop in the same way and these may thus be clustered in a single operation.

We examined the possibility of clustering a simplified representation of loop conformation by using C^{α} pseudo-torsion angles instead of true torsions. While the technique worked for the majority of loops, one example in an antibody CDR-L3 loop was mis-clustered as the information content of the simplified representation was insufficient. The method was, therefore, rejected and true torsion angles were used.

Each item to be clustered is considered to be a vector in multi-dimensional space. For our purpose, the vectors consist of sines and cosines of the ϕ, ψ and ω angles for each residue in the loop. Thus a loop of maximum length ten residues is represented by a 60-dimensional vector. We have used Ward's minimum variance method (Ward, 1963) to perform the cluster analysis. This method attempts to minimise a sum-of-squares term that describes the loss of information resulting from merging clusters. At each step in the agglomeration, the algorithm considers combining every possible pair of clusters and the union that results in the minimum loss of information is accepted. This is defined as an error sum-of-squares:

$$E_{ss} = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (x_i - \bar{x})^2$$

where x_i is the ith vector, $\bar{x}$ is the mean of the vectors in the cluster, n_j is the number of members of cluster j and N is the number of clusters. This error term is an absolute value that is linearly dependent on the number of dimensions of the vectors being clustered. Dividing E_{ss} by the number of dimensions in the vectors gives us a mean square value $\bar{E}_{ss}$ (formally equivalent to the square of the familiar RMS), which may be compared directly between clustering of different loops.

Empirically, we have defined an $\bar{E}_{ss}$ value of 0.06 as a cutoff for defining true clusters that correspond to visually different loop conformations. This is approximately equivalent to an RMS deviation on torsion angles of 14°, i.e. If $\sqrt{\langle \Delta^2_{(\phi,\psi,\omega)} \rangle}$ is greater than 14°, the loops would fall into separate clusters.

Clustering in internal co-ordinate space has a number of advantages. In particular, loops of different lengths can be clustered in a single operation. In addition, there is no requirement to perform multiple least-squares fits of structures to one another (thus increasing efficiency), or to make an arbitrary selection of a structure against which to fit.

However, it was observed that visually similar conformations of low α-carbon RMS deviation were frequently classified into separate clusters. In the main this results from peptide flips changing ϕ/ψ torsion angles while not affecting the positions of α-carbon atoms. However, it is also possible for relatively small changes in atom position to be accompanied by relatively large changes in torsion angle. This situation is best visualised by imagining four atoms with a torsion angle close to 0°. In one case the atoms may be, in sequence, below-the-plane, above, below, above, while in the other case they are above-the-plane, below, above, below. The distance between equivalent atoms (and, therefore, the RMS deviation) would be small compared with the relatively large difference in torsion angle.

Therefore, to avoid this effect, a post-clustering in Cartesian space was applied. For each cluster of a given loop length, the representative conformation (defined as that closest to the centroid of a cluster in torsion space) is least-squares fitted (in Cartesian space) on α-carbon atoms to all the other representatives. When the RMS deviation is less than 1.0 Å, the maximum distance between equivalent α-carbon atoms is less than 1.5 Å and the maximum distance between equivalent β-carbon atoms is less than 1.9 Å, the clusters are merged. Thus different clusters are characterised by an RMS deviation on α-carbon atoms of more than 1.0 Å, a distance of at least 1.5 Å between at least one pair of equivalent α-carbon atoms, or a distance of at least 1.9 Å between at least one pair of equivalent β-carbon atoms.

The selection of clustering and post-clustering thresholds is a subjective decision, but in so doing, we ensured that all the canonical classes for antibody hypervariable loops defined by Chothia *et al*. (1989) fell into different clusters and made careful choices about the effect on side-chain positions and therefore, in the case of antibodies, the topography of the antigen combining site.

Derivation of sequence templates

Having defined conformational clusters, we wish to be able to identify the cluster into which a novel sequence will fit. To facilitate this modelling procedure, we have identified key structurally determining residues and sequence templates for each conformation.

