Preface

The exquisite binding specificity of antibodies has made them valuable tools from the laboratory to the clinic. Since the description of the murine hybridoma technology by Köhler and Milstein in 1975, a phenomenal number of monoclonal antibodies have been generated against a diverse array of targets. Some of these have become indispensable reagents in biomedical research, while others were developed for novel therapeutic applications. The attractiveness of antibodies in this regard is obvious—high target specificity, adaptability to a wide range of disease states, and the potential ability to direct the host's immune system for a therapeutic response. The initial excitement in finding Paul Ehrlich's "magic bullet," however, was met with widespread disappointment when it was demonstrated that murine antibodies frequently elicit the human anti-murine antibody (HAMA) response, thus rendering them ineffective and potentially unsafe in humans. Despite this setback, advances in recombinant DNA techniques over the last 15–20 years have empowered the engineering of recombinant antibodies with desired characteristics, including properties to avoid HAMA. The ability to produce bulk quantities of recombinant proteins from bacterial fermentation also fueled the design of numerous creative antibody constructs. To date, the United States Food and Drug Administration has approved more than 10 recombinant antibodies for human use, and hundreds more are in the development pipeline. The recent explosion in genomic and proteomic information appears ready to deliver many more disease targets amenable to antibody-based therapy. Without doubt, the continued use of antibodies in the 21st century is ensured by virtue of their powerful recognition properties and, as now demonstrated, by their successful partnership with protein engineering.

Antibody Engineering: Methods and Protocols presents cutting-edge techniques in antibody engineering research. In Part I, popular resources for antibody sequence analysis are described, together with in-depth discussions on antibody structural modeling. A directory summarizing useful websites relevant to antibody engineering is also included. Part II presents protocols for antibody lead generation from the cloning of immunoglobulin genes to the selection and generation of human recombinant antibodies by molecular display technologies and transgenic animals. For well-characterized murine antibodies with clinical potential, humanization by CDR grafting offers a proven solution to minimizing HAMA, while sparing the additional efforts in generating a completely new human antibody entity. Procedures are also described on reformatting anti-

body leads into monovalent, multivalent, and bispecific binding fragments for a wide range of in vivo applications. Part III focuses on the expression and optimization of antibody leads. Traditional antibody expression systems such as bacterial and mammalian cell culture are described, followed by more recent developments in insect cell cultures and transgenic plants. The use of plants is particularly important as it provides the scope for the mass production of antibodies at a fraction of the cost compared to conventional systems, hence making therapeutic antibodies more economical. Besides lead expression, chapters are also devoted to the in vitro affinity maturation of recombinant antibodies using phage display and a rational approach in the design of minimally immunogenic antibodies. Finally, Part IV details state-of-the-art technologies for the characterization of antigen-binding affinity and specificity. Some novel applications of recombinant antibodies in radioimmunotargeting, cancer immunotherapy, drug abuse, and the emerging field of proteomics are also presented. Although *Antibody Engineering: Methods and Protocols* cannot cover every facet of antibody engineering research, it is hoped that these chapters will provide the antibody engineer with the fundamental techniques upon which further imaginative technologies can be developed.

I am indebted to all the authors for their expert contributions. In addition, I thank my colleagues at the Laboratory of Molecular Biology for discussions and support, John Walker for editorial guidance, and Humana Press for publishing this book.

