Chapter 3
Protein Sequence and Structure Analysis of Antibody Variable Domains

Andrew C.R. Martin

3.1 Introduction

The protocols described here provide methods for computational analysis of antibody sequence and structure. With the availability of the World Wide Web, many online analysis tools have been made available, and URLs for these are cited throughout the text. The author has provided a Web site at http://www.bioinf.org.uk/abs/labmanual/ containing links to all the tools described here.

3.1.1 Brief Review of Early Work on Sequence Variability and Antibody Structure

Porter (1959) first proposed the four-chain model for antibodies consisting of two light chains and two heavy chains, linked by disulphide bonds. The structure of antibodies has been reviewed in detail by a number of authors (Alzari et al. 1988; Padlan 1994; Searle et al. 1994), while the structural basis of antibody/antigen interactions has also been reviewed extensively (Padlan 1977; Mariuzza et al. 1987; Davies et al. 1990; Wilson and Stanfield 1993). The major points will be covered here briefly.

Edelman and Gall (1969) analysed sequences of IgG chains and identified homologous regions, which they proposed to be related to domains of specific function (Edelman 1970). Wu and Kabat (1970) examined the sequences of the region proposed to form the variable domain and identified "hypervariable" segments within that domain which, they proposed, formed the actual antigen

A.C.R. Martin
Institute of Structural and Molecular Biology, Darwin Building, University College London, Gower Street, London, WC1E 6BT, UK
e-mail: andrew@bioinf.org.uk

R. Kontermann and S. Dübel (eds.), *Antibody Engineering Vol. 2*,
DOI 10.1007/978-3-642-01147-4_3,

combining site. These were termed "complementarity determining regions" (CDRs), and they suggested that CDRs are supported on a framework formed by the rest of the variable domain.

IgG is the best studied of the immunoglobulin classes, and electron micrographs revealed the "Y" shape (Valentine and Green 1967). The structure of antibodies is divided into variable regions able to bind to a virtually infinite range of substrates and constant regions able to perform a given set of common functions for all antibodies within a class (IgG, IgM, etc.).

Light chains consist of V_L and C_L domains, while heavy chains consist of V_H, C_H1, C_H2 and C_H3 domains (IgM and IgE have an additional C_H4 domain). Various fragments are generated artificially or by proteolytic digestion:

Fv	V_H/V_L dimer
Fab	A single arm of the "Y", consisting of a V_H,C_H1/V_L,C_L dimer (from Papain cleavage)
$F(ab')_2$	The two Fab arms of the "Y" joined by the disulphide(s) between the heavy chains (from Pepsin cleavage)
Fc	The stem consisting of C_H2,C_H3/C_H2,C_H3 (from Papain cleavage)

The first X-ray crystal structure of a Fab fragment was solved in 1973 (Poljak et al. 1973), and it showed that the hypervariable regions corresponded approximately to structural loops. The anti-lysozyme antibody D1.3 was the first antibody crystal structure to be solved complexed with antigen (Amit et al. 1985), confirming the role of the CDRs in binding antigen.

Variable and constant domains consist of two twisted antiparallel β-sheets, which form a β-sandwich. Constant regions have three and four stranded sheets, while variable regions have a further two short strands forming sheets of four and five strands. The two sheets are linked by a conserved disulphide bond and are inclined by approximately 30° to one another (Chothia and Janin 1981), varying by up to 18° in variable domains and up to 10° in constant domains (Lesk and Chothia 1982).

Packing between the V_L/V_H domains can vary between antibodies. Recent analysis of more than 500 antibody structures in the author's group has shown a mean packing angle of –45.6° with a standard deviation of 3.36° and an overall observed range of approximately 30° (Abhinandan and Martin, in preparation). In addition to variation in V_L/V_H domain packing, the "elbow angle" describes the flexibility between the V_L/V_H and C_L/C_H1 pseudo-diads. The angle between the arms of the "Y" is variable as a result of a flexible hinge region (deleted in IgM). The role of flexibility in antigen binding is reviewed by Huber and Bennett (1987).

3.1.2 Linking Sequence and Structure

Once a structure has been solved for a protein in a homologous family, only one simple ingredient is needed in order to link the sequences of other homologous family members to that structure: a standardised numbering scheme. In this way, one always

knows that, for example, residue number 35 is at the start of a β-strand. Insertions in the sequence relative to that standard numbering scheme are given numbers such as 27A, while deletions are accommodated by simply skipping numbers.