Initially, as well as the residues within the loops, we identified all positions that make side-chain contacts with every member of any one of the conformational clusters. These ''sequence environments'' may be used predictively, but they are complex and over-restrictive; a simpler subset of ''key'' residues has previously been used effectively (Chothia *et al*., 1986, 1989). In order to identify such a subset of key residues from the sequence environment, we select positions having the following properties in every example within a cluster: (1) buried hydrophobic residues within the loop; (2) framework residues that make hydrophobic contacts with residues identified in the previous step; (3) residues in the loop that make side-chain hydrogen bonds with the framework; (4) residues in the loop that make side-chain hydrogen bonds with the loop's backbone; (5) residues in the framework that make side-chain hydrogen bonds with the loop; (6) proline or glycine residues in the loop when conserved throughout clusters with at least two members or *cis*-proline residues in one-member clusters; (7) other residues conserved throughout clusters with at least four members. Buried residues were identified using the Lee & Richards (1971) solvent-accessibility algorithm while hydrogen bonds were defined according to the simple criteria of Baker & Hubbard (1984): where a hydrogen atom position can be defined, the maximum hydrogen to acceptor distance is 2.5 Å; when the exact position of the hydrogen atom cannot be defined (for example in serine residues where the C^{β}-O^{γ} bond is freely rotatable), the maximum donor to acceptor distance is

3.5 Å; in both cases, the angle at the acceptor atom and (where available) the hydrogen atom is 90 to 180°.

This procedure to identify key residues is valid for clusters with multiple members, but can be applied to single structures to identify residues that appear to be important in defining conformation, although more positions will be defined than really necessary. As more structures become available with the same conformation, the algorithm ''learns'' which residue are truly important in determining conformation.

Acknowledgements

A.C.R.M. thanks to UK Medical Research Council for support.

References

Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.

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

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

Brady, R. L., Edwards, D. J., Hubbard, R. E., Jiang, J. S., Lange, G., Roberts, S. M., Todd, R. J., Adair, J. R., Emtage, J. S., King, D. J. & Low, D. C. (1992). Crystal structure of a chimeric Fab' fragment of an antibody binding tumor-cells. *J. Mol. Biol.* **227**, 253–264.

Bruccoleri, R. E. & Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, **26**, 137–168.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **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., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M. Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883.

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). Structural repertoire of the human V_H segments. *J. Mol. Biol.* **227**, 799–817.

Claessens, M., Vancutsem, E., Lasters, I. & Wodak, S. (1989). Modeling the polypeptide backbone with spare parts from known protein structures. *Protein Eng.* **2**, 335–345.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, National Biomedical Research Foundation, Silver Spring, Washington DC.

Fidelis, K., Stern, P. S., Bacon, S. & Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein-structure. *Protein Eng.* **7**, 953–960.

Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L. & Levinthal, C. (1986). Predicting antibody hypervariable loop conformations II: minimization and molecular dynamics studies of McPC603 from many randomly generated loop conformations. *Proteins: Struct. Funct. Genet.* **1**, 342–362.

Golinelli-Pimpaneua, B., Gigant, B., Bizebard, T., Navaza, J., Saludjian, P., Zemel, R., Tawfik, D. S., Eshhar, Z., Green, B. S. & Knossow, M. (1994). Crystal structure of a catalytic antibody Fab with esterase-like activity. *Structure*, **2**, 175–183.

Gregory, D. S. (1992). Prediction, design and characterisation of metal binding sites in antibodies. D. Phil. thesis, University of Oxford.

Gregory, D. S., Martin, A. C. R., Cheetham, J. C. & Ress, A. R. (1993). The prediction and characterization of metal binding sites in proteins. *Protein Eng.* **6**, 29–35.

Guddat, L. W., Shan, L., Anchin, J. M., Linthicum, D. S. & Edmundson, A. B. (1994). Local and transmitted conformational changes on complexation of an anti-sweetener Fab. *J. Mol. Biol.* **236**, 247–274.

