VIR.II: a new interface with the antibody sequences in the Kabat database

Maria del Carmen Ramirez-Benitez a, Gabriel Moreno-Hagelsieb b, Juan Carlos Almagro a,c,*

a Instituto de Biotecnologia, UNAM, Mexico
b Programa de Biologia Computacional, Centro de Investigación sobre Fijación de Nitrógeno, UNAM, Mexico
c Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C3-168, P.O. Box 19024, Seattle, WA 98109-1024, USA

Received 9 April 2001; accepted 31 August 2001

Abstract

The Kabat database is the source of information par excellence on antibody sequences. In 1995, we developed an interface with the Kabat database, called VIR. VIR has been very useful in conducting studies aiming to find structure-function relationships in antibodies. Here we report a new version adapted to the World Wide Web, called VIR.II. VIR.II allows searches by type of chain (VH or VL), by species, and by specificity. The species are selected using a pulldown menu, whereas the specificities can be selected from a list containing the unique specificities reported in the Kabat database. These facilities avoid mistakes and redundancies in the searches. Another feature, and probably the most important one, is that VIR.II introduces a classification of specificities in terms of the chemical and biochemical nature of the antigen, like anti-protein, anti-peptide, anti-hapten, etc. This classification has been useful in discovering patterns in the antigen-binding site of antibodies that correlate with the type of antigen the antibody interacts with. To illustrate this, while showing the capabilities of VIR.II, we analyze all the murine anti-peptide and anti-protein antibody sequences compiled as of July, 2000 in the Kabat database. © 2001 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Immunoglobulin; Complementarity determining regions; CDRs; Antigen-binding site; Amino acid sequences; Nucleotide sequences

1. Introduction

Antibodies recognize a seemingly unlimited number of antigens with exquisite specificity. The first antibody sequence was determined in the mid-1960s (Hilschmann and Craig, 1965). In the early 1970s, Wu and Kabat compiled all the complete and partial sequences for antibodies published at that time; 77 in total (Wu and Kabat, 1970). That constituted the first version of the Kabat database (Johnson and Wu, 2000). The analysis of such information, prior to the determination of the three dimensional structure of an antibody (Poljak et al., 1972), identified the loca-
tion of the so-called complementarity determining regions (CDR’s): the regions responsible for binding antigen in this family of proteins.

The number of antibody sequences in the Kabat database has increased exponentially (http://www.ibt.unam.mx/vir/VIR/stat_aim_vir.html), amounting as July 2000 to 19,382 sequences. These sequences have been isolated from more than 70 species and as many as 7989 of them have the specificity annotated [4547 from V_H and 3442 from V_L (kappa + lambda)]. Thus, the Kabat database is the source of information par excellence for studies aiming to find correlations between patterns in antibody sequences and the specificity and/or the species.

To access the ever-increasing and valuable information gathered in the Kabat database two interfaces have been developed and are currently available in the World Wide Web (WWW): SeqhuntII (Johnson et al., 1995) and Kabatman (Martin, 1996).

SeqhuntII (http://immuno.bme.nwu.edu/seqhunt.html) allows searches through the annotations and sequences (Johnson and Wu, 2000) and the output is a fixed-line length record of 80 characters per line. This output format constrains the user to see one sequence at once. Therefore, to find patterns in the antibody sequences raised against a given specificity or isolated from a particular species, the user needs to run many queries, and the information has to be put together by using home-written programs. There is an electronic mail server (seqhunt2@immuno.bme.nwu.edu), which is considerably more flexible, though it does not offer ease of use.

