PROTEINS: Structure, Function, and Genetics 25:130–133 (1996)

TECHNICAL REPORT

Accessing the Kabat Antibody Sequence Database by Computer

Andrew C.R. Martin
Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT, United Kingdom

ABSTRACT **The Kabat antibody sequence database has for many years been the primary site for depositing sequence information on antibodies and other proteins of immunological interest. The chief drawback of this database has been that it has only been available in the form of a printed book (Kabat et al., *Sequences of Proteins of Immunological Interest*, 1991). These data have recently become available on the global computer Internet, but no method of searching the data has, as yet, been provided. Here, the development of a specialized database program for accessing the antibody data is described. This database software has been made accessible over the World Wide Web, together with a program which allows a novel antibody sequence to be tested against the Kabat sequence database, to identify unusual features of an antibody sequence which may represent cloning artifacts or sequencing errors.**

Key words: immunoglobulin, sequence comparison, cloning, complementarity determining regions, alignment

INTRODUCTION

Kabat, et al.[1] have, for some years, published sequence data for immune-related proteins in the form of a book, but as the number of sequences determined increases, the need for electronic distribution becomes more and more necessary.

Many sequence databases (e.g., SWISS-PROT,[2] PIR,[3] GenBank,[4] and NRL-3D[5]) are available and contain sequence data for antibodies. However, only the Kabat database presents the antibody sequence data in a standard alignment with the numbering scheme introduced by Kabat et al. Thus while the alternative sequence databases do have powerful query languages, they do not allow one to define queries based on the additional information stored in the Kabat database. For example, one might wish to select all the antibody sequences having an 11 residue CDR-L1 with a proline at position L29.

The Kabat data are available on the global computer Internet by File Transfer Protocol (FTP) from ftp://ncbi.nlm.nih.gov/repository/kabat/fixlen. Although these files contain all the alignment information, the format, while convenient for visual inspection, is somewhat difficult to use by computer and no software is currently supplied to derive data, or make selections, from the files. KabatMan ("Kabat Manager") has been written as a specialized database program to fill this gap. It assembles the data from the Kabat "dump" files and allows searches to be performed on the data. Standard PIR sequence files may be created and other information (such as CDR lengths) may be obtained while allowing many constraints to be placed on the selection of sequences.

The "World Wide Web" is a hypertext system which runs on the global computer Internet. Hypertext documents allow the user to click on a highlighted word or phrase and have the relevant part of the document, or another document, displayed. The World Wide Web takes this one stage further by allowing hypertext links between documents which may be on different computers anywhere in the world. In addition, documents may have "forms," areas into which data may be entered. The KabatMan database program, together with an antibody sequence testing program, has been made accessible via a World Wide Web forms interface.

METHODS

Establishing the Database

When the KabatMan software is first run, it reads the Kabat dump files and rejects any non-antibody sequences (T-cell receptors, etc.) Correlations are

Received July 13, 1995; revision accepted November 20, 1995.
Address reprint requests to Andrew C.R. Martin, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, United Kingdom.

made between heavy and light chains coming from the same species, having the same name and sharing at least one author in the references. The resulting data are written to a file for future use by KabatMan.

The Kabat numbering scheme allows only limited insertions at certain sites. It does not define names for additional residues. The Kabat database may insert extra residues between, for example, H100G and H100H. For simplicity, the KabatMan database program simply labels insertion regions with letters alphabetically throughout the insertion region. This means that there is a slight deviation from the Kabat numbering scheme when very long loops occur (especially in CDR-H3). If, in future, the maintainers of the Kabat sequence database define a numbering scheme for these additional insertions, KabatMan will be modified to support the new scheme.

