2

The Kabat Database and a Bioinformatics Example

George Johnson and Tai Te Wu

1. Introduction

In 1969, Elvin A. Kabat of Columbia University College of Physicians and Surgeons and Tai Te Wu of Cornell University Medical College began to collect and align amino acid sequences of human and mouse Bence Jones proteins and immunoglobulin (Ig) light chains. This was the beginning of the *Kabat Database.* They used a simple mathematical formula to calculate the various amino acid substitutions at each position and predict the precise locations of segments of the light-chain variable region that would form the antibody-combining site from a variability plot ***(1)***. The *Kabat Database* is one of the oldest biological sequence databases, and for many years was the only sequence database with alignment information.

The *Kabat Database* was available in book form free to the scientific community starting in 1976 ***(2)***, with an updated second edition released in 1979 ***(3)***, third edition in 1983 ***(4)***, fourth edition in 1987 ***(5),*** and fifth printed edition in 1991 ***(6)***. Because of the inclusion of amino acid as well as nucleotide sequences of antibodies, T-cell receptors for antigens (TCR), major histocompatibility complex (MHC) class I and II molecules, and other related proteins of immunological interest, it became impossible to provide printed versions after 1991. In that same year, George Johnson of Northwestern University created a website to electronically distribute the database located temporarily at:

http://kabatdatabase.com

During the following decade, the *Kabat Database* had grown more than five times. Thanks to the generous financial support from the National Institutes of Health, access to this website had been free for both academic and commercial use.

With the completion of the human genome project as well as several other genome projects, scientific emphasis has gradually shifted from determining

From: *Methods in Molecular Biology, Vol. 248: Antibody Engineering: Methods and Protocols*
Edited by: B. K. C. Lo © Humana Press Inc., Totowa, NJ

more sequences to analyzing the information content of the existing sequence data. With regard to the *Kabat Database,* the collection and alignment of amino acid and nucleotide sequences of proteins of immunological interest has been progressing side-by-side with the ability to determine structure and function information from these sequences, from its very start.

1.1. Historical Analysis and Use

After the pioneering work of Hilschmann and Craig ***(7)*** on the sequencing of three human Bence Jones proteins, many research groups joined the effort of determining Ig light chain amino acid sequences. By 1970, there were 77 published complete or partial Ig light chain sequences: 24 human κ-I, 4 human κ-II, 17 human κ-III, 10 human λ-I, 2 human λ-II, 6 human λ-III, 5 human λ-IV, 2 human λ-V, 2 mouse κ-I, and 5 mouse κ-II proteins ***(1)***. The invariant Cys residues were aligned at positions 23 and 88, the invariant Trp residue positioned at 35, and the two invariant Gly residues at positions 99 and 101. To align the variable region of kappa and lambda light chains, single-residue gaps were placed at positions 10 and 106A. Longer gaps were introduced between positions 27 and 28 (27A, 27B, 27C, 27D, 27E, and 27F) and between 97 and 98 (97A and 97B), which was later changed to between 95 and 96 (95A, 95B, 95C, 95D, 95E and 95F). A similar alignment technique with a different numbering system was introduced for the Ig heavy-chain variable regions ***(8)***. The invariant Cys residues were located at positions 22 and 92, the Trp residue at position 36, and the two invariant Gly residues at positions 104 and 106.

The most important discovery to come from alignment of the Ig heavy- and light-chain sequences was the location of segments forming the antibody-combining site, known as the complementarity (initially called hypervariable)-determining regions (CDRs). Since different antibodies bind different antigens, numerous amino acid substitutions occur in these segments, leading to large, calculated variability values. The first variability plot of the 77 complete and partial amino acid sequences of human and mouse light chains showed three distinct peaks of variability, located between positions 24 to 34, 50 to 56, and 89 to 97 ***(1)***. Three similar peaks were discovered in heavy chains at positions 31 to 35, 50 to 65, and 95 to 102. These six short segments were hypothesized to form the antigen-binding site and were designated as CDRL1, CDRL2, CDRL3 for light chains, and CDRH1, CDRH2, and CDRH3 for heavy chains, respectively.

Initial Ig three-dimensional (3D) X-ray diffraction experiments suggested that the six binding-site segments were indeed physically located on one side of the Ig macromolecule. Final verification of this theoretical prediction came after the development of hybridoma technology ***(9)***. An anti-lysozyme monoclonal antibody F_{ab} fragment was co-crystallized with lysozyme ***(10)***, and the

combined 3D structure was determined by X-ray diffraction analysis. Several amino acid residues in each of the six CDRs of the antibody were found to be in direct contact with the antigen. As theoretically predicted, antibody specificity thus resided exclusively in the CDRs. During the past decade, designer antibodies have been constructed genetically by selecting these CDRs for their affinity for the target antigen.

By comparing the amino acid sequences of the CDRs as well the stretches of sequence that connect them, known as framework regions (FR), Kabat and Wu hypothesized that the Ig variable regions were assembled from short genetic segments ***(11,12)***. This hypothesis was verified experimentally by Bernard et al. ***(13)*** with the discovery of the J-minigenes, reminiscent of the switch peptide proposed by Milstein ***(14)***. The D-minigenes were soon identified as another component of the heavy-chain variable region ***(15,16)***. In addition, the idea of gene conversion ***(17)*** was proposed as a possible mechanism of antibody diversification, and appears to play a central role in chickens ***(18)***, and to a varying extent in humans, rabbits, and sheep.