Haynes, M. R., Stura, E. A., Hilvert, D. & Wilson, I. A. (1994). Routes to catalysis: structure of a catalytic antibody and comparison with its natural counterpart. *Science*, **263**, 646–652.

Herron, J. N., He, X. M., Mason, M. L., Voss, E. W. & Edmundson, A. B. (1989). 3-Dimensional structure of a fluorescein Fab complex crystallized in 2-methyl-2,4-pentanediol. *Proteins: Struct. Funct. Genet.* **5**, 271–280.

Herron, J. N., He, X. M., Ballard, D. W., Blier, P. R., Pace, P. E., Bothwell, A. L. M., Voss, E. W. & Edmundson, A. B. (1991). An autoantibody to single-stranded DNA – comparison of the 3-dimensional structures of the unliganded Fab and a deoxynucleotide Fab complex. *Proteins: Struct. Funct. Genet.* **11**, 159–175.

Jedrzejas, M. J., Miglietta, J., Griffin, J. A. & Luo, M. (1995). Structure of a monoclonal anti-ICAM-1 antibody R6.5 Fab fragment at 2.8 Ångström resolution. *Acta Crystallog. sect. D*, **51**, 380–385.

Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N. & Sheriff, S. (1993). 26-10 Fab-digoxin complex – affinity and specificity due to surface complementarity. *Proc. Natl Acad. Sci. USA*, **90**, 10310–10314.

Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*. 5th edit., U. S. Department of Health and Human Services, National Institutes for Health, Bethesda, MD.

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

Leszczynski, J. F. & Rose, G. D. (1986). Loops in globular proteins: a novel category of secondary structure. *Science*, **234**, 849–855.

Love, R. A., Villafranca, J. E., Aust, R. M., Nakamura, K. K., Jue, R. A., Major, J. G., Radhakrishnan, R. & Butler, W. F. (1993). How the anti-(metal chelate) antibody CHA255 is specific for the metal-ion of its antigen – X-ray structures for 2 Fab' hapten

