IMGT® databases (Marie-Paule LEFRANC)

http://www.imgt.org

ELIXIR survey 2008

IMGT/LIGM-DB

13. Please provide a short description of unique content

IMGT/LIGM-DB is the IMGT® comprehensive database of immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences from human and other vertebrate species. It was created in 1989 by Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire LIGM (CNRS and Université Montpellier 2), Montpellier, France and is the oldest and the largest database of IMGT®. IMGT/LIGM-DB includes all germline (non-rearranged) and rearranged IG and TR genomic DNA (gDNA) and complementary DNA (cDNA) sequences published in generalist databases. IMGT/LIGM-DB allows searches from the Web interface according to biological and immunogenetic criteria through five distinct modules depending on the user interest. For a given entry, nine types of display are available including the IMGT flat file, the translation of the coding regions and the analysis by the IMGT/V-QUEST tool. IMGT/LIGM-DB distributes expertly annotated sequences. The annotations hugely enhance the quality and the accuracy of the distributed detailed information. They include the sequence identification, the gene and allele classification, the constitutive and specific motif description, the codon and amino acid numbering, and the sequence obtaining information, according to the main concepts of IMGT-ONTOLOGY. They represent the main source of IG and TR gene and allele knowledge stored in IMGT/GENE-DB and in the IMGT reference directory.

14. Please provide a short description of biological utility

Owing to the complexity of the IG and TR molecular genetics that is unique to the vertebrate genomes, IMGT/LIGM-DB has to deal with (i) large germline (non-rearranged) genomic DNA (gDNA) sequences, which may involve a complete locus from several hundred kilobases to one (or more) megabase(s); (ii) rearranged gDNA sequences resulting from the recombination of V (variable), D (diversity) and J (joining) genes (V-J genes and V-D-J genes); and (iii) rearranged V-J-C (constant) and V-D-J-C complementary DNA (cDNA designated as ‘mRNA’ in generalist databases) sequences. The complexity is further enhanced by the characteristics of the loci and chain types in the different species (reviewed in the IMGT Repertoire) and by the mechanisms of diversity such as combinatorial diversity, N diversity, somatic hypermutation and gene conversion. Thus, the detailed sequence annotation is a huge and complex task which requires the interpretation of DNA rearrangements and recombination, of sequence polymorphisms, of nucleotide deletions and insertions at the V-J and V-D-J junctions and, for IG, of somatic hypermutations.

15. Please provide a short description of scientific impact

Annotations rely on the accuracy and the coherence of IMGT-ONTOLOGY, the first ontology in the field of immunogenetics which has allowed to set up the rules for standardized sequence identification, gene and allele classification, constitutive and specific motif description, amino acid numbering and sequence obtaining information. The unique source of IMGT/LIGM-DB nucleotide sequences is EMBL. Prior to being entered in IMGT/LIGM-DB, IG and TR sequences must be submitted to EMBL, GenBank or DDBJ, in order to get a unique accession number which is also the entry identifier in IMGT/LIGM-DB. IMGT/LIGM-DB flatfiles are available by anonymous FTP servers at CINES (ftp://ftp.cines.fr/IMGT/), at EBI (ftp://ftp.ebi.ac.uk/pub/databases/imgt/), and at IGH (ftp://ftp.igh.cnrs.fr/pub/IMGT/) and from many SRS (Sequence Retrieval System)

sites. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (e.g. CINES, EBI and Institut Pasteur). IMGT/LIGM-DB data can also be retrieved through Web services. IMGT/LIGM-DB data are cross-referenced in the EMBL databank, in IMGT/GENE-DB which allows to link gene entries with the corresponding genomic reference sequences and with the known expressed cDNAs, and in IMGT/PRIMER-DB in order to display the oligonucleotide primers within the sequences.

IMGT/GENE-DB

13. Please provide a short description of unique content

IMGT/GENE-DB is the comprehensive IMGT genome database for immunoglobulin (IG) and T cell receptor (TR) genes from human, mouse and from other vertebrates (rat and rabbit). IMGT/GENE-DB is the international reference for the IG and TR gene nomenclature and works in close collaboration with the HUGO Nomenclature Committee, Mouse Genome Database and genome committees for other species. IMGT/GENE-DB allows a search of IG and TR genes by locus, group and subgroup, which are CLASSIFICATION concepts of IMGT-ONTOLOGY. Short cuts allow the retrieval gene information by gene name or clone name. Direct links with configurable URL give access to information usable by humans or programs. An IMGT/GENE-DB entry displays accurate gene data related to genome (gene localization), allelic polymorphisms (number of alleles, IMGT reference sequences, functionality, etc.) gene expression (known cDNAs), proteins and structures (Protein displays, IMGT Colliers de Perles). It provides internal links to the IMGT sequence databases and to the IMGT Repertoire Web resources, and external links to genome and generalist sequence databases. IMGT/GENE-DB manages the IMGT reference directory used by the IMGT tools for IG and TR gene and allele comparison and assignment, and by the IMGT databases for gene data annotation.

