The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

IMGT, the international ImMunoGeneTics information system®, http://imgt.cines.fr: the reference in immunoinformatics

Marie-Paule Lefranc[a,b], Véronique Giudicelli[a], Chantal Ginestoux[a] and Denys Chaume[a]

[a] *Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier II, UPR CNRS 1142, Institut de Génétique Humaine, IGH, Montpellier, France*
[b] *Institut Universitaire de France*

Abstract

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr), is a high quality integrated information system specializing in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) and related proteins of the immune system of human and other vertebrates, created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM), at the Université Montpellier II, CNRS, Montpellier, France. IMGT is the global reference in immunogenetics and immunoinformatics and provides a common access to standardized data which include nucleotide and protein sequences, oligonucleotide primers, gene maps, genetic polymorphisms, specificities, 2D and 3D structures. IMGT includes three sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB hosted at EBI, IMGT/PRIMER-DB), one genome database (IMGT/GENE-DB), one 3D structure database (IMGT/3Dstructure-DB), Web resources comprising 8000 HTML pages ("IMGT Marie-Paule page") and interactive tools for sequence (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/Allele-Align, IMGT/PhyloGene) and genome (IMGT/GeneSearch, IMGT/GeneView, IMGT/LocusView) analysis. IMGT data are expertly annotated according to the rules of the IMGT Scientific chart, based on the IMGT-ONTOLOGY concepts. IMGT tools are particularly useful for the analysis of the IG and TR repertoires in physiological normal and pathological situations. IMGT has important applications in medical research (repertoire analysis in autoimmune diseases, AIDS, leukemias, lymphomas, myelomas), biotechnology related to antibody engineering (phage displays, combinatorial libraries) and therapeutic approaches (graft, immunotherapy). IMGT is freely available at http://imgt.cines.fr.

Keywords:

Database; IMGT; Ontology; Immunoinformatics; Medical Informatics; Immunoglobulin; T cell receptor; MHC; HLA; Antibody

1. Introduction

The molecular synthesis and genetics of the Immunoglobulin (IG) and T cell Receptor (TR) chains is particularly complex and unique as it includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in human) located on different chromosomes (four in human), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (for review [1,2]). The number of potential protein forms of IG and TR is almost unlimited. Owing to the complexity and high number of published sequences, data control and classification and detailed annotations are a very difficult task for the generalist databanks such as EMBL, GenBank and DDBJ. These observations were the starting point of IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [3] created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM), at the Université Montpellier II, CNRS, Montpellier, France.

IMGT is the global reference in immunogenetics and immunoinformatics. It is a high quality integrated information system, specializing in IG, TR, MHC and related proteins of the immune system of human and other vertebrates, which consists of three sequence databases, one genome database, one 3D structure database, Web resources ("IMGT Marie-Paule page") and interactive tools for sequence and genome analysis. A graphic diagram of the IMGT structure and components is available in "Information on IMGT" at http://imgt.cines.fr. IMGT expertly annotated data are described according to the IMGT-ONTOLOGY concepts [4] and to the rules of the IMGT Scientific chart. IMGT interactive tools are particularly useful for the analysis of the IG and TR repertoires in physiological and pathological situations. By its easy data distribution, IMGT has important implications in medical research (repertoire analysis in autoimmune diseases, AIDS, leukemias, lymphomas, myelomas), biotechnology related to antibody engineering (phage displays, combinatorial libraries) and therapeutic approaches (grafts, immunotherapy). IMGT is freely available at http://imgt.cines.fr.

2. IMGT Databases

IMGT Sequence databases

IMGT/LIGM-DB is a comprehensive database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, created in 1989 by LIGM, Montpellier, France, on the Web since July 1995 [3]. In February 2003, IMGT/LIGM-DB contained 66,909 nucleotide sequences of IG and TR from 105 species. IMGT/LIGM-DB data are provided with a user friendly interface. The Web interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge in a computing language. Selection is displayed at the top of the resulting sequences pages, so the users can check their own queries. Users have the possibility to modify their request or consult the results with a choice of nine possibilities. IMGT/LIGM-DB data are also distributed by anonymous FTP servers at CINES (ftp://ftp.cines.fr/IMGT/) and EBI (ftp://ftp.ebi.ac.uk/pub/databases/imgt/) and from many SRS (Sequence Retrieval System) sites [3]. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI, IGH, INFOBIOGEN, Institut Pasteur, etc.).

