Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains

K. R. Abhinandan and Andrew C. R. Martin*

Institute of Structural and Molecular Biology,
Division of Biosciences,
University College London, Gower Street,
London WC1E 6BT, United Kingdom

*Corresponding author

Running title: Analysis of Antibody Numbering

Keywords: Automated numbering; Kabat numbering; Chothia numbering; antibody sequences; computer software

Abstract

In analysing protein sequence and structure, standardized numbering schemes allow comparison of features without explicit alignment. This has proved particularly valuable in the case of antibodies. The most widely used schemes (Kabat: sequence-based; Chothia: structure-based) differ only in the numbering of the Complementarity Determining Regions (CDRs). We have analyzed the numbered annotations in the widely-used Kabat database and found that approximately 10% of entries contain errors or inconsistencies. Further analysis of sequence alignments in the context of structure suggest that the sites of the insertions in some framework regions in the Kabat and Chothia schemes are incorrect. We therefore propose a corrected version of the Chothia scheme which is structurally correct throughout the CDRs and frameworks. To perform this analysis, we have developed, and made available, a tool for the automatic application of Kabat, Chothia and modified-Chothia numbering schemes and have carefully benchmarked the performance of this tool.

Introduction

Antibodies are amongst the most important classes of proteins involved in the adaptive immune system. Together with T-cell receptors, they provide a robust system of defense against infection caused by foreign bodies (Berek and Milstein, 1987). Both families of proteins adopt the immunoglobulin fold (Amzel and Poljak, 1979). Antibodies are capable of binding to a virtually infinite set of antigens, generally with high specificity and affinity. Recently there has been a huge resurgence of interest in using Antibodies in the treatment of human disease. 206 antibodies underwent clinical trials between 1980 and 2005 (Reichert and Valge-Archer, 2007) and it is estimated that more than 400 monoclonal antibodies are currently undergoing clinical trials and almost a third of all drugs under development are monoclonal antibodies.

Antibodies are all-beta proteins consisting of two identical light chains and two identical heavy chains. The light chains contain 2 immunoglobulin domains (V_L and C_L) while the heavy chains have 4 or 5 depending on the heavy-chain class (V_H and C_H1–4). The heavy chains pair in C_H2–4, while the light and heavy chains come together to form the arms of a Y-shaped structure, each arm being known as a Fab. In the pairing of light and heavy chains, the two variable domains (V_H and V_L) dimerize to form the Fv fragment which contains the antigen combining site. The ability of the variable domains to fold independently of the constant domains has been demonstrated through the construction of single chain Fvs that bind antigens with affinity similar to that of the whole Fab (Hanisch *et al.*, 1989). The six loops that form the antigen combining site are hypervariable and are termed Complementarity Determining Regions (CDRs) (Wu and Kabat, 1970). Despite their hypervariability, the loops have been shown to adopt a restricted set of conforma-

tions based on the presence of certain residues at key positions in the CDRs and the neighbouring framework regions (Al-Lazikani *et al.*, 1997; Chothia and Lesk, 1987).

Having a standardized numbering scheme for closely related proteins is of enormous benefit. It allows one to refer to particular positions in the sequence by number and know that this position will be structurally equivalent across structures. For example, Zuccotto *et al.* (1998) numbered dihydrofolate reductase (DHFR) sequences from *E. coli* and trypanosomes with respect to the human sequence. In the case of antibodies, knowing, for example, that residue H50 is always the N-terminal residue of the CDR-H2 loop has allowed easy definition of features within the antibody Fv. Similarly, it has enabled the description of key residues that define the conformation of the CDRs (Al-Lazikani *et al.*, 1997) and residues that are important in 'humanizing' antibodies through CDR grafting (Riechmann *et al.*, 1988). Indeed the concept of a standard numbering scheme for antibodies has proved so popular that there are now four such schemes!

A standardized numbering scheme for antibodies was first introduced by Kabat (Kabat *et al.*, 1983). This numbering scheme was derived on the basis of sequence alignments when no structural information for antibodies was available. Chothia *et al.* (1987) examined the variable domains of antibody structures and showed that the sites of insertions and deletions (indels) in CDR-L1 and CDR-H1 suggested by Kabat on the basis of sequence were not structurally correct. This led to the introduction of the Chothia numbering scheme. Unfortunately in 1989, their numbering scheme was erroneously changed (Chothia *et al.*, 1989), but in 1997 the structurally correct numbering scheme originally proposed in 1987 was reintroduced (Al-Lazikani *et al.*, 1997). In both Kabat and Chothia schemes, the numbering is based on the most common sequence lengths and insertions are accommodated with insertion letters (e.g. 30a).

