Briefings in Bioinformatics Advance Access published April 19, 2008

doi:10.1093/bib/bbn014

IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics

Lefranc Marie-Paule, Véronique Giudicelli, Laetitia Regnier and Patrice Duroux

Submitted: 9th January 2008; Received (in revised form): 13th February 2008

Abstract

IMGT®, the international ImMunoGeneTics information system (http://imgt.cines.fr), is the reference in immunogenetics and immunoinformatics. IMGT standardizes and manages the complex immunogenetic data that include the immunoglobulins (IG) or antibodies, the T cell receptors (TR), the major histocompatibility complex (MHC) and the related proteins of the immune system (RPI), which belong to the immunoglobulin superfamily (IgSF) and the MHC superfamily (MhcSF). The accuracy and consistency of IMGT data and the coherence between the different IMGT components (databases, tools and Web resources) are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets relying on seven axioms, 'IDENTIFICATION', 'DESCRIPTION', 'CLASSIFICATION', 'NUMEROTATION', 'LOCALIZATION', 'ORIENTATION' and 'OBTENTION', that postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and that the way they are obtained has to be determined. These axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope. These axioms have been essential for the conceptualization of the molecular immunogenetics knowledge and for the creation of IMGT. Indeed all the components of the IMGT integrated system have been developed, based on standardized concepts and relations, thus allowing IMGT to bridge biological and computational spheres in bioinformatics. The same axioms can be used to generate concepts for multi-scale level approaches at the molecule, cell, tissue, organ, organism or population level, emphasizing the generalization of the application domain. In that way the Formal IMGT-ONTOLOGY represents a paradigm for the elaboration of ontologies in system biology.

Keywords: *IMGT; immunoinformatics; immunogenetics; ontology; information system; systems biology*

INTRODUCTION

IMGT®, the international ImMunoGeneTics information system (http://imgt.cines.fr) [1], is the international reference in immunogenetics and immunoinformatics. Created in 1989 at the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) by Marie-Paule Lefranc (Université Montpellier 2 and CNRS) in Montpellier, France, IMGT provides a high-quality integrated knowledge resource, specialized in the immunoglobulins (IG)

Corresponding author. Marie-Paule Lefranc, IMGT, 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; E-mail: Marie-Paule.Lefranc@igh.cnrs.fr

Marie-Paule Lefranc, PhD, is Professor at the Université Montpellier 2 and Head of the Laboratoire d'ImmunoGénétique Moléculaire at the Institut de Génétique Humaine CNRS. She is Director of IMGT, the international ImMunoGeneTics information system (http://imgt.cines.fr) that she created in 1989, at Montpellier, France.

Véronique Giudicelli, PhD, Engineer in Bioinformatics, has developed IMGT/GENE-DB, the IMGT genome database, and IMGT/V-QUEST, the IMGT tool for the analysis of the IG and TR sequences. She is in charge of the bioinformatic developments of IMGT, the international ImMunoGeneTics information system.

Laetitia Regnier, Engineer in Bioinformatics is in charge of the improvement of the annotations in IMGT/LIGM-DB and of the quality assurance in IMGT, the international ImMunoGeneTics information system.

Patrice Duroux, PhD, Research Engineer in informatics, is in charge of the computing management of IMGT, the international ImMunoGeneTics information system, at the Laboratoire d'ImmunoGénétique Moléculaire, Institut de Génétique Humaine CNRS, Montpellier, France.

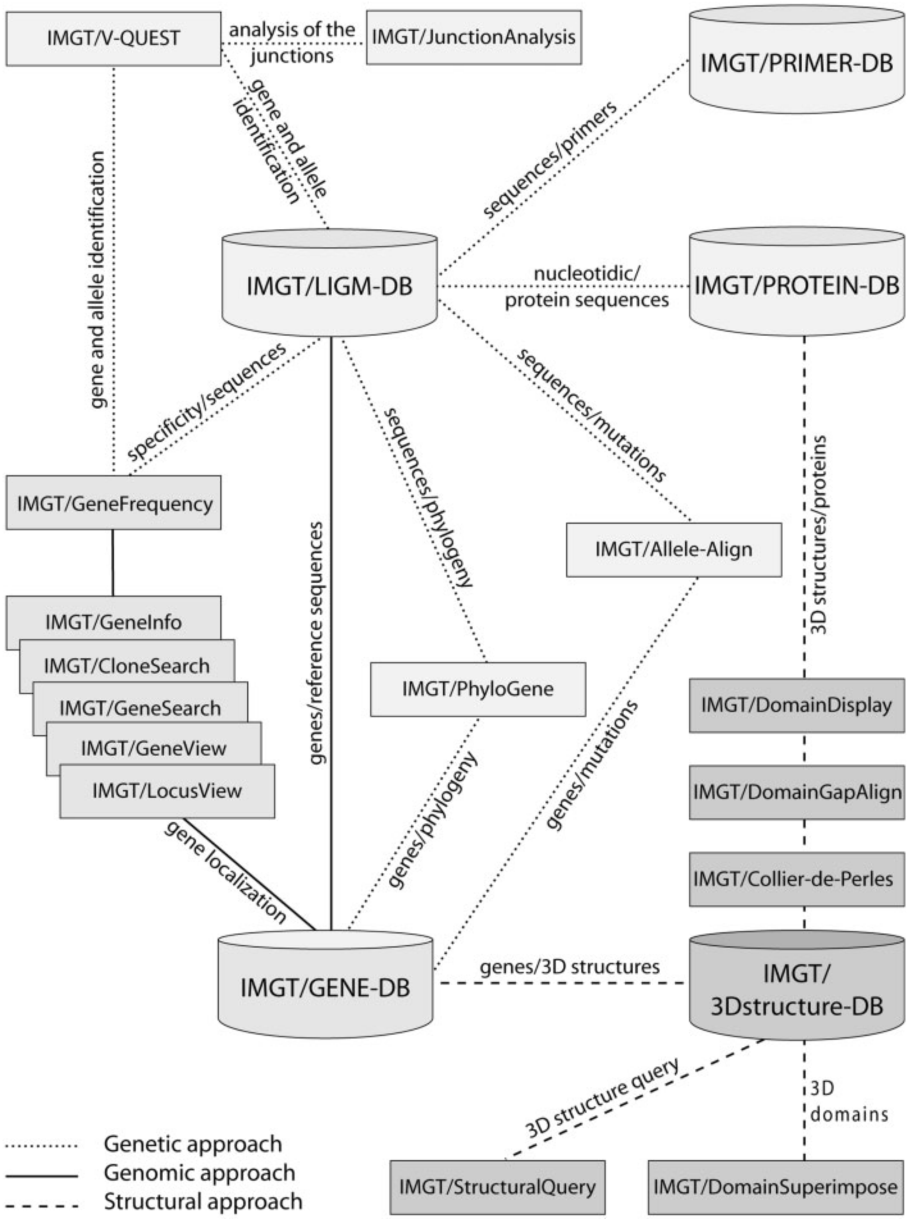

Figure 1: IMGT databases and tools with their interactions according to the genetic, genomic and/or structural approaches (http://imgt.cines.fr).

or antibodies, T cell receptors (TR), major histocompatibility complex (MHC) of human and other vertebrates and related proteins of the immune system (RPI), which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF) of any species. The IMGT information system consists of databases (three of sequences, one of genes and one of three-dimensional (3D) structures) and interactive tools for sequence, genome and 3D structure analysis, which interact together according to genetic, genomic and structural approaches [2] (Figure 1). Moreover, IMGT provides Web resources comprising more than 10 000 HTML pages of synthesis (IMGT Repertoire), knowledge (IMGT Scientific chart, IMGT Biotechnology page, IMGT Medical page, IMGT Veterinary page, IMGT Education and IMGT Index) and external links (IMGT Immunoinformatics page, IMGT Bloc-notes and IMGT other accesses) [1, 2].