The database uses a simple functional data model (FDM). As well as allowing data stored directly in the database tables to be accessed, an FDM database allows access to derived data. This reduces the required storage size allowing information to be calculated when it is requested and is completely transparent to the user (i.e., stored and derived data are accessed in exactly the same way). For example, the CDRs of the antibodies may be extracted using the KabatMan database. They are not stored explicitly in the database (as would be necessary using a normal flat file or relational database). Instead, a function is stored which allows the CDRs to be derived from the sequence data stored for the whole antibody Fv. Similarly, the lengths of the CDRs are not stored directly in the database tables; they are calculated from the sequence using a function in the database. The reduced storage requirement of using an FDM model means that all data can be stored in the computer's main memory reducing the requirements for optimizing the search methods; most queries are answered in less than 2 seconds. Some 9.5MBytes of memory are used by the program with the January 1995 release of the Kabat sequence database.

Querying the Database

Data in the database is accessed using a query language similar to the SELECT statement of the common database language "Structured Query Language" (SQL).[6] The main deviation between the language used here and SQL is that sub-clauses within the WHERE clause are combined in Reverse Polish Notation. This is described in detail below. Also, since the database has only one table, there is no FROM clause. No other SQL statements, or clauses of the SELECT statement, are currently supported.

In addition, a SET statement exists which allows one to select the loop definitions and apply a homology cutoff, rejecting very similar sequences from the results of a search.

The "SELECT" statement allows one to view the name of the antibody, the antigen, the sequences and lengths of the six CDRs, the light chain class, the animal source, the publication references, the amino acid at a specified residue, the sequences of light and heavy chains, and the Chothia canonical class for a loop. In addition, the sequence may be displayed in PIR sequence file format.

A "WHERE" clause filters the data allowing only sequences which match particular requirements to be selected. The same data fields used in the SELECT statement may be used and, in addition, one can test that the antibody is complete (i.e., has both light and heavy chains). Tests may be made for substring matches as well as numeric comparisons (equal, not-equal, greater-than, less-than, greater-than-or-equal, less-than-or-equal).

Sub-clauses within the WHERE clause are combined in "Reverse Polish Notation" (RPN). RPN is a stack-based method for performing operations on data. A stack is a data structure in which items can be added or removed only from the top of the stack; items cannot be inserted or removed from lower down the stack. RPN is used on some Hewlett-Packard scientific calculators and in computer languages such as PostScript and FORTH. Its advantages over normal "infix" notation are (a) avoidance of ambiguity and the need for parentheses, and (b) simplicity of implementation. It would clearly be possible to add code to the program which converts infix notation to RPN. However, current users of the database have found RPN simple and flexible to use.

Selections resulting from each WHERE subclause are placed on a stack and logical operators act on this stack to combine the results. Logical operators (AND, OR, and NOT) remove one (or more) items from the top of the stack and replace them with other items. For example, if the top two items on the stack are the values TRUE and FALSE, the OR operator would remove these top two items and place a single value (the Boolean OR of these items, in this case, the value TRUE) onto the stack.

Figure 1 shows some simple example queries and truncated results.

Sequence Testing—A Specialized Query

Given the sequence of a new antibody, it is frequently useful to search for any unusual features. This may be an aid to experimentalists who wish to look for potential cloning artifacts or sequencing errors. Unusual features which are established as genuine are likely to be critical to the binding of the antibody.

To test a new sequence, it is simply necessary for each position in the sequence, to use the database to examine the frequency of occurrence of the amino acids from the new sequence. However, with a typical antibody Fv having approximately 230 residues, this is a long and tedious process.

a)

```
SELECT name, antigen, length(l1), length(l2), length(l3),
       length(h1), length(h2), length(h3)
WHERE  antigen != ''
       complete = true AND
```

Results:

```
8E10'CL, ANTI-DNA, IGG HYBRIDOMA, 12, 7, 10, 5, 17, 14
RF-TS7'CL, ANTI-HUMAN/RABBIT IGG RHEUMATOID FACTOR, 14, 7, 7, 5, 17, 6
UC'CL, ANTI-MYELIN/SSDNA, 17, 7, 9, 5, 17, 14
mAb113'CL, ANTI-IGG FC FRAGMENT RHEUMATOID FACTOR, 12, 7, 9, 5, 17, 17
.....
Number of hits = 1163
```

b)

```
SELECT pir
WHERE  source   includes "mouse"
       antigen  includes "lysozyme" AND
       complete =        true       AND
```