Since then, two further schemes have been introduced. The IMGT numbering scheme (Lefranc *et al.*, 2003) tries to unify numbering for antibody light and heavy chains with T-cell receptor α and β chains. However, since IMGT is predominantly a DNA database, the numbering does not extend beyond the part of the sequence encoded by the V-gene fragment. The AHo numbering scheme (Honegger and Plückthun, 2001) extends the IMGT numbering scheme by introducing corrections for longer loops in CDR-L1 and multiple sites for indels. Unlike the IMGT scheme, it also extends throughout CDR-3 and into framework region 4. Both IMGT and AHo schemes accommodate indels at specified sites by providing enough numbers that all expected insertions can be accommodated and numbered without insertion letters. Nonetheless, it is possible that unusual antibodies with extremely long insertions will be identified in future which cannot be numbered using these schemes. While a common scheme for light and heavy chains and T-cell receptors has a certain elegance, the practical applications are less obvious. It remains true that immunologists mostly continue to use the Kabat scheme while those interested in structural analysis generally use the Chothia scheme.

The Kabat databank (Johnson and Wu, 2001) remains a popular and important resource for antibody sequence data even though the public version of the data has not been updated since June 2000. This is largely because of the annotation of the sequences with the standard Kabat numbering. The KabatMan (Martin, 1996) database available online at `http://www.bioinf.org.uk/abs/` has provided point-and-click and written-query based access to these data enabling queries such as 'give me the sequences of antibodies known to bind lysozyme and having a serine at H23'.

In this paper, we have examined the annotations with standard numbering that are present in the Kabat databank and have found that approximately 10% of sequences have an error in the (manually applied) numbering. To perform this analysis we have created a software tool (*AbNum*) that applies the Kabat or Chothia numbering in an automatic and reliable manner. The numbering generated by the *AbNum* program was compared with the numbering appearing in Kabat. Where differences were identified, the numbering was examined manually to determine whether the error was in *AbNum* or in Kabat. After several rounds of refinement of the software, we determined that all errors appear to be in Kabat. In the main, these are inconsistent use on indel sites in framework regions, in particular in HFR3. Further analysis of the errors, and of the sites of indels, suggests corrections to the indel sites and in particular that the indel site at H82 in the Chothia scheme (as well as the Kabat scheme) should be located at H72 such that insertions and deletions occur at the structurally correct positions in the frameworks as well as in the CDRs. A server allowing an antibody sequence or structure to be numbered automatically has been made available at `http://www.bioinf.org.uk/abs/abnum/`.

Table I: Benchmarking the performance of *AbNum* through comparison with the Kabat database annotations.

Chain type	Total number of sequences	Not numbered	Do not match Kabat	Sample size	Error (%)†	
					Kabat	*AbNum*
Light chain complete	794	1	111	50	50/50 (14%)	0/50 (0.12%)
Light chain truncated	3044	30	326	40	40/40 (10.7%)	0/40 (1%)
Heavy chain complete	2641	19	206	50	50/50 (7.85%)	0/50 (0.72%)
Heavy chain truncated	1272	27	452	39	39/39 (10.7%)	0/39 (2.12%)

†The percentages reported in the last two column are estimated error percentages based on the sample set examined manually, where inability to number a sequence is treated as an error. Note that no errors were detected in those sequences that could be numbered.

Results

Identifying errors in Kabat and benchmarking the numbering program

In order to identify errors in the Kabat databank and to assess the performance of our numbering program, *AbNum*, we extracted antibody sequences and their Kabat numbering from the July 2000 release of the Kabat database using KabatMan. (No pre-numbered databank containing Chothia numbering is available.) Four test datasets were prepared based on chain type (light or heavy) and type of sequence (complete or truncated). Sequences were considered 'complete' where the sequence annotations in Kabat included the first and last residues in the variable domain (L1 and L109 in the light chain, and H1 and H113 in the heavy chain). All other sequences were regarded as truncated (see Table I).

AbNum was applied to sequences in the four datasets described above. In evaluating performance of *AbNum*, sequences that could not be numbered were regarded as errors. For the other sequences, the *AbNum* numbering results were compared with the numbering annotations in the Kabat database. In the Kabat databank, we found the application of the Kabat numbering is highly inconsistent for residues L106–L111 in light chains and H100–H101 (including insertions at H100–H100a,b,c, etc.) in the heavy chain. For ease of comparison, residues in these zones were excluded from examination.

Sequences where the *AbNum* numbering matched the Kabat database numbering were regarded as being correctly numbered. For the other cases where mismatches occurred, a random sample of approximately 40–50 sequences was selected and manually examined to determine whether the error was in the *AbNum* numbering, or in the Kabat database. These data were then extrapolated to estimate the overall error percentages for the Kabat database and *AbNum* which were calculated as:

$$E_k = \frac{e_k \times N_m}{N_s} \times \frac{100}{N_T} \tag{1}$$

and

$$E_a = \left(U_a + \frac{(e_a \times N_m)}{N_s}\right) \times \frac{100}{N_T} \tag{2}$$

respectively, where E_k is the estimated percentage of errors in Kabat, E_a is the estimated percentage of errors in *AbNum*, e_k and e_a are the number of errors identified in Kabat and *AbNum* respectively in a sample of N_s sequences, U_a is the number of sequences that *AbNum* was unable to number, N_m is the total number of mismatches between *AbNum* and Kabat and N_T is the total number of sequences.