The accuracy and the consistency of the IMGT data, as well as the coherence between the different IMGT components (databases, tools and Web resources) are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and

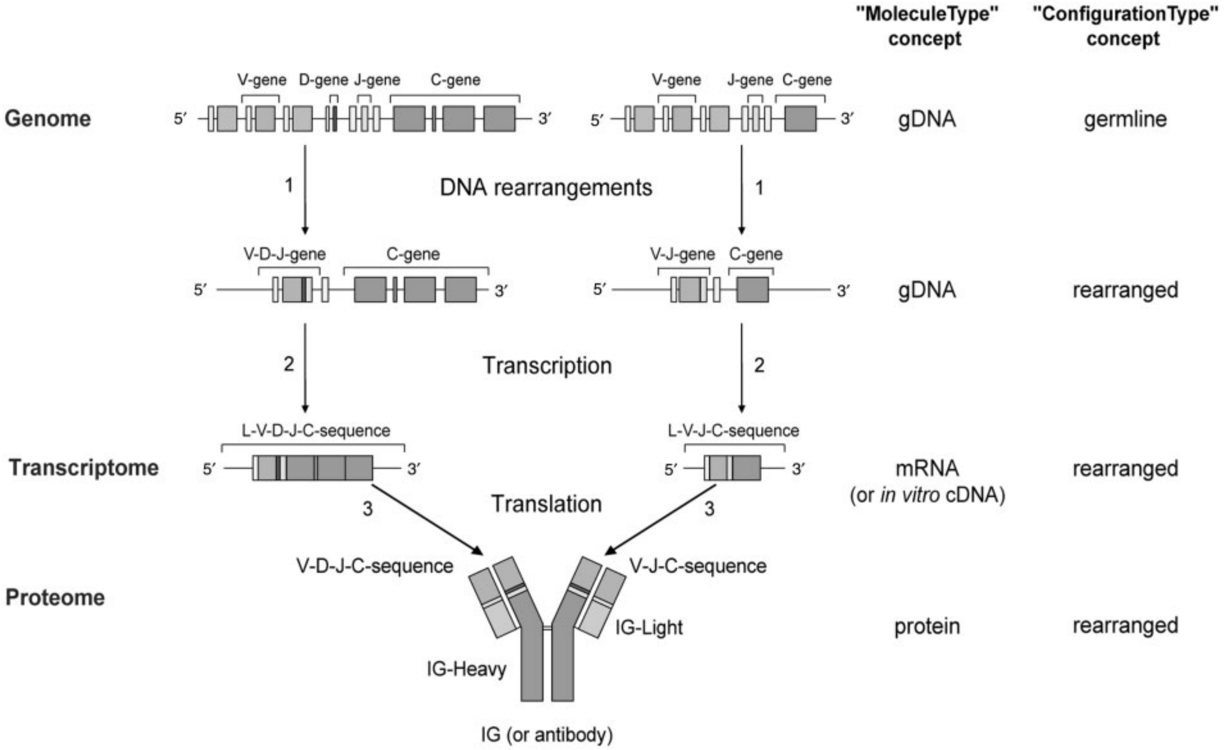

Figure 2: An example of biological knowledge at the molecular level: the synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies [5]. (**1**) DNA rearrangements (is_rearranged_into); (**2**) Transcription (is_transcribed_into); (**3**) Translation (is_translated_into).

immunoinformatics [3]. The standardization rules, defined in the IMGT Scientific chart, are based on the IMGT-ONTOLOGY concepts that were developed to solve the complexity of the immunogenetics knowledge and that, interestingly, were defined in an unprecedented approach, and before the term 'ontology' became commonly used in biology and bioinformatics [3]. The IMGT-ONTOLOGY concepts are generated from the seven axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope: 'IDENTIFICATION', 'DESCRIPTION', 'CLASSIFICATION', 'NUMEROTATION', 'LOCALIZATION', 'ORIENTATION' and 'OBTENTION' [4]. These axioms postulate that the approach to manage biological data and to represent knowledge in biology comprises various facets [4].

Immunogenetics, the science that studies the genetics of the immune responses, has shown a considerable expansion in biomedical fields since the last decades. It has highlighted the complex mechanisms by which B cells and T cells are at the origin of the extreme diversity of antigen receptors, the main actors of the adaptive immune responses, that comprise the IG or antibodies, and the TR (10^{12} different IG and 10^{12} different TR per individual, in humans) [5, 6]. These mechanisms include in particular DNA rearrangements [7] and, for the IG, somatic hypermutations [5, 6]. The IMGT-ONTOLOGY axioms and concepts have allowed the representation, at the molecular level, of immunogenetics knowledge related to the genome, transcriptome, proteome, genetics and 3D structures. In this article, we review some of the major IMGT-ONTOLOGY concepts that allow to bridge the gap between the biological and computational spheres in bioinformatics.

IDENTIFICATION AXIOM: IMGT STANDARDIZED KEYWORDS

An example of biological knowledge

The IG synthesis is a key example of biological knowledge. An IG or antibody is composed of two identical IG-Heavy chains associated with two identical IG-Light chains, kappa or lambda (Figure 2). In humans, the IG heavy genes (locus IGH), IG kappa genes (locus IGK) and IG lambda genes (locus IGL) are located on the chromosomes 14 (14q32.3), 2 (2p11.2) and 22 (22q11.2), respectively.