IMGT/MHC-DB, hosted at EBI, comprises a database of the human MHC (HLA) allele sequences, developed by Cancer Research and ANRI, UK, on the Web since December 1998 [5], and a database of MHC class II sequences from non human primates (NHP), curated by BPRC, The Netherlands, on the Web since April 2002.

IMGT/PRIMER-DB is an oligonucleotide primer database for IG and TR, developed by LIGM, Montpellier and EUROGENTEC, Belgium, on the Web since July 2002.

IMGT Genome and Structure databases

IMGT/GENE-DB is the first IMGT genome database which allows a search per gene name, created by LIGM, on the Web since February 2003.

IMGT/3Dstructure-DB is a database which provides the IMGT gene and allele identification and Colliers de Perles of IG, TR, MHC and related proteins with known 3D structures, created by LIGM, on the Web since November 2001 [6]. In February 2003, IMGT/3Dstructure-DB contained 596 atomic coordinate files.

3. IMGT Web Resources

IMGT Web resources ("IMGT Marie-Paule page") comprise 8000 HTML pages in the following sections : "IMGT Repertoire", "IMGT Index", "IMGT Scientific chart", "IMGT Bloc-notes", "IMGT Education" and "IMGT Aide-mémoire".

IMGT Repertoire

IMGT Repertoire is the global Web Resource in ImMunoGeneTics for the IG, TR, MHC and related proteins of the immune system of human and other vertebrates, based on the "IMGT Scientific chart" [3]. IMGT Repertoire provides an easy-to-use interface to carefully and expertly annotated data on the genome, proteome, polymorphism and structural data of the IG, TR, MHC and related proteins. Only titles of this large section are quoted here. Genome data include chromosomal localizations, locus representations, locus description, gene tables, lists of genes and links between IMGT, HUGO, GDB, LocusLink and OMIM, correspondence between nomenclatures. Proteome and polymorphism data are represented by protein displays, alignments of alleles, tables of alleles, allotypes. Structural data comprise 2D graphical representations or Colliers de Perles, FR-IMGT and CDR-IMGT lengths, and 3D representations [3,6,7]. This visualization permits rapid correlation between protein sequences and 3D data retrieved from the Protein Data Bank PDB. Other data comprise: (i) phages, (ii) probes used for the analysis of IG and TR gene rearrangements and expression, and Restriction Fragment Length Polymorphism (RFLP) studies, (iii) data related to gene regulation and expression: promoters, primers, cDNAs, reagent monoclonal antibodies, etc., (iv) genes and clinical entities: translocations and inversions, humanized antibodies, monoclonal antibodies with clinical indications, (v) taxonomy of vertebrate species present in IMGT/LIGM-DB, (vi) immunoglobulin superfamily: gene exon-intron organization, protein displays, Colliers de Perles and 3D representations of V-LIKE and C-LIKE domains.

IMGT Index

IMGT Index is a fast way to access data when information has to be retrieved from different parts of the IMGT site [3]. For example, "allele" provides links to the IMGT Scientific chart rules for the allele description, and to the IMGT Repertoire Alignments of alleles and Tables of alleles.

IMGT Scientific chart rules and IMGT-ONTOLOGY concepts

IMGT Scientific chart provides the controlled vocabulary and the annotation rules for data and knowledge management of the IG, TR, MHC and related proteins of the immune system of human and other vertebrates [3,4]. IMGT has developed a formal specification of the terms to be used in the domain of immunogenetics and bioinformatics to ensure accuracy, consistency and coherence in IMGT. This has been the basis of IMGT-

ONTOLOGY [4], the first ontology in the domain, which allows the management of the immunogenetics knowledge for human and other vertebrate species. IMGT Scientific chart rules are based on the five concepts defined in IMGT-ONTOLOGY: IDENTIFICATION, DESCRIPTION, CLASSIFICATION, NUMEROTATION and OBTENTION.