Kabatman allows searches using an SQL-like query language (http://www.bioinf.org.uk/abs/kabatman.html) and the output is a list of sequences aligned following the Kabat conventions (Kabat et al., 1991). Sequences having a given pattern in the CDRs, for instance the CDR-L1 of 11 residues and a proline at position L29, can be obtained by using Kabatman (Martin, 1996). That is not possible when using SeqhuntII. Moreover, the URL http://www.bioinf.org.uk/abs/simkab.html provides access to a simple point-and-click version of Kabatman, allowing user-friendly searches. However, Kabatman is limited to work with amino acid sequences and the information at the nucleotide level is indispensable for determining with precision the genetic mechanisms that originated a particular antibody, i.e. the germline V, D, J gene combination, the addition and/or deletion of nucleotides at the V–J, V–D or D–J junctions and the putative somatic mutations (see for example IMGT/V-QUEST at http://imgt.cines.fr).

Early in 1995, we developed an interface to manage the antibody sequences available in the Kabat database (Almagro et al., 1995). Our interface, developed to run on PCs, was called VIR (V variable domains of the Immune system Receptors). During the last 6 years, VIR has been very useful in conducting studies that have discovered structure–function relationship in antibodies (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996 Almagro et al., 1997, 1998). This, together with the fact that VIR solved several of the SeqhuntII and Kabatman limitations, stimulates the implementation of a new version of VIR. The new version, designed to be accessible through the WWW, has been called VIR.II (http://www.ibt.unam.mx/vir/cgi/vir_searchform.cgi).

2. Main features of VIR.II

VIR.II allows searches of amino acid and nucleotide antibody sequences. The sequences from a given species are selected using a pulldown menu. This facility avoids mistakes and redundancies in the searches.

VIR.II also introduces a classification of the specificities in terms of the chemical and biochemical nature of the antigen. This classification has allowed the finding of rules to correlate the primary structure of an antibody with its capability to recognize different kinds of antigens (Vargas-Madrazo et al., 1995).

The output is a sequence alignment with a header showing the Kabat numbering schedule (Kabat et al., 1991). The sequences are aligned following the conventions on placements of deletions/insertions at CDRs proposed by Kabat et al. (1991). Each sequence is identified by the Kabat
VIR.II searches through databases created in DBM (database management) format (Glover and Humphrey, 1996). The DBM files are generated starting from the information contained in the dump files of the Kabat database (ftp://ncbi.nlm.nih.gov/repository/kabat/fxlen/). These files are downloaded after each Kabat database update.

The interface (http://www.ibt.unam.mx/vir/cgi/vir_searchform.cgi) consists of three sections (Fig. 1). The first section is used to select the kind of sequence (amino acid or nucleotide), the type of chain (VH, Vkappa or Vlambda) and the species. The type of chain and the species are selected in pulldown menus.

In the case of species the pulldown menu is generated by searching for unique species in the raw information downloaded from the dump files of the Kabat database. This preprocessing of the information avoids mistakes and redundancies in subsequent searches, since some entries in the Kabat database have the species annotated in

---

**VIR.II**

- **General**
  - Type of Sequence
  - Type of Chain
  - Species

- **Specificity**
  - Fine Specificity
  - Gross Specificity

- **Others**
  - Completeness
  - Counterpart

- **Amino acid**
- **Nucleotide**

- **VH**
- **Vkappa**
- **Vlambda**

- **Arctic Char**
- **Armenia hamster**
- **Atlantic cod**

---

Fig. 1. Structure of VIR.II (http://www.ibt.unam.mx/vir/cgi/vir_searchform.cgi).
colloquial terms like ‘frog’, whereas others use the scientific nomenclature: ‘Xenopus laevis’. Furthermore, the pulldown menu is an inventory of the species in itself, which offer quick access to the number and the species that originated the anti-species in itself, which offer quick access to the more, the pulldown menu is an inventory of the cities is generated from the raw information downloaded from the dump files of the Kabat database. The list is available to the user, but in this case, the query should be written or pasted in a text box from the list. This was so designed to allow more flexibility in the search. Also, we implemented the facilities ‘OR’ and ‘AND’. By using the list and the ‘OR’ and ‘AND’ facilities in a search, the user can construct strings that could take into account all the variants of a given specificity; including misspellings and redundancies.