Benny K. C. Lo

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

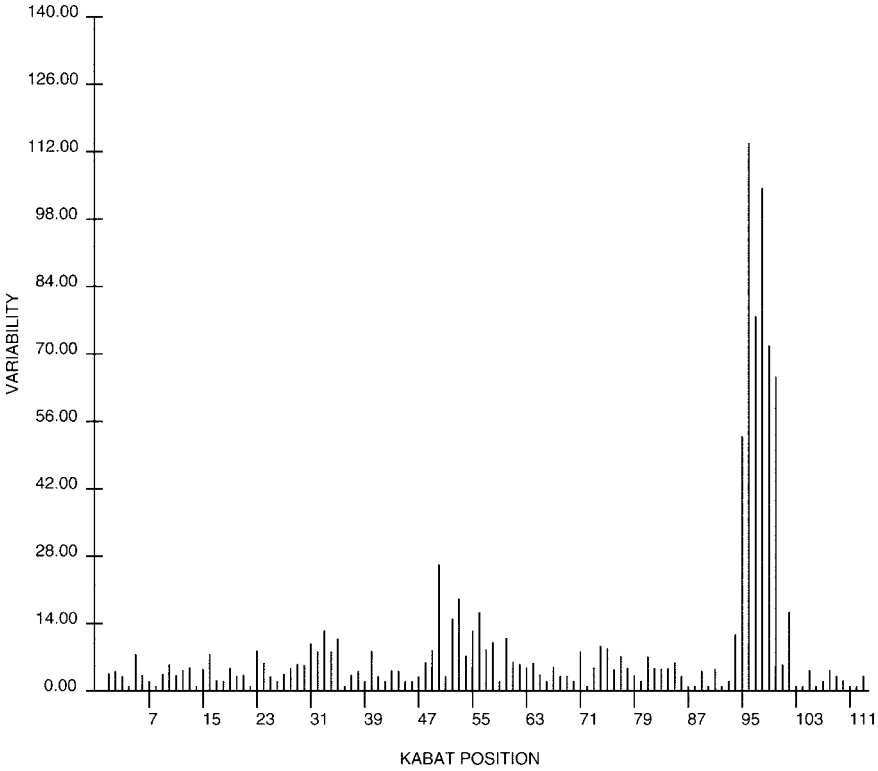

Fig. 1. Variability plot for human anti-DNA heavy-chain variable region.

analyses are possible from the data stored in the *Kabat Database*, as shown in **Table 3.**

In the following section, a current bioinformatics example is illustrated, using the uniquely aligned data contained in the *Kabat Database.*

2. Kabat Database Bioinformatics Example: HIV gp120 V3-loop and Human CDRH3 Amino Acid Sequences

The human immunodeficiency virus (HIV) has intrigued the scientific community for several decades. It is a retrovirus with two copies of RNA as its genetic material. Upon infecting humans, HIV uses its reverse-transcriptase molecules to convert its RNA into DNA, which are in turn transported into the nucleus and incorporated into the host chromosomes of CD4+ T cells. Although the infected individual produces antibodies against the initial viral strain, not all viruses can be eliminated because of the integration of its genetic material into the host cells. Gradually, the viral-coat proteins change in sequence, rendering the host's antibodies less effective. Eventually, acquired

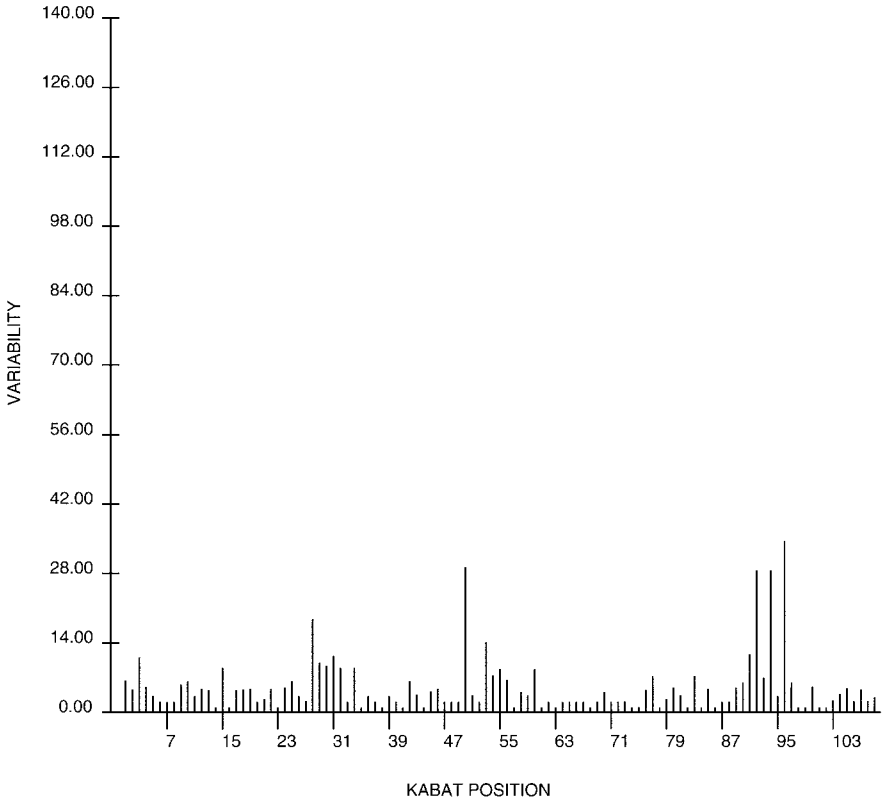

Fig. 2. Variability plot for human anti-DNA kappa light-chain variable region.

immunodeficiency syndrome (AIDS) develops with a latent period of approx 10 ± 3 yr. Because of this, HIV is classified as a lentivirus or slow virus.

