Nucleic Acids Research, 2003, Vol. 31, No. 1 **307–310**
DOI: 10.1093/nar/gkg085

IMGT, the international ImMunoGeneTics database®

Marie-Paule Lefranc*

Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine, Montpellier, Institut Universitaire de France, France

Received October 21, 2002; Accepted October 22, 2002

ABSTRACT

The international ImMunoGeneTics database® (IMGT) (http://imgt.cines.fr), is a high quality integrated information system specializing in Immunoglobulins (IG), T cell Receptors (TR) and Major Histocompatibility Complex (MHC) of human and other vertebrates, created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM), at the Université Montpellier II, CNRS, Montpellier, France. IMGT 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, IMGT/PRIMER-DB), one genome database (IMGT/GENE-DB) with different interfaces (IMGT/GeneSearch, IMGT/GeneView, IMGT/LocusView), one 3D structure database (IMGT/3Dstructure-DB), Web resources comprising 8000 HTML pages ('IMGT Marie-Paule page') and interactive tools for sequence analysis (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/Allele-Align, IMGT/PhyloGene). IMGT data are expertly annotated according to the rules of the IMGT Scientific chart, based on IMGT-ONTOLOGY. 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 (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.

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 the international ImMunoGeneTics database®, (IMGT) (http://imgt.cines.fr) (3–10) created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM), at the Université Montpellier II, CNRS, Montpellier, France.

IMGT is a high quality integrated information system specializing in IG, TR and MHC of human and other vertebrates which consists of three sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB, IMGT/PRIMER-DB), one genome database (IMGT-GENE-DB) with different interfaces (IMGT/GeneSearch, IMGT/GeneView, IMGT/LocusView), one 3D structure database (IMGT/3Dstructure-DB), Web resources ('IMGT Marie-Paule page') and interactive tools for sequence analysis (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/Allele-Align, IMGT/PhyloGene). IMGT expertly annotated data and 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 (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.

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 (3–8). In September 2002, IMGT/LIGM-DB contained 63 566 nt sequences of IG and TR from 105 species (9,10). IMGT/

*IMGT, the international ImMunoGeneTics database, Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France. Tel: +33 499619965; Fax: +33 499619901; Email: lefranc@ligm.igh.cnrs.fr

LIGM-DB data are provided with a user friendly interface (3). 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 (5). 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 Sequence Retrieval System (SRS) sites (3). IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI, IGH, INFOBIOGEN, Institut Pasteur, etc.).

IMGT/MHC-DB comprises a database of the human MHC allele sequences (IMGT/MHC-HLA, developed by Cancer Research and ANRI, London, UK, on the Web since December 1998) (11), databases of MHC class II sequences from non human primates (IMGT/MHC-NHP, curated by BPRC, The Netherlands) and from felines and canines (IMGT/MHC-FLA and IMGT/MHC-DLA, curated by the Faculty of Veterinary Science, Liverpool, UK), 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.

IMGT genome and structure databases

IMGT/GENE-DB is a database which allows a search per gene name. The interactive interfaces of IMGT/GeneSearch, IMGT/GeneView and IMGT/LocusView are available for the human IG, TR and MHC genes and loci and for the mouse TRA/TRD genes and locus.

IMGT/3Dstructure-DB is a database which provides the IMGT gene and allele identification and Colliers de Perles of IG, TR and MHC with known 3D structures, created by LIGM, on the Web since November 2001 (12). In September 2002, IMGT/3Dstructure-DB contained 596 atomic coordinate files.

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 and MHC of human and other vertebrates, based on the ‘IMGT Scientific chart’. 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 and MHC (5,8). 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 (4–6,8,12).

IMGT Index

IMGT Index is a fast way to access data when information has to be retrieved from different parts of the IMGT site (8).

IMGT Scientific chart

IMGT Scientific chart provides the controlled vocabulary and the annotation rules and concepts defined by IMGT for the identification, the description, the classification and the numerotation of the IG and TR data of human and other vertebrates (5,8,13).

Concept of identification: standardized keywords. IMGT standardized keywords for IG and TR include general keywords, indispensable for the sequence assignments, and specific keywords, more specifically associated to particularities of the sequences or to diseases (3).

Concept of description: standardized sequence annotation. 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. Annotation of sequences with these labels constitutes the main part of the expertise (3).