Ideally, such schemes are designed in the light of both large amounts of sequence information and multiple structures. Insertion sites (i.e. residue 27A, etc.) are placed only in loop regions (or form β-bulges) and have structural meaning such that topologically equivalent residues in these loops get the same label.

3.1.3 The Kabat Numbering Scheme

The Kabat numbering scheme is the most widely adopted standard for numbering the residues in an antibody in a consistent manner. However, the scheme does have problems.

The numbering scheme was developed solely from somewhat limited sequence data. Unfortunately, the position at which insertions are placed in CDR-L1 and CDR-H1 does not match the structural insertion position. Thus, topologically equivalent residues in these loops do not get the same number.

The second problem is that the numbering adopts a rigid specification. For example, in the potentially very long CDR-H3, insertions are numbered between residue H100 and H101 with letters up to K (i.e. H100, H100A,. . ., H100K, H101). If there are more residues than that, there is no standard way of numbering them. Such situations occur at other positions too. The raw Kabat data files in these cases simply state what the additional insertions are and where they occur, but do not assign numbers to them.

The numbering throughout the chains is shown in Table 3.1.

3.1.4 The Chothia Numbering Scheme

The Chothia numbering scheme is identical to the Kabat scheme with the exception of CDR-L1 and CDR-H1, where the insertions are placed at the structurally correct positions. This means that topologically equivalent residues in these loops do get the same label.

There are two disadvantages. First, the Kabat scheme is so widely used that some confusion can arise. Second, Chothia et al. erroneously changed their numbering scheme in their 1989 Nature paper (Chothia et al. 1989) such that insertions in CDR-L1 are placed after residue L31 rather than L30. A visual examination of the conformations of CDR-L1 loops shows that L30 is the correct position. Chothia's group returned to using residue L30 as the insertion site in CDR-L1 in their 1997 paper on CDR conformation (Al-Lazikani et al. 1997).

The structurally correct version of the Chothia numbering (as used before 1989 and since 1997) throughout the chains is shown in Table 3.2.

Table 3.1 Kabat numbering scheme

Light chain																		
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27												
27A	27B	27C	27D	27E	27F														
								28	29	30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95				
95A	95B	95C	95D	95E	95F														
																96	97	98	99
100	101	102	103	104	105	106													
106A																			
							107	108	109										
Heavy chain																			
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35				
35A	35B																		
																36	37	38	39
40	41	42	43	44	45	46	47	48	49	50	51	52							
52A	52B	52C																	
													53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
80	81	82																	
82A	82B	82C																	
			83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
100																			
100A	100B	100C	100D	100E	100F	100G	100H	100I	100J	100K									
	101	102	103	104	105	106	107	108	109	110	111	112	113						

Table 3.2 Chothia numbering scheme

Light chain																		
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29	30									
30A	30B	30C	30D	30E	30F														
											31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95				
95A	95B	95C	95D	95E	95F											96	97	98	99
100	101	102	103	104	105	106													
106A																			
							107	108	109										
Heavy chain																			
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29	30	31								
31A	31B																		
												32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49	50	51	52							
52A	52B	52C																	
													53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
80	81	82																	
82A	82B	82C																	
			83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
100																			
100A	100B	100C	100D	100E	100F	100G	100H	100I	100J	100K									
	101	102	103	104	105	106	107	108	109	110	111	112	113						

3.1.5 The Abhinandan (Updated Chothia) Numbering Scheme

The Chothia numbering scheme provides structurally correct numbering in the CDRs, but does not examine the framework regions. An updated version has recently been described by Abhinandan and Martin (2008) which provides for structurally correct insertion/deletion sites in the frameworks as well as the CDRs. The most significant difference is the insertion site in the third heavy framework region which appears after residue H72 rather than H82 and the introduction of an insertion/deletion site in CDR-L2 at L52. This scheme is illustrated in Table 3.3 where standard locations of deleted residues are indicated with parentheses.

3.1.6 Other Numbering Schemes

Two other numbering schemes have also proven popular. The immunogenetics (IMGT) database (Giudicelli et al. 2006) has introduced a numbering scheme (Lefranc et al. 2003) which unifies numbering across antibody lambda and kappa light chains, heavy chains and T-cell receptor chains (see http://imgt.cines.fr/textes/IMGTScientificChart/Numbering/IMGTnumbering.html). The scheme avoids the use of insertion codes for all but the least common very long insertions. While the published version places all insertions at the ends of CDRs (and is therefore not structurally correct), recent changes in their V-QUEST software have rectified this (Brochet et al. 2008).