For example, we have found four different definitions for the specificity ‘ANTI-(4-HYDROXY-3-NITROPHENYL) ACETYL’:

1. ANTI-(4-HYDROXY-3-NITRO-PHENYL) ACETYL (‘-’ after ‘NITRO’),
2. ANTI-(4-HYDROXY-3-NITRO-PHENY) ACETYL (no ‘L’ after ‘PHENY’),
3. ANTI-(4-HYDROXY-3-NITRO-PHENYL) ACETYL (space before ‘ACETYL’) and
4. ANTI-(4-HYDROXY-3-NITRO-PHENYL) ACETYL (no space before ‘PHENYL’).

When the Kabat database (as of July, 2000) was queried with VIR.II and these definitions were used, sets of different sequences were obtained. Definition 1 gave 10 sequences, definition 2 gave 75 sequences, definition 3 gave 248 sequences and definition 4 gave 263 sequences. To obtain all of them, a search using all of the definitions and the facility ‘OR’ was conducted. An alternative is to use the substring ‘ANTI-(4-HYDROXY-3-NITRO’, which is common to all the definitions. The hapten ‘ANTI-(4-HYDROXY-3-NITROPHENYL) ACETYL’ is not an exception, there are many more examples.

Within this section, VIR.II introduces a classification of antigens in seven groups (see Fig. 1). These groups have been called gross specificities (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996). By using a similar classification, different types of antigen-binding sites have been discovered, some of them with preference for the recognition of proteins, peptides, haptens, carbohydrates, nucleic acids, and so on, and others with multi-specific capabilities (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996).

There is no way to classify the specificities in gross specificities automatically. Once a number of antigens were classified, the process was achieved in a semi-automatic way. Today, in each update, the information downloaded from the Kabat database is compared with the specificities that have been classified previously. If a new sequence is specific, for instance for lysozyme, which already exits in the database as anti-protein, then it is classified as anti-protein. Otherwise, a file with exceptions is generated and the specificity is classified de novo. Obviously, the process is more automatic as more antigens are classified. The list of antigens, as classified in gross specificities, can be consulted in the URL: http://www.ibt.unam.mx/vir/VIR/gross_specificity.html.

The third section of VIR.II contains two options that have been designed to filter the search. The first option has been denoted ‘completeness’ to differentiate it from the definition of a complete antibody of Kabatman (http://www.bioinf.org.uk/abs/simkab.html). Completeness allows selection of sequences with a certain percentage of undefined positions in a given segment. This kind of filtering is very useful when considering that a mere 20% of the se-

<table>
<thead>
<tr>
<th>Gross specificity</th>
<th>CDR-L1 length</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-protein</td>
<td>359</td>
<td>265</td>
</tr>
<tr>
<td>Anti-peptide</td>
<td>16</td>
<td>87</td>
</tr>
<tr>
<td>Total</td>
<td>375</td>
<td>352</td>
</tr>
</tbody>
</table>

Table 1 Contingency table of the CDR-L1 length distribution, as grouped in short and long loops, and the sequences classified as anti-protein and anti-peptide antibodies.
Fig. 2. Distribution of the CDR-L1 lengths in the murine anti-protein and anti-peptide antibody sequences available in the Kabat database.

quences (3850 out of 19,382) available in the Kabat database are 100% complete.

The other option, called ‘counterpart’, is equivalent to the option ‘complete’ in Kabatman. This allows searches of partners. If the user selects, for instance $V_H$ chains, then by choosing counterpart, all the $V_L$ chains (kappa and lambda) that share the same name of the $V_H$ chains, will be obtained. This filter is indispensable when patterns in the antigen-binding site as a whole ($V_H + V_L$) are wanted.

