Immunogenetic sequence annotation based on IMGT-ONTOLOGY

Joumana Jabado-Michaloud, Géraldine Folch, Fatena Bellahcene, François Ehrenmann, Patrice Duroux, Véronique Giudicelli and Marie-Paule Lefranc

IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier 2, Institut de Génétique Humaine, IGH, CNRS UPR 1142, 141 rue de la Cardonille, F-34396 Montpellier cedex 05, France

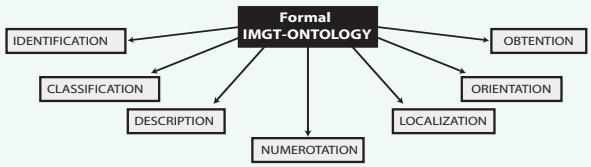
Im Muno Gene Tics Information system®
http://www.imgt.org

IMGT/LIGM-DB [1] is the first and the largest IMGT® database [2] in which are managed, analysed and annotated more than 136,000 immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences from human and 235 other vertebrate species (April 2009). The expert annotation of these sequences and the added standardized knowledge are based on IMGT-ONTOLOGY, the first ontology developed in the field of immunogenetics and immunoinformatics [3]. The annotation of immunogenetic sequences requires important expertise, owing to the unusual structure (non-classical exon/intron structure) of the IG and TR genes and characteristic chain synthesis owing to DNA V-J and V-D-J rearrangements. The way to annotate these sequences depends on the molecular type (gDNA, mRNA, cDNA or protein) and the configuration type (germline or rearranged), and if sequences from the concerned species are present or not in the IMGT reference directory sets. IMGT/V-QUEST [5] and internal tools (IMGT/Automat, IMGT/LIGMotif, IMGT/BLAST and IMGT/DomainGapAlign) were developed. The first step in annotation allows to identify the chain type (for instance IG-Heavy) and to assign standardized keywords (IDENTIFICATION axiom). The second step is the classification of IG and TR genes and alleles (CLASSIFICATION axiom). The third step is the description (DESCRIPTION axiom) of the V, D, J and C genes and alleles with specific standardized labels. There are more than 590 IMGT standardized labels from which 64 have been entered in Sequence Ontology (SO). The delimitation of the FR-IMGT and CDR-IMGT lengths and the positions of conserved amino acids based on the IMGT unique numbering (NUMEROTATION axiom) allow to bridge the gap between sequences and 3D structures [6]. The complete annotation of immunogenetic germline (V, D, J) and C sequences is followed by the update of the IMGT Repertoire (IMGT Gene tables, Alignments of alleles, Protein displays, Colliers de Perles, etc.), IMGT® gene database (IMGT/GENE-DB) and IMGT reference directory sets of the IMGT® tools (IMGT/V-QUEST, IMGT/JunctionAnalysis and IMGT/DomainGapAlign).

[1] Giudicelli, V. et al., Nucleic Acids Res., 34, D781-784 (2006).
[2] Lefranc, M.-P. et al., Nucleic. Acids Res., 37, D1006-1012 (2009).
[3] Giudicelli, V. and Lefranc, M.-P., Bioinformatics, 15, 1047-1054 (1999).
[4] Duroux, P. et al., Biochimie, 90, 570-583 (2008).
[5] Brochet X. et al. Nucleic Acids Res. 36,W503-508 (2008).
[6] Lefranc, M.-P., et al., Briefings in Bioinformatics, 9(4):263-275 (2008).

Formal IMGT-ONTOLOGY axioms

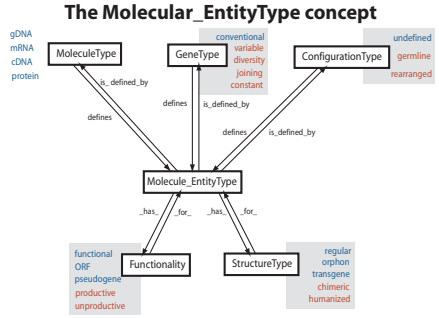
IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) is based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics [1]. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on seven axioms of the formal IMGT-ONTOLOGY or IMGT- Kaleidoscope [2]. Each axiom gives rise to a set of concepts. The concepts of identification, classification, description and numerotation are particularly used for the immunogenetic sequence annotation.



[1] Giudicelli, V. and Lefranc, M.-P., Bioinformatics, 15, 1047-1054 (1999).
[2] Duroux, P. et al., Biochimie, 90, 570-583 (2008).

1 The IDENTIFICATION axiom

The Molecular_EntityType concept



The IDENTIFICATION axiom has generated the concepts of identification which provide the terms and rules to identify an entity, its processes and its relations in IMGT®. They provide the IMGT standardized keywords.

The "Molecule_EntityType" concept, shown as an example, is defined by the "MoleculeType", "GeneType" and "ConfigurationType" concept and has relations with the "Functionality" and "StructureType" concepts. It includes 19 instances (L-V-gene, L-V-D-J-gene….).