Results:

```
>P1;HYHEL-
HYHEL-10 - (MOUSE) mouse
DIVLTQSPATLSVTPGNSVSLSCRASQSIGNNLHWYQQKS
HESPRLLIKYASQSISGIPSRFSGSGSGTDFTLSINSVET
EDFGMYFCQQSNSWPYTFGGGTKLE*
DVQLQESGPSLVKPSQTLSLTCSVTGDSITSDYWSWIRKF
PGNRLEYMGYVSYSGSTYYNPSLKSRISITRDTSKNQYYL
DLNSVTTEDTATYYCANWDGDYWGQGT*
.....
Number of hits = 4
```

c)

```
SELECT name, l1
WHERE  len(l1)  = 11
       res(L29) = "P" AND
```

Results:

```
HIL, SANALPNQYAY
BEN-27'CL, SGDALPNQYAY
ITC63B'CL, SGDALPKQYSY
KIR, SGDALPNQYAY
.....
Number of hits = 18
```

Fig. 1. Example queries using the KabatMan database program. **a:** Find the name, antigen, and CDR lengths for all complete antibodies where the antigen is known; **b:** Get the sequences in PIR format of all complete mouse antibodies which bind to lysozyme, **c:** Find the name and CDR-L1 sequence of all antibodies with 11 residue CDR-L1s and a proline at Kabat position L29.

A program has been written which takes the sequence of an antibody Fv and calculates the standard Kabat numbering for the sequence using an alignment procedure. The program then writes a script for the KabatMan database and tests each position in turn. Any residues occuring in less than 1% of the sequences in the database are displayed. Again, this program is accessible through the World Wide Web.

For simplicity and speed, the supplied sequence is aligned with a single consensus antibody sequence rather than performing a multiple alignment. This can cause problems with the positioning of gaps and is very sensitive to the exact consensus sequence. Gaps are provided in the consensus sequence to allow CDRs with maximum lengths equal to the maximum lengths observed in the January 1995 Kabat sequence database.

The alignment must account for the fact that either an Fv or an Fab could be supplied and that the end point of the supplied sequence for an Fv may vary. If the supplied sequence is <75% of the length of the full Fab chain, then the consensus is truncated at the end of the Fv and the test sequence is truncated 13 residues after the sequence FGxGT (occurring after residue 90) in the light chain and at 11 residues after the sequence WGxG (also occurring after residue 90) in the heavy chain. These sequences are almost invariant, appearing just after CDR-3. If they are not found, a warning is issued to indicate that the sequences could not be truncated.

Alignment is performed using a simple Needleman and Wunsch algorithm[7] using the Dayhoff-78 mutation matrix[8] normalized at a value of 10. Values for symbols in the consensus such as "!" (used to indicate any hydrophobic residue) have been obtained by approximate averaging of the contributing residues. More refined mutation matrices are now available, but the Dayhoff-78 matrix performs well in this application. The gap penalty is set to 15.

If the consensus sequences have expanded in length, the program reports an error. This results from loops longer than allowed by the consensus sequence and numbering, or from very unusual sequences.

The positioning of insertions within the loops must follow the standard Kabat numbering scheme. Normal automatic alignment procedures will not do this; for example in CDR-H2 the insertions at 52 will actually go into positions 52C, then 52B, then 52A rather than vice versa. Thus the aligned sequence is "re-packed" to place insertions which occur in the loops in the standard Kabat positions. Within loop regions, the residues are numbered from the N-terminus until an insertion code label is met. Numbering then proceeds from the C-terminus until all residues have a label, or an insertion code is met. Finally, any remaining residues are labeled with their insertion code labels from N- to C-terminus.

EXAMPLE APPLICATIONS

The database program allows simple access to the Kabat database of antibody sequences and provides a facility which allows a novel antibody sequence to be tested against the Kabat database.