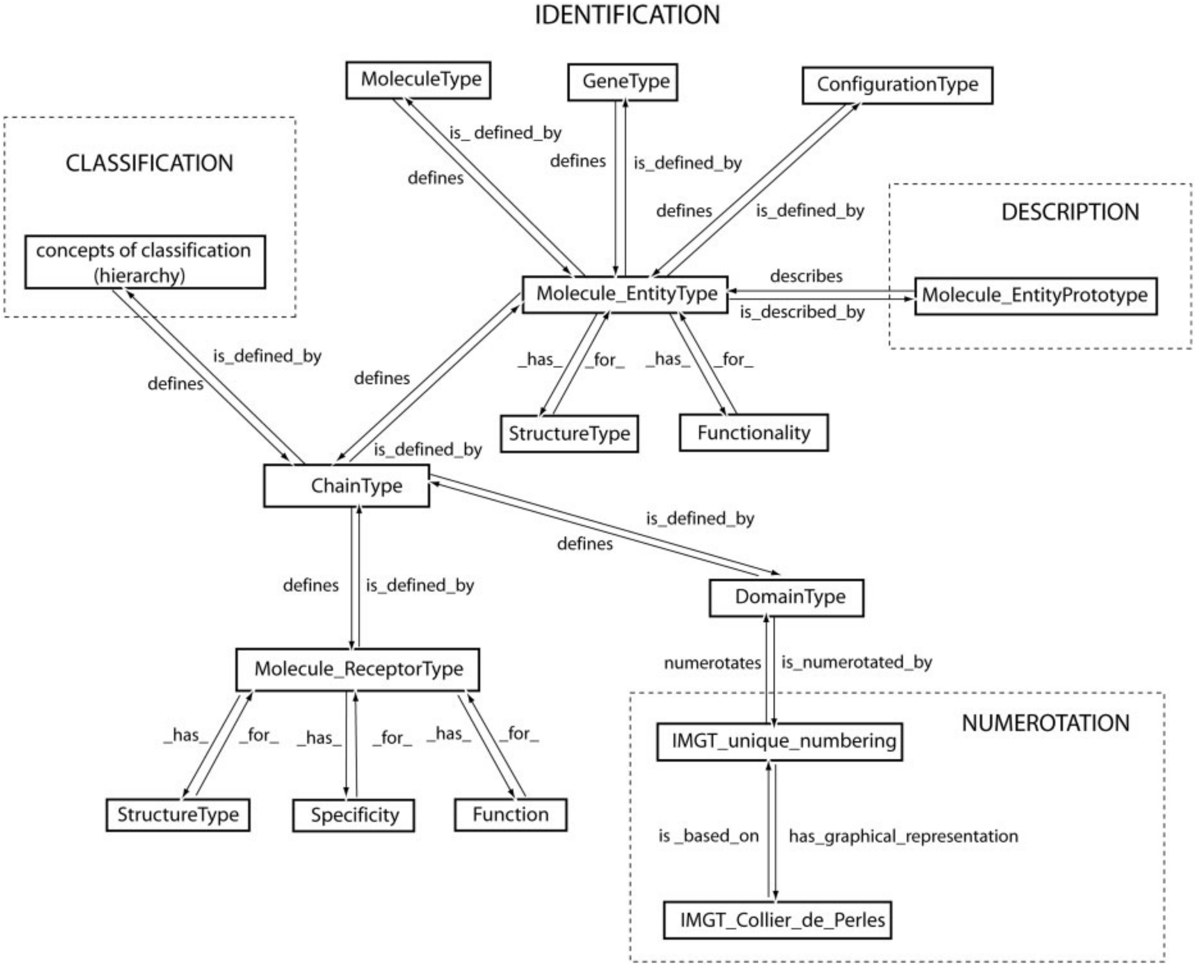

Figure 3: Main IMGT-ONTOLOGYconcepts of identification generated from the IDENTIFICATION axiom and their relations with concepts generated from the DESCRIPTION, CLASSIFICATION and NUMEROTATION axioms, at the molecular level.

The synthesis of an IG requires rearrangements of the IGH, IGK and IGL genes during the differentiation of the B lymphocytes [5, 6]. In the human genome (genomic DNA or gDNA), four types of genes code the IG (and TR): variable (V), diversity (D), joining (J) and constant (C) genes. The configuration of the V-gene, D-gene and J-gene is identified as 'germline' (Figure 2), the configuration of the C-gene is 'undefined'. During the differentiation of the B lymphocytes in the bone marrow, the genomic DNA is rearranged first in the IGH locus, and then in the IGK and IGL loci. The rearrangements in the IGH locus lead to the junction of a D-gene and a J-gene to form a D-J-gene, and then to the junction of a V-gene to the D-J-gene to form a V-D-J-gene. The rearrangements in the IGK or IGL loci lead to the junction of a V-gene and a J-gene to form a V-J-gene. The configuration of these genes is identified as 'rearranged'. After transcription and maturation of the pre-messenger by splicing, the messenger RNA (mRNA) L-V-D-J-C-sequence and L-V-J-C-sequence (L for leader) are obtained and then translated into the heavy chain (IG-Heavy) and the light chain (IG-Light) of an IG (or antibody) (Figure 2).

Concepts of identification

In a first step, biological objects, processes and relations require to be identified in order to be entered into IMGT. The IDENTIFICATION axiom [4] has generated the IMGT-ONTOLOGY concepts of identification that were necessary to define the IMGT standardized keywords and their relations [3, 8]. At the molecular level, four major concepts were defined: 'Molecule_Entity Type', 'Molecule_ReceptorType', 'ChainType' and 'DomainType' (Figure 3).

'Molecule_EntityType' concept

In Figure 2, 10 molecular entities can be identified: V-gene, D-gene, J-gene, C-gene, V-D-J-gene, V-J-gene, L-V-D-J-C-sequence, L-V-J-C-sequence, V-D-J-C-sequence and V-J-C-sequence. They represent instances of the 'Molecule_EntityType' concept. This concept that includes 21 instances, allows to identify any coding molecule of the genome, transcriptome and proteome. These instances are defined by the instances of three other concepts of identification: 'MoleculeType' ('gDNA', 'mRNA', 'cDNA', 'protein'), 'GeneType' ['conventional', 'variable' (V), 'diversity' (D), 'joining' (J) and 'constant' (C)], and 'ConfigurationType' ('undefined' for conventional and C genes, 'germline' for unrearranged V, D and J genes and 'rearranged' for V, D and J genes after DNA rearrangements [5, 6]). Three instances, 'gene', 'nt-sequence' and 'AA-sequence', respectively, identify the gDNA, mRNA and protein ('MoleculeType') of a conventional gene ('GeneType') in undefined configuration ('ConfigurationType'). The nt-sequence instance is also valid for cDNA. Eighteen instances identify the IG and TR, including the 10 ones shown in Figure 2. For example, the instance 'V-gene' identifies a 'gDNA' containing a 'V' gene, in 'germline' configuration. The instance 'L-V-J-C-sequence' identifies a sequence of 'mRNA' or 'cDNA' containing 'V', 'J' and 'C' genes, with V and J in 'rearranged' configuration. The eight instances not shown in Figure 2 correspond to partial rearrangements or to sterile transcripts [4].