Standardized keywords

2 The CLASSIFICATION axiom

The CLASSIFICATION axiom generates the concepts of classification, they allow to classify and name the genes and their alleles. The genes which code the IG and TR belong to highly polymorphic multigenic families. A major contribution of IMGT-ONTOLOGY was to set the principles of their classification and to propose a standardized nomenclature[1,2].

[1] Lefranc, M.-P. and Lefranc, G., The Immunoglobulin FactsBook (2001)
[2] Lefranc, M.-P. and Lefranc, G., The T cell receptor FactsBooks (2001)

Standardized nomenclature

IMGT FLAT-FILE

```
ID   M18809 IMGT/LIGM annotation : by annotators; genomic DNA; HUM; 1194 BP.
AC   M18809;
DT   15-MAY-1995 (Rel. 2, arrived in LIGM-DB )
DT   27-FEB-2007 (Rel. 200709-2, Last updated, Version 7)
XX
DE   Human Ig rearranged H-chain gene (WS1) V-region (VDJ4).  ;
DE   genomic DNA; rearranged configuration; Ig-Heavy; regular; functionality
DE   productive; group IGHV; subgroup HV5.
XX
KW   antigen receptor; Immunoglobulin superfamily (IgSF);
KW   Immunoglobulin (IG); IG-Heavy; IG-Heavy-Mu; variable; diversity;
KW   joining; gDNA; rearranged; V-D-J-gene.
XX
OS   Homo sapiens (human)
OC   cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa;
.....
OC   Homo/Pan/Gorilla group; Homo.
XX
FH   Key             Location/Qualifiers
FH
FT   L-V-D-J-GENE    1..1194
FT                   /db_xref="taxon:9606"
FT                   /map="14q32.33"
FT                   /organism="Homo sapiens"
FT                   /mol_type="genomic DNA"
FT   5'UTR           1..198
FT   L-PART1         199..244
FT                   /protein_id="AAA36053.1"
FT                   /translation="MGSTAILALLLAVLQ"
FT   INIT-CODON      199..201
FT   DONOR-SPLICE    244..246
FT   V-INTRON        245..327
FT   ACCEPTOR-SPLICE 325..329
FT   V-D-J-EXON      328..717
FT                   /codon_start=3
FT                   /translation="VCAEVQLVQSGAEVKKPGESLKISCKGSGYSFSTYWIAWLR
FT                   QMPGKGLEWMGIIYAGDSETRYTPSFRGQVTISADKSTSTAYLQWSSLKASDTAM
FT                   YYCARLEGRGYTGYALPYYFDYWGQGPLVTVSS"
FT   V-EXON          328..632
FT                   /codon_start=3
FT                   /translation="VCAEVQLVQSGAEVKKPGESLKISCKGSGYSFSTYWIAWLR
FT                   QMPGKGLEWMGIIYAGDSETRYTPSFRGQVTISADKSTSTAYLQWSSLKASDTAM
FT                   YYCAR"
FT   L-PART2         328..338
FT                   /codon_start=3
FT                   /translation="VCA"
FT   V-D-J-REGION    339..717
FT                   /translation="EVQLVQSGAEVKKPGESLKISCKGSGYSFSTYWIAWLRQMP
FT                   GKGLEWMGIIYAGDSETRYTPSFRGQVTISADKSTSTAYLQWSSLKASDTAMYYC
FT                   ARLEGRGYTGYALPYYFDYWGQGPLVTVSS"
FT   V-D-REGION      339..671
FT                   /putative_limit="3' side"
FT                   /translation="EVQLVQSGAEVKKPGESLKISCKGSGYSFSTYWIAWLRQMP
FT                   GKGLEWMGIIYAGDSETRYTPSFRGQVTISADKSTSTAYLQWSSLKASDTAMYYC
FT                   ARLEGRGYTGYALPY"
FT   V-REGION        339..632
FT                   /allele="IGHV5-51*01"
FT                   /gene="IGHV5-51"
FT                   /CDR_length="[8.8.19]"
FT                   /putative_limit="3' side"
FT                   /translation="EVQLVQSGAEVKKPGESLKISCKGSGYSFSTYWIAWLRQMP
FT                   GKGLEWMGIIYAGDSETRYTPSFRGQVTISADKSTSTAYLQWSSLKASDTAMYYC
FT                   AR"
FT   FR1-IMGT        339..413
FT                   /AA_IMGT="AA 1 to 26,AA 10 is missing"
FT                   /translation="EVQLVQSGAEVKKPGESLKISCKGS"
FT   1st-CYS         402..404
```

1 IDENTIFICATION: Keywords
genomic DNA=MoleculeType
rearranged=ConfigurationType
regular=StructureType
productive=Functionality
Homo sapiens=Taxon
Ig-Heavy=ChainType
variable, diversity, joining=GeneType

