Nucleic Acids Research, 2001, Vol. 29, No. 1 205–206

Kabat Database and its applications: future directions

George Johnson* and Tai Te Wu

Departments of Biochemistry, Molecular Biology and Cell Biology, and Biomedical Engineering, Northwestern University, Evanston, IL 60208, USA

Received September 7, 2000; Accepted September 11, 2000

ABSTRACT

The Kabat Database was initially started in 1970 to determine the combining site of antibodies based on the available amino acid sequences. The precise delineation of complementarity determining regions (CDR) of both light and heavy chains provides the first example of how properly aligned sequences can be used to derive structural and functional information of biological macromolecules. This knowledge has subsequently been applied to the construction of artificial antibodies with prescribed specificities, and to many other studies. The Kabat database now includes nucleotide sequences, sequences of T cell receptors for antigens (TCR), major histocompatibility complex (MHC) class I and II molecules, and other proteins of immunological interest. While new sequences are continually added into this database, we have undertaken the task of developing more analytical methods to study the information content of this collection of aligned sequences. New examples of analysis will be illustrated on a yearly basis. The Kabat Database and its applications are freely available at http://immuno.bme.nwu.edu.

INTRODUCTION

A number of analytical tools have been included in our website (1) for the Kabat database and its applications (2). We are continually developing new methods to study this collection of aligned sequences. Every year, we shall try to illustrate the importance of such methods for one specific example. In the present paper, we would like to use pair-wise comparison (3–6) of nucleotide sequences to study the evolutionary history of antibody heavy chain V-genes among human, mouse, chicken and shark. The situation for multi-gene families is very complicated (7,8).

Since there are relatively few available nucleotide sequences for chicken and shark, we have restricted human and mouse sequences to rheumatoid factors and anti-DNA antibodies. The sample sizes would be of similar orders of magnitude for these six groups: (i) 80 human rheumatoid factors; (ii) 52 human anti-DNA antibodies; (iii) 32 mouse rheumatoid factors; (iv) 167 mouse anti-DNA antibodies; (v) all 45 chicken sequences; and (vi) all 34 shark sequences. Only complete and distinct sequences are used in our analysis. For heavy chain variable region V-genes, they include codons 1–94, according to the Kabat numbering system (1).

RESULTS

Among all human rheumatoid factor heavy chain V-gene sequences, which are complete and distinct, the minimum difference is one base. However, totally unexpectedly, the maximum difference is 147 bases (Table 1) for a stretch of <300 nt. In the case of other proteins, such differences would have suggested that the two sequences are unrelated. Without a large collection of precisely aligned sequences, we would have never discovered this interesting finding.

Similarly, among V-gene sequences of human anti-DNA antibody heavy chain, the minimum difference is again 1 base while the maximum is 156 bases. For mouse rheumatoid factors, they are 1 and 140, respectively; and for mouse anti-DNA antibodies, they are 1 and 151. For all chicken heavy chain V-gene sequences, however, they are 1 and 50, respectively. This is most likely due to the existence of only one functional heavy chain V-gene, with the others being generated by gene conversion (9). For all shark sequence, the minimum and maximum are 1 and 167, respectively.

The minimum and maximum nucleotide differences between heavy chain V-gene sequences among different species are also summarized in Table 1. While the minimum difference is much larger than that for sequences from the same species, i.e. 1 base, the maximum differences are not substantially larger. For example, between human and mouse rheumatoid factors, the minimum is 45 base differences, while the maximum is 148 bases, similar to the value among all human rheumatoid factor sequences, i.e. 147 bases. Between human rheumatoid factor sequences and all shark sequences, the minimum difference is substantially increased to 114 bases. But the maximum only goes up to 186 bases.

This database of sequences can provide references to future sequence studies. In addition, as illustrated here, analytical methods as applied to properly aligned sequence collections can open up new areas of research.

