A database of immunoglobulins with integrated tools: DIGIT

Chailyan, Anna; Tramontano, Anna; Marcatili, Paolo

Published in:
Nucleic acids research

Link to article, DOI:
10.1093/nar/gkr806

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
A database of immunoglobulins with integrated tools: DIGIT

Anna Chailyan1, Anna Tramontano1,2,* and Paolo Marcatili1*

1Department of Physics, Sapienza University of Rome, P.le A. Moro, 5-00185, Italy and 2Center for Life Nano Science, Sapienza, Istituto Italiano di Tecnologia, Viale Regina Elena, 291-00161, Rome (I), Italy

Received August 14, 2011; Accepted September 13, 2011

ABSTRACT

The DIGIT (Database of Immunoglobulins with Integrated Tools) database (http://biocomputing.it/digit) is an integrated resource storing sequences of annotated immunoglobulin variable domains and enriched with tools for searching and analyzing them. The annotations in the database include information on the type of antigen, the respective germline sequences and on pairing information between light and heavy chains. Other annotations, such as the identification of the complementarity determining regions, assignment of their structural class and identification of mutations with respect to the germline, are computed on the fly and can also be obtained for user-submitted sequences. The system allows customized BLAST searches and automatic building of 3D models of the domains to be performed.

INTRODUCTION

Successful recognition of foreign antigens by antibodies (or immunoglobulins) is crucial for the defense of an organism against pathogens and strictly depends upon the enormous diversity of the sequences and structures of these molecules. At the same time, these molecules play an exceptionally important role in diagnosis, therapy and biotechnology applications.

The effective usage of antibodies in all these applications demands knowledge and understanding of their sequence and structural properties in order to study the molecular basis of their specificity, their ‘evolutionary’ history within the organism and to be able to modify them as in humanization experiments or in the design of combinatorial libraries.

There are several resources aimed at providing an integrated view of the sequences and structures of antibodies, each with advantages and disadvantages. The most renowned one is the Kabat database (1), which has been the textbook (and originally was indeed released as such) for immunologists. Unfortunately, this is now only available at a cost and is not regularly updated. The Abysis portal (2) provides some of the needed services, such as the possibility of querying the database by accession number, antigen, author name, reference, year of first publication, chain type (lambda or heavy or both), species, etc., but is limited to amino acid sequences only and cannot be used for nucleotide sequences. The Vbase2 database (3) is limited to human and mouse germline sequences and, most importantly, has not been updated since 2006. IMGT (4) is a database of fully annotated sequences of immunoglobulins and T-cell receptors from human and other vertebrates (150 species). It does not provide sequence-searching tools for amino acid sequences nor it includes information on light and heavy-chain pairing of the entries.

To overcome some of the shortcomings of the systems described above and the problems that we ourselves faced when analyzing real life cases (5–9), we took advantage of our long-lasting experience in immunoglobulin sequence and structure analysis and structural prediction (8,10–19) and developed the DIGIT (Database of Immunoglobulins with Integrated Tools) system.

The annotations in our database include information on the type of antigen, the respective germline sequences and on pairing information between light and heavy chains.

The user can query the database using the antigen type, source organism, accession number, chain type (heavy, lambda and kappa) or free text (disease, process, etc.) with the option of selecting only complete immunoglobulins (i.e. cases where both the correctly paired light and heavy-chain sequences are available).

Other annotations are computed on the fly (and therefore can also be obtained for user-submitted sequences), for example:

(1) numbering of the sequence according to the Kabat–Chothia numbering scheme (20);
are called ‘canonical structures’. The canonical structure main-chain conformations of the hypervariable regions variable regions (17,18,21). The commonly occurring unusual main chain conformations, are primarily respon- 
sibility for the main-chain conformations of the hyper- 
packing, hydrogen bonding or the ability to assume 
identification of relatively few residues that, through their 
binding sites has been extensively studied leading to the 
classification of two light and two identical heavy chains. These are 
variable and a constant one; form the light 
chain, while one variable domain and three constant 
domains usually form the heavy chain. Most of the diver- 
sity of the variable domains resides in three regions from 
each chain, called the hypervariable or CDRs. These are 
named according to the chain they belong to and the order 
they appear in the sequence (L1, L2, L3, H1, H2 and H3). The regions between the CDRs in the variable domain are 
called the framework regions (FW). Immunoglobulin light 
chains are classified as kappa or lambda according to their 
serological and sequence properties.