Table I shows the results of the benchmarking and comparison. All discrepancies in the *AbNum* numbering and Kabat database annotations can be attributed to errors in the manual Kabat numbering. Every sequence that can be numbered by *AbNum* appears to be numbered accurately.

Errors in numbering of CDR-H2 and HFR3

The single most common observed error in the numbering in the Kabat databank was found to be in CDR-H2 and HFR3. The Kabat definition for HFR3 is from H66 to H94, a standard length of 29 residues. However, the vast majority of sequences in Kabat have an insertion of three residues in HFR3 following residue H82 (i.e. H82a,b,c) as shown in Figure 1a. There are a small number of sequences that do not contain this insertion, but because this situation is rare, the majority of these are erroneously

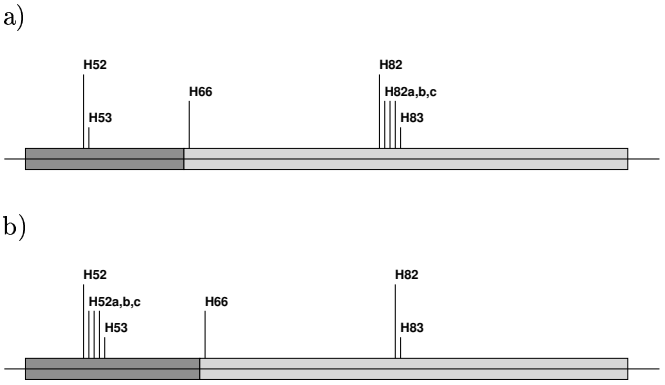

Figure 1: Numbering in CDR-H2 and HFR3. a) The numbering observed in the majority of heavy chains includes an insertion in the Kabat numbering at H82a,b,c. b) We have identified 74 sequences numbered in Kabat with the insertion at H82 which should instead have the insertion in CDR-H2 at residue H52.

annotated in Kabat as containing the 3-residue insertion in HFR3 whereas the residues should be inserted in CDR-H2 at position H52a,b,c (Figure 1b). In total, we have identified 74 sequences in Kabat where the end of the CDR-H2 and the start of HFR3 have been annotated incorrectly.

A modified numbering scheme to accommodate insertions and deletions in framework regions of antibodies

As described above, the two most widely used numbering schemes for antibodies are the Kabat and the Chothia schemes. The Kabat numbering scheme (Kabat *et al.*, 1983) was based on sequence alignments and placed insertions where they occurred in sequence. Chothia and co-workers (Chothia and Lesk, 1987; Al-Lazikani *et al.*, 1997) examined structures of antibodies and proposed a numbering scheme correcting the positions of insertions at the structural level rather than at the sequence level. However, only the CDRs were included in this analysis; framework regions were not examined.

In order to determine whether additional changes in the numbering should be made in the framework to introduce insertions at structurally correct locations, a list of antibody structures was extracted from SACS (Allcorn and Martin, 2002). Light chain and heavy chain sequences from 561 structures were then extracted from the SEQRES records of the PDB files. These sequences were numbered using the Chothia scheme with *AbNum* and the

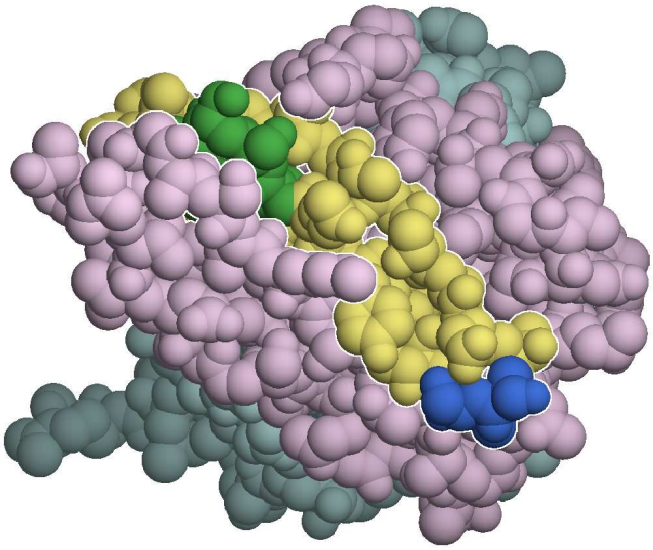

Figure 3: Space-filled representation of the variable domains of an antibody. The light chain is shown in blue-grey and the heavy chain in pink. HFR3 is highlighted by white borders and coloured yellow while residues which would form the insert if the insert is placed at the standard position of H82 are shown in green and the residues which would form the insert if the insert position is at H72 are shown in blue. This illustrates that having the insert at H72 results in three exposed residues having to be deleted compared with the rather buried site at H82. The exact position of the residues involved is shown in more detail in Figure 5c.

numbering was patched into the PDB files. The sequence of every framework region was extracted and analyzed for deviations from the standard lengths described in Kabat (Wu and Kabat, 1970). Structures whose framework region lengths differed from the standard were fitted using ProFit (see Methods). In some cases, framework regions are known to have differing lengths in the sequence data, but structures with unusual length regions are not available. In these cases, four or five structures were chosen and fitted with one another to see whether certain positions in the region appear to be more flexible in structure and therefore likely to accommodate indels.