The "Aho" numbering scheme was introduced by Honegger and Plückthun (2001). This can be considered as a more structurally correct version of the IMGT scheme. Insertions and deletions, rather than growing uni-directionally, are placed symmetrically around a key position. Furthermore, length variations in CDR-1 and CDR-2 are accounted for by a single gap position in all other schemes, whereas the Aho scheme has two locations at which gaps or insertions may be introduced.

While these schemes, in particular the Aho scheme, have distinct advantages over the earlier schemes, it has been hard for them to gain acceptance because the Kabat and Chothia schemes are so well established.

3.1.7 CDR Definitions

Table 3.4 illustrates the main definitions of the CDRs which are commonly in use:

- The Kabat definition is based on sequence variability and is the most commonly used.
- The Chothia definition is based on the location of the structural loop regions.

Table 3.3 Abhinandan (updated Chothia) numbering scheme

Light chain																			
0	1	2	3	4	5	6	7	8	9	(10)	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29	(30)									
30A	30B	30C	30D	30E	30F														
											31	32	33	34	35	36	37	38	39
40																			
40A	(41)	42	43	44	45	46	47	48	49	50	51	(52)							
52A	52B	52C	52D	52E															
													53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	(68)											
68A	68B	68C	68D	68E	68F	68G	68H												
									69	70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	(95)				
95A	95B	95C	95D	95E	95F														
						96	97	98	99	100	101	102	103	104	105	106	107		
107A																			
								108	109										
Heavy chain																			
0	1	2	3	4	5	6	7	(8)											
8A	8B	8C	8D																
									9	10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29	30	(31)								
31A	31B																		
												32	33	34	35	36	37	38	39
40	41	(42)	43	44	45	46	47	48	49	50	51	(52)							
52A	52B	52C																	
													53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69	70	71	72							
72A	72B	72C																	
													73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
(100)																			
100A	100B	100C	100D	100E	100F	100G	100H	100I	100J	100K									
	101	102	103	104	105	106	107	108	109	110	111	112	113						

Table 3.4 Different definitions of the CDRs

Loop	Kabat	AbM	Chothia	Contact
L1	L24–L34	L24–L34	L24–L34	L30–L36
L2	L50–L56	L50–L56	L50–L56	L46–L55
L3	L89–L97	L89–L97	L89–L97	L89–L96
H1 (Kabat numbering)	H31–H35B	H26–H35B	H26–H32...34	H30–H35B
H1 (Chothia numbering)	H31–H35	H26–H35	H26–H32	H30–H35
H2	H50–H65	H50–H58	H52–H56	H47–H58
H3	H95–H102	H95–H102	H95–H102	H93–H101

Note that in their 1997 paper, Chothia's group has used the AbM definition for CDR-H2

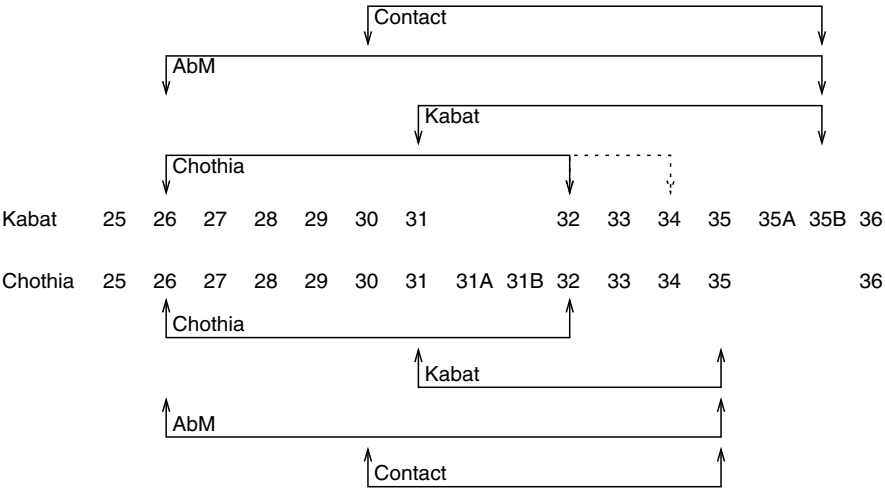

Fig. 3.1 Boundaries of different definitions of CDR-H1 using the Kabat and Chothia numbering schemes

- The AbM definition is a compromise between Kabat and Chothia definitions based on that used by Martin et al. (1989) and used by Oxford Molecular's AbM antibody modelling software.
- The contact definition was introduced by MacCallum et al. (1996) and is based on an analysis of available complex crystal structures. This definition is likely to be the most useful for people wishing to perform mutagenesis to modify the affinity of an antibody since these are residues which take part in interactions with antigen.