Immunoglobulin sequences are usually numbered accord- 
ing to a common scheme (Kabat–Chothia) aimed at assigning the same number to topologically equivalent residues (20). This is a widely adopted standard for number- 
bering the residues of antibodies in a consistent manner. The relationship between the amino acid sequences of immunoglobulins and the 3D structures of their antigen binding sites has been extensively studied leading to the identification of relatively few residues that, through their packing, hydrogen bonding or the ability to assume unusual main chain conformations, are primarily respon- 
sible for the main-chain conformations of the hypervariable regions (17,18,21). The commonly occurring main-chain conformations of the hypervariable regions are called ‘canonical structures’. The canonical structure 
definitions can be effectively used to predict the structure of immunoglobulin variable domains (12,14).

DIGIT SYSTEM OVERVIEW
DIGIT consists of four main modules:

(2) identifications of the complementarity determining regions (CDRs) in the sequence and of the framework regions;
(3) assignment of the canonical structures for the CDRs (21);
(4) identification of mutations with respect to the germline;
(5) automatic link to our 3D modeling tool for immuno- 
globulin variable domains (14); and
(6) sequence searching that, given the input immuno- 
globulin sequence of interest (amino acid or nucleo- 	tide sequence of heavy-chain variable domain sequence; light-chain variable domain sequence or both), retrieves the closest sequences (sorted accord- 
ing to the E-value or percentage of sequence identity).

We believe that this is a much-needed resource as the information that it contains is either absent from any other database or can only be obtained by browsing several sites, most of which is not regularly updated and we are convinced that DIGIT will be extremely useful to researchers interested in immunology as well as to scientists performing experiments such as antibody humaniza-
tion, stabilization and functionalization.

IMMUNOGLOBULIN VARIABLE DOMAIN STRUCTURE AND NOMENCLATURE
Immunoglobulins are glycoproteins specifically binding to 
one or a few closely related antigens. All immunoglobulins have a four-chain structure as their basic unit. They are composed of two identical light chains (L) and two ident- 
tical heavy chains (H) held together by inter-chain disul-
fide bonds and by non-covalent interactions. Two domains, a variable and a constant one, form the light 
chain, while one variable domain and three constant 
domains usually form the heavy chain. Most of the diver-
sity of the variable domains resides in three regions from 
each chain, called the hypervariable or CDRs. These are 
named according to the chain they belong to and the order 
they appear in the sequence (L1, L2, L3, H1, H2 and H3). The regions between the CDRs in the variable domain are 
called the framework regions (FW). Immunoglobulin light 
chains are classified as kappa or lambda according to their 
serological and sequence properties.

Immunoglobulin sequences are usually numbered accord- 
ing to a common scheme (Kabat–Chothia) aimed at assigning the same number to topologically equivalent residues (20). This is a widely adopted standard for num-
bering the residues of antibodies in a consistent manner. The relationship between the amino acid sequences of immunoglobulins and the 3D structures of their antigen binding sites has been extensively studied leading to the identification of relatively few residues that, through their packing, hydrogen bonding or the ability to assume unusual main chain conformations, are primarily respon-
sible for the main-chain conformations of the hypervariable regions (17,18,21). The commonly occurring main-chain conformations of the hypervariable regions are called ‘canonical structures’. The canonical structure 
definitions can be effectively used to predict the structure of immunoglobulin variable domains (12,14).

DIGIT SYSTEM OVERVIEW
DIGIT consists of four main modules:

(2) identifications of the complementarity determining regions (CDRs) in the sequence and of the framework regions;
(3) assignment of the canonical structures for the CDRs (21);
(4) identification of mutations with respect to the germline;
(5) automatic link to our 3D modeling tool for immuno- 
globulin variable domains (14); and
(6) sequence searching that, given the input immuno- 
globulin sequence of interest (amino acid or nucleo-
tide sequence of heavy-chain variable domain sequence; light-chain variable domain sequence or both), retrieves the closest sequences (sorted according to the E-value or percentage of sequence identity).

We believe that this is a much-needed resource as the information that it contains is either absent from any other database or can only be obtained by browsing several sites, most of which is not regularly updated and we are convinced that DIGIT will be extremely useful to researchers interested in immunology as well as to scientists performing experiments such as antibody humaniza-
tion, stabilization and functionalization.

