MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

IMGT, the International ImMunoGeneTics Database: a New Design for Immunogenetics Data Access

Véronique Giudicelli[a], Denys Chaume[b], Gérard Mennessier[c], Hans-Helmar Althaus[d], Werner Müller[d], Julia Bodmer[e], Ansar Malik[f] and Marie-Paule Lefranc[a]

[a]Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UMR 5535 (CNRS, Université Montpellier II),1919 route de Mende, 34293 Montpellier Cedex 5, France, lefranc@ligm.crbm.cnrs-mop.fr[b], CNUSC, 950 avenue de Saint Priest, BP 7229, 34184 Montpellier Cedex 4, France,[c] ESA 5032 (CNRS, Université Montpellier II), Place Eugène Bataillon, 34095 Montpellier Cedex 5, France,[d] IFG, Universität zu Köln, Weyertal 121, 50931 Köln, Germany,[e] ICRF, CIL, Institute of Molecular Medicine, Oxford OX3 9DU, UK,[f] EMBL-EBI, Hinxton Hall, Hinxton CB10 1RQ, UK

Abstract

IMGT, the international ImMunoGeneTics database [1, 2] is an integrated database specializing in Immunoglobulins (Ig), T-cell receptors (TcR) and MHC molecules of all vertebrate species, created by Marie-Paule Lefranc, University of Montpellier, CNRS, Montpellier, France (Nucleic Acids Research, Database issue, Vol 26, January 1998). IMGT includes three databases: LIGM-DB (for Ig and TcR), MHC/HLA-DB and IMGT/PRIMER-DB (an Ig, TcR and MHC-related primer database), the last two in development. IMGT comprises expertly annotated sequences and alignment tables. LIGM-DB contains more than 24.000 Immunoglobulin and T cell Receptor sequences from 81 different species. MHC/HLA-DB contains class I and class II Human Leucocyte Antigen alignment tables. An IMGT tool, DNAPLOT, developed for Ig, TcR and MHC sequence analysis, is also available. IMGT goals are to establish a common data access to all immunogenetics data, including nucleotide and protein sequences, oligonucleotide primers, gene maps and other genetic data of Ig, TcR and MHC molecules, from all species, and to provide a graphical user friendly data access. IMGT has important implications in medical research (repertoire in autoimmune diseases, AIDS, leukemias, lymphomas), therapeutical approaches (antibody engineering), genome diversity and genome evolution studies. In this paper, we describe our approach for the data modelisation, the automation of the annotation procedure and control of data quality in LIGM-DB database. IMGT is freely available on the CNUSC WWW server at Montpellier: ***http://imgt.cnusc.fr: 8104*** *(contact: Denys.Chaume@cnusc.fr) and on the EBI servers: http://www.ebi.ac.uk/imgt (contact: malik@ebi.ac.uk) and ftp.ebi.ac.uk/pub/databases/imgt. LIGM-DB users are encouraged to report errors or suggestions to giudi@ligm.crbm.cnrs-mop.fr. IMGT initiator and coordinator: Marie-Paule Lefranc, lefranc@ligm.crbm.cnrs-mop.fr. (fax: +33 (0)4 67 04 02 31)*

Keywords

Immunogenetics; Database; Immunoglobulin; T cell receptor; Major Histocompatibility Complex; Sequence Annotation; Motif Search; Sequence Alignment

Introduction

The immune system has evolved to protect individuals against pathogenic viruses, micro-organisms and parasites. It is vital therefore that individuals have a normal immune system. Normal immune responses depend on the ability to recognize foreign molecules or antigens on the potential pathogen in order to eliminate the source of the antigen. The molecules involved in the recognition of antigens are encoded by the immunoglobulin superfamily, and this includes immunoglobulins (Ig), T cell receptors (TcR) and Major Histocompatibility Complex (MHC). In humans, the latter is referred to as the Human Leukocyte Antigen (HLA) system.