complexes with different metals in the chelate. *Biochemistry*, **32**, 10950–10959.
MacCallum, R. M., Martin, A. C. R. & Thornton, J. M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol*. **262**, 732–745.
Martin, A. C. R., Cheetham, J. C. & Rees, A. R. (1989). Modelling antibody hypervariable loops: a combined algorithm. *Proc. Natl Acad. Sci. USA*, **86**, 9268–9272.
Martin, A. C. R., Toda, K., Stirk, H. & Thornton, J. M. (1995). Long loops in proteins. *Protein Eng.* **8**, 1093–1101.
Milner-White, E. J. (1986). Classification of β-hairpin turns. *Biochem. Soc. Trans.* **14**, 877.
Milner-White, E. J. & Poet, R. (1986). Four classes of β-hairpins in proteins. *Biochem. J.* **240**, 289–292.
Milner-White, E. J. & Poet, R. (1987). Loops, bulges, turns and hairpins in proteins. *Trends Biochem. Sci.* **12**, 189–192.
Padlan, E. A., Silverton, E. W., Sheriff, S., Cohen, G. H., Smithgill, S. J. & Davies, D. R. (1989). Structure of an antibody antigen complex – crystal structure of the HyHel-10 Fab-lysozyme complex. *Proc. Natl Acad. Sci. USA*, **86**, 5938–5942.
Pedersen, J. T. (1993). Molecular modelling of antibody combining sites. PhD thesis, University of Bath.
Prasad, L., Sharma, S., van Donselaar, M., Quail, J. W., Lee, J. S., Waygood, E. B., Wilson, K. S., Dauter, Z. & Delbaere, L. T. J. (1993). Evaluation of mutagenesis for epitope mapping: structure of an antibody-protein antigen complex. *J. Biol. Chem.* **268**, 10705–10708.
Reczko, M., Martin, A. C. R., Bohr, H. & Suhai, S. (1995). Prediction of hypervariable CDR-H3 loop structures in antibodies. *Protein Eng.* **9**, 389–395.
Rooman, M. J., Wodak, S. J. & Thornton, J. M. (1989). Amino-acid sequence templates derived from recurrent turn motifs in proteins – critical-evaluation of their predictive power. *Protein Eng.* **3**, 23–27.
Saul, F. A. & Poljak, R. J. (1992). Crystal-structure of human-immunoglobulin fragment Fab NEW refined at 2.0 Ångström resolution. *Proteins: Struct. Funct. Genet.* **14**, 363–371.
Schiffer, M., Ainsworth, C., Xu, Z. B., Carperos, W., Olsen, K., Solomon, A., Stevens, F. J. & Chang, C. H. (1989). Structure of a 2nd crystal form of Bence-Jones protein Loc – strikingly different domain associations in 2 crystal forms of a single protein. *Biochemistry*, **28**, 4066–4072.
Searle, S. J., Pedersen, J. T., Henry, A. H., Webster, D. A. & Rees, A. R. (1994). Antibody structure and function. In *Antibody Engineering* (Borreback, C. A. K., ed.), pp. 3–51, Oxford University Press, Oxford, UK.
Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smithgill, S. J., Finzel, B. C. & Davies, D. R. (1987). 3-Dimensional structure of an antibody-antigen complex. *Proc. Natl Acad. Sci. USA*, **84**, 8075–8079.
Sibanda, B. L. & Thornton, J. M. (1985). β-Hairpin families in globular proteins. *Nature*, **316**, 170–174.
Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). Conformation of β-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* **206**, 759–777.
Staunton, D., Jones, A. E. & Rees, A. R. (1993). Creation of a metal-binding site within the antibody combining site. *Protein Eng.* **6** (Special Supplement), 93.
Strong, R. K., Campbell, R., Rose, D. R., Petsko, G. A., Sharon, J. & Margolies, M. N. (1991). 3-Dimensional structure of murine anti-para-azophenylarsonate Fab-36-71. 1. X-ray crystallography, site-directed mutagenesis, and modelling of the complex with hapten. *Biochemistry*, **30**, 3739–3748.
Sutcliffe, M. J. (1988). An automated approach to the systematic model building of homologous proteins. PhD thesis, University of London, Birkbeck College.
Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). Knowledge based modelling of homologous proteins. Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.
Swayne, D. F., Cook, D. & Buja, A. (1991). *User's Manual for XGobi, a Dynamic Graphics Program for Data Analysis Implemented in the X Window System*. Bellcore Technical Memorandum.
Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 208–229.
Tomlinson, I. A., Cox, J. P. L., Gherardi, E., Lesk, A. M. & Chothia, C. (1995). The structural repertoire of the human V_κ domain. *EMBO J.* **14**, 4628–4638.
Tramontano, A., Chothia, C. & Lesk, A. M. (1989). Structural determinants of the conformations of medium-sized loops in proteins. *Proteins: Struct. Funct. Genet.* **6**, 382–394.
Tramontano, A., Chothia, C. & Lesk, A. M. (1990). Framework residue-71 is a major determinant of the position and conformation of the 2nd hypervariable region in the V_H domains of immunoglobulins. *J. Mol. Biol.* **215**, 175–182.
Unger, R., Harel, D., Wherland, S. & Sussman, J. L. (1989). A 3D building-blocks approach to analyzing and predicting structure of proteins. *Proteins: Struct. Funct. Genet.* **5**, 355–373.
Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.* **58**, 236–244.
Wintjens, R., Rooman, M. J. & Wodak, S. J. (1996). Automatic classification and analysis of αα-turn motifs in proteins. *J. Mol. Biol.* **255**, 235–253.

Edited by F. E. Cohen

(Received 4 June 1996; received in revised form 22 August 1996; accepted 29 August 1996)