Note that the end of the Chothia CDR-H1 loop, when numbered using the Kabat numbering convention, varies between H32 and H34 depending on the length of the loop as illustrated in Fig. 3.1.

3.2 Procedure

The following subprotocols describe a number of aspects of sequence/structure analysis.

3.2.1 Accessing Kabat Sequence Data

Since the first edition of this book was published, maintenance of the Kabat sequence data has ceased and EMail query access and web resources are no longer available from Johnson, Kabat and Wu's web site. The last publicly available update of the data was released in April 2000. Updates continued on a for-fee basis until October 2003 when the datasets were closed. The 2003 dataset together with analysis tools may be obtained on payment of a licence fee from http://www.kabatdatabase.com/.

However, the last public release of the data remains a useful resource, primarily because the data can be downloaded with the standard Kabat numbering applied to the sequences. Unfortunately recent analysis in the author's group has shown that up to 10% of entries have errors or inconsistencies in the numbering (Abhinandan and Martin 2008).

The data may be accessed in a number of ways described as follows:

3.2.1.1 Download the Raw Data

The raw Kabat sequence data may be downloaded for local analysis from either:

- ftp://ftp.ncbi.nlm.nih.gov/repository/kabat/
- ftp://ftp.ebi.ac.uk/pub/databases/kabat/

The most up-to-date raw data are in the fixlen subdirectory with a 1999 FASTA format dump in the fasta_format subdirectory, and analysed data from 1996 in PostScript format.

3.2.1.2 KabatMan

KabatMan (Martin 1996) is a specialised database for the analysis of Kabat antibody sequence data. It may be queried using a language similar to SQL ("structured query language" – a standard for querying relational databases) or via a point and click interface at http://www.bioinf.org.uk/abs/simkab.html.

While the software provided by Johnson, Wu and Kabat (available from http://www.kabatdatabase.com/ for a license fee) is largely aimed at finding all the information about a single antibody, KabatMan is more suited to global analysis of the antibody data. It also provides a more direct link with structural information

by allowing searches to specify individual amino acids or the contents of one of the six CDRs. For example, one could find all the antibodies which bind to DNA, but do not contain arginine in CDR-H3 using the query:

```
SELECT      Name
WHERE       Antigen INC 'dna'
            h3 <> '' AND
            h3 INC 'R' NOT AND
```

The "SELECT" statement specifies which results are to be returned (in this case the name of the antibody). The "WHERE" statement first specifies that the antigen should be DNA, then specifies that the sequence of CDR-H3 should not be blank (i.e. unknown) and finally specifies that CDR-H3 should not include the letter "R" (i.e. arginine). More examples are given in the paper that describes KabatMan and in the on-line help.

When using KabatMan to find details of a specific antibody, it is important not to overspecify the antibody. For example if you want to find details of the mouse anti-lysozyme antibody HYHEL-5, then the name, HYHEL-5 uniquely identifies the antibody – there is no point in specifying the antigen, class or animal source.

3.2.1.3 Abysis

Abysis is a new resource integrating Kabat sequence data with IMGT and PDB data and is described in Section 3.2.4.

3.2.1.4 SUBIM

Deret et al. (1995) have written a program, SUBIM, for the analysis of Kabat sequence data. Many of the data access functions of this program are also available in KabatMan (described earlier). The program must be downloaded and installed locally, and it has no form of graphical user interface. Since the original download site is no longer available, the author has made the software available at http://www.bioinf.org.uk/abs/subim.tar.gz.

3.2.2 Accessing IMGT Sequence Data

IMGT (Giudicelli et al. 1997, 2006) is a databank in which the large quantity of DNA sequence data for antibodies (and other proteins of the immune system) has been extracted from the EMBL databank. The data in IMGT are updated on a regular basis.

The chief advantage of IMGT is the volume of data it contains and the rate at which these are updated. The main disadvantage compared with the Kabat data is that the sequence information cannot currently be downloaded with numbering applied in a straightforward manner. IMGT is primarily a DNA sequence databank, and the raw data ftp://ftp.ebi.ac.uk/pub/databases/imgt/ligm/ are stored in an EMBL-like format with protein sequences present as translations. For some entries, protein sequences of regions (such as CDRs and frameworks) are provided separately together with the IMGT-numbered range they represent.