The program has been used extensively in-house for the analysis of residue distributions in antibody CDRs (Martin, A.C.R., Nesbeth, D., and Thornton, J.M., manuscript in preparation) and for testing of sequences before modeling. For example, one cloned antibody, Yol (Little, M., Dübel, S., and Breitling, F., personal communication) which showed much poorer binding than the original monoclonal, showed a deletion from H14 to H16 in the first heavy framework region. This is seen in no other antibody in the Kabat database and is likely to be a cloning artifact responsible for the reduced binding.

Similarly, in an anti-cardiolipin antibody, UK4

(Kalsi, J., and Isenberg, D., personal communication), unusual features in the sequence helped the experimentalists to identify errors in their preliminary sequencing which have subsequently been corrected.

Analysis of antibody sequence features such as the analysis of CDR-H3 length distribution performed by Wu et al.[9] become extremely simple to perform using this software. A query of the form:

```
SELECT  name
WHERE   length(h3) = 7
        source includes "human" AND
```

would select all the human heavy chains with CDR-H3 of length 7 and would report the number of hits. Adding the WHERE sub-clause:

```
        antigen !=''                AND
```

would report those examples where the antigen is known. Repeating these queries for different lengths of CDR-H3 allows the distribution of CDR-H3 lengths to be analyzed.

AVAILABILITY

This work has made the Kabat database of antibody sequence information available in an easy-to-search computer-accessible form over the World Wide Web using the uniform resource locator (URL):

http://www.biochem.ucl.ac.uk/~martin/abs/kabatman.html

An alternative simple-to-use point and click interface to KabatMan which alleviates the need to learn the KabatMan query language, but is restricted to simple queries is available using the URL:

http://www.biochem.ucl.ac.uk/~martin/abs/simkab.html

The sequence testing facility, which allows antibody sequences to be tested against the current Kabat database to identify unusual features is available through the URL:

http://www.biochem.ucl.ac.uk/~martin/abs/seqtest.html

NOTE ADDED IN PROOF

Since submitting this paper, Johnson and Wu have implemented a Web interface to their data. Their interface, however, is aimed at performing very different types of searches. It may be accessed using the URL:

http://immuno.bme.nwu.edu/

ACKNOWLEDGMENTS

I thank George Johnson and Tai Te Wu, maintainers of the Kabat sequence database, for their help with making the latest version of the database available and for incorporating suggestions for modifications and improvements to the data format. The United Kingdom Medical Research Council are thanked for support.

REFERENCES

1. Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S., Foeller, C. "Sequences of Proteins of Immunological Interest." 5th edit. Bethesda, MD: U.S. Department of Health and Human Services, 1991.
2. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank: Current status. Nucleic Acids Res. 22: 3578–3580, 1994.
3. George, D.G., Barker, W.C., Mewes, H.-W., Pfeiffer, F., Tsugita, A. The PIR-international protein sequence database. Nucleic Acids Res. 22:3569–3573, 1994.
4. Benson, D.A., Boguski, M., Lipman, D.J., Ostell, J. GenBank. Nucleic Acids Res. 22:3441–3444, 1994.
5. Namboodiri, K., Pattabiraman, N., Lowrey, A., Gaber, B.P., George, D.G., Barker, W.C. NRL-3D—A sequence-structure database. Biophys. J. 57:A406, 1990.
6. ISO-ANSI. The ISO-ANSI SQL2 standard—ANSI Document X3H2-88-72/ISO Document DBL CPH-2. American National Standards Institute, 1430 Broadway, New York, NY 10018, USA, 1988.
7. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:444–453, 1970.
8. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure." Vol. 5, Suppl. 3. Dayhoff, M.O. (ed.). Silver Spring, Washington, DC: National Biomedical Research Foundation, 1978:345–352.
9. Wu, T.T., Johnson, G., Kabat, E.A. Length distribution of CDRH3 in antibodies. Proteins 16:1–7, 1993.