Of particular note is the indel in HFR3. As stated above, the vast majority of sequences in Kabat have an insertion of three residues in HFR3 following residue H82. Unfortunately no crystal structures of antibodies are available with the residues at H82a,b,c absent. However analysis of the sequence and structure of HFR3 indicates that position H82 is unlikely to accommodate

a)

```
         4444444444 555   5555555666666 66667777777777888   888888899999
         0123456789 012abc3456789012345 67890123456789012abc345678901234
         ...HFR2--> <------CDR-H2-----> <------------HFR3-------------->
axo1     QAPGKGLEWV LRFHSGRNPPQYASEAVKG RVTASTDSSSCYMQMNSL--KTEDTGIYYCAR
mab113   QAPGQGLEWM GRINP-NTGGTNSAQKFQG RVTMTRDTSISTAYMELSNLRSDDTAMYSCAR
```

b)

```
         4444444444 555   5555555666666 6666777   7777777888888888899999
         0123456789 012abc3456789012345 6789012abc345678901234567890123456
         ...HFR2--> <------CDR-H2-----> <------------HFR3-------------->
axo1     QAPGKGLEWV LRFHSGRNPPQYASEAVKG RVTASTDS--SSCYMQMNSLKTEDTGIYYCAR
mab113   QAPGQGLEWM GRINP-NTGGTNSAQKFQG RVTMTRDTSISTAYMELSNLRSDDTAMYSCAR
```

Figure 2: Alignment between antibodies axo1 and mab113. a) with the position of the insertion placed at the conventional position after H82 b) with the insertion placed after H72.

insertions. A pairwise sequence alignment between the antibodies axo1 (Patel and Hsu, 1997) and mab113 (Mantovani *et al.*, 1993), as shown in Figure 2, suggests that H72 is the likely position of the 3-residue indel. Figure 3 shows the space-filled representation of the Fv region of an antibody. Residues that would be numbered H72 and H82 are indicated and it can be seen that H82a,b,c are relatively buried while H72a,b,c are on the surface making it more likely that these residues would be deleted. This is further corroborated by the analysis of Honegger and Plückthun (2001) who aligned sequences of both light and heavy chains of antibodies and α and β chains of T-cell receptors. The resulting AHo numbering scheme suggests that the heavy chain has a 2-residue insertion with respect to the light chain at position H72.

Similar analysis of indels in other framework regions suggests refinements to their locations and Table II compares the results of our analysis with the Kabat standards for the positions of insertions and deletions in each of the framework regions. For LFR1 (Kabat definition L1 to L23) which has a standard length of 23 residues, a structure with 22 residues (PDB Code **2vit** (Fleury *et al.*, 1998)) was found. **2vit** also has an LFR4 (Kabat definition L98 to L110) length of 13 residues compared with the standard length of 12 residues. We fitted the LFR1 and LFR4 regions of **2vit** to that of **12e8** (Trakhanov *et al.*, 1999) which has standard lengths in these regions. For the remaining regions however, no structures with unusual framework region lengths exist.

The fitted structures of light and heavy chain framework regions are shown in Figures 4 and 5 respectively. Structures shown are **12e8** (Trakhanov *et al.*, 1999), **1a0q** (Buchbinder *et al.*, 1998) and **2vit** (Fleury *et al.*, 1998). The sites of insertions and deletions according to the Kabat numbering scheme and our proposed structurally correct sites are indicated.

We believe that the standard Kabat position of the indel in CDR-L2 (L54) is incorrect. Chothia did not consider indels in this loop as all available structures were of the same length. Unfortunately that remains the case today and the loops are relatively conserved in structure. However we consider that, by analogy with CDR-H2, L52 is probably the correct site and this is supported by the AHo numbering scheme.

Discussion

From our analysis of antibody variable-region structures, we have found that approximately 10% of sequences in the manually annotated Kabat database have errors in the numbering. Given the fact that the publicly available Kabat data have not been updated since June 2000, the availability of reliable numbering is the key reason that people still use these data. The major alternative source of antibody sequence data (IMGT) does not provide numbered sequence files.

Table II: Comparison of the Kabat indels with the structurally corrected indels in the frameworks.

Region Name	Kabat definition (Standard length)	Length range Min–Max	Kabat position	Structural insertion	Structural deletion
LFR1	L1–L23 (23)	22–23	L10	–	L10
LFR2	L35–L49 (15)	14–16	L39	L40	L41
LFR3	L57–L88 (32)	31–40	L66	L68	L68
LFR4	L98–L110 (12)	12–13	L106	L107	–
HFR1	H1–H30 (30)	29–34	H6	H8	–
HFR2	H36–H49 (14)	13–14	†	–	H42
HFR3	H66–H94 (29)	30–34	H82	H72	–

†Only three sequences in Kabat have a deletion in HFR2 and the site of the deletion is inconsistent (H36 in entry **000873**; H38 in entry **003348**; H42 in entry **033822**).