IMGT provides a number of services via the web:

- http://www.imgt.org/ The main IMGT server
- http://www.ebi.ac.uk/imgt/ European Bioinformatics Institute (EBI) site which provides access to the main site and to a Sequence Retrieval Service (SRS) interface.
- http://www.imgt.org/IMGT_vquest/share/textes/ A set of tools for analysis of antibody and T-cell receptor nucleotide sequences.

The main IMGT server (http://www.imgt.org/) provides a hierarchical interface to the data allowing one to home in on a particular sequence. This is not suited to global analysis of the data.

3.2.3 Accessing Antibody Structure Data

Antibody structures are available from the Protein Databank (Berman et al. 2000) (PDB, http://www.rcsb.org/). The problem, however, is to identify them. Sequence search methods will also find related sequences such as T-cell receptors, while keyword searches may lead to missing or spurious entries. The SACS database (Allcorn and Martin 2002) (http://www.bioinf.org.uk/abs/sacs/) uses a careful set of keyword and sequence tests followed by a final manual confirmation, to identify antibody structures in the PDB. It is updated approximately every 6 months.

3.2.4 Abysis: Integrated Access to Sequence and Structure Data

Abysis (http://www.bioinf.org.uk/abysis/) is a new integrated web-accessible database which integrates sequence data obtained from Kabat, IMGT and the PDB with structural data from the PDB. All sequences are numbered using the Kabat and Chothia standards and the system is designed to allow easy expansion to other schemes. Numbering is applied using the author's automated numbering system (Abhinandan and Martin 2008) to ensure accuracy and consistency.

The interface allows similar searches to KabatMan, but with a rather more extensive set of options. Results can be displayed in a table, or as XML, and sequences may be displayed as FASTA files or in the style of the Kabat book

(Kabat et al. 1991). Abysis also allows BLAST searches to be performed against the sequences and enables plots of distributions of residues at given positions and of the lengths of CDR and framework regions.

3.2.5 Assigning Subgroups

Deret's SUBIM program (Deret et al. 1995) which may be downloaded from http://www.bioinf.org.uk/abs/subim.tar.gz allows the assignment of the subgroup of new human sequences by comparison of the N-terminal 15 residues with consensus sequences determined by Kabat et al. (1991).

Sophie Deret has kindly made this part of the SUBIM program available to the author who has made this accessible via a Web page (http://www.bioinf.org.uk/abs/hsubgroup.html). In addition, the functionality has been incorporated into KabatMan to allow human subgroup assignment for sequences in the Kabat data.

3.2.6 Identifying the CDRs

This protocol describes how to identify the CDRs (Kabat definition) by examining the sequence. Of course there are always (minor) exceptions to these rules, so the word "always" should be interpreted with care! For example, CDR-L2 is always seven residues, but antibody NEW (Protein Databank code: 7FAB, http://www.rcsb.org/) has a deletion in this region. This also means that the position of the start of CDR-L3 is no longer 33 residues after the end of CDR-L2.

CDR-L1

- *Start* approximately residue 24
- *Residue before* is always C
- *Residue after* is always W. Typically WYQ, but also, WLQ, WFQ, WYL
- *Length* 10–17 residues

CDR-L2

- *Start* always 16 residues after the end of CDR-L1
- *Residues before* generally IY, but also, VY, IK, IF
- *Length* always seven residues

CDR-L3

- *Start* always 33 residues after end of CDR-L2
- *Residue before* is always C
- *Residues after* always FGXG
- *Length* 7–11 residues

CDR-H1

- *Start* approximately residue 31 (always 9 after a C) (Chothia/AbM definition starts 5 residues earlier)
- *Residues before* always CXXXXXXXX
- *Residues after* always W. Typically WV, but also WI, WA
- *Length* 5–7 residues (Kabat definition); 7–9 residues (Chothia definition); 10–12 residues (AbM definition)

CDR-H2

- *Start* always 15 residues after the end of Kabat/AbM definition of CDR-H1
- *Residues before* typically LEWIG, but a number of variations
- *Residues after* K[RL]IVFT[AT]SIA (where residues in square brackets are alternatives at that position)
- *Length* Kabat definition 16–19 residues (AbM definition and most recent Chothia definition ends seven residues earlier; earlier Chothia definition starts two residues later and ends nine earlier).

CDR-H3

- *Start* always 33 residues after end of CDR-H2 (always three after a C)
- *Residues before* always CXX (typically CAR)
- *Residues after* always WGXG
- *Length* 4–24 residues.

3.2.7 Screening New Antibody Sequences

Given a new antibody sequence, one is likely to wish to assign families and subgroups using the tools described earlier. An additional facility is available at http://www.bioinf.org.uk/abs/seqtest.html to identify unusual features in the sequence.