Several specific drugs have been synthesized during recent years to treat HIV infection and AIDS. They include reverse-transcriptase inhibitors, protease inhibitors, and fusion inhibitors. However, these drugs have serious side effects, and most are very expensive, making the cost of treatment prohibitive in countries with a large percentage of HIV-positive patients. For years, the ideal solution has been to develop an inexpensive vaccine. Unfortunately, because of the rapid changes of its envelope coat proteins, especially gp120, HIV strains cannot be singled out as candidates for vaccine. Many research laboratories around the world have undertaken the task of sequencing gp120, and these sequences have been stored on two websites:

http://ncbi.nlm.nih.gov and http://www.lanl.gov

Figure 3 shows a variability plot for the 302 nearly complete sequences of HIV-1 stored at the latter site. For comparison, a variability plot of 138

Table 3
Partial Listing of Bioinformatics Studies Performed Using the Kabat Database

Subject	Summary
Binding Site Prediction	The CDRs of Ig heavy and light chains were predicted from variability calculations made over the sequence alignments ***(1,8)***.
Antibody Humanization	It is possible to identify the most similar framework regions between the mouse antibody and all existing human antibodies stored in the database ***(30)***.
Gene Count Estimation	From the existing sequences, it is possible to estimate the total number of human and mouse V-genes for antibody light and heavy chains, as well as TCR alpha and beta chains ***(31,32)***.
MHC Class I gene assortment	The known sequences of human MHC class I sequences suggest that their a1 and a2 regions can be assorted ***(33)***.
TCR CDR3 length distribution	The lengths of CDR3s in antibodies and TCRs have distinct features ***(34,35)***. In the case of TCR alpha and beta chains, their CDR3 lengths follow a narrow and random distribution. That may be a result of the relatively fixed size and shape of the processed peptide in the groove of MHC class I or II molecules. On the other hand, although the TCR gamma chain CDR3 lengths are similarly distributed, those of TCR delta chains exhibit a bimodal distribution ***(35)***. TCR delta chains with shorter CDR3s may be MHC-restricted, although those with longer CDR3s MHC-unrestricted.
Antibody and TCR evolution	Possible mechanisms of antibody and TCR evolution can also be investigated by comparing aligned sequences from different species ***(36,37)***.
Designer Antibodies	More specific/potent antibodies may be designed using the preferred CDR lengths calculated from database sequences against the same antigen ***(34)***.
Autoimmunity	Similarities between non-self antigens such as influenza virus and Ig autoantibodies have been found. Certain antigens may help initially trigger autoimmunity, and certain antibody clones may help to stimulate the autoimmune response ***(36)***.

aligned human influenza virus A hemagglutinin amino acid sequences is shown in **Fig. 4.**

Based on various studies, the V3-loop has been singled out for vaccine development. Although the V3-loop has the least amount of variation among

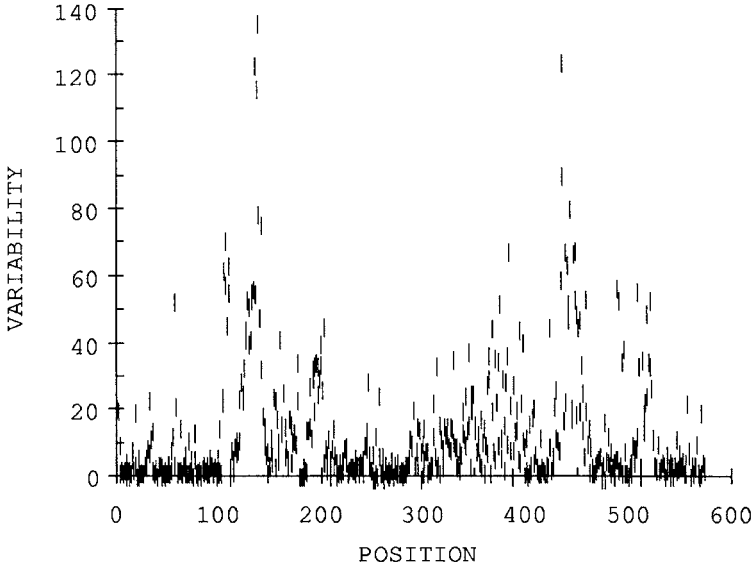

Fig. 3. Variability plot for HIV-1 gp120.