Table III: Summary of the revised Chothia numbering scheme which takes into account correct structural sites for indels in the framework regions†.

Segment	Residues
LFR1	1 2 3 4 5 6 7 8 9 (10) 11 12 13 14 15 16 17 18 19 20 21 22 23
CDR-L1	24 25 26 27 28 29 (30) a b c d e f 31 32 33 34
LFR2	35 36 37 38 39 40 a (41) 42 43 44 45 46 47 48 49
CDR-L2	50 51 (52) a b c d e 53 54 55 56
LFR3	57 58 59 60 61 62 63 64 65 66 67 (68) a b c d e f g h 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88
CDR-L3	89 90 91 92 93 94 (95) a b c d e f 96 97
LFR4	98 99 100 101 102 103 104 105 106 107 a 108 109 110
HFR1	1 2 3 4 5 6 7 8 a b c d (9) 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
CDR-H1	(31) a b 32 33 34 35
HFR2	36 37 38 39 40 41 (42) 43 44 45 46 47 48 49
CDR-H2	50 51 (52) a b c 53 54 55 56 57 58 59 60 61 62 63 64 65
HFR3	66 67 68 69 70 71 72 a b c 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94
CDR-H3	95 96 97 98 99 (100) a b c d e f g h i j k ... 101 102
HFR4	103 104 105 106 107 108 109 110 111 112 113

†Structurally correct sites of deletions are shown in parentheses and CDR regions are as defined by Kabat.