It is simply necessary to enter the amino acid sequence of your Fv fragment (one or both chains). Optionally you may include the whole Fab fragment, but only the Fv portion will be tested.

The tool aligns the provided sequence with a standard sequence in order to assign standard Kabat numbering and then uses the KabatMan database to identify unusual amino acids (i.e. those occurring in less than 1% of the data in the database). This allows the identification of potential cloning artefacts and sequencing errors. If unusual features are verified as being correct, then these residues are likely to be critical to the specificity of the antibody. The method is described in detail at http://www.bioinf.org.uk/abs/seqmethod.html.

The results need to be examined carefully. A typical sequence has 1–2 "unusual" residues. Very unusual sequence features and loops longer than anything observed in the current Kabat database may cause the alignment stage to fail causing errors in the Kabat numbering. Errors can also occur at the C-terminus of the chains. These problems can lead to residues being flagged as "unusual" when they are not.

Thus, if more than two or three unusual residues are seen (especially outside the CDRs and if insertions/deletions are observed), the first step is to ensure that the alignment and assignment of Kabat numbering is correct (check the features described in the section "Identifying the CDRs"). If all is judged correct with the alignment, the clone should be checked. If it is confirmed that no cloning errors have occurred, then it is likely that these are key features of the antibody.

3.2.8 Assigning Canonicals

The analysis by Chothia and co-workers introduced the concept of canonical conformations for the CDRs. It was proposed that these conformations were influenced by just a small number of residues either in the CDRs or in the framework regions which pack with them. This allows a direct prediction of three-dimensional conformation from sequence.

Chothia and co-workers have published around ten papers describing this analysis, but unfortunately do not provide a summary of the required amino acids to assign canonical classes. Table 3.5 attempts to summarise the data from their publications together with results from Martin and Thornton (1996). Chothia numbering of the sequences is used throughout.

3.2.9 Modelling Antibodies

There are various approaches to modelling antibodies, but it is widely accepted that methods based on Chothia's analysis of CDR canonical conformations provide the best results where they can be applied.

Any modelling procedure involves the following steps.

3.2.9.1 Build the Framework

Antibody crystal structures from the Protein Databank (PDB, http://www.rcsb.org/) are searched to identify the most similar light and heavy chains. These are identified separately. If the best match for the light chain is La (paired with Ha) and the best match for the heavy chain is Hb (paired with Lb) then the structure of La is least squared fitted to Lb and chains La and Hb are retained, deleting Ha and Lb. In this way, the V_L/V_H packing angle is inherited from Lb/Hb. To inherit the packing angle

Table 3.5 Key residues which define the Chothia canonical classes

CDR-L1									
Class	1	2	3	4	5	6	5λ	6λ	7λ
Length	10	11	17	16	15	12	13	14	14
L2	I	I	I	VIL	I	N			
L25	A	A	S	S	A	A	G	G	S
L29	VIL	VIL	VIL	L	V	V			
L30	–						I	I	V
L30D				G					
L33	ML	VIL	L	L	L	L	V	V	A
L71	YF	YF	YF	F	F	Y	A	A	A
CDR-L2									
Class	1								
Length	7								
L34	N								
CDR-L3									
Class	1	2	3	4	5	4λ	5λ		
Length	9	9	8	7	10	9	11		
L90	QNH	Q	Q	Q	Q				
L94		P					DNG		
L95	P	L	P						
L97	T	T		–	T	IV	VG		
CDR-H1									
Class	1	2	3						
Length	10	11	12						
H24	TAVGS	VF	VFG						
H26	G	G	G						
H29	IFLVS	IL	VIL						
H34	IVMWTL	WC	WV						
H94	RKGSHNTA	HR	HR						
CDR-H2									
Class	1	2	3	4					
Length	9	10	10	12					
H54			SGND	KS					
H55	GD	GS	GS	Y					
H71	RHVI	VALT	R	R					

For CDR-H1, Chothia et al. suggest that residue H27 is also a key residue, but Martin and Thornton did not find this residue influencing the conformation. Similarly, for CDR-H2, Chothia et al. identify residue H52A as a key residue in determining the conformation of Classes 2 and 3, but Martin and Thornton found that this residue does not influence the conformation

from La/Ha, the fitting is performed on the heavy chains. The choice is relatively arbitrary, and it may be worth constructing two models.

Structural fitting is best performed using a program such as ProFit (http://www.bioinf.org.uk/software/).