IMMUNOGLOBULIN VARIABLE DOMAIN STRUCTURE AND NOMENCLATURE
Immunoglobulins are glycoproteins specifically binding to 
one or a few closely related antigens. All immunoglobulins have a four-chain structure as their basic unit. They are composed of two identical light chains (L) and two ident-
tical heavy chains (H) held together by inter-chain disul-
fide bonds and by non-covalent interactions. Two domains, a variable and a constant one, form the light 
chain, while one variable domain and three constant 
domains usually form the heavy chain. Most of the diver-
sity of the variable domains resides in three regions from 
each chain, called the hypervariable or CDRs. These are 
named according to the chain they belong to and the order 
they appear in the sequence (L1, L2, L3, H1, H2 and H3). The regions between the CDRs in the variable domain are 
called the framework regions (FW). Immunoglobulin light 
chains are classified as kappa or lambda according to their 
serological and sequence properties.

Immunoglobulin sequences are usually numbered accord- 
ing to a common scheme (Kabat–Chothia) aimed at assigning the same number to topologically equivalent residues (20). This is a widely adopted standard for number-
bering the residues of antibodies in a consistent manner. The relationship between the amino acid sequences of immunoglobulins and the 3D structures of their antigen binding sites has been extensively studied leading to the identification of relatively few residues that, through their packing, hydrogen bonding or the ability to assume unusual main chain conformations, are primarily respon-
sible for the main-chain conformations of the hypervariable regions (17,18,21). The commonly occurring main-chain conformations of the hypervariable regions are called ‘canonical structures’. The canonical structure 
definitions can be effectively used to predict the structure of immunoglobulin variable domains (12,14).

DIGIT SYSTEM OVERVIEW
DIGIT consists of four main modules:

(2) identifications of the complementarity determining regions (CDRs) in the sequence and of the framework regions;
(3) assignment of the canonical structures for the CDRs (21);
(4) identification of mutations with respect to the germline;
(5) automatic link to our 3D modeling tool for immuno-
globulin variable domains (14); and
(6) sequence searching that, given the input immuno-
globulin sequence of interest (amino acid or nucleo-
tide sequence of heavy-chain variable domain sequence; light-chain variable domain sequence or both), retrieves the closest sequences (sorted according to the E-value or percentage of sequence identity).

We believe that this is a much-needed resource as the information that it contains is either absent from any other database or can only be obtained by browsing several sites, most of which is not regularly updated and we are convinced that DIGIT will be extremely useful to researchers interested in immunology as well as to scientists performing experiments such as antibody humaniza-
tion, stabilization and functionalization.
chains reported an unambiguous identifier after the keywords ‘clone’, ‘sample’ or ‘isolate’.

The name of the antigen was retrieved by searching in the NCBI description field the words following the ‘anti’ term. The type of antigen was attributed using a vocabulary developed on purpose.

A manual analysis of the automatic assignment on a few hundreds immunoglobulins stored in the database showed that only a handful of assignments were incorrect, and these were mainly due to ambiguous description fields in the NCBI entry.

**DB SEARCH TOOL**

The database can be queried by chain type (heavy, lambda, kappa), source organism, antigen, accession number or through a free text search.

The output of the DB search operation includes the corresponding NCBI identifier, a description of the antigen, the organism source and the reference to the original article.

Fields are linked to the corresponding NCBI entry, to the PUBMED record for the article and to a summary page reporting the genomic locus, the NCBI definition of the entry, the organism, the reference to the original article, the sequences of the light and heavy chain, the isotype of the lambda chain (kappa or lambda), the Kabat–Chothia numbering and, if available, the antigen and the antigen type (protein, peptide, carbohydrate, hapten and small molecule).

A toggle button can be used to select an entry for further analysis (see ‘Analysis Tools’ section).

**BROWSE TOOL**

This is a simple interface for browsing the content of the database and selecting an entry to analyze. The user can select the chain type or (optionally) the progressive number of the entry from which to start. The output page contains a list of the first 25 retrieved entries and of buttons to move to the next or previous 25 entries.

**SEQUENCE SEARCH TOOL**

The user can provide the amino acid or nucleotide sequence(s) of a light or heavy chain or both.

