IMGT, the international ImMunoGeneTics database

Marie-Paule Lefranc*, Véronique Giudicelli, Chantal Ginestoux, Julia Bodmer[1], Werner Müller[2], Ronald Bontrop[3], Marc Lemaitre[4], Ansar Malik[5], Valérie Barbié and Denys Chaume[6]

Laboratoire d'ImmunoGénétique Moléculaire, Université Montpellier II, UPR CNRS 1142 IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France, [1]ICRF, Oxford, UK, [2]IFG, Köln, Germany, [3]BPRC, Rijswijk, The Netherlands, [4]EUROGENTEC S.A., Seraing, Belgium, [5]EMBL Outstation EBI, Hinxton, UK and [6]CNUSC, Montpellier, France

Received September 25, 1998; Revised and Accepted October 7, 1998

ABSTRACT

IMGT, the international ImMunoGeneTics database (http://imgt.cnusc.fr:8104), is a high-quality integrated database specialising in Immunoglobulins (Ig), T cell Receptors (TcR) and Major Histocompatibility Complex (MHC) molecules of all vertebrate species, created in 1989 by Marie-Paule Lefranc, Université Montpellier II, CNRS, Montpellier, France (lefranc@ligm.igh.cnrs.fr). IMGT comprises three databases: LIGM-DB, a comprehensive database of Ig and TcR, MHC/HLA-DB, and PRIMER-DB (the last two in development); a tool, IMGT/DNAPLOT, developed for sequence analysis and alignments; and expertised data based on the IMGT scientific chart, the IMGT repertoire. By its high quality and its easy data distribution, IMGT has important implications in medical research (repertoire in autoimmune diseases, AIDS, leukemias, lymphomas), therapeutic approaches (antibody engineering), genome diversity and genome evolution studies. IMGT is freely available at http://imgt.cnusc.fr:8104

INTRODUCTION

The international ImMunoGeneTics (IMGT) database (1), created in 1989 by Marie-Paule Lefranc (Université Montpellier II, CNRS, Montpellier, France; lefranc@ligm.igh.cnrs.fr), is a high quality integrated database specialising in Immunoglobulins (Ig), T cell Receptors (TcR) and Major Histocompatibility Complex (MHC) molecules of all vertebrate species (IMGT home page at http://imgt.cnusc.fr:8104). IMGT comprises alignment tables and expertly annotated sequences, and consists of three databases: (i) LIGM-DB, a comprehensive database of Ig and TcR from human and other vertebrates, (ii) MHC/HLA-DB, and (iii) PRIMER-DB (an Ig, TcR and MHC-related primer database), the two latter are currently being developed.

In order to provide the highest data quality, rules have been set up by LIGM, which constitute the IMGT scientific chart (2). Based on these rules, the IMGT repertoire has become the reference in immunogenetics, by providing expertised data such as Ig and TcR germline gene tables, alignments of alleles, protein displays, Colliers de Perles, 3D representations, etc. IMGT scientific chart and IMGT repertoire are freely available at the IMGT Marie-Paule page from http://imgt.cnusc.fr:8104

IMGT/LIGM-DB proposes two new access media: the first one is a set of URLs to get direct links to the IMGT server, the second one is an IMGT/API that allows Java™ programmers to remotely access and integrate LIGM-DB data in other computer environments.

IMGT/LIGM-DB ORGANIZATION AND CONTENT

LIGM-DB development is mainly based on a relational model organization. The database is maintained with SYBASE as relational DBSM (Data Base System Manager).