The sidechains of the framework are then replaced using automated processes available in molecular graphics programs, or software such as CONGEN (Bruccoleri and Karplus 1987). Sidechains are generally built using the "Maximum Overlap

Protocol" where the atom positions are inherited from the parent wherever possible, and where not possible, conformations are taken from a rotamer library.

3.2.9.2 Build the CDRs

The methods described earlier are used to identify canonical classes. This is generally possible for four or five of the six CDRs. CDR-H3 is too variable to be classified at present (it may become possible once sufficient structures are available and a number of authors have begun to suggest limited sets of rules). Antibody crystal structures from the Protein Databank are then searched to find CDRs of the correct canonical classes with maximum sequence identity to the sequence to be modelled. These loops are then attached to the framework either manually using molecular graphics, or by using a least squares fitting program, such as ProFit, to fit the three residues on either side of the loop itself (i.e. within the framework region).

CDRs which cannot be built using canonicals may be built by conformational search using CONGEN, by searching the PDB for loops of the same length and with similar distance between the attachment points to the framework, or by combined methods such as CAMAL (Martin et al. 1989). Such loops can never be built with a high degree of confidence in their accuracy.

Sidechains of the CDRs are then built as described earlier.

3.2.9.3 Refinement and Assessment

Finally, the model may be refined by energy minimisation using a package such as GROMOS, CHARMM or DISCOVER, and structural quality should be assessed using a program such as ProCheck (http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html) or WhatCheck (http://swift.cmbi.ru.nl/gv/whatcheck/). Both programs may be accessed via an online server at http://www.jcsg.org/scripts/prod/validation1.cgi.

3.2.9.4 Automated Methods

The earlier description details the stages that are necessary in a manual modelling protocol. As an alternative, a number of automated procedures are available. Two of these are general automatic modelling programs and may be used to generate models in a quick and simple manner. However, they do not take advantage of the special properties of antibodies. The third is a program specially designed for the automated modelling of antibodies.

- *MODELLER* is a general purpose automated protein modelling program (http://www.salilab.org/modeller/). As such, it is able to produce reasonably reliable models of structures given just a sequence or a sequence aligned to one or more templates from the Protein Databank. However, since it is not designed

specifically for antibody modelling, it does not make use of Chothia's canonical analysis and will not model the CDRs as accurately. The software must be downloaded and run locally on a Unix type computer system.

- *SwissModel* is another general purpose automated protein modelling program which is accessible over the Web and does not need to be installed and run locally (http://swissmodel.expasy.org/). The quality of models is generally not as high as those created by MODELLER, and the same caveats about not being antibody-specific apply.
- *WAM* is a web server specifically designed for modelling antibodies (http://antibody.bath.ac.uk/). The methodology is based on AbM which was a commercial program available from Oxford Molecular. Being antibody-specific, it automates the manual procedure described earlier taking account of canonical structures and using the CAMAL modelling method described by Martin et al. (1989) for modelling those loops that cannot be built using canonicals. Academics may submit five sequences a month while commercial use requires payment of a license fee.
- *Rosetta Antibody Modelling* is another web server specifically designed for modelling antibodies (http://antibody.graylab.jhu.edu/). The method uses a CDR grafting technique for the light chain CDRs, CDR-H1 and CDR-H2. CDR-H3 is then built using a combination of fragment insertion and low-resolution moves. This is followed by sidechain rebuilding, refinement of V_H/V_L packing and minimisation (Sivasubramanian et al. 2009).

3.2.10 Analysis Tools Applied to Humanisation

Sequence/structure analysis programs such as Abysis and KabatMan can be applied to problems such as humanisation as well as being used for general analysis. For example, Saldanha et al. (1999) used KabatMan to identify a residue that restored the binding of a humanised antibody.

In brief, the humanisation protocol was as follows. Mouse CDRs were grafted onto human frameworks with highest sequence identity. Residues in the framework were then considered for "back-mutation" to restore full binding. First, key residues (identified by Chothia's canonical analysis) were identified in the framework and back-mutated to those seen in the original mouse antibody. KabatMan was then used to identify residues in the human framework, which are particularly unusual in mouse frameworks, even though they may be remote from the combining site. Nine such positions were identified and these were examined on a computer model. Seven of these were conservative and one was a surface residue. However, position 9 in the light chain was unique to the human kappa IV subclass and only seen in one of 1,848 mouse kappa sequences. Back-mutation of this residue to that seen in the mouse sequence completely restored binding.