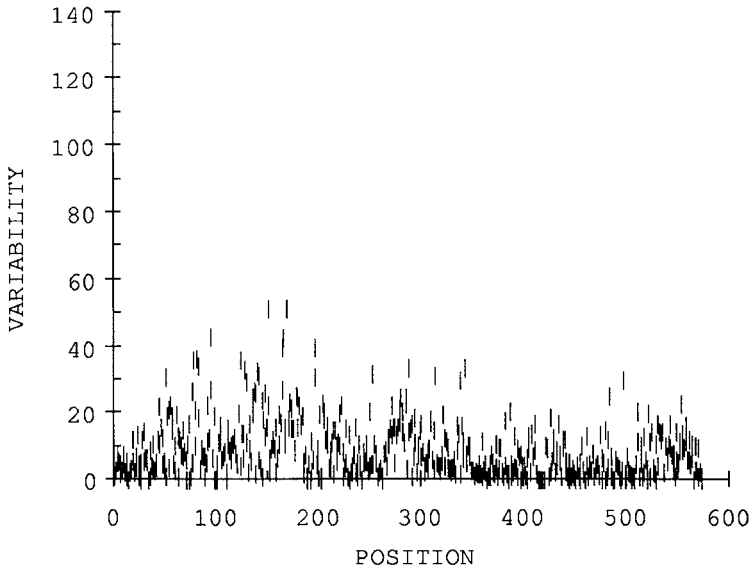

Fig. 4. Variability plot for influenza virus A hemagglutinin.

the five V-loops, there are still many different sequences from various strains of HIV. How these different sequences are related to the pathogenesis and progression of HIV infection is unclear. Longitudinal analysis of sequences of the V3-loop as the disease progresses is of vital importance in understanding the

changes that occur during infection, so that an effective vaccine can be developed. Unfortunately, there is only one published report for a 10-yr sequence analysis, and in that case, the authors were unable to describe how the V3-loop amino acid sequences are related to disease progression ***(27)***.

When HIV infects a person, its gp120 is a foreign protein and the patient produces antibodies toward this foreign antigen. However, once the HIV gene is integrated into the host chromosome, as in various human endogenous retroviruses, the gp120 becomes a self-protein. This transition from foreign to self usually cannot occur instantaneously, but as it occurs the host will have increasing difficulty producing effective antibodies. Indeed, initial antibodies from patients who are infected with HIV are usually ineffective in binding HIV at later stages of the disease.

The V3-loop has been described as being located on the surface of gp120. One way for the gp120 to become less antigenic would be for the virus to replace portions of the exposed V3-loop with segments of the host chromosome. Although any human protein could serve this purpose, we investigate the possibility that human CDRH3 regions are being used. CDRH3 is particularly attractive, because they can assume many possible configurations and they are on the surface of normal human proteins.

To locate matches between the V3-loop and CDRH3, the *Kabat Database* is uniquely useful. BLAST *(http://www.ncbi.nlm.nih.gov)* has recently allowed matches of short amino acid sequences, and eMOTIF *(http://emotif.stanford.edu/emotif/)* can be used to search for various length sequences. However, both programs use sequence databases containing large numbers of HIV-1 sequences and relatively few antibody heavy-chain variable region sequences. A search for short V3-loop sequences at these two websites usually results in a listing of other V3-loop sequences, and few, if any, CDRH3 sequences. By using the *SEQHUNTII* program, we picked the human heavy-chain variable regions and searched for all penta-peptides in the sequences of V3-loops determined in the 10-yr longitudinal study. The result of matching is listed in **Table 4.**

The initial number of matches is gradually reduced over the years, until the CD4+ T-cell count drops below 200. At that time, the number of matches increases dramatically. The match number appears to closely correlate with the number of HIV RNA molecules in the patient's blood. For example, after treatment, the number of matches drops to zero, along with a reduction in the plasma HIV RNA number. Subsequently, after 10 yr of HIV infection, the number of matches begins to creep up again.