The system performs one or more Blast searches with default parameters. If the input consists of a light and

---

Figure 1. Schematic view of the options provided by DIGIT. After selecting an entry obtained by either browsing or searching the database or directly submitted, the user can retrieve the Kabat–Chothia numbering of the sequence, the canonical structures of the CDRs, the mutations with respect to the germline and the 3D model of the molecule. Note that the input can be either a nucleotide or an amino acid sequence. The system provides the possibility of printing and saving the results as well as of aligning the sequences of the displayed entries.
heavy chain, the database is searched with the light chain, with the heavy chain and with both. The latter result is useful, for example, for humanization experiments or for selecting a modeling template.

In both cases, for example, it might be convenient to select the chains of the same immunoglobulin as templates for both the light and heavy chain rather than chains coming from different antibodies, since even minor differences in the interface can lead to a different packing geometry and affect the topology of the antigen binding site (11).

In the description below, we will assume that the user wishes to search the database with a complete variable domain including both the light and heavy chains. The only difference from the case when only one of the two chains is used as input is that, obviously, only the results for the selected chain are available in this case.

**ANALYSIS TOOLS**

Upon completion of the search, the user can access several pages using tabs labeled ‘Blast L’, ‘Blast H’, ‘Blast L+H’, ‘Kabat numbering’, ‘Canonical structures’, ‘Features’, ‘Mutations’ and ‘3D model’ (Figure 2).

The first three options provide the Blast results for the light and heavy chain and those obtained by concatenating the two chains and searching a dataset where paired light and heavy-chain sequences are concatenated as well.

These pages contain the NCBI sequence identifier of the retrieved sequence(s), the definition(s), the reference, the antigen, the percentage of sequence identity and the E-value as reported by BLAST. Each column can be used for sorting the results and columns can be moved around using the mouse.

The Kabat numbering page reports the alignment of the input sequence with the commonly used Kabat–Chothia numbering scheme (20). The canonical structures (describing the main chain conformation of the CDRs) are also reported together with the length of the CDRs.

The Features page is a summary reporting the sequences of the various regions of the molecule separately in the order they appear in the sequence (Framework 1, CDR1, Framework 2, CDR2, Framework 3, CDR3 and Framework 4).

The ‘Mutations’ page includes information about the mutations with respect to the germline of the selected sequence. If the input sequence was provided as a nucleotide sequence, both the nucleotide and the corresponding amino acid mutations are shown. The Blossum62 score for the amino acid variation is also shown below each mutation.

**Figure 2.** The output page of the DB search tool. The left panel lists previous jobs the results of which can be retrieved. The current job is enclosed in a box. The various tabs provide access to the corresponding results. Each of the entries can be used as starting point for a new analysis by clicking the ‘A’ button. The results can be printed, saved or filtered according to user-defined keywords. The multiple sequence alignment for all the displayed entries, obtained by clicking the ‘Alignment’ button, is displayed on a new page.
Finally, the user can directly obtain a 3D model of the input sequence through the PIGS immunoglobulin modeling tool (14).

In all cases when a list of sequences is displayed on the page, it is possible to obtain an alignment of their sequences in a new window. If both light and heavy chains are listed on the page (as for example in a Blast ‘L+H’ output), the alignment will include both light and heavy-chain sequences separated by the ’/’ symbol.

CONCLUSIONS

Several biomedical and biotechnological projects need to take advantage of a detailed understanding of the immunoglobulin features. Humanization and combinatorial library design require an analysis of the properties of the framework of the molecules and of their similarity with other antibodies from the same or a different organism, immunologically based diagnosis can be helped by comparing the sequences of different immunoglobulins as well as by understanding their mutation patterns with respect to the germline and by inspecting their known or predicted 3D structures.

In all these cases, the light and heavy chains of the antibody cannot be treated as separate entities (as is the case in sequence databases). Furthermore, the analyses have to be easy to perform within a single site and reported in a language and with an organization reflecting the flowchart of the design of an experiment.

We believe that the DIGIT system described here meets all these requirements and can become a ‘one stop shop’ for the biomedical and biotechnological community interested in immunoglobulins.

FUNDING

King Abdullah University of Science and Technology (KAUST) (Award number KUK-I1-012-43); FIRB Proteomica, Ministry of Health grant (Contract No. Onc_Ord 25/07); IIT SEED project. Funding for open access charge: King Abdullah University of Science and Technology (KAUST) (Award number KUK-I1-012-43).

REFERENCES