4. Example of application

Studies conducted with VIR have revealed features at the antigen-binding site of antibodies that correlate with the specificity (Vargas-Madrazo et al., 1995; Lara-Ochoa et al., 1996; Almagro et al., 1997, 1998). We have found that antibodies with a short CDR-L1 preferentially recognize large antigens such as proteins (Vargas-Madrazo et al., 1995). In contrast, antibodies with a long CDR-L1 have been found to bind smaller molecules such as peptides (Vargas-Madrazo et al., 1995). To illustrate these findings, while showing the potential use of VIR.II, in this section we analyze the lengths of the CDR-L1 in the murine anti-peptide and anti-protein antibody sequences available in the Kabat database (as of July, 2000).

We choose amino acid as kind of sequence, $\text{Vkappa}$ as chain type and mouse as species. Then, we choose anti-peptide in gross specificity. Finally, we set 24 as the beginning position and 34 as the ending one (according to the definition of CDR-L1 of Kabat et al., 1991). The completeness was defined as 100%. We found 103 anti-peptide sequences. In the case of anti-protein antibodies, we just changed to anti-protein in gross specificity and found 624 sequences.

Fig. 2 shows the distribution of the CDR-L1 lengths of the anti-protein and anti-peptide sequences. The lengths of the CDR-L1 follow a bimodal distribution. No sequence has a CDR-L1 of 13 residues, which defined the boundary between a short and a long CDR-L1. The proportion of short and long CDR-L1 loops in anti-protein antibodies is similar: 58 and 42%, respectively. Anti-peptide sequences are biased towards the use of a long CDR-L1 (84% of the sample).
To test whether this bias is related to the classification of the sequences in anti-protein and anti-peptide antibodies, a $2 \times 2$ contingency table (Table 1) was set up. The $X^2$ gave 30.231 and $X^2_{0.001,1}$ is 10.828 (Zar, 1999), meaning that the bias in the length of the CDR-L1 is related to the classification of the sequences in anti-peptide and anti-protein antibodies ($P \ll 0.01$). A similar analysis could be conducted for the remaining CDRs, in particular for the CDR-H3, where biases in the length of this loop may be related to the specificity (Kabat and Wu, 1991). Also, a study of the combination of lengths in CDRs in antibodies that interact with different gross specificities could be conceived.

It is worth mentioning that this kind of analysis is difficult to achieve by using Kabatman. With that interface many queries, one for each CDR length, should be run and then the sequences should be classified in terms of gross specificities. This latter task, as mentioned above, is not a trivial one since the classification of antigens within groups is not a straightforward process.

5. Conclusion

Antibodies constitute a paradigm of molecular recognition since they bind to a virtually infinite number of antigens with exquisite specificity. To decipher the molecular basis of such a feature, several thousands of antibody sequences have been obtained in the last 40 years. As a consequence, several specialized immunogenetic databases have been created (Brusic et al., 2000; Düböl, 2000). However, despite this accumulation of information, its classification and analysis, it is yet not possible to predict the specificity of a given antibody sequence, even if much is known about the antigen structure.

One of the main features of VIR.II is that the amino acid and nucleotide sequences available at the Kabat database can be queried by using a classification of the antigens in terms of gross specificities. This classification may be the seed of a taxonomy of antigens. Actually, by using this classification, patterns at the antigen-binding site have emerged (Vargas-Madrazo et al., 1995; MacCallum et al., 1996).

In the section ‘example of application’, we showed how by using VIR.II, a long CDR-L1 may be associated to antibodies that bind to peptides. Such a rule is still quite general and rudimentary. However, we envision that systematic analyses of the ever-increasing and valuable information available in the Kabat database, via VIR.II, would serve to generate rules of increasing degree of complexity. Such rules probably allow, in the near future, better (fine-tuning) predictions of the recognition features of antibodies.

Acknowledgements

The authors thank T. Buss for the critical revision of the manuscript. Ramirez-Benitez was supported by CONACyT. This work was partially supported by the grants of DGAPA-UNAM IN223399 and CONACyT 33352-N.

References


Martin, A.C.R., 1996. Accessing the kabat antibody sequence database by computer PROTEINS: structure, function and genetics. 25, 130–133.