A possible explanation for this finding is that the presence of CDRH3 penta-peptides in the V3-loop reduces its antigenicity. Such mutant HIV would bind existing anti-HIV antibodies in the patient less effectively, becoming more pathogenic. Based on this observation, the use of amino acid or nucleotide sequences of V3-loop as a vaccine would not be very efficient.

Table 4
Longitudinal Study of HIV gp120 V3-Loop Sequence Variations

Sample	Months after Infection	Sequence of V3-loop determined	Matches in human CDRH3	CDR4+ T-cells	HIV RNA per mL of plasma
A1	0	10	6		230
A2	12	10	3		230
A2b	27	7	0	427	2,300
A3	42	5	0	277	230
A4	70	3	0	186	230
A5	94	12	21	156	23,000
treatment	97				
A6	110	12	0	248	2,300
A7	118	12	1	212	2,300

An effective vaccine would most likely be made from an area of the exposed surface that does not contain high variability, as indicated in **Fig. 3.** There are several segments of seven or more nearly invariant amino acid residues in HIV gp120, in contrast to influenza virus hemagglutinin. Nearly invariant residues are defined as those that occur more than about 95% of the time at a particular position ***(1)***. They are located at the following positions (numbering including the precursor region) in the C1, C2, or C5 region of gp120:

Segment #	Position #	Sequence
I	4 to 14	WVTVYYGVPVW
II	23 to 30	LFCASDA
III	44 to 50	ACVPTDP
IV	225 to 231	PIPIHYC
V	261 to 267	VQCTHGL
VI	269 to 282	PVVSTQLLL-NGSL
VII	538 to 545	ELYKYKVV

Some of the adjacent residues occur more than 90% of the time. Furthermore, segments II and III and segments VI and V form disulfide bonds. Segment VI is only one residue away from segment V, and that residue is either K or R most of the time. Segment I is near the N-terminal and segment VII near the C-terminal, and they are physically located near each other in the folded structure of gp120 ***(28)***. If these segments are indeed located on the surface of gp120, we may then suggest that segment I linked to segment VII—with linkers consisting of repeats of GGGGS, segment II disulfide bounded to segment

III, and segment IV S-S bounded to segment V joined to segment VI with an intervening residue of K or R—should be used as possible peptide vaccine candidates. Additional residues that occur more than 90% of the time may also be included in these segments, suggesting the following three possible peptides:

```
WVTVYYGVPVWGGGSGGGSDNWRSELYKYKVV,

          LFCASDAK
             |
       WATHACVPTDP, and

        PIPIHYC
              |
           VQCTHGIKPVVSTQLLLNQSL.
```

In contrast, for influenza virus hemagglutinin amino acid sequences, no such segments of seven or more residues are found.

3. Future Directions

As previously discussed, during the past few years a substantial decline in the number of published sequences of proteins of immunological interest has occurred. With the shift in focus from brute-force data collection to in-depth analysis and "data mining" by various researchers, well-characterized data sets have become extremely important. Each entry in the database inherently contains a large amount of bioinformatic analysis such as alignment information, the relationship between gene sequence and protein sequence, and coding region designation. These relationships prove most valuable in allowing researchers to ask more intuitive, abstract questions than would be possible with most unaligned, raw sequence databases. We continue to locate, annotate, and align sequences found in the published literature. Periodically, the database and website are updated to reflect inclusion of the new data. Corrections of errors found in the sequence data by us and by database users are constantly made, ensuring the collection's accuracy. We continue to explore new ways of relating the database entries, such as incorporating links to journal abstracts, links to 3D structural information, and germline gene assignment.

We continue to create and develop software programs for performing various analyses of the data. We are in the process of converting many tools we have used into Java and adding graphical interfaces. Two major groupings of tools are currently being created: the first to update and extend the current entry retrieval tools (such as SeqhuntII), and the second to perform distribution analyses on entire groups of sequences (such as variability). Java tools for locating sequences based on pattern matching, length distribution of a specified region, positional

examination of a codon or residue, and sequence length have been developed and are undergoing testing. Many of the studies we have performed on the database require tools for grouping and analyzing collections of sequences rather than each one individually. We are developing a Java interface for creating distributions based on position (used most frequently for calculating variability), region length (used in length distribution analyses), and sequence pattern (used in gene count estimations and various homology studies). Together, these powerful interfaces will allow researchers to quickly perform many complex bioinformatics studies on the aligned sequence data and combine their results.

4. Conclusion

The fundamental reason for creating and maintaining most sequence databases is to study and correlate a protein's primary sequence structure with its 3D structure. Although there are many proteins with known 3D structures, there are probably two orders of magnitude more proteins with known amino acid or nucleotide sequences. In the 1950s, Anfinsen proposed and summarized in his 1973 paper ***(29)*** that the primary sequence of a protein should determine its 3D folding. Unfortunately, we still do not know how to decipher this information.

In the long run, the Kabat Database must be self-sustained. However, the transition from a free NIH-supported database to a self-sustaining format will take time and continued investigator interest. For example, it is hoped that the rapid development of therapeutic antibody techniques, using chimeric or humanized approaches, will eventually lead to the *de novo* synthesis of designer antibodies. Thus, immunotherapy for cancers and viral infections may rely heavily on the *Kabat Database* collections.

We will also rely on users to suggest to us what basic immunological ideas, what computer programs, and which types kinds of structure and function information will be of importance for future studies in this central problem in biomedicine. This feedback from users is of primary importance to the existence of the *Kabat Database.*

References

1. Wu, T. T. and 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.
2. Kabat, E. A., Wu, T. T., and Bilofsky, H. (1976) *Variable Regions of Immunoglobulin Chains.* Bolt Beranek and Newman Inc., Cambridge, MA.
3. Kabat, E. A., Wu, T. T., and Bilofsky, H. (1979) *Sequences of Immunoglobulin Chains.* NIH Publication No. 80–2008, Bethesda, MD.
4. Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Miller, M., and Perry, H. (1983) *Sequences of Proteins of Immunological Interest.* NIH Publication No. 369–847, Bethesda, MD.

5. Kabat, E. A., Wu, T. T., Reid-Miller, M., Perry, H., and Gottesman, K. (1987) *Sequences of Proteins of Immunological Interest,* 4th ed., U. S. Govt. Printing Off. No. 165–492, Bethesda, MD.
6. Kabat, E. A., Wu, T. T., Perry, H., Gottesman, K., and Foeller, C. (1991) *Sequences of Proteins of Immunological Interest,* 5th ed., NIH Publication No. 91–3242, Bethesda, MD.
7. Hilschmann, N., and Craig, L. C. (1965) Amino acid sequence studies with Bence Jones proteins. *Proc. Natl. Acad. Sci. USA* **53,** 1403–1409.
8. Kabat, E. A. and Wu, T. T. (1971) Attempts to locate complementarity-determining residues in the variable portions of light and heavy chains. *Ann. NY Acad. Sci.* **190,** 382–393.
9. Kohler, G. and Milstein, C. (1975) Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256,** 495–497.
10. Amit, A. G., Mariussa, R. A., Phillips, S. E., and Poljak, R. J. (1986) Three-dimensional structure of antigen-antibody complex at 2.8 A resolution. *Science* **233,** 747–753.
11. Wu, T. T., Kabat, E. A., and Bilifsky, H. (1975) Similarities among hypervariable segments of immunoglobulin chains. *Proc. Natl. Acad. Sci. USA* **72,** 5107–5110.
12. Kabat, E. A., Wu, T. T., and Bilofsky, H. (1978) Variable region genes for immunoglobulin framework are assembled from small fragments of DNA—a hypothesis. *Proc. Natl. Acad. Sci. USA* **75,** 2429–2433.
13. Bernard, O., Hozumi, N., and Tonegawa, S. (1978) Sequences of mouse light chain genes before and after somatic changes. *Cell* **15,** 1133–1144.
14. Milstein, C. (1967) Linked groups of residues in immunoglobulin chains. *Nature* **216,** 330–332.
15. Early, P., Huang, H., Davis, M., Calame, K., and Hood, L. (1980) An Immunoglobulin heavy chain variable gene is generated from three segments of DNA: VH, DH, and JH. *Cell* **19,** 981–992.
16. Sakano, H., Maki, R., Kurosawa, Y., Roeder, W., and Tonegawa, S. (1980) Two types of somatic recombinations are necessary for the generation of complete heavy chain genes. *Nature* **286,** 676–683.
17. Baltimore, D. (1981) Gene conversion: some implications for immunoglobulin genes. *Cell* **24,** 592–594.
18. Reynaud, C., Anquez, V., Dahan, A., and Weill, J. (1985) A single rearrange event generates most of the chicken immunoglobulin light chain diversity. *Cell* **40,** 283–291.
19. Desiderio, S. V., Yancopoulos, G. D., Paskind, M., Thomas, E., Boss, M. A., Landau, N., et al. (1984) Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. *Nature* **311,** 752–755.
20. Sleckman, B. P., Gorman, J. R., and Alt, F. W. (1996) Accessibility control of antigen-receptor variable-region gene assembly: role of cis-acting elements. *Annu. Rev. Immunol.* **14,** 459–481.
21. Kabat, E. A. and Wu, T. T. (1991) Indentical V-region amino acid sequences and segments of sequences in antibodies of different specificities: relative contributions