For precisely aligned amino acid sequences of Ig heavy-chain variable regions, CDRH3 is defined as the segment from position 95 to position 102, with possible insertions between positions 100 and 101. The CDRH3-binding loop is the result of the joining of the V-genes, D-minigenes, and J-minigenes. This intriguing process has been studied extensively ***(19,20)***, and suggests the CDRH3 plays a unique role in conferring fine specificity to antibodies ***(21,22)***. Indeed, a particular amino acid sequence of CDRH3 is almost always associated with one unique antibody specificity. The CDRH3 sequences within the *Kabat Database* have further been analyzed by their length distributions ***(23)***, for which the length distributions of 2,500 complete and distinct CDRH3s of human, mouse, and other species were found to be more-or-less in agreement with the Poisson distribution. Interestingly, the longest mouse CDRH3 had a length of 19 amino acid residues, and that of human had 32 residues, and only one of them was shared by both species ***(24)***, suggesting that CDRH3 may be species-specific.

Because of the subtle differences between the variable regions of the Ig light and heavy chains, their alignment position numberings are independent. For example, in light chains, the first invariant Cys is located at position 23 and CDRL1 is from position 24 to 34—e.g., immediately after the Cys residue. However, in heavy chains, the invariant Cys is located at position 22 and CDRH1 is from position 31 to 35—e.g., eight amino residues after that Cys. Because of this important difference, the Kabat numbering systems are separate for Ig light and heavy chains. Attempts to combine these two numbering systems into one in other databases have resulted in the presence of many gaps and confusions. Similarly, variable regions of TCR alpha, beta, gamma, and

Table 1
FRs and CDRs of Antibody and TCR Variable Regions

FR or CDR	V_L	V_H	V_α	V_β	V_γ	V_δ
FR1	1–23	1–22	1–22	1–23	1–21	1–22
CDR1	24–34	31–35B	23–33	24–33	22–34	23–34A
FR2	35–49	36–49	34–47	34–49	35–49	35–49
CDR2	50–56	50–65	48–56	50–56	50–59	50–57
FR3	57–88	66–91	57–92	57–94	60–95	58–89
CDR3	89–97	95–102	93–105	95–107	96–107	90–105
FR4	98–107	103–113	106–116	108–116A	108–116C	106–116

delta chains are aligned using different numbering systems. The alignments are summarized in **Table 1,** with the locations of CDRs indicated.

1.2. Current Analysis and Use

There are approx 25,000 unique yearly logins to the website of the *Kabat Database* by immunologists and other researchers around the world. The website is designed to be simple to use by those who are familiar with computers and those who are not. A description of the tools currently available is shown in **Table 2.** We encourage researchers who use the database to share their suggestions for improving the access and searching tools.

A common but extremely important question asked by researchers is whether a new sequence of protein of immunological interest has been determined before and stored in the database. Without asking this simple question, one may encounter the following situation: a heavy-chain V-gene from goldfish was sequenced ***(25)*** and found to be nearly identical to some of the human V-genes. Subsequently, the authors suggested that it might be of human origin, possibly because of the extremely sensitive amplification method used in the study and minute contamination of the sample by human tissue.

Another common use of the database is to confirm the reading frame of an immunologically related nucleotide sequence. Comparing short segments of sequence with stored database sequences can easily identify inadvertent omission of a nucleotide in the sequencing gel. Of course, if the missing nucleotide is real, this can suggest the presence of a pseudogene. Researchers also use the website to calculate variability for groupings of similar sequences of interest. For example, the variability plots of the variable regions of the Ig heavy and light chains of human anti-DNA antibodies are shown in **Figs. 1** and **2.** These two plots seem to indicate that CDRH3 may contribute most to the binding of DNA.

In many instances, investigators would like to identify the germline gene that is closest to their gene of interest, as well as the classification of that par-

Table 2
Listing of Tools Available on the Kabat Database Website

Tool	Description
Seqhunt II	The *SeqhuntII* tool is a collection of searching programs for retrieving sequence entries and performing pattern matches, with allowable mismatches, on the nucleotide and amino acid sequence data. The majority of fields in the database are searchable—for example, a sequence's journal citation. Matching entries may be viewed as HTML files or downloaded and printed. Pattern matching results show the matching database sequence aligned with the target pattern, with differences highlighted.
Align-A-Sequence	The Align-A-Sequence tool attempts to programmatically align different types of user-entered sequences. Currently kappa and lambda Ig light-chain variable regions may be aligned using the program.
Subgrouping	The Subgrouping tool takes a user-entered sequence of either Ig heavy, kappa, or lambda light-chain variable region and attempts to assign it a subgroup designation based on those described in the 1991 edition of the database. In many cases the assignment is ambiguous because of a sequence's similarity to more than one subgroup.
Find Your Families	The Find Your Family tool attempts to assign a "family" designation to a user-entered sequence. The user-entered target sequence is compared to previously assembled groupings of sequences, based on sequence homology. Please note that the assigned family number is arbitrary, since the groupings usually change as new data is added to the database.
Current Counts	Current amino acid, nucleotide, and entry counts may be made for various groupings of sequences.
Variability	Variability calculations may be made over a user-specified collection of sequences. The distributions used to calculate the variability are also available for viewing and printing. Variability plots can be customized for scale, axis labels, and title, or downloaded for printing.

ticular gene to a specific family or subgroup. *SEQHUNT* ***(26)*** can pinpoint the sequence available in the database with the least number of amino acid or nucleotide differences.

The previous examples represent most of the current uses of the *Kabat Database* by immunologists and other scientists. However, many more detailed

DOCKET ALARM

Explore Litigation Insights

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

Real-Time Litigation Alerts

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

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

Advanced Docket Research

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

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

Analytics At Your Fingertips

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

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

API

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

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

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

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

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