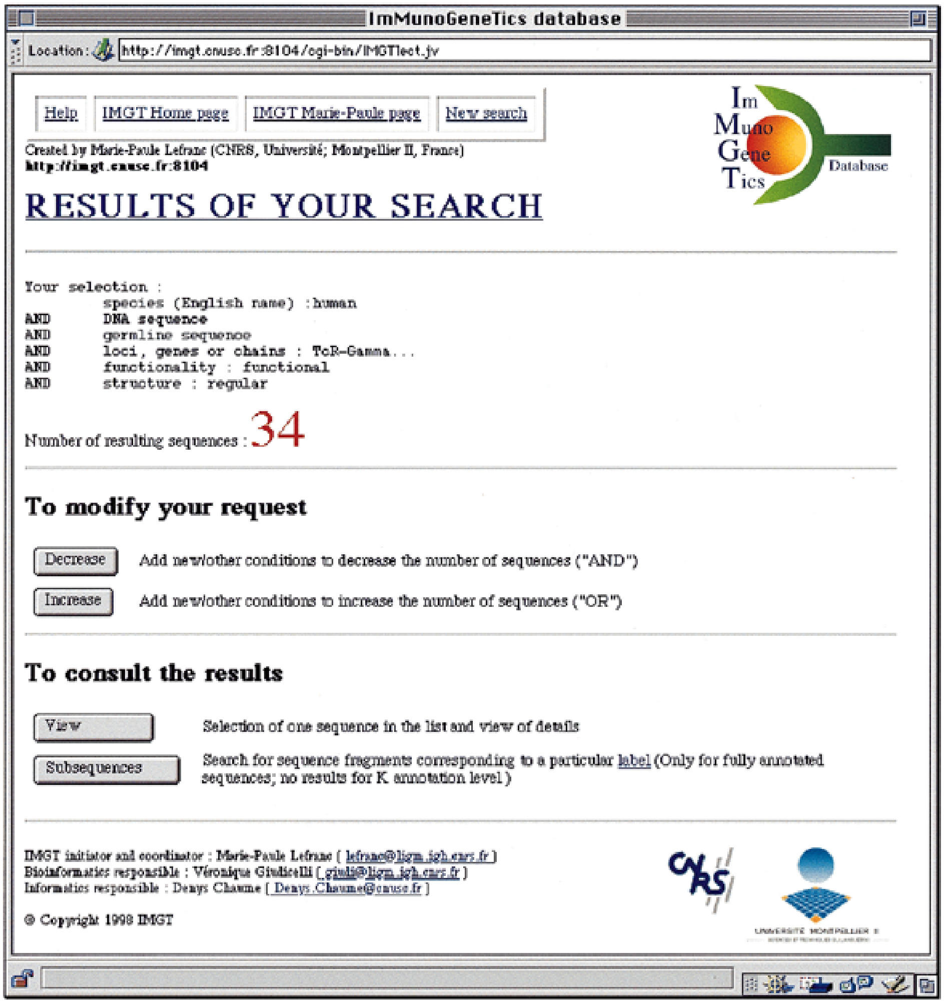
In September 1998, LIGM-DB contained 27 125 nucleic acid sequences of Ig or TcR from 81 species, and translation for fully annotated sequences. IMGT sequences are identified by the EMBL/GenBank/DDBJ (3–5) accession number. All LIGM-DB information is available through the following search criteria: catalogue, accession number, mnemonic, definition, length etc.; taxonomy, nucleic acid type, loci, genes or chains, functionality, structure, specificity, etc.; keywords; annotation labels; references. Selection is displayed at the top of the resulting sequences pages, so the users can check their own queries (Fig. 1). Users have the possibility to modify their request or consult the results (Fig. 1). They can (i) add new conditions to increase or decrease the number of resulting sequences, (ii) view details concerning the selected sequences and choose among eight possibilities: IMGT annotations, IGMT flat file, coding regions with protein translation, catalogue and external references, sequence in dump format, sequence in FASTA format, sequence with three reading frames, EMBL flat file, or (iii) search for sequence fragments corresponding to a particular label (Fig. 1).