of VH and VL genes, minigenes and CDRs to binding of antibody combining sites. *J. Immunol.* **147,** 1709–1819.
22. Wu, T. T. (1994) From esoteric theory to therapeutic antibodies. *Appl. Biochem. Biotechnol.* **47,** 107–118.
23. Wu, T. T., Johnson, G., and Kabat, E. A. (1993) Length distribution of CDRH3 in antibodies. *Proteins* **16,** 1–7.
24. Wu, T. T. (2001) *Analytical Molecular Biology.* Kluwer Academic Publishers, Norwell, MA.
25. Wilson, M. R., Middleton, D., and Warr, G. W. (1988) Immunoglobulin heavy chain variable region gene evolution: structure and family relations of two genes and a pseudogene in a teleost fish. *Proc. Natl. Acad. Sci. USA* **85,** 1566–1570; and (1989) Erratum. *Proc. Natl. Acad. Sci. USA* **86,** 3276.
26. Johnson, G., Wu, T. T., and Kabat, E. A. (1995) SEQHUNT, a program to search aligned nucleotide and amino acid sequences, in *Antibody Engineering Protocols* (Paul, S., ed.), Humana Press, Totowa, NJ, pp. 1–15.
27. Janssens, W., Nkengasong, J., Heyndricks, L. van der Auwera, G., Vereecken, K., Coppens, S., et al. (1999) Intrapatient variability of HIV type I group O ANT70 during a 10-year follow-up. *AIDS Res. Hum. Retrovir.* **15,** 1325–1332.
28. Wyatt, R., Kwong, P. D., Desjardins, E., Sweet, R. W., Robinson, J., Hendrickson, W. A., et al. (1998) The antigen structure of HIV gp120 envelope glycoprotein. *Nature* **393,** 705–711.
29. Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science* **181,** 223–230.
30. Wu, T. T. and Kabat, E. A. (1992) Possible use of similar framework region amino acid sequences between human and mouse immunoglobulins for humanizing mouse antibodies. *Mol. Immunol.* **29,** 1141–1146.
31. Johnson, G. and Wu, T. T. (1997a) A method of estimating the numbers of human and mouse immunoglobulin V-genes. *Genetics* **145,** 777–786.
32. Johnson, G. and Wu, T. T. (1997b) A method of estimating the numbers of human and mouse T cell receptor for antigen alpha and beta chain V-genes. *Immunol. Cell Biol.* **75,** 580–583.
33. Johnson, G. and Wu, T. T. (1998a) Possible assortment of a1 and a2 regiuon gene segments in human MHC class I molecules. *Genetics* **149,** 1063–1967.
34. Johnson, G. and Wu, T. T. (1998b) Preferred CDRH3 lengths for antibodies with defined specificities. *Int. Immunol.* **10,** 1801–1805.
35. Johnson, G. and Wu, T. T. (2000a) Kabat database and its applications: 30 years after the first variability plot. *Nucleic Acids Res.* **28,** 214–218.
36. Johnson, G. and Wu, T. T. (2000b) Matching amino acid and nucleotide sequences of mouse rheumatoid factor CDRH3-FRH4 segments to other mouse antibodies with known specificities. *Bioinformatics* **16,** 941–943.
37. Johnson, G. and Wu, T. T. (2001) Kabat database and its applications: future directions. *Nucleic Acids Res.* **29,** 205–206.