Scientists over a number of years have been rapidly sequencing the DNA which encodes the molecules of the immune system. Presently more than 24.000 gene sequences have been determined, and this number will double over the next two years. The molecular synthesis and genetics of the Ig and TcR chains is particularly complex [3,4], and the generalist databases cannot adequately label or annotate these sequences. Moreover, there is no way to search exhaustive and efficient links between pathologies and sequence components. This makes the core data held at these generalist databases difficult to use or interpret for researchers and clinicians. Obviously, a specialist database was required to add value, and to link it to other biological databases.

The international ImMunoGeneTics (IMGT) database [2] was created in 1992 by Marie-Paule Lefranc (CNRS, Université Montpellier II, Montpellier, France, lefranc@ligm.crbm.cnrs-mop.fr). IMGT (**http://imgt.cnusc.fr: 8104**) comprises alignment tables and expertly annotated sequences and consists of three databases: LIGM-DB for Ig and TcR, MHC/HLA-DB and PRIMER-DB (an Ig, TcR and MHC-related primer database), these last two are currently in development. The database is developed by LIGM (Montpellier, France), in collaboration with EMBL-EBI (Hinxton, UK), ICRF (Oxford, UK), IFG (Köln, Germany), BPRC (Rijswijk, The Netherlands) and EUROGENTEC (Seraing, Belgium).

A first objective of IMGT is to be able to cope with the enormous flow of new data. IMGT has standardized the data description and set up a basic semantics which will allow a direct WWW submission of the annotations by the authors. A

second objective is the improvement in the automation of the annotation procedure. To reach that goal, several tools, currently in development, will be available for the users on the WWW interface. Checking redundancy, coherence and integrity of data, with an external international network of experts, insures a high quality in IMGT databases. The model is not limited to any single species and enables the collection and annotation of immunogenetic information for all species.

Methods

Material, languages and general tools

IMGT/LIGM-DB is a relational database managed by the Sybase RDBMS as many other biological databases. Sybase has been chosen for its robustness and its available tools that allow to check and maintain data consistency. It is running on an IBM RISK6000 server. The main user access is a Web interface, which evolves with the HTML language to keep it friendly to use.

Design of the database: a systematized approach

Sub-Systems

Several sub-systems have been defined to take into account the content and the evolution of the IMGT application. These include biological knowledge, research progress, data model evolution, new software tools and computer constraints. More particularly, an Administration sub-system, an Internet Interface sub-system and a Data sub-system were developed to manage, query and update the application.

Modules

The description and management of the biological sequences are included in several modules, which have overlaps with the different sub-systems. Each module comprises:

- biological knowledges and sequence linked data stored in relational tables that are part of the Data sub-system,
- data processing software programmes (see below Data integrity)
- computer and/or human interfaces (human interfaces are included in the Internet Interface sub-system).

As a first example, the **annotation module** has been designed to manage an efficient and complete storage of annotations (keywords, definition and delimitation of sequence subregions) of Ig and TcR sequences and to allow users to make search according to the annotation related criteria. Data analysis was performed by expert annotators from LIGM in order to standardize, define and label all data involved in the sequence annotations. Knowledge tables were then established which contain the list of allowed standardized keywords to describe the Ig or TcR sequences and the definition of all structural and functional subregions that could compose them. 177 feature labels are necessary to cover all cases of combination in Ig and TcR sequences, whereas only 7 of them are available in EMBL [5], GENBANK [6] or DDBJ [7]. LIGM-DB keywords, label list, label definitions and representation are available at URL http://imgt.cnusc.fr: 8104. Annotation of sequences with these labels constitutes the main part of the expertise. Information useful either to check the data consistency or to help annotators is added in these tables. For each subregion, it is indicated if it can be translated, and if so, which is the corresponding translation frame by default. In a second set of tables are recorded the links between the sequences and the data defined in the knowledge tables. These tables are used to allow users to select sequences according to keywords and label subregions.

A second module example, the **gene and allele module,** has been developed by the same approach. The set of knowledge tables contains the IMGT nomenclature for human mapped Ig and TcR gene names and the correspondence with previous existing nomenclatures. For each human gene are recorded the identified alleles with indication of the position of the mutations and their characteristics (if they are silent or if they induce an amino acid change, or if they are deletions or insertions). Functionality, according to IMGT definition, is indicated for each germline gene and its alleles. Each human gene or gene allele is associated to a nucleic reference sequence which has been chosen for its completeness and/or its anterior date of publication. In the second set of tables are recorded the genes which have been identified in the germline, rearranged or cDNA nucleic sequences using alignment tools (see below). These data are provided to users for searches in the database. Step by step, this module will be completed with data from other species once Ig and TcR loci will be mapped and sequenced.

Additional modules are in development for other biological or clinical aspects that are not yet efficiently searched: this is the case for pathologies in which Ig and TcR are involved as leukemia and lymphoma, infectious diseases, AIDS, autoimmune diseases, such as rheumatoid arthritis.... This approach will be particularly useful since there is no way to make consistent search according to disease criteria in generalist sequence databases: the current list of keywords enabling to extract disease-related sequences in IMGT will be completed. Sequence search in diseases, recognition site specificities and links to structural data are already available for part of the IMGT sequences. We propose in the future to provide alignment tables to which a new sequence could be compared and classified. This development will first concern leukemia and lymphoma, autoimmune diseases as rheumatoid arthritis, and anti-HIVspecific sequences.

Results and discussion

Enhancement of IMGT data acquisition: WWW direct submission.

Sequence annotation, a time consuming and an elaborate step, is the limiting factor for the addition of expert data onto the nucleotide sequence information. Moreover, the publication of novel Ig, TcR and MHC sequences continues at an ever increasing pace and the development of automated sequencing techniques suggests that this trend will continue for the foreseeable future. To overcome this major problem, we are developing, in collaboration with EMBL-EBI, a WWW interface for direct submission of the IMGT keywords and annotations by the sequence submitters. According to this view, IMGT data acquisition com-

prises two steps:

Acquisition of core data. For each sequence entry, the core data consist of the sequence data, the citation information (bibliographical references) and the taxonomic data. The acquisition of the core data from EMBL [5] to IMGT is currently done by mail. This procedure will be upgraded by using communication protocols for data transfer between IMGT and ORACLE (EMBL).

Acquisition of annotation data submitted by the authors. This totally new aspect of IMGT will allow the integration of new sequences and annotations directly from the submitting authors. This approach is now possible due to the standardization of IMGT data. The analysis and specification of a Web questionnaire for the authors have been developed and provide the authors with the content of knowledge tables that will guide them to identify and delimit the subregions of their sequences with IMGT labels according to the receptor type (Ig or TcR) and the type of sequence (germline, rearranged DNA, or cDNA). Some context-related help facility will be improved for submission of many similar sequences or submission of short sequences as the junctions between the V-REGION and the J-REGION. Before entering in the database, author submissions will be checked by an expert annotator who will be in charge of contacting the author in case of discrepancy.

Improvement in the automation of the annotation procedure

Currently there is a considerable backlog of sequences waiting to be fully annotated in IMGT/LIGM-DB; this situation is exacerbated by 300 new or updated sequences arriving each week. Clearly the keyword and feature label assigament represents the limiting factor in this procedure, nevertheless IMGT/LIGM-DB standardized keywords have been assigned to all entries (24000), and more than 9.400 are fully annotated. Since August 1996, the content of IMGT/LIGM-DB closely follows the one of EMBL with the advantage of being deleted of sequences wrongly assigned to Ig or TcR.

In order to speed up the annotation procedures, semi-automatic annotation software has been developed (LIGMotif, validation module, DNAPLOT), in collaboration with LIGM, CNUSC and IFG. Since annotations of biological data would never be totally automatic, these programs are used as an aid by the IMGT annotators at LIGM. These tools still need verification by experts but the aim is to reduce the time required for annotations to a minimum. This model of sequence submission to primary databases and control, currently developed by IMGT, can be later extended to other specialized and integrated databases.

LIGMotif: search for conserved motifs in Ig and TcR

As an aid for the annotation procedure, the LIGMotif written in portable C language has been developed. The algorithms used are specific for the search of Ig and TcR patterns in DNA sequences. These algorithms are based on the delimitation rules established after extensive research by LIGM. Motifs of interest for the Leader, Variable, Diversity and Joining regions can be searched for in the germline or rearranged DNA and cDNA. The output is an ASCII file containing features associated to sequences, plus a number of on line information which is useful for the annotator to decide which solution(s) might be correct. New rules are defined as the scientific knowledge of the immunogenetic sequences becomes available. The tool will be integrated into the validation module described below. This will make it easy to use during the annotation process by the annotators: validation module + LIGMotif will become a single integrated tool, available through a single user friendly interface. The basic LIGMotif tool has been extended to include more generalized algorithms to include both DNA and PROTEIN characteristics such as hydrophobicity and similarity. This tool, designated as BIOMOTIF, is described in the EMBL-EBI BioCatalog.at http://www.ebi.ac.uk/biocat/biocat.html

DNAPLOT: software development for nucleotide and protein sequence alignments

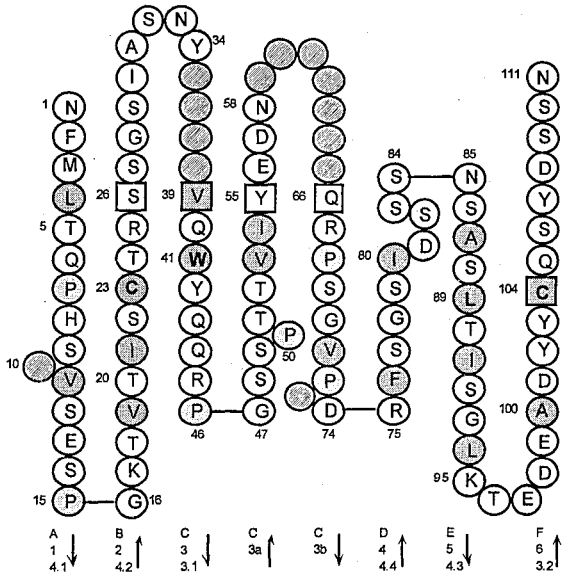
DNAPLOT is an alignment tool, part of IMGT, which uses sets of adequate sequences to build, display, maintain and search nucleotide sequence alignment tables. The programme, developed by IFG, can be downloaded from http://www.genetik.uni-koeln.de/dnaplot/dnaplot.html. A version adapted to Ig and TcR sequences, IMGT/DNAPLOT is available from the IMGT home page http://imgt.cnusc.fr: 8104. The aim is to provide an easy-to-use and a fast tool for research. These alignments are very important for verification of new sequences, their annotation and also in the design of experimental procedures in sequence analysis. IMGT/DNAPLOT currently performs two tasks: (i) the analysis of functional germline variable sequences to identify the gene and/or the subgroups and (ii) the analysis of rearranged variable sequence to identify the V-GENE, the D-SEGMENT (for IgH and TcR beta and delta chains) and J-SEGMENT involved in the rearrangements. Searches can be done related to Ig, TcR and soon to MHC gene alignments using IMGT reference sequence data. In case of Ig and TcR genes, the rearrangements are analysed and dissected into individual sequence blocks. However, DNAPLOT can be applied to all multiple sequence alignments and new search pages will be added. The use of this tool together with LIGMotif and the validation module improves the quality and speeds the incorporation of new sequence data in the IMGT database up.

Innovations in data integrity and IMGT quality

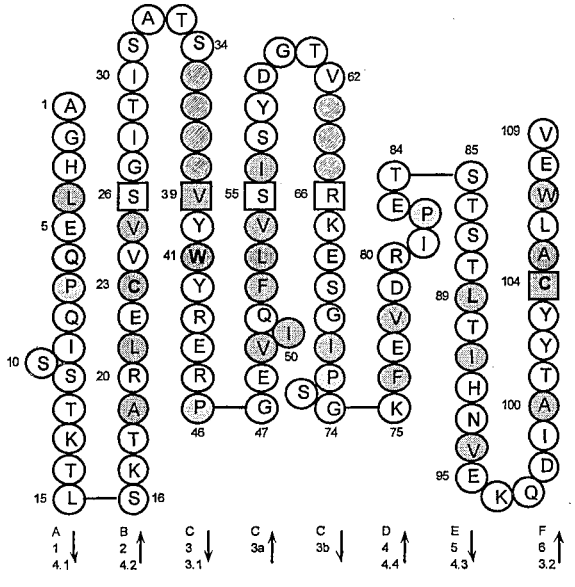
Amino acids are shown in the one-letter abbreviation. Hydrophobic aminoacids (hydropathy index with positive value) and tryptophan (W) found at a given position in more than 50% of analysed Ig and TcR sequences are shown in grey. Arrows indicate the direction of the beta sheets and their different designation in Ig or TcR structure. Hatched circles correspond to missing positions according to IMGT numbering. The CDR-IMGT are limited by aminoacids shown in squares, which belong to neighbouring FR-IMGT.

The IMGT unique numbering.

A uniform numbering system for Ig and TcR sequences of all species has been established by Marie-Paule Lefranc to facilitate sequence comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor



a- human germline IGLV6-57 gene



a- human germline TRGV9 gene

Figure 4 - (IExamples of graphical representation of the IMGT unique numbering for Ig and TcR V-REGION available at the IMGT Marie-Paule page from http://imgt.cnusc.fr: 8104.

a - Human germline IGLV6-57 gene

b- Human germine TRGV9 gene

Amino acids are shown in the one-letter abbreviation. Hydrophobic aminoacids (hydropathy index with positive value) and tryptophan (W) found at a given position in more than 50% of analysed Ig and TcR sequences are shown in grey. Arrows indicate the direction of the beta sheets and their different designation in Ig or TcR structure. Hatched circles correspond to missing positions according to IMGT numbering. The CDR-IMGT are limited by aminoacids shown in squares, which belong to neighbouring FR-IMGT.

(Ig or TcR), the chain type or the species (Figure 1). This numbering results from the analysis of more than 5000 Ig and TcR variable region sequences of vertebrate species from fish to .human. It takes into account and combines the definition of the framework (FR) and complementarity determining region (CDR) [8], structural data from X-ray diffraction studies [9], and the characterization of the hypervariable loops [10]. In the IMGT numbering, conserved amino acids from frameworks always have the same number whatever the Ig or TcR variable sequence, and whatever the species they come from. As example: cysteine 23 (in FR1), tryptophan 41 (in FR2), leucine 89 and cysteine 104 (in FR3). Tables and graphs are available on the WWW interface at the IMGT Marie-Paule page from http:// imgt.cnusc.fr: 8104

Internal cross-references

IMGT contains sequences in different biological forms (e.g. alleles, germline, rearranged or cDNA). These sequences of different biological states are cross-referenced to each other. Tables of germline genes and schematic representations of the human loci are available at the IMGT Marie-Paule page from http://imgt.cnusc.fr: 8104.

Data integrity

Control of data coherence or data integrity has been introduced step by step in LIGM-DB according to the semantic evolution of the data model. Data integrity is important for the database management especially when data input is from more than one site (the core data come from EMBL, the annotations from the authors or from LIGM). It is checked, for example, that each entry is associated to IMGT standardized keywords, and that there is coherence between keywords of an annotated entry and the labels of its subregions. This step is essential to maintain the quality of the database since biological knowledge is always improving. When new rules for describing the data appear, the existing rules are updated accordingly. This allows new entries to be correctly annotated. For the backlog, we implement automatic procedures that regularly verify the coherence of the data and select pools of sequences to be updated. This avoids the introduction of data discrepancies.

IMGT data access and data distribution

One of the major objectives of IMGT was to provide the immunologists with a user friendly interface. The interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge in a computing language. According to this view, the current interface has been developed in WWW client-server architecture (development of interaction WWW-SYBASE) that allows the users to get easily connected from any type of platform (PC, Macintosh, workstation) using freeware such as Netscape. All LIGM-DB information is available through the following search criteria:

- taxonomy, receptor type, functionality, specificity
- sources as genes, clones etc.(currently in development)
- keywords

- LIGM-DB labels
- accession number, mnemonic, definition, length etc.
- bibliographical references

Since July 1995, IMGT/LIGM-DB is currently available on the CNUSC WWW server at Montpellier from **http:// imgt.cnusc.fr: 8104.**

To facilitate the integration of IMGT data into applications developed by other laboratories, we are currently building an Application Programming Interface to access the database and its software tools. This API includes: a set of URL links to access biological knowledge data (keywords, labels, nomenclature,), a set of URL links to access all data related to one given sequence, a set of JAVA™ class packages to select and retrieve data from an appropriate IMGT server using an Object Oriented approach.

IMGT/ LIGM-DB is also available from a number of different sites as part of the SRS-WWW servers (EMBL-EBI UK, INFOBIOGEN France, and EMBnet Nodes). The data from IMGT are distributed to the end user by other established methods already used at the EMBL-EBI. This includes: distribution of CD-ROM, EBI WWW server (http://www.ebi.ac.uk/imgt), and EBI anonymous FTP server: (ftp.ebi.ac.uk).

From January 1996 to December 1997, the IMGT WWW server at Montpellier has been accessed by >15700 sites with an average of 2500 request a week.

Acknowledgements

We thank Chantal Busin (LIGM) for the IMGT WWW interface design and Valérie Barbié (EUROGENTEC) for the development of IMGT/PRIMER-DB.We are deeply grateful to Géraldine Folch, Nathalie Pallarès, Manuel Ruiz and Dominique Scaviner who are the present LIGM-DB annotators. We thank Steven Marsh, Ronald Bontrop and Marc Lemaitre for helpful discussions. IMGT is funded by European Union's BIOMED1 and BIOTECH programmes, the CNRS (Centre National de la Recherche Scientifique) and MENRT (Ministère de l'Education Nationale, de la Recherche et de la Technologie). Subventions have been received from ARC (Association de Recherche sur le Cancer), ARP (Association de Recherche sur la Polyarthrite), FRM (Fondation pour la Recherche Médicale) and the Région Languedoc-Roussillon.

References

[1] Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, Bontrop R, Lemaitre M, Malik A, and Lefranc MP. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research* 1997: 25, 206-211

[2] Lefranc M.-P., Giudicelli V., Busin C., Bodmer J., Müller W., Bontrop R., Lemaitre M., Malik A., and Chaume D. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research* 1998:26 (in press)

[3] Honjo T, and Alt FW. *(eds)* Immunoglobulin genes . Academic Press, 1995; pp.3-443

[4] Lefranc MP. Organization of the human T-cell receptor genes. *Eur. Cytokine Network* 1990: 1, 121-130

[5] Stoesser G, Sterk P, Tuli MA, Stoehr PJ, and Cameron GN. The EMBL Nucleic Sequence Database. *Nucleic Acids Research* 1997: 25, 7-13

[6] Benson DA, Boguski MS, Lipman DJ, and Ostell J. GenBank *Nucleic Acids Research* 1997: 25, 1-6

[7] Tateno Y, and Gojobori T. DNA Data Bank of Japanin age of information. *Nucleic Acids Research* 1997: 25, 14-17

[8] Kabat EA, Wu TT, Perry HM, Gottesman KS, and Foeller C. Sequences of Proteins of Immunological Interest. NIH Publication n.91-3242, National Institutes of Health, Bethesda, US 1991: pp91-3242

[9] Satow Y., Cohen G.H., Padlan E.A., and Davies D.R. Phosphocholine binding immunoglobulin Fab McPC603 *J. Mol. Biol* 1986: 190, 593-604

[10] Chothia C, and Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* 1987: 196:901-917

Address for correspondence

Professor Marie-Paule Lefranc
Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UMR 5535 (CNRS, Université Montpellier II), 1919 route de Mende 34293 Montpellier Cedex 5, France
lefranc@ligm.crbm.cnrs-mop.fr
tel: +33 (0)4 67 61 36 34, Fax: +33 (0)4 67 04 02 31