IMGT SCIENTIFIC CHART AND REPERTOIRE FOR DATA INTEGRITY AND IMGT QUALITY

IMGT standardized keywords, labels and prototypes

IMGT standardized keywords for Ig and TcR have been assigned to all entries. 177 feature labels are necessary to describe all

*To whom correspondence should be addressed. Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: lefranc@ligm.igh.cnrs.fr



Created by Marie-Paule Lefranc (CNRS, Université; Montpellier II, France)
http://imgt.cnusc.fr:8104

RESULTS OF YOUR SEARCH

```
Your selection :
          species (English name) :human
AND       DNA sequence
AND       germline sequence
AND       loci, genes or chains : TcR-Gamma...
AND       functionality : functional
AND       structure : regular
```

Number of resulting sequences : 34

To modify your request

Decrease Add new/other conditions to decrease the number of sequences ("AND")

Increase Add new/other conditions to increase the number of sequences ("OR")

To consult the results

View Selection of one sequence in the list and view of details

Subsequences Search for sequence fragments corresponding to a particular label (Only for fully annotated sequences; no results for K annotation level)

IMGT initiator and coordinator : Marie-Paule Lefranc (lefranc@ligm.igh.cnrs.fr)
Bioinformatics responsible : Véronique Giudicelli (giudi@ligm.igh.cnrs.fr)
Informatics responsible : Denys Chaume (Denys.Chaume@cnusc.fr)

© Copyright 1998 IMGT

Figure 1. Example of a ‘Results’ screen from IMGT/LIGM-DB, http://imgt.cnusc.fr:8104 . Note the search selection at the top of the page and the possibility, for the users, either to modify the request or to consult directly the results.

structural and functional subregions that compose Ig and TcR 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 expertize. Knowledge tables have been established to record and standardize theoretical and experimental research. Their content is under the responsibility of the IMGT coordinator.

Prototypes represent the organizational relationship between labels (6) and give information on the order and the expected length (in number of nucleotides) of the labels (7). Prototypes can apply to general configuration of Ig or TcR, independently of the chain type, the species or any other parameters like functionality. However, prototypes may also be established for very precise cases when sequence characteristics are clearly established (7).

IMGT reference sequences

IMGT reference sequences have been defined based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence, mapped sequence. They are listed in the germline gene tables of the IMGT repertoire (8–10). IMGT reference sequences are crucial for the high quality of IMGT/DNAPLOT results (1,11), available from http://imgt.cnusc.fr:8104 . The set of sequence fragments used for the IMGT/DNAPLOT tool can be downloaded in FASTA format from the IMGT repertoire at the IMGT Marie-Paule page from http://imgt.cnusc.fr:8104

IMGT gene name nomenclature

The objective is to provide immunologists and geneticists with a unique nomenclature per locus which will allow extraction and comparison of data for the complex B and T cell antigen receptor molecules, whatever the species. IMGT nomenclature for Ig and TcR genes of all species follows the Human Gene Mapping Nomenclature rules. This has been applied as early as 1988, for the human IGL and IGH loci (12,13), and 1989 for all the genes of the human TRG locus (14,15). Correspondence between nomenclatures are described in tables. An exhaustive and standardized list of human Ig and TcR gene names is available from the IMGT repertoire at the IMGT Marie-Paule page.

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 antigene receptor (Ig or TcR), the chain type or the species (1,16). In the IMGT unique numbering, conserved amino acids from frameworks (FR) always have the same number whatever the Ig or TcR 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). Correspondence between numberings is available at the IMGT Marie-Paule page.

FR-IMGT and CDR-IMGT regions

The IMGT unique numbering has allowed to redefine the limits of the framework (FR) and complementary determining regions (CDR). 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. Framework amino acids (and codons) located at the same position in different sequences can be compared without requiring sequence alignments. This also holds for amino acids belonging to CDR-IMGT of same length. Tables of FR-IMGT and CDR-IMGT lengths are available from the IMGT repertoire (http://imgt.cnusc.fr:8104).