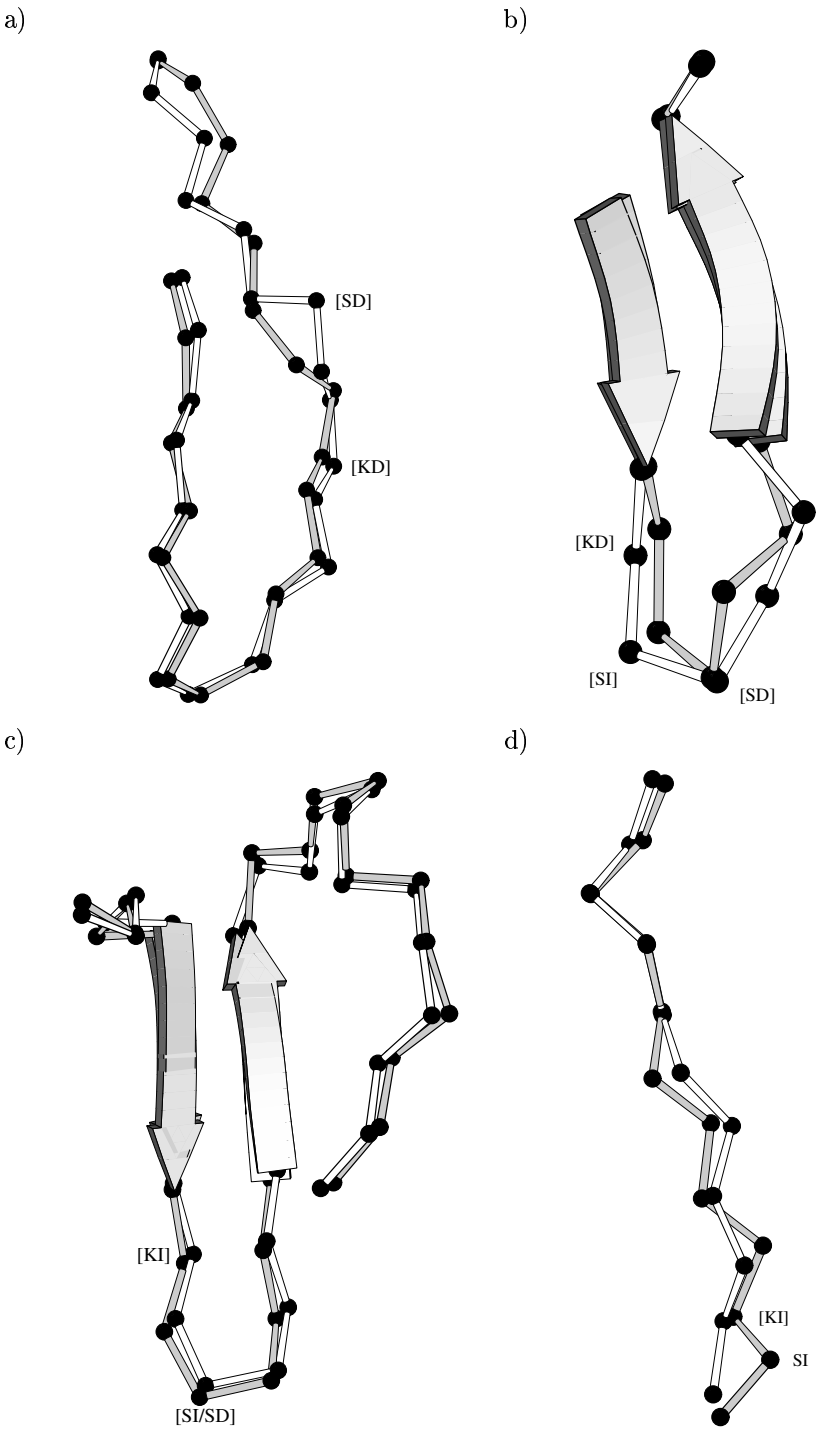

Figure 4: Superimposition of light chain framework regions. a) LFR1 (12e8 and 2vit), b) LFR2 (12e8 and 2vit), c) LFR3 (12e8 and 1a0q) and d) LFR4 (12e8 and 1a0q). The site of an insertion or deletion as defined by Kabat is shown with [KI] or [KD] respectively with actual inserted residues (as per the Kabat numbering scheme) indicated by KI. The suggested structurally correct deletion positions are indicated with [SD] while the structurally correct site of an insertion follows a residue marked [SI] and actual inserted residues are indicated with SI.

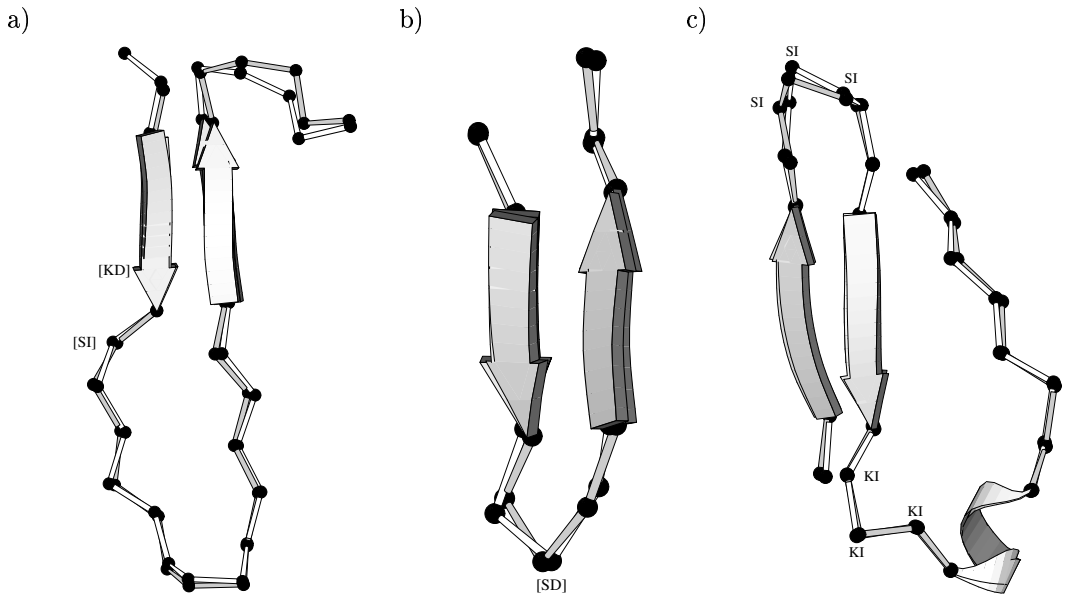

Figure 5: Superimposition of heavy chain framework regions. a) HFR1 (12e8 and 1a0q), b) HFR2 (12e8 and 1a0q) and c) HFR3 (12e8 and 1a0q). Annotations are as in Figure 4.

We have been able to suggest corrections to the positions of insertions and deletions in the framework region in comparison to the standard Kabat locations used in both the Kabat and Chothia numbering schemes. We have therefore proposed a new numbering scheme (summarized in Table III) that extends the Chothia analysis to correct the positions of indels in the framework regions.

The *AbNum* numbering program has been thoroughly tested and benchmarked and can be used to apply numbering schemes to antibody sequences with a very high level of accuracy. *AbNum* was able to number 99% of sequences and we believe that in all cases discrepancies from the manual numbering in the Kabat databank resulted from errors in the Kabat databank and not in our program. Simply by supplying different data files, Kabat and Chothia numbering schemes can be applied as can the scheme we propose here — a modification of the Chothia scheme to give structurally correct indels in the framework regions. Thus the program can be used reliably to apply standard numbering schemes to sequences in IMGT enhancing the usefulness of this resource.

Experimental Procedures

Sequences of antibodies were extracted from the July 2000 release of the Kabat database (Johnson and Wu, 2001) (the last public release). Raw data were obtained from `ftp://ftp.ebi.ac.uk/pub/databases/kabat/fixlen/` and read into KabatMan (Martin, 1996) which was used to query and extract the sequence data. Programs for analysis were written using C and PERL. Structures of antibodies were extracted from the PDB (Berman *et al.*, 2000) using SACS (Allcorn and Martin, 2002). Fitting was performed using the McLachlan algorithm (McLachlan, 1982) as implemented in the program ProFit (Martin, A.C.R., unpublished, available at `http://www.bioinf.org.uk/software/profit/`).
Structures were examined using RasMol (Sayle and Milner-White, 1995) and images were generated with QTree (Martin, A.C.R., unpublished, available at `http://www.bioinf.org.uk/software/qtree/`) and MolScript (Kraulis, 1991).