Once a 'Molecule_EntityType' concept instance has been identified, it is possible to define two properties: its structure and its functionality, defined by the instances of two other concepts of identification. The 'StructureType' concept allows to identify entities with a classical organization ('regular'), from those that have been modified either naturally *in vivo* ('orphon', 'processed orphon', 'unprocessed orphon', 'unspliced', 'partially spliced', etc.) or artificially *in vitro* ('chimeric', 'humanized', 'transgene', etc.). The 'Functionality' concept includes five instances, three of them, 'functional', 'ORF' (open reading frame) and 'pseudogene' identify the functionality of 'Molecule_EntityType' instances in undefined or germline configuration (conventional genes, C genes, germline V, D and J genes), whereas the two others 'productive' and 'unproductive' identify the functionality of 'Molecule_EntityType' instances in rearranged configuration (rearranged V, D and J genes, fusion genes resulting from translocations or obtained by biotechnology and/or molecular engineering).

'Molecule_ReceptorType' concept

The 'Molecule_ReceptorType' concept identifies the type of protein receptor, defined by its chain composition [4]. Thus, IG is an instance of the 'Molecule_ReceptorType' concept, defined as comprising four chains, two IG-Heavy and two IG-Light chains, identical two by two and covalently linked (Figure 4).

Once a 'Molecule_ReceptorType' concept instance has been identified, it is possible to define its properties such as structure, specificity and function by instances of other concepts of identification. Thus, instances of the 'Specificity' concept identify the antigen recognized by an antigen receptor (IG or TR). The instances of that concept (several hundreds at the present time) can be connected on the one hand, with the 'Epitope' concept that identifies the part of the antigen recognized by the antigen receptor and, on the other hand, with the 'Paratope' concept that identifies the part of the antigen receptor (IG or TR), which recognizes and binds to the antigen [9]. Instances of the 'Function' concept identify the dual function of the antigen receptors [5].

'ChainType' and 'DomainType' concepts

The two IG-Heavy and two IG-Light chains of an IG are instances of the 'ChainType' concept (Figure 2). The 'ChainType' concept identifies the type of chain. It is one of the most important concepts of identification for the standardization of genome, transcriptome and proteome data in system biology. Indeed, an instance of the 'ChainType' concept is defined not only by an instance of the 'Molecule_EntityType' (for instance V-J-sequence) but also by an instance of a concept of classification and by a definition in domains, which bridges the gap with 3D structures.

By its relation with the concepts of classification, the 'ChainType' concept contains a hierarchy of concepts that identify the chain type at different levels of granularity. The finest level of granularity, the 'GeneLevelChainType' concept, identifies the chain type by reference to the gene(s), which code(s) the chain (reciprocal relations 'is_coded_by' and 'codes'). The number of instances of the 'GeneLevelChainType' concept depends on the number of functional genes and ORF per haploid

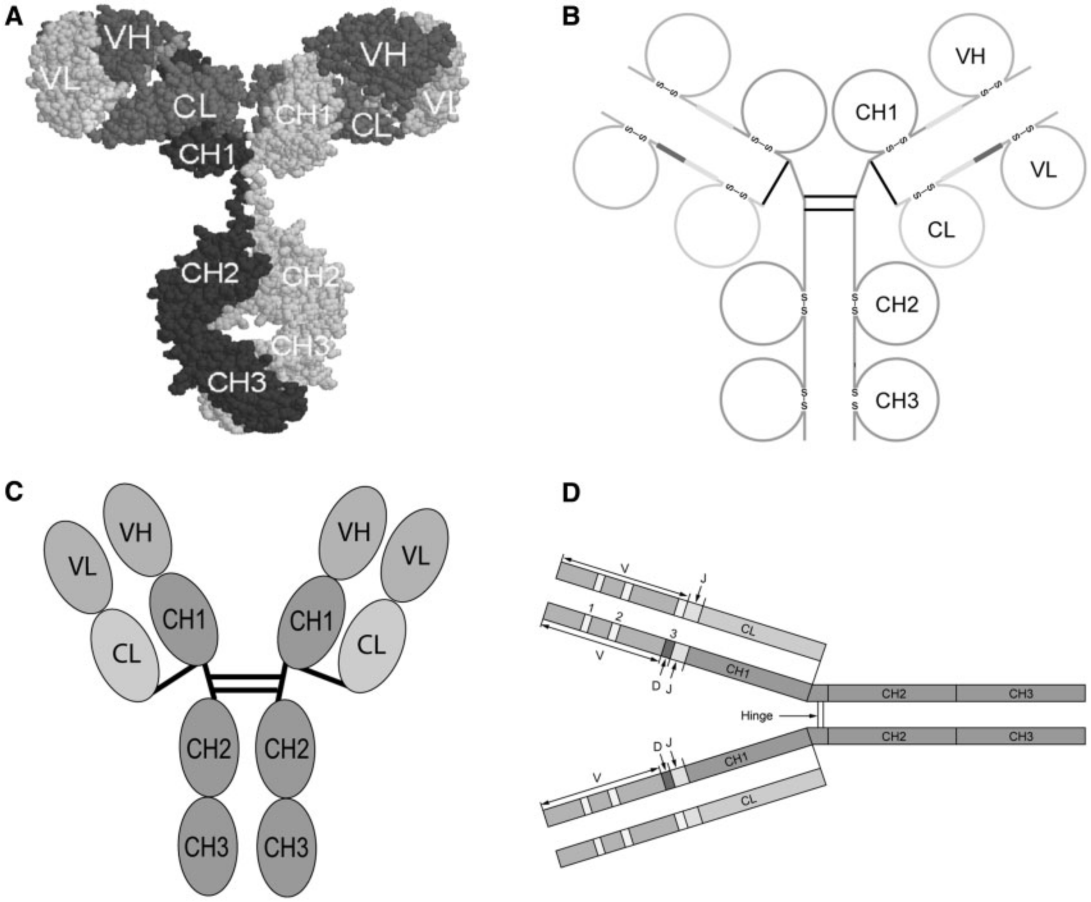

Figure 4: Concepts of identification. The different representations (**A**), (**B**), (**C**) and (**D**) of an IG or antibody (here, an IgG) allow to identify the 'ChainType' (IG-Heavy, IG-Light) and the 'DomainType' (VH, CH1, CH2, CH3, VL, CL), and in (**D**), the 'Molecule_EntityType' (V-J-C-sequence, V-D-J-C-sequence) (IDENTIFICATION axiom). The IG-Heavy of an IgM or an IgE would have a CH4 and no hinge. Based on (**D**), a schematized Y shape is frequently used to represent an IG.

genome in a given species (in the case of the IG and TR, it is the number of functional and ORF constant genes, which is taken into account). If only the functional genes are considered, the instances of this concept correspond to the isotypes.

A chain type instance can also be defined by its constitutive structural units ('DomainType' concept). A domain is a chain subunit characterized by its 3D structure and, by extension, its amino acid sequence and the nucleotide sequence that encodes it. The 'DomainType' concept may theoretically comprise many instances, but so far only the instances that have been carefully characterized by LIGM have been entered in IMGT-ONTOLOGY. The 'DomainType' concept has currently three instances, V type domain (variable domains of the IG and TR and V-like domains of other IgSF proteins), C type domain (constant domains of the IG and TR and C-like domains of other IgSF proteins) and G type domain (groove domains of the MHC and G-like domains of other MhcSF proteins) [10–14].