IMGT mutation and allele polymorphism description

The IMGT unique numbering has allowed a standardized IMGT description of mutations and the description of allele polymorphisms and somatic hypermutations. These mutations and allelic polymorphisms are described by comparison to the germline IMGT reference sequence (allele *01). Based on these criteria, alignments of alleles (1) and tables of alleles (8–10) have been set up for the coding region of the germline genes and are available from the IMGT repertoire.

Protein displays, Colliers de Perles and 3D representations

Protein displays, and 2D graphical representations designated as Colliers de Perles (1) are provided for all the human germline variable regions of Ig and TcR, with FR-IMGT and CDR-IMGT delimitations. Conserved hydrophobic amino acids are highlighted. The most recent 2D representations available at the IMGT WWW interface from http://imgt.cnusc.fr:8104 were generated with the IMGT/Colliers de Perles tool, developed by Gérard Mennessier (LPM, Montpellier, France). Note that this IMGT/Collier de Perles representation is also of great interest for all sequences belonging to the V-set of the Ig superfamily including non-rearranging sequences in vertebrates (Xenopus CTXg1, human CD4, etc.) and in invertebrates (drosophila amalgam, drosophila fasciclin II, etc.) (1,6). A standardization of 3D representations of Ig and TcR rearranged variable regions, in which CDR loops are delimited according to the IMGT numbering is being developed. Examples are available at the IMGT Marie-Paule page.

INNOVATIONS IN DATA COHERENCE AND DATA DISTRIBUTION

IMGT data coherence

Control of coherence in IMGT combines data integrity control and biological data evaluation (7).

Data integrity control. An administration system has been set up to coordinate the entry flow (7). This ensures coherence of the IMGT data which come from multiple sources: core data from EMBL, annotations from LIGM experts, and annotations from direct submissions by the authors.

Biological data evaluation. Knowledge tables have been built to record controled vocabulary and rules defined in the IMGT scientific chart (7). Knowledge tables allow (i) to control data coherence with the IMGT rules and prototypes, (ii) to identify potential sources of incoherences, (iii) to propose solutions in case of errors, and (iv) to take into account the evolution of knowledge issued from the immunogenetic research.

An ontology for Immunogenetics is currently built by IMGT in order to provide a semantic repository which will be of great help to increase interoperability between specialist and generalist databases.

IMGT data distribution

Since July 1995, IMGT/LIGM-DB has been available on the WWW server of CNUSC Montpellier at the IMGT home page http://imgt.cnusc.fr: 8104 . IMGT provides the immunologists with an easy to use and friendly interface.

From January 1996 to October 1998, IMGT WWW server at Montpellier was accessed by more than 30 800 sites, with an average of 5000–5500 requests a week. IMGT data are also distributed by the CNUSC anonymous FTP server (ftp://imgt.cnusc.fr/pub/IMGT), by EBI (distribution of CD-ROM, network fileserver: netserv@ebi.ac.uk, and anonymous FTP server: ftp.ebi.ac.uk). IMGT/ LIGM-DB is available from many SRS sites.

To facilitate the integration of IMGT data into applications developed by other laboratories, we have built an Application Programming Interface to access the database and its software tools (7). This API includes: a set of URL links to access biological knowledge data (keywords, labels, nomenclature, etc.), a set of URL links to access all data related to one given sequence. A set of Java™ class packages allows application developers to select and retrieve data from an appropriate IMGT server using an Object Oriented approach and the Java/RMI protocol in a CORBA like architecture (7). The complete list of information can be found at the IMGT informatics page from http://imgt.cnusc.fr:8104

ELECTRONIC AND MAILING ADDRESSES

IMGT home page: http://imgt.cnusc.fr:8104 (IMGT contact lefranc@ligm.igh.cnrs.fr).

Anonymous FTP servers: ftp://imgt.cnusc.fr/pub/IMGT (contact Denys.Chaume@cnusc.fr), ftp.ebi.ac.uk (contact malik@ebi.ac.uk).