The strategy used for numbering is summarized in Figure 6. A set of anchor points in the sequence (chosen as the starts and ends of framework regions) was defined and the numbering was then filled in based around these locations. To locate these anchor points, residue propensities for the 6 residues at the start and end of each framework

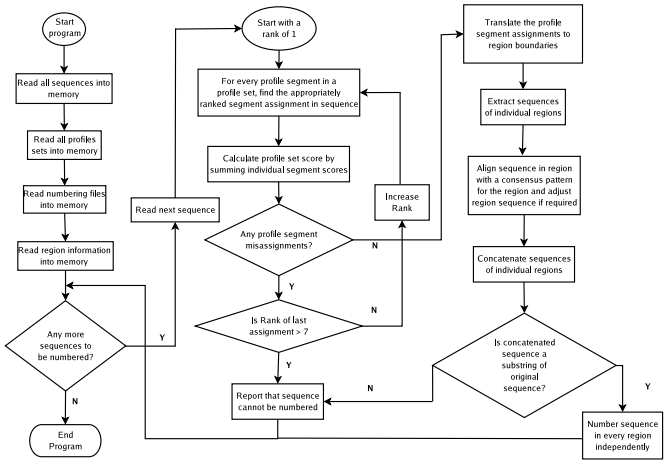

Figure 6: Flowchart of the numbering program. A 'profile segment' is a profile of 6 residues that matches the end of a framework region. The term 'region' is used to refer either to one of the seven framework regions (LFR1, HFR3, etc.), or to a CDR (CDR-L1, CDR-L2, ... CDR-H3).

region were extracted from the Kabat database using KabatMan to form *profile segments*. Initially, we created 3 *profile sets* (a set of profile segments) classified on the basis of chain: heavy chain, lambda light chain and kappa light chain. However, we found that a significant proportion of sequences could not be numbered because of profile segment mis-assignments. Thus the profiles were made more specific and 32 profile sets were generated on the basis of human subgroup classes and species of origin of the sequences. Human subgroup classes were defined in the 1994 version of the Kabat database and sequences were divided into families based on amino acid identity where members of a family differ by 12 amino acids or less (Deret *et al.*, 1995). This led to the creation of 16 profiles (6 λ, 4 κ and 6 heavy) split by subgroup. An alternative set of 16 profiles was created split on the species of origin (6 λ, 6 κ and 4 heavy). Because errors were identified in the numbering contained in Kabat (especially in the region corresponding to the anchor point at the start of HFR3), the profiles were updated after renumbering the Kabat sequences using the initial profiles defined from the Kabat data.

Table IV gives the positions that were used to define the anchor points for each of the profile sets. To number a sequence, every profile set is scanned against the sequence and the score for a profile set is calculated by

Table IV: Kabat positions used in defining the profiles, and minimum and maximum observed lengths of the seven regions in the light and heavy chain in the Kabat databank.

	Profile location		Length range	
Region name	Start	End	Min	Max
LFR1	L1–L6	L18–L23	22	23
CDR-L1			7	17
LFR2	L35–L40	L44–L49	14	16
CDR-L2			5	12
LFR3	L57–L62	H66–H71	30	40
CDR-L3			4	18
LFR4	L98–L103	L104–L109	12	13
HFR1	H1–H6	H20–H25	29	34
CDR-H1			1	13
HFR2	H36–H41	H44–H49	12	14
CDR-H2			10	23
HFR3	H66–H71	H89–H94	28	34
CDR-H3			2	30
HFR4	H103–H108	H108–H113	10	12

summing the scores of the best individual profile segment assignments. Based on the best profile segment assignments, anchor points in the sequence are fixed and the sequence in every region is extracted and numbered independently. Profile mis-assignments are detected when they are out of order (for example, when the best LFR2-end profile assignment is positioned after the best LFR3-start profile assignment), or the distance between a pair of profiles falls outside pre-set limits. These limits were set after the distribution of region lengths in the Kabat database was examined manually (Table IV). It must be envisaged that it may be necessary to extend these limits to accommodate unusually long sequences. However, this would require cautious modification to ensure that sequences are not numbered incorrectly.

A ranking scheme was introduced to cope with profile mis-assignments. When a profile mis-assignment is detected on the basis of profile order and separation, the best seven profile set assignments are examined in turn to see if the correct match can be found. If not, it is reported that the sequence cannot be numbered. Once profile assignments are completed, the sequence of every region is extracted.

In some regions, the Kabat numbering does not appear to impose a fixed site for indels. For instance, in HFR2 (Heavy chain framework region 2) the deletion appears to be placed at the position most likely based on sequence alignment. In these cases, an alignment is per-

formed between the sequence and a consensus pattern in that region. Alignment with a consensus sequence also eliminates any residues that may not be part of the variable region.

Numbering is applied to the sequence in every region based on one of the following rules:

1. *Deletions made before the position of insertion.* For example, the Kabat definition for the region CDR-L2 is L50 to L56 which gives it a standard length of 7 residues. A maximum length of 12 residues and minimum length of 5 residues have been observed for this region. The position of insertion according to the Kabat standard is L54 (L54a, L54b, L54c, etc.). Deletions are placed before the position of insertion (L54). For example, in the case of a 5-residue CDR-L2, residues L53 and L54 are deleted.

2. *Deletions made after the position of insertion.* For example in CDR-L1, whose Kabat definition is L24 to L34, the standard length is 11 residues. A maximum length of 17 residues and minimum length of 7 residues have been observed in this region. Insertions are placed at position L27 according to the Kabat standard. Deletions are placed after the position of insertion (L27). For a 7-residue CDR-L1, residues L28, L29, L30, and L31 are deleted.

3. *Residues are numbered sequentially.* In the heavy chain framework region 4, residues are numbered sequentially as there is no clear location for indels in this region.

Acknowledgments

This work was supported by the Dorothy Hodgkin Postgraduate Award funded by the BBSRC and GlaxoSmithKline to KRA.

References

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

Allcorn, L. C., Martin, A. C. R., 2002. SACS–self-maintaining database of antibody crystal structure information. Bioinformatics 18, 175–181.

Amzel, L. M., Poljak, R. J., 1979. Three-dimensional structure of immunoglobulins. Annu Rev Biochem 48, 961–997.

Berek, C., Milstein, C., 1987. Mutation drift and repertoire shift in the maturation of the immune response. Immunol. Rev. 96, 23–41.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., Westbrook, J., 2000. The Protein Data Bank and the challenge of structural genomics. Nature: Struct. Biol. 7 Suppl, 957–959.

Buchbinder, J. L., Stephenson, R. C., Scanlan, T. S., Fletterick, R. J., 1998. A comparison of the crystallographic structures of two catalytic antibodies with esterase activity. J. Mol. Biol. 282, 1033–1041.

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., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., 1989. Conformations of immunoglobulin hypervariable regions. Nature (London) 342, 877–883.

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. Computer Applications in the Biosciences 11, 435–439.

Fleury, D., Wharton, S. A., Skehel, J. J., Knossow, M., Bizebard, T., 1998. Antigen distortion allows influenza virus to escape neutralization. Nature: Struct. Biol. 5, 119–123.

Hanisch, F. G., Uhlenbruck, G., Egge, H., Peter-Katalinić, J., 1989. A B72.3 second-generation-monoclonal antibody (CC49) defines the mucin-carried carbohydrate epitope Gal beta(1-3) [NeuAc alpha(2-6)]GalNAc. Biol. Chem. Hoppe Seyler 370, 21–26.

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.

Johnson, G., Wu, T. T., 2001. Kabat Database and its applications: Future directions. Nuc. Ac. Res. 29, 205–206.

Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Miller, M., Perry, H. 1983. Sequence of Proteins of Immunological interest. (National Institutes of Health, Bethesda).

Kraulis, P. J., 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24, 946–950.

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.

Mantovani, L., Wilder, R. L., Casali, P., 1993. Human rheumatoid B-1a (CD5+ B) cells make somatically hypermutated high affinity IgM rheumatoid factors. J. Immunol. 151, 473–488.

Martin, A. C., 1996. Accessing the Kabat antibody sequence database by computer. Proteins: Struct., Funct., Genet. 25, 130–133.

McLachlan, A. D., 1982. Rapid comparison of protein structures. Acta Crystallogr. A38, 871–873.

Patel, H. M., Hsu, E., 1997. Abbreviated junctional sequences impoverish antibody diversity in urodele amphibians. J. Immunol. 159, 3391–3399.

Reichert, J. M., Valge-Archer, V. E., 2007. Development trends for monoclonal antibody cancer therapeutics. Nat. Rev. Drug Discov. 6, 349–356.

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

Sayle, R. A., Milner-White, E. J., 1995. RASMOL: biomolecular graphics for all. Trends Biochem. Sci. 20, 374–374.

Trakhanov, S., Parkin, S., Raffaï, R., Milne, R., Newhouse, Y. M., Weisgraber, K. H., Rupp, B., 1999. Structure of a monoclonal 2E8 Fab antibody fragment specific for the low-density lipoprotein-receptor binding region of apolipoprotein E refined at 1.9 A. Acta Crystallogr. D55, 122–128.

Wu, T. T., Kabat, E. A., 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.

Zuccotto, F., Martin, A. C. R., Laskowski, R. A., Thornton, J. M., Gilbert, I. H., 1998. Dihydrofolate reductase: a potential drug target in trypanosomes and leishmania. J. Comp. Aided Molec. Design 12, 241–257.