DESCRIPTION AXIOM: IMGT STANDARDIZED LABELS

In a second step, once entered in IMGT, biological objects, processes and relations require to be described. The DESCRIPTION axiom [4] has generated the IMGT-ONTOLOGY concepts of description that were necessary to define the IMGT standardized labels (written in capital letters) and their relations [3, 8].

'Molecule_EntityPrototype' concept

Each one of the 21 instances of the 'Molecule_EntityType' concept (IDENTIFICATION axiom) is linked to an instance of the 'Molecule_Entity Prototype' concept (DESCRIPTION axiom),

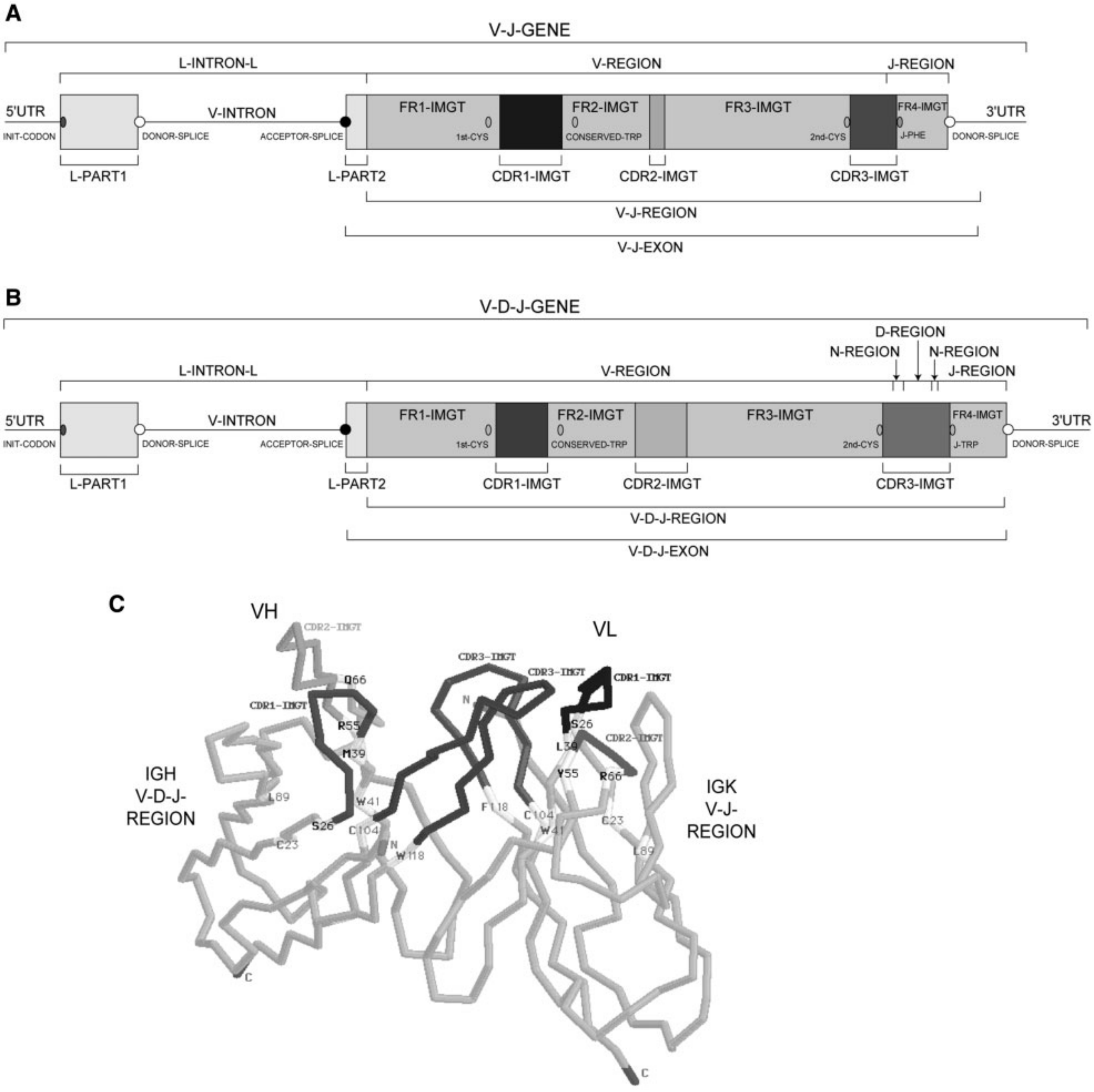

Figure 5: Concepts of description. (**A**) V-J-GENE and (**B**) V-D-J-GENE instances of the 'Molecule_EntityPrototype' concept graphically representated with their labels (DESCRIPTION axiom). Twenty-seven labels and 10 relations are necessary and sufficient for a complete description of these instances [4]. (**C**) VH and VL domains encoded by the V-D-J-REGION (of V-D-J-GENE) and by the V-J-REGION (of V-J-GENE).

by the reciprocal relations 'is_described_by' and 'describes'. Each instance of the 'Molecule_Entity Prototype' concept can be described with its constitutive motifs, which belong to other concepts of description. Figure 5 shows as examples the graphical representation of the V-D-J-GENE and V-J-GENE instances with their constitutive motifs and their encoded VH and VL domains.

A set of 10 relations is necessary and sufficient to compare the localization of the motifs of an instance of the concept 'Molecule_EntityPrototype' (Table 1). These relations are part of the concepts of localization (LOCALIZATION axiom) (IMGT Index, http://imgt.cines.fr).

The ontology for sequences and 3D structures has been the focus of IMGT for many years. More than 500 standardized labels were defined [221 for the nucleotide sequences (http://imgt.cines.fr/cgi-bin/IMGTlect.jv?query=7) [15] and 285 for the 3D structures [9] (http://imgt.cines.fr/textes/IMGT

Table 1: Relations for sequence description (LOCALIZATION axiom)

Relation	Reciprocal relation
'adjacent_at_its_5_prime_to'	'adjacent_at_its_3_prime_to'
'included_with_same_5_prime_in',	'includes_with_same_5_prime',
'included_with_same_3_prime_in',	'includes_with_same_3_prime',
'overlaps_at_its_5_prime_with'	'overlaps_at_its_3_prime_with'
'included_in'	'includes'

ScientificChart/SequenceDescription/IMGT3Dlabeldef.html)]. Prototypes represent the organizational relationship between labels and give information on the order and expected length (in number of nucleotides) of the labels [3]. This provides rules to verify the manual annotation, and to design automatic annotation tools such as IMGT/LIGMotif and IMGT/Automat [16]. Annotation of sequences and 3D structures with these labels constitutes the main part of the expertise. Interestingly, 64 IMGT specific labels defined for nucleotide sequences have been entered in the newly created Sequence Ontology (SO) (http://song.sourceforge.net/) [17] to describe specific IG and TR gene organization (http://imgt.cines.fr/textes/IMGTindex/ontology.html).