IMGT Initiator and Coordinator: Marie-Paule Lefranc, IMGT, the International ImMunoGeneTics database, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142,

IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France. Tel: +33 (0)4 99 61 99 65; Fax: +33 (0)4 99 61 99 01; Email: lefranc@ligm.igh.cnrs.fr

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.cnusc.fr:8104

ACKNOWLEDGEMENTS

We are deeply grateful to Gérard Lefranc for helpful discussion, Gérard Mennessier (LIGMotif and IGMT/Colliers de Perles), Hans-Helmar Althaus (DNAPLOT), Johanne Abad, Sylvaine Artero, Valérie Contet, Géraldine Folch, Christèle Martinez, Manuel Ruiz and Dominique Scaviner (LIGM-DB), James Robinson, Steven Marsh and Natasja de Groot (MHC/HLA-DB). IMGT is funded by the European Union's BIOTECH programme BIO4CT96-0037, the CNRS (Centre National de la Recherche Scientifique), and the MENRT (Ministère de l'Education Nationale, de la Recherche et de la Technologie). Subventions have been received from ARC (Association pour la Recherche sur le Cancer), ARP (Association de Recherche sur la Polyarthrite), FRM (Fondation pour la Recherche Médicale), Ligue Nationale contre le Cancer, and the Région Languedoc-Roussillon.

REFERENCES

1 Lefranc,M.-P., Giudicelli,V., Busin,C., Bodmer,J., Müller,W., Bontrop,R., Lemaitre,M., Malik,A. and Chaume,D. (1998) *Nucleic Acids Res*., **26**, 297–303.
2 Lefranc,M.-P. (1998) *Exp. Clin. Immunogenet*., **15**, 1–7.
3 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res*., **26**, 8–15.
4 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F.F. (1998) *Nucleic Acids Res*., **26**, 1–7.
5 Tateno,Y., Fukami-Kobayashi,K., Miyazaki,S., Sugawara,H. and Gojobori,T. (1998) *Nucleic Acids Res*., **26**, 16–20.
6 Giudicelli,V., Chaume,D., Bodmer,J., Müller,W., Busin,C., Marsh,S., Bontrop,R., Lemaitre,M., Malik,A. and Lefranc,M.-P. (1997) *Nucleic Acids Res*., **25**, 206–211.
7 Giudicelli,V., Chaume,D. and Lefranc,M.-P. (1998) Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, ISBM-98, 59–68.
8 Pallarès,N., Frippiat,J.-P., Giudicelli,V. and Lefranc,M.-P. (1998) *Exp. Clin. Immunogenet*., **15**, 8–18.
9 Barbié,V. and Lefranc,M.-P. (1998) *Exp. Clin. Immunogenet*., in press.
10 Martinez,C. and Lefranc,M.-P. (1998) *Exp. Clin. Immunogenet*., in press.
11 Giudicelli,V., Chaume,D., Mennessier,G., Althaus,H.H., Müller,W., Bodmer,J., Malik,A. and Lefranc,M.-P. (1998) Proceedings of the Ninth World Congress on Medical Informatics, MEDINFO' 98, 351–355.
12 Ghanem,N., Dariavach,P., Bensmana,M., Chibani,J., Lefranc,G. and Lefranc,M.-P. (1988) *Exp. Clin. Immunogenet*., **5**, 186–195.
13 Bensmana,M., Huck,S., Lefranc,G. and Lefranc,M.-P. (1988) *Nucleic Acids Res*., **16**, 3108.
14 Lefranc,M.-P. and Rabbitts,T.H. (1989) *Trends Biochem. Sci*., **14**, 214–218.
15 Lefranc,M.-P. and Rabbitts,T.H. (1990) *Res. Immunol*., **141**, 615–618.
16 Lefranc,M.-P. (1997) *Immunol. Today*, **18**, 509.