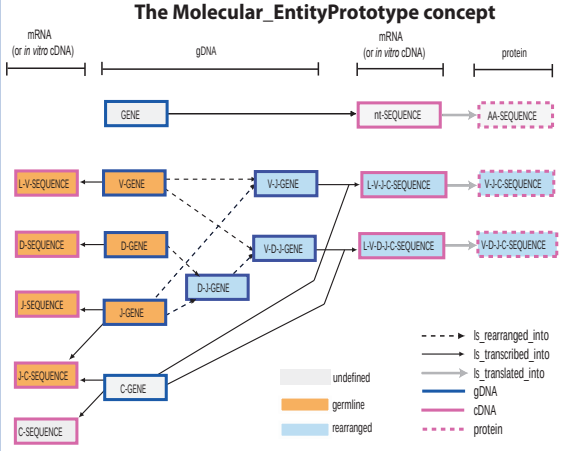
2 CLASSIFICATION: Nomenclature
IGHV=Group
IGHV5=Subgroup
IGHV5-51=Gene
IGHV5-51*01=Allele

3 DESCRIPTION:Labels
L-V-D-J-SEQUENCE=Entity
V-D-J-REGION=ComposedRegion
V-REGION=CoreRegion
FR1-IMGT=SubRegion
1st-CYS=ConservedAminoAcid

4 NUMEROTATION
V [8.8.19]=V-REGION CDR lengths
1 to 26, AA 10 is missing=AA IMGT numbering

3 The DESCRIPTION axiom

The Molecular_EntityPrototype concept

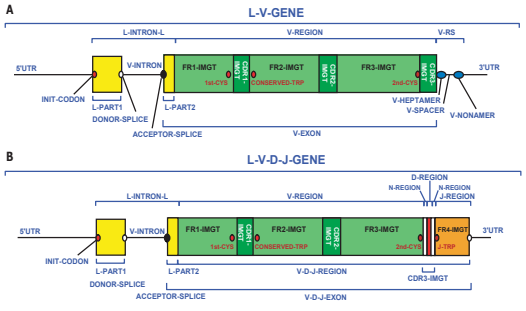


The DESCRIPTION axiom has generated the concepts of description which allow the description of any instance in IMGT®. The instances of the concepts of description correspond to IMGT standardized labels (available in the IMGT Scientific chart), 242 for the nucleotide sequences and 349 for the 3D structures.

Three instances "GENE", "nt-SEQUENCE" and "AA-SEQUENCE" correspond to conventional genes while the 16 other instances are specific of the IG and TR. The concept instances for mRNA are also valid for *in vitro* cDNA. The first column correspond to 'sterile transcript' instances.

Standardized labels

L-V-GENE and L-V-D-J-GENE prototypes



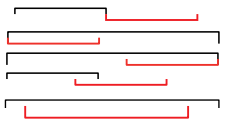
Each instance of the Moleular_EntityPrototype concept has a graphical representation of prototype.

Twenty-five labels and ten relations are necessary and sufficient for a complete description of the L-V-GENE and L-V-D-J-GENE of the Molecular_EntityPrototype concept.

Relations between IMGT labels

Relation	Reciprocal relation
"adjacent_in_its_3_prime_to"	"adjacent_in_its_5_prime_to"
"includes_with_same_5_prime",	"is_included_with_same_5_prime_in",
"includes_with_same_3_prime",	"is_included_with_same_3_prime_in",
"overlaps_in_its_3_prime_with"	"overlaps_in_its_5_prime_with"
"includes"	"is_included_in"

Relations between IMGT labels



Ten relations between labels are necessary and sufficient for a complete description of an instance of a Molecular_EntityPrototype concept.

4 The NUMEROTATION axiom

The NUMEROTATION axiom and the concepts of numerotation determine the principles of a unique numbering for a domain (sequences and 3D structures). The "IMGT_unique_numbering" concept is illustrated by the "IMGT_Collier_de_Perles" concept which allows graphical representation in two dimensions (2D) of amino acid sequences of V type [1], C type [2] or G type [3] domains.

[1] Lefranc, M.-P. et al., Dev. Comp. Immunol., 27, 55-77 (2003)
[2] Lefranc, M.-P. et al., Dev. Comp. Immunol., 29, 185-203 (2005)
[3]Lefranc, M.-P. et al., Dev. Comp. Immunol., 29, 917-938 (2005)

IMGT unique numbering

IMGT founder and director: Marie-Paule Lefranc (Marie-Paule.Lefranc@igh.cnrs.fr)
Bioinformatics manager: Véronique Giudicelli (Veronique.Giudicelli@igh.cnrs.fr)
Computer manager: Patrice Duroux (Patrice.Duroux@igh.cnrs.fr)
Interface design: Chantal Ginestoux (Chantal.Ginestoux@igh.cnrs.fr)

IMGT® is funded by CNRS, MESR and EU ImmunoGrid IST-2004-028069