References

Abhinandan KR, Martin ACR (2008) Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. Mol Immunol 45:3832–3839

Abhinandan KR, Martin ACR (in preparation) Analysis and prediction of VH/VL packing in antibodies

Al-Lazikani B, Lesk AM, Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. J Mol Biol 273:927–948

Allcorn LC, Martin ACR (2002) SACS-self-maintaining database of antibody crystal structure information. Bioinformatics 18:175–181

Alzari PM, Lascombe M-B, Poljak RJ (1988) Three-dimensional structure of antibodies. Annu Rev Immunol 6:555–580

Amit AG, Mariuzza RA, Phillips SEV, Poljak RJ (1985) Three-dimensional structure of an antigen–antibody complex at 6 Å resolution. Nature (London) 313:156–158

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucl Acids Res 28:235–242

Brochet X, Lefranc M-P, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucl Acids Res 36:W503–W508

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

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

Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ (1989) Conformations of immunoglobulin hypervariable regions. Nature (London) 342:877–883

Davies DR, Padlan EA, Sheriff S (1990) Antibody–antigen complexes. Annu Rev Biochem 59:439–473

Deret S, Maissiat C, Aucouturier P, Chomilier J (1995) SUBIM: A program for analysing the Kabat database and determining the variability subgroup of a new immunoglobulin sequence. Comput Appl Biosci 11:435–439

Edelman GM (1970) The covalent structure of a human γ-immunoglobulin: XI functional implications. Biochemistry 9:3197–3205

Edelman GM, Gall WE (1969) The antibody problem. Annu Rev Biochem 38:415–466

Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, Bontrop R, Marc L, Malik A, Lefranc MP (1997) IMGT, the international ImMunoGeneTics database. Nucl Acids Res 25:206–211

Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P (2006) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucl Acids Res 34:D781–D784

Honegger A, Plückthun A (2001) Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. J Mol Biol 309:657–670

Huber R, Bennett WS (1987) Antibody–antigen flexibility. Nature (London) 326:334–335

Kabat EA, Wu TT, Perry HM, Gottesman KS, Foeller C (1991) Sequences of proteins of immunological interest, 5th edn. U.S. Department of Health and Human Services, National Institutes for Health, Bethesda, MD

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55–77

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

MacCallum RM, Martin AC, Thornton JM (1996) Antibody–antigen interactions: Contact analysis and binding site topography. J Mol Biol 262:732–745

Mariuzza RA, Phillips SEV, Poljak RJ (1987) The structural basis of antigen–antibody recognition. Annu Rev Biophys Bioeng 16:139–159

Martin ACR (1996) Accessing the Kabat antibody sequence database by computer. Proteins Struct Funct Genet 25:130–133

Martin ACR, Thornton JM (1996) Structural families in homologous proteins: Automatic classification, modelling and application to antibodies. J Mol Biol 263:800–815

Martin ACR, Cheetham JC, Rees AR (1989) Modelling antibody hypervariable loops: A combined algorithm. Proc Natl Acad Sci USA 86:9268–9272

Padlan EA (1977) The structural basis for the specificity of antibody–antigen reactions and structural mechanisms for the diversification of antigen-binding specificities. Quant Rev Biophys 10:35–65

Padlan EA (1994) Anatomy of the antibody molecule. Mol Immunol 31:169–217

Poljak RJ, Amzel LM, Avey HP, Chen BL, Phizackerley RP, Saul F (1973) The three-dimensional structure of the Fab′ fragment of a human immunoglobulin at 2.8 Å resolution. Proc Natl Acad Sci USA 70:3305–3310

Porter RR (1959) The hydrolysis of rabbit γ-globulin and antibodies with crystalline papain. Biochem J 73:119–127

Saldanha JW, Martin AC, Léger OJ (1999) A single backmutation in the human κIV framework of a previously unsuccessfully humanized antibody restores the binding activity and increases the secretion in cos cells. Mol Immunol 36:709–719

Searle SJ, Pedersen JT, Henry AH, Webster DM, Rees AR (1994) Antibody structure and function. In: Borreback CAK (ed) Antibody engineering. Oxford University Press, London, pp 3–51

Sivasubramanian A, Sircar A, Chaudhury A, Gray JJ (2009) Toward high-resolution homology modeling of antibody Fv regions and application to antibody–antigen docking. Proteins 74:497–514

Valentine RC, Green NM (1967) Electron microscopy of an antibody–hapten complex. J Mol Biol 27:615–617

Wilson IA, Stanfield RL (1993) Antibody–antigen interactions. Curr Opin Struct Biol 3:113–118

Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 132:211–250