IDENTIFICATION concept: standardized keywords. IMGT standardized keywords for IG, TR and MHC include general keywords, indispensable for the sequence assignments, and specific keywords, more specifically associated to particularities of the sequences or to diseases [3].

DESCRIPTION concept: standardized labels and annotations. 177 feature labels are necessary to describe all structural and functional subregions that compose IG and TR sequences, whereas only seven of them are available in EMBL, GenBank or DDBJ and none in PDB. Annotation of sequences with these labels constitutes the main part of the expertise [3]. Levels of annotation have been defined, which allow the users to query sequences in IMGT/LIGM-DB even though they are not fully annotated. Prototypes represent the organizational relationship between labels and give information on the order and expected length (in number of nucleotides) of the labels [3].

CLASSIFICATION concept: standardized gene nomenclature. The CLASSIFICATION concept has been used to set up a unique nomenclature of human IG and TR genes, which was approved by the Human Genome Organization (HUGO) Nomenclature Committee (HGNC) in 1999 [1,2] and has become the community standard. The complete list of the human IG and TR gene names is available in Genew (UK), the Genome DataBase GDB (Canada), LocusLink at NCBI (USA) and GeneCards (Israël). IMGT reference sequences have been defined for each allele of each gene [1,2].

NUMEROTATION concept: the IMGT unique numbering. A uniform numbering system for IG and TR sequences of all species has been established to facilitate sequence comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor (IG or TR), the chain type, or the species [7]. In the IMGT numbering, conserved amino acids from frameworks always have the same number whatever the IG or TR variable sequence, and whatever the species they come from. As examples: Cysteine 23 (in FR1), Tryptophan 41 (in FR2), Leucine 89 and Cysteine 104 (in FR3) [1,2]. The IMGT unique numbering represents a big step forward in the analysis of the IG and TR sequences of all vertebrate species. It has allowed (i) a standardized description of the allele polymorphisms [1,2] and of the IG somatic hypermutations, and (ii) the redefinition of the limits of the FR and CDR of the IG and TR variable domains [6]. The FR-IMGT and CDR-IMGT lengths become in themselves crucial information which characterize variable regions belonging to a group, a subgroup and/or a gene [7]. Moreover, it gives insight into the structural configuration of the domains and opens interesting views on the evolution of these sequences, since this numbering has been applied with success to all the sequences belonging to the V-set and C-set of the immunoglobulin superfamily [7].

OBTENTION concept. The OBTENTION concept is a set of standardized terms that precise the origins of the sequence (the 'origin concept') and the conditions in which the sequences were obtained (the 'methodology concept').

Other IMGT Web sections

IMGT Bloc-notes provides numerous hyperlinks towards the Web servers specializing in immunology, genetics, molecular biology and bioinformatics (associations, collections, companies, databases, immunology themes, journals, molecular biology servers, resources, societies, tools, etc.) [3]. IMGT Education is a section which provides useful biological resources for students. It includes figures and tutorials (in english and/or in french) on the IG and TR variable and constant domain 3D structures, the molecular genetics of immunoglobulins, the regulation of IG gene transcription, B cell differentiation and activation, etc. IMGT Aide-mémoire provides useful information such as genetic code, splicing sites, amino acid structures, restriction enzymes sites, etc.

4. IMGT Interactive Tools

IMGT/V-QUEST

IMGT/V-QUEST (V-QUEry and STandardization) is an integrated software for IG and TR. This tool, easy to use, analyses an input IG or TR germline or rearranged variable nucleotide sequences [3]. IMGT/V-QUEST results comprise the identification of the V, D and J genes and alleles and the nucleotide alignment by comparison with sequences from the IMGT reference directory, the delimitations of the FR-IMGT and CDR-IMGT based on the IMGT unique numbering, the protein translation of the input sequence, the identification of the JUNCTION and the V-REGION Collier de Perles. The set of sequences from the IMGT reference directory, used for IMGT/V-QUEST, can be downloaded in FASTA format from the IMGT site.