DISCUSSION

The minimum nucleotide differences between any two antibody heavy chain V-gene sequences among different species may be used to estimate the time of evolutionary divergence of these species. For example, V-genes of human rheumatoid factor differ from those of mouse anti-DNA antibody by 38 bases, and from those of mouse rheumatoid factor by 45 bases

*To whom correspondence should be addressed. Tel: +1 847 491 7849; Fax: +1 847 491 4928; Email: johnson@immuno.bme.nwu.edu

Table 1. Minimum and maximum nucleotide differences among antibody heavy chain V-gene sequences from various species and different specificities

	Human RF	Human anti-DNA	Mouse RF	Mouse anti-DNA	Chicken (all)	Shark (all)
human RF	1/147	1/161	45/148	38/149	89/156	114/186
human anti-DNA		1/156	45/158	39/163	89/167	115/187
mouse RF			1/140	1/151	97/154	118/185
mouse anti-DNA				1/151	97/158	115/185
chicken (all)					1/50	121/185
shark (all)						1/167

(Table 1). The usual estimation for the time of divergence of human and mouse is considered to be about 100 million years ago, and may be related to the 38–45 base differences. If we assume a linear relationship between the time of divergence and the number of base differences, the time of divergence of chicken from human can be calculated from the values listed in Table 1 to be around 198–234 million years ago. Similarly, chicken may have diverged from mouse around 216–255 million years ago. Roughly, chicken has diverged from human and mouse about 225 million years ago.

As to shark, its V-genes of antibody heavy chains have diverged from human around 253–303 million years ago, from mouse around 256–311 million years ago and from chicken around 269–318 million years ago. Taken together, shark has diverged from human, mouse and chicken about 286 million years ago. Our estimations are in reasonably good agreement with others.

Furthermore, the maximum base difference between any two V-genes of antibody heavy chains, i.e. 187 (Table 1), can probably provide an estimation for the time at which the antibody V-genes first appeared. Our method gives a value of around 416–492 million years ago.

As clearly illustrated here, properly aligned nucleotide sequences of V-genes of antibody heavy chains in the Kabat database can provide useful information about their evolutionary past. A totally unexpected finding is that within the same species of human, mouse or shark, two sequences can differ by ~150 or more bases in a stretch of <300 nt. In chicken, where gene conversion is the basic mechanism of generating antibody diversity, the maximum difference is only 50 bases. This result seems to suggest that some of the V-genes may have remained functional during the entire evolutionary history. There may be little or no selective pressure to eliminate them.

CONCLUSION

The Kabat database is not only a sequence database, but also incorporates vital aspects of the biology of the immune system. Various analytical methods have been and will be developed to study the structure and function relations of proteins of immunological interest. As illustrated here, analysis of the entire updated database will be of vital importance in the future.

Electronic addresses

http://immuno/bme/nwu/edu
seqhunt2@immuno.bme.nwu.edu

Citing the Kabat database

Authors using this updated database may cite this paper together with the electronic address.

ACKNOWLEDGEMENT

Supported in part by NIH Grant 5 R24 AI25616-11.

REFERENCES

1. Johnson,G. and Wu,T.T. (2000) Kabat database and its applications: 30 years after the first variability plot. *Nucleic Acids Res.*, **28**, 214–218.
2. 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.
3. Johnson,G. and Wu,T.T. (1997) A method of estimating the numbers of human and mouse immunoglobulin V-genes. *Genetics*, **145**, 777–786.
4. Johnson,G. and Wu,T.T. (1997) A method of estimating the numbers of human and mouse T cell receptors for antigen alpha and beta chain V-genes. *Immunol. Cell Biol.*, **75**, 580–583.
5. Johnson,G. and Wu,T.T. (1997) Profile of numbers of sequence differences among V-genes coding for the variable regions of T cell receptor for antigen alpha and beta chains. *J. Mol. Evol.*, **44**, 253–257.
6. Johnson,G. and Wu,T.T. (1998) Possible assortment of a1 and a2 region gene segments in human MHC class I molecules. *Genetics*, **149**, 1063–1067.
7. Marchalonis,J.J., Schluter,S.F., Bernstein,R.M. and Hohman,V.S. (1998) Antibodies of sharks: revolution and evolution. *Immunol. Rov.*, **166**, 103–122.
8. Litman,G.W., Anderson,M.K. and Rast,J.P. (1999) Evolution of antigen binding receptors. *Ann. Rev. Immunol.*, **17**, 109–147.
9. Reynaud,C., Dahan,A., Anquez,V. and Weill,J.-C. (1989) Somatic hyperconversion diversifies the single VH gene of the chicken with a high incidence in the D region. *Cell*, **59**, 171–183.