Concept of classification: standardized IG and TR gene nomenclature. The objective is to provide immunologists and geneticists with a standardized nomenclature per locus and per species which will allow extraction and comparison of data for the complex B and T cell antigen receptor molecules. The concepts of classification have been used to set up a unique nomenclature of human IG and TR genes, which was approved by Human Genome Nomenclature Committee (HGNC), the Human Genome Organization (HUGO) in 1999 (1,2). The complete list of the human IG and TR gene names was entered by the IMGT Nomenclature Committee in the Genome DataBase (GDB), Canada, and in LocusLink at NCBI, USA, and is available from the IMGT site (1,2). IMGT reference sequences have been defined for each allele of each gene based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence, mapped sequence (5). They are listed in the germline gene tables of the IMGT Repertoire. The protein displays show translated sequences of the alleles (*01) of the functional or ORF genes (1,2).

Concept of numerotation: 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 (14). 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 (1,2,4,5) 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. 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 (14). 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 (14).

Other IMGT web sections

IMGT Bloc-notes provides numerous hyperlinks towards the Web servers specializing in immunology, genetics, molecular biology and bioinformatics (15). IMGT Education and IMGT Aide-mémoire provide useful information for students (figures, tutorials).

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 (8,16). 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 (16). 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. Several hundreds of junction sequences can be analysed simultaneously.

Other IMGT 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-ONTOLOGY AND IMGT INTEROPERABILITY

IMGT-ONTOLOGY

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. 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 (13), the first ontology in the domain, which allows the management of the immunogenetics knowledge for all vertebrate species. Control of coherence in IMGT combines data integrity control and biological data evaluation.

IMGT interoperability

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 160 000 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.

CONCLUSION

The information provided by IMGT is of much value to clinicians and biological scientists in general (9,10). IMGT is designed to allow a common access to all immunogenetics data, and a particular attention is given to the establisment of cross-referencing links to other databases pertinent to the users of IMGT.

CITING IMGT

Authors who make use of the information provided by IMGT should cite this article as a general reference for the access to and content of IMGT, and quote the IMGT home page URL, http://imgt.cines.fr.

ACKNOWLEDGEMENTS

IMGT is funded by the European Union's 5th PCRDT programme (QLG2-2000-01287), the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Education Nationale et de la Recherche.

REFERENCES

1. Lefranc,M.-P. and Lefranc,G. (2001) *The Immunoglobulin FactsBook*. Academic Press, London, UK, p. 458.
2. Lefranc,M.-P. and Lefranc,G. (2001) *The T cell receptor FactsBook*. Academic Press, London, UK, p. 398.
3. Giudicelli,V., Chaume,D., Bodmer,J., Müller,W., Busin,C., Marsh,S., Bontrop,R., Lemaitre,M., Malik,A. and Lefranc,M.-P. (1997) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **25**, 206–211.
4. Lefranc,M.-P., Giudicelli,V., Busin,C., Bodmer,J., Müller,W., Bontrop,R., Lemaitre,M., Malik,A. and Chaume,D. (1998) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **26**, 297–303.
5. Lefranc,M.-P., Giudicelli,V., Ginestoux,C., Bodmer,J., Müller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbié,V. and Chaume,D. (1999) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **27**, 209–212.
6. Ruiz,M., Giudicelli,V., Ginestoux,C., Stoehr,P., Robinson,J., Bodmer,J., Marsh,S.G., Bontrop,R., Lemaitre,M., Lefranc,G., Chaume,D. and Lefranc,M.-P. (2000) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **28**, 219–221.

7. Lefranc,M.-P. (2000) IMGT ImMunoGeneTics database. *International Bioforum*, **4**, 98–100.
8. Lefranc,M.-P. (2001) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **29**, 207–209.
9. Lefranc,M.-P. (2002) IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. *Dev. Comp. Immunol.*, **26**, 697–705.
10. Lefranc,M.-P. (2003) IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis, http://imgt.cines.fr. *Leukemia*, in press.
11. Robinson,J., Malik,A., Parham,P., Bodmer,J.G. and Marsh,S.G.E. (2000) IMGT/HLA Database—a sequence database for the human major histocompatibility complex. *Tissue Antigens*, **55**, 280–287.
12. Ruiz,M. and Lefranc,M.-P. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, **53**, 857–883.
13. Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for Immunogenetics: IMGT-ONTOLOGY. *Bioinformatics*, **12**, 1047–1054.
14. Lefranc,M.-P., Pommié,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2002) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, in press.
15. Lefranc,M.-P. (2000) Web sites of Interest to Immunologists. *Current Protocols in Immunology.* J. Wiley and Sons, New York, USA, A.1J.1–A.1J.33.
16. Lefranc,M.-P. (2003) IMGT, the international ImMunoGeneTics database®, http://imgt.cines.fr. In Lo,B.K.C. (ed.), *Antibody Engineering Protocols*, 2nd Edn. Human Press, Totowa NJ, USA, in press.