14. Please provide a short description of biological utility

IMGT/GENE-DB has been implemented to provide an easy and common access to standardized and expertly annotated IG and TR gene and allele data and knowledge. The first task of IMGT was to define a reference sequence for each individual gene and allele, based on the IMGT ‘gene’ and ‘allele’ concepts. IMGT/GENE-DB, which currently contains human, mouse, rat and rabbbit IG and TR genes, is the international reference for the IG and TR gene nomenclature. The molecular genetics of the IG and TR genes is so complex and unique in the genome of vertebrates that a specific gene database was required to manage all their characteristics. Indeed, the synthesis of IG and TR chains involves multigene families from four different gene types: variable (V), diversity (D), joining (J) and constant (C), each one with unique characteristics. These genes are organized in hundreds of cassettes, as in fish, or in large clusters from several hundred kilobases to one (or more) megabase(s), as in mouse and human. IG and TR genes that belong to same subgroup may be highly similar in their coding sequence, but at the same time, highly polymorphic (e.g. 13 allelic forms have been sequenced for the human IGHV2-70 gene), with alleles displaying different functionalities. The presence of many pseudogenes in the loci, and the frequency of the polymorphisms by gene insertion and deletion in these multigene families, add an additional level of complexity. Although most human IG and TR genes were sequenced and characterized independently from and before the completion of the Human Genome Project, the classification and the characterization of the IG and TR genes remain a big challenge in the analysis of the genome.

15. Please provide a short description of scientific impact

The human IMGT gene names were approved by the Human Genome Organisation (HUGO)

Nomenclature Committee (HGNC) in 1999, and entered in IMGT/GENE-DB, Genome DataBase GDB (Canada), LocusLink at NCBI (USA) and GeneCards. Reciprocal links exist between IMGT/GENE-DB and the generalist nomenclature (HGNC Genew) and genome databases (GDB, LocusLink and Entrez Gene at NCBI, and GeneCards). The mouse IG and TR gene names with IMGT reference sequences were provided by IMGT to HGNC and to the Mouse Genome Database (MGD) in July 2002. The IMGT/GENE-DB data are used by other IMGT databases (IMGT/PRIMER-DB, IMGT/3D structure-DB) and tools (IMGT/V-QUEST, IMGT/JunctionAnalysis, etc.). The dynamic interactions are currently implemented through IMGT-Choreography based on IMGT-ONTOLOGY and using IMGT-ML Web services.

IMGT/PRIMER-DB

13. Please provide a short description of unique content

IMGT/PRIMER-DB is the IMGT® oligonucleotide (primer) database for the immunoglobulins (IG) and T cell receptors (TR). IMGT/PRIMER-DB, developed by Laboratoire d'ImmunoGénétique Moléculaire LIGM has been on the Web since February 2002. IMGT/PRIMER-DB provides standardized information on oligonucleotides or primers of immunoglobulins and T cell receptors from human and other vertebrate species. IMGT/PRIMER-DB contains information on primers and combinations of primers described as 'sets' [primers sharing identical properties (species, group and orientation)] and 'couples' [sets of opposite orientation for which IMGT/LIGM-DB sequences are known (or expected)]. Primers, Sets and Couples are described in IMGT Primer cards, IMGT Set cards and IMGT Couple cards, respectively. An IMGT Primer is an oligonucleotide described by comparison to an IMGT/LIGM-DB reference sequence, according to the standardized rules of the IMGT Scientific chart, based on the IMGT-ONTOLOGY axioms and concepts.

14. Please provide a short description of biological utility

IMGT/PRIMER-DB provides immunoglobulin (IG) and T cell receptor (TR) primers for PCR gene amplification, combinatorial library constructions, scFv, plage display and microarray biotechnologies.

15. Please provide a short description of scientific impact

The IG and TR primers provided by IMGT/PRIMER-DB are particularly useful for the studies on the expression of immunoglobulin and T cell receptor in normal and pathological situations and for biotechnologies.

IMGT/3Dstructure-DB

13. Please provide a short description of unique content

IMGT/3Dstructure-DB is the IMGT® 3D structure database which comprises IG, TR, MHC, IgSF, MhcSF and RPI with known 3D structures. IMGT/3Dstructure-DB contains atomic coordinate files extracted from the Protein Data Bank (PDB) which are renumbered according to the standardized IMGT unique numbering. The IMGT/3Dstructure-DB cards provide chain details with IMGT annotations (receptor, chain and domain description with IMGT labels, assignment of IMGT gene and allele names, domain delimitations and amino acid positions according to the IMGT unique numbering, and IMGT Colliers de Perles on one layer and two layers), contact analysis, downloadable renumbered IMGT/3Dstructure-DB flat files, vizualisation tools (Jmol and QuickPDB), and external links. IMGT Residue@Position cards provide detailed information on the inter- and intra-domain contacts at each residue position, based on the IMGT unique numbering. The contacts are described per domain (intra- and inter-domain

contacts) and annotated in terms of IMGT® labels (chain and domain), positions (IMGT unique numbering), backbone or side-chain implication.

14. Please provide a short description of biological utility

IMGT/3Dstructure-DB is widely used by clinicians and biological scientists from both academic and industrial laboratories, in diverse research domains. IMGT/3Dstructure-DB is used in structural evolution of the IgSF and MhcSF proteins, in biotechnology related to antibody engineering (single chain Fragment variable (scFV), phage displays, combinatorial libraries, chimeric, humanized and human antibodies) and therapeutical approaches (grafts, immunotherapy and vaccinology).

15. Please provide a short description of scientific impact

IMGT Colliers de Perles built from IMGT/3Dstructure-DB data are particularly useful for antibody engineering to map a sequence on the domain conserved topology, to compare with germline sequences, to provide standardized framework and complementarity determining region delimitations, to visualize hydrogen bonds in known 3Dstructures, to localize amino acids involved in ADCC (FcR binding) and to localize amino acids located in N-glycosylation sites. The IMGT unique numbering and gene standardization provides a great help in large scale sequence-structure studies and more generally in protein engineering.