IMGT/JunctionAnalysis

IMGT/JunctionAnalysis is a tool, complementary to IMGT/V-QUEST, which provides a thorough analysis of the V-J and V-D-J junctions of IG and TR rearranged genes [3]. IMGT/JunctionAnalysis identifies the D-GENE and allele involved in the IGH, TRB and TRD V-D-J rearrangements by comparison with the IMGT reference directory, and delimits precisely the P, N and D regions. Results from IMGT/JunctionAnalysis are more accurate than those given by IMGT/V-QUEST regarding the D-GENE identification. Indeed, IMGT/JunctionAnalysis works on shorter sequences (JUNCTION), and with a higher constraint since the identification of the V-GENE and J-GENE and alleles is a prerequisite to perform the analysis. Several hundreds of junction sequences can be analysed simultaneously.

Other IMGT sequence and genome analysis tools

IMGT/Allele-Align allows the comparison of two alleles highlighting the nucleotide and amino acid differences. IMGT/PhyloGene is an easy to use tool for phylogenetic analysis of IMGT standardized reference sequences. IMGT/GeneSearch, IMGT/GeneView and IMGT/LocusView are tools providing an interactive interface for genes and loci of human IG, TR and MHC and mouse TRA/TRD.

5. IMGT Web access

Since July 1995, IMGT has been available on the Web at http://imgt.cines.fr. IMGT provides the biologists with an easy to use and friendly interface. Since January 2000, the IMGT WWW Server at Montpellier was accessed by more than 180,000 sites. IMGT has an exceptional response with more than 120,000 requests a month. Two thirds of the visitors are equally distributed between the European Union and the United States. To facilitate the integration of IMGT data into applications developed by other laboratories, we have built an Application Programming Interface (API) to access the database and its software tools (see "IMGT Informatics page" at http://imgt.cines.fr). This API includes: a set of URL links to access biological knowledge data (keywords, labels, functionalities, list of gene names, etc.), a set of URL links to access all data related to one given sequence, a set of JAVA^{TM} class packages to select and retrieve data from an appropriate IMGT server using an Object Oriented approach.

IMGT distributes high quality data with an important incremental value added by the IMGT expert annotations, according to the rules described in the IMGT Scientific chart. Control of coherence in IMGT combines data integrity control and biological data evaluation.

The information provided by IMGT is of much value to clinicians and biological scientists in general [3]. IMGT is designed to allow a common access to all immunogenetics data, and a particular attention is given to the establishment of cross-referencing links to other databases pertinent to the users of IMGT.

6. Citing IMGT

Users of IMGT are encouraged to cite [3] and to quote the IMGT home page URL, http://imgt.cines.fr when referring to IMGT in a publication.

7. Acknowledgments

We thank Gérard Lefranc for helpful discussion and Valérie Thouvenin-Contet for help in the preparation of the manuscript. We are deeply grateful to the IMGT team for its expertise and constant motivation and specially to our curators for their hard work and enthusiasm. IMGT is a registered mark of Centre National de la Recherche Scientifique (CNRS). IMGT is funded by the European Commission 5th PCRDT programme (QLG2-2000-01287), the CNRS, the Ministère de l'Education Nationale et de la Recherche.

8. References

[1] Lefranc M-P, and Lefranc G. *The Immunoglobulin FactsBook.* Academic Press, London, UK, ISBN:012441351X, 458 pages, 2001.

[2] Lefranc,M.-P. and Lefranc,G. *The T cell receptor FactsBook.* Academic Press, London, UK, ISBN:0124413528, 398 pages, 2001.

[3] Lefranc M-P. IMGT, the international ImMunoGeneTics database. *Nucl Acids Res* 2003: 31: 307-310.

[4] Giudicelli V, and Lefranc M-P. Ontology for Immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* 1999: 12: 1047-1054.

[5] Robinson J, Malik A, Parham P, Bodmer JG, and Marsh SGE. IMGT/HLA Database - a sequence database for the human major histocompatibility complex. *Tissue Antigens* 2000: 55: 280-287.

[6] Ruiz M, and Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics DOI 10.1007/s00251-001-0408-6. *Immunogenetics* 2002: 53: 857-883.

[7] Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, and Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 2002: 27: 55-77.

9. Address for correspondence

Marie-Paule Lefranc, IMGT, the international ImMunoGeneTics information system®, LIGM, UPR CNRS 1142, IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France;
Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: lefranc@ligm.igh.cnrs.fr; IMGT, http://imgt.cines.fr