The 'Molecule_EntityPrototype' concept is fundamental in IMGT-ONTOLOGY as relations between its instances allow the representation of the knowledge related to the complex mechanisms of IG and TR gene rearrangements and chain synthesis [5, 6]. The relation 'is_rearranged_into' is specific to the synthesis of the IG and TR. The relations 'is_transcribed_into' and 'is_translated_into' are general for molecular biology. These three relations allow the organization of the various instances of the 'Molecule_EntityPrototype' concept during the synthesis of the IG and the TR, and in a more general way for the expression of any protein. They allow in addition, by more specific relations, to take into account the alternative transcripts, the protein isoforms and the post-translational modifications.

CLASSIFICATION AXIOM: IMGT STANDARDIZED GENE NAMES

In IMGT, biological objects, processes and relations require to be classified. The CLASSIFICATION axiom [4] has generated the IMGT-ONTOLOGY concepts of classification that were necessary to set up the IMGT standardized nomenclature for the genes and alleles [3, 8].

Group, subgroup, gene and allele

The objective was initially to provide immunologists and geneticists with a standardized nomenclature per locus and per species, which allows extraction and comparison of data for the complex B and T cell antigen receptor molecules, most of the instances of the 'Molecule_EntityType' concept containing more than one gene (three for V-J-C, four for V-D-J-C) (Figure 2). Moreover, the IG and TR genes belong to highly polymorphic multigene families. A major contribution of IMGT-ONTOLOGY was to set the principles of their classification and to propose a standardized nomenclature [5, 6] (Figure 6).

The 'Group' concept classifies a set of genes that belong to the same multigene family, within the same species or between different species. For the IG and TR, the set of genes is identified by an instance of the 'GeneType' concept (V, D, J or C). The 'Subgroup' concept classifies a subset of genes that belong to the same group, and which, in a given species, share at least 75% of identity at the nucleotide sequence level (and in the germline configuration for the V, D and J genes). The 'Gene' concept classifies a unit of DNA sequence that can be potentially transcribed and/or translated (this definition includes the regulatory elements in 5′ and 3′, and the introns, if present). The instances of the 'Gene' concept are gene names. In IMGT-ONTOLOGY, a gene name is composed of the name of the species (instance of the Taxon 'Species' concept) and of the gene symbol, for example, *Homo sapiens* IGHV1-2. By extension, orphons and pseudogenes are also instances of the 'Gene' concept. The 'Allele' concept classifies a polymorphic variant of a gene. The instances of the 'Allele' concept are allele names. Alleles identified by the mutations of the nucleotide sequence are classified by reference to allele *01. Full description of mutations and allele name designations are currently recorded for the core sequences (V-REGION, D-REGION, J-REGION, C-REGION). They are reported in Alignment tables, in IMGT Repertoire (http://imgt.cines.fr) and in IMGT/GENE-DB [18].

International IMGT gene nomenclature

The IMGT gene nomenclature was approved at the international level by Human Genome Organisation (HUGO) Nomenclature Committee HGNC in

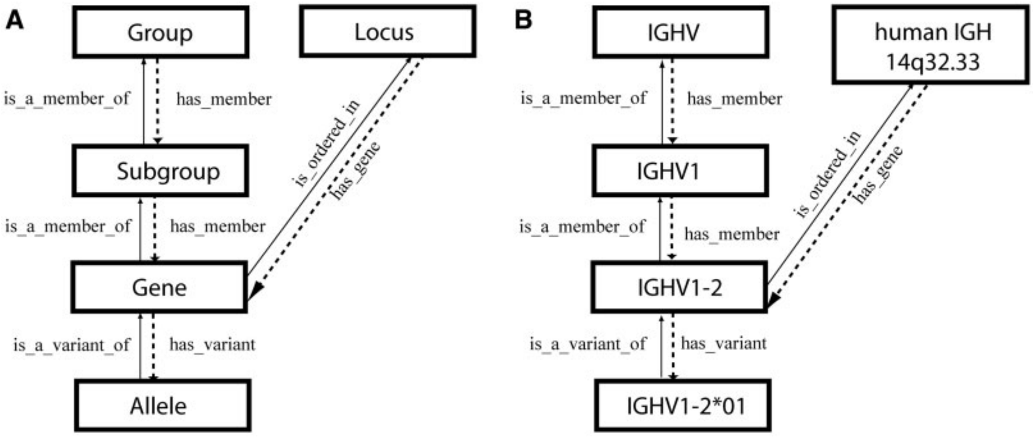

Figure 6: Concepts of classification for gene and allele nomenclature (CLASSIFICATION axiom). (**A**) Hierarchy of the concepts of classification and their relations. (**B**) Examples of concept instances for each concept of classification. The concepts instances are associated with an instance of the 'Taxon' concept (IDENTIFICATION axiom), and more precisely for the 'Gene' and 'Allele' concepts with an instance of the 'Species' concept (here, *Homo sapiens*). The 'Locus' concept is a concept of localization (LOCALIZATION axiom).

1999 [19]. The IMGT IG and TR gene names are the official references for the genome projects and, as such, were entered in Genome Database (GDB) and in LocusLink at the National Center for Biotechnology Information (NCBI) in 1999–2000, in Entrez Gene (NCBI) [20] when this gene database superseded LocusLink and in IMGT/GENE-DB [18]. The IMGT IG and TR gene names are the official references for the World Health Organization-International Union of Immunological Societies (WHO-IUIS) Nomenclature Subcommittee for IG and TR [21, 22] and for the genome projects and, as such, have been integrated in the MapViewer at NCBI and, in 2006, in the Ensembl server at the European Bioinformatics Institute (EBI).

IMGT reference sequences and analysis tools

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 published sequence, longest sequence, mapped sequence. They are listed in the germline gene tables of the IMGT Repertoire (http://imgt.cines.fr) and are available in IMGT/GENE-DB [18]. The IMGT Protein displays show the translated sequences of the alleles *01 of the functional or ORF genes [5, 6]. IMGT/DomainGapAlign identifies the closest IG and TR (V, J and C), MHC and RPI genes from amino acid sequences [9]. The IMGT/V-QUEST tool allows to identify the closest germline V and J genes and alleles from user IG and TR nucleotide sequences by comparison with the IMGT reference directory sets, whereas the IMGT/JunctionAnalysis tool analyses precisely the V-J and V-D-J junctions and identifies the closest D genes and alleles [23, 24]. These tools are widely used for the IG and TR repertoire analysis in normal and pathological situations [25]. They are recommended by the European Research Initiative on Chronic lymphocytic leukemia CLL (ERIC) for IGHV mutational status analysis in CLL [26].

NUMEROTATION AXIOM: IMGT UNIQUE NUMBERING

In IMGT, biological objects, processes and relations require to be numerotated. The NUMEROTATION axiom [4] has generated the IMGT-ONTOLOGY concepts of numerotation, and among them, the flagship of IMGT: the IMGT unique numbering [10–14] and its two-dimensional (2D) graphical representation or IMGT Collier de Perles [5, 6, 27, 28] (Figure 7). These concepts were recently reviewed [29] and are only briefly described here.

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),

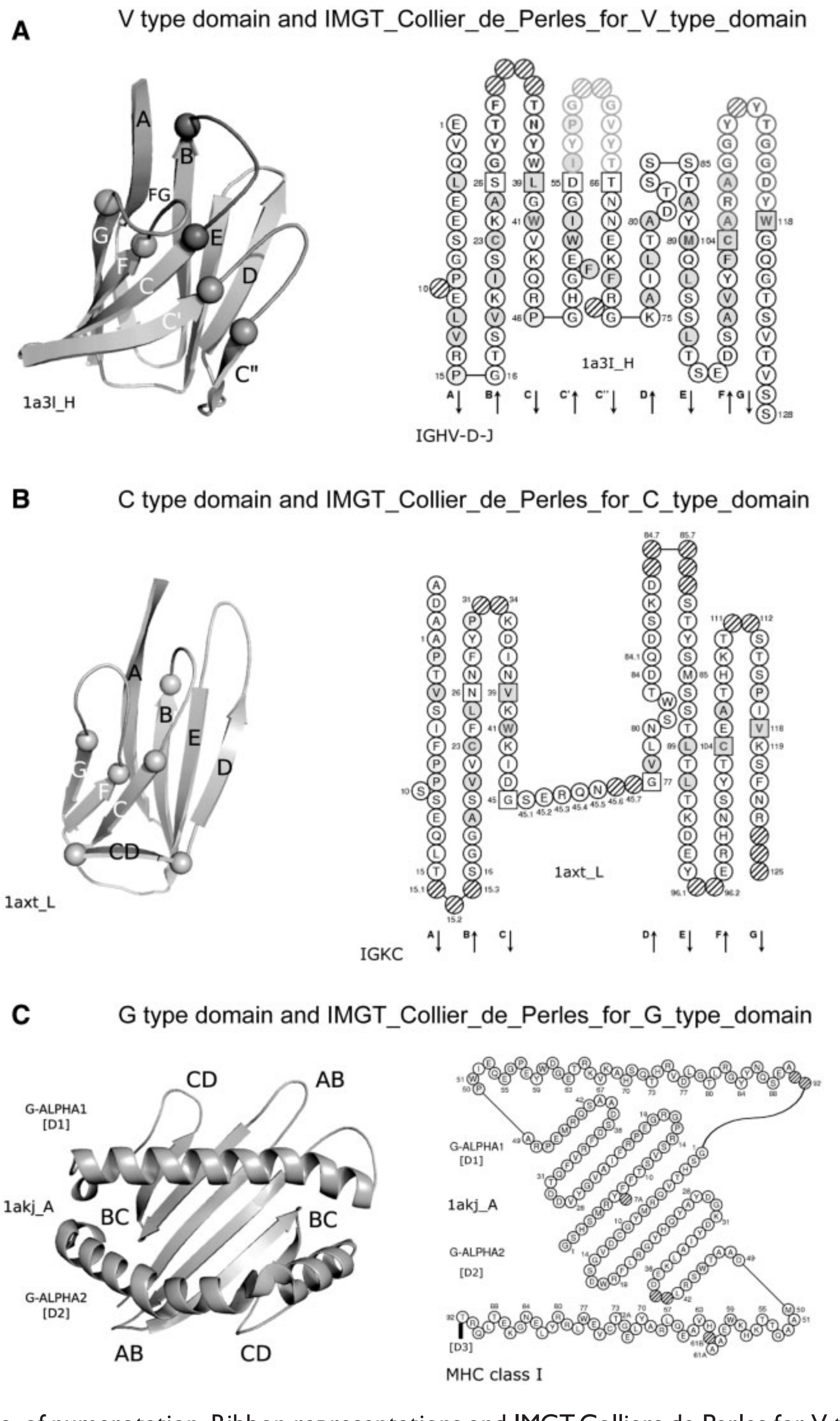
A
V type domain and IMGT_Collier_de_Perles_for_V_type_domain
A
B
FG
G
E
F
C
D
C'
C"
1a3l_H
1a3I_H
IGHV-D-J
B
C type domain and IMGT_Collier_de_Perles_for_C_type_domain
A
B
E
G
F
D
C
CD
1axt_L
1axt_L
IGKC
C
G type domain and IMGT_Collier_de_Perles_for_G_type_domain
CD
AB
G-ALPHA1 [D1]
1akj_A
BC
BC
G-ALPHA2 [D2]
AB
CD
G-ALPHA1 [D1]
1akj_A
G-ALPHA2 [D2]
[D3]
MHC class I

Figure 7: Concepts of numerotation. Ribbon representations and IMGT Colliers de Perles for V type, C type and G type domains, based on the IMGTunique numbering (NUMEROTATION axiom). (**A**) V type IgSF domain [12]. (**B**) C type IgSF domain [13]. (**C**) G type MhcSF domain [14]. Amino acids are shown in the one-letter abbreviation. Hatched circles correspond to missing positions according to the IMGTunique numbering.

the chain type or the species [10–14]. Structural and functional domains of the IG and TR chains comprise the V-DOMAIN (9-strand beta-sandwich framework), which corresponds to the V-J-REGION or V-D-J-REGION and is coded by two or three genes [5, 6] and the constant domain or C-DOMAIN (7-strand beta-sandwich framework), which corresponds to the C-REGION (CL of IG-Light) or is part of it (CH1, CH2 and CH3 of IG-Heavy) (Figure 4). In the IMGT unique numbering, conserved amino acids from frameworks always have the same number whatever the IG or TR domain, and whatever the species they come from. As examples: cysteine 23 (B-STRAND), tryptophan 41 (C-STRAND), leucine (or other hydrophobic amino acid) 89 (E-STRAND) and cysteine 104 (F-STRAND) [12, 13].

The IMGT unique numbering has been extended to the V-LIKE-DOMAINs and to the C-LIKE-DOMAINs of IgSF proteins other than IG and TR [12, 13]. More recently, the IMGT unique numbering has also been defined for the groove domain or G-DOMAIN (four beta-strands and one alpha-helix) of the MHC class I and II chains [14, 30, 31] and for the G-LIKE-DOMAINs of MhcSF proteins other than MHC [14]. The IMGT unique numbering has made possible standardized comparisons of the sequences and structures of the V type (V-set) and C type (C-set) IgSF domains [32–34] and of the G type (G-set) MhcSF domains [35, 36], whatever the species, the receptor or the chain type.

IMGT Collier de Perles

IMGT Colliers de Perles can be drawn from user amino acid sequences, after being gapped according to the IMGT unique numbering (for example with IMGT/DomainGapAlign), using the IMGT/Collier-de-Perles tool on the IMGT Web site at http://imgt.cines.fr. The IMGT Colliers de Perles are used in antibody humanization design based on complementarity determining regions (CDR) grafting, to precisely define the CDR-IMGT to be grafted, and in antibody engineering [37, 38]. IMGT Colliers de Perles statistical profiles for the human expressed IGHV, IGKV and IGLV repertoires were established [39], they help to identify potential immunogenic residues at given positions in chimeric or humanized antibodies [37]. In IMGT/3Dstructure-DB, IMGT Colliers de Perles give access to the IMGT Residue@Position cards, which provide the atom contact types and atom contact categories [9]. They bridge the gap between linear amino acid sequences and 3D structures, as illustrated by the display of hydrogen bonds (for V and C type domains) [9] and pMHC contact sites (for G type domains) [9, 30, 31].

IMPACT AND PERSPECTIVES

The concepts of IMGT-ONTOLOGY led to the setting up of international standards: gene and allele nomenclature by the WHO-IUIS/IMGT Nomenclature Committees, and standardized monoclonal antibody definitions by the WHO-International Nonproprietary Names (INN) Programme. The IMGT IG and TR genes names are used in the genome databases at NCBI (Entrez Gene), EBI (Ensembl) and Wellcome Trust Sanger Institute (UK). Many research laboratories, including pharmaceutical companies, journals, European and American networks and scientific societies advocate the use of the IMGT databases and tools, as they provide the international standards for nomenclature, description and numerotation (IMGT Scientific chart, http://imgt.cines.fr), based on the IMGT-ONTOLOGY concepts [3, 4, 40]. In December 2007, the Antibody Society has acknowledged IMGT as the molecular informatics framework in antibody engineering and antibody humanization. Indeed, IMGT standardized results represent a crucial step in the evaluation of homologies between monoclonal antibodies and their closest germline human counterparts and hence their possible immunogenicity [37].

Beyond these direct applications, the IMGT concepts are used for the exchange and the sharing of knowledge in very diverse academic and industrial fields of research: (i) fundamental and medical research (repertoire analysis of the IG antibody sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukaemias, lymphomas, myelomas), (ii) veterinary research (IG and TR repertoires in farm and wild life species), (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) molecular biotechnology related to antibody engineering and antibody therapeutics (scFv, phage displays, combinatorial libraries, chimeric, humanized and human antibodies), (vi) diagnostics (clonalities, detection and follow-up of residual diseases) and (vii) therapeutical approaches (grafts, immunotherapy, vaccinology).

The concepts of IMGT-ONTOLOGY are available, for the users of IMGT and the biologists in general, in natural language in IMGT Scientific chart (http://imgt.cines.fr) [8], and have been formalized for programming purpose in IMGT-ML (XML Schema) [41]. IMGT-ONTOLOGY is being implemented in Protégé and OBO-Edit to facilitate the export in formats such as OWL, and to link, whenever possible, the concepts of

IMGT-ONTOLOGY to those of other ontologies in biology such as the Gene Ontology (GO) [42], and in immunology, such as the Immunome Epitope database and Analysis Resource (IEDB) [43] and other Open Biomedical Ontologies (OBO) (http://obo.sourceforge.net).

The huge amount of immunological experimental data continues to grow exponentially and necessitates to be managed and analysed computationally. This is the goal of immunoinformatics, a science that implements the bioinformatics methodologies to answer these needs. At the same time, standardized representation of genomic, genetic, proteomic and 3D structural data is required to organize immunogenetics knowledge towards system biology and for the modelling and a better understanding of the immune system. As the same axioms can be used to generate concepts for multi-scale level approaches, the Formal IMGT-ONTOLOGY represents a paradigm for the elaboration of ontologies in system biology, which requires to identify, to describe, to classify, to numerotate, to localize, to orientate and to determine the obtaining and evolution of biological knowledge from molecule to population, in time and space. In that perspective, IMGT-ONTOLOGY represents a key component in the elaboration and setting up of standards of the European ImmunoGrid project (http://www.immunogrid.org/) whose aim is to define the essential concepts for modelling of the immune system.

Key Points

- The Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope comprises seven axioms that postulate that biological objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated and that the way they are obtained has to be determined.
- The 'IDENTIFICATION' axiom of the Formal IMGT-ONTOLOGY and the generated concepts of identification allowed to define the IMGT standardized keywords and their relations.
- The 'DESCRIPTION' axiom of the Formal IMGT-ONTOLOGY and the generated concepts of description allowed to define the IMGT standardized labels and their relations.
- The 'CLASSIFICATION' axiom of the Formal IMGT-ONTOLOGY and the generated concepts of classification allowed to set up the IMGT standardized nomenclature for the genes and alleles that is the international reference.
- The 'NUMEROTATION' axiom of the Formal IMGT-ONTOLOGY and the generated concepts of numerotation allowed to set up the IMGT unique numbering for V, C and G type domains and their graphical representation or IMGT Collier de Perles.

Acknowledgements
We thank Chantal Ginestoux for the figures. We are grateful to Gérard Lefranc for helpful discussion and to the IMGT team for its expertise and constant motivation. This work was funded by Centre National de la Recherche Scientifique CNRS; Ministère de l'Enseignement Supérieur et de la Recherche MESR (ACI-IMPBIO IMP82-2004, Université Montpellier 2 Plan Pluri-Formation); Région Languedoc-Roussillon; Agence Nationale de la Recherche ANR BYOSIS (ANR-06-BYOS-0005-01); European Union ImmunoGrid project (IST-2004-0280069).

References

1. Lefranc M-P, Giudicelli V, Kaas Q, *et al*. IMGT, the international ImMunoGeneTics information system®. *Nucleic Acids Res* 2005;**33**:D593–7.
2. Lefranc M-P, Clément O, Kaas Q, *et al*. IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol* 2005;**5**:45–60.
3. Giudicelli V, Lefranc M-P. Ontology for Immunogenetics: IMGT-ONTOLOGY. *Bioinformatics* 1999;**15**:1047–54.
4. Duroux P, Kaas Q, Brochet X, *et al*. IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie* 2007 Sep 11. [Epub ahead of print.]
5. Lefranc M-P, Lefranc G. *The Immunoglobulin FactsBook*. London: Academic Press, 2001;1–458.
6. Lefranc M-P, Lefranc G. *The T Cell Receptor FactsBook*. London: Academic Press, 2001;1–398.
7. Sakano H, Huppi K, Heinrich G, *et al*. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* 1979;**280**:288–94.
8. Lefranc M-P, Giudicelli V, Ginestoux C, *et al*. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics, http://imgt.cines.fr. *In Silico Biol* 2004;**4**:17–29.
9. Kaas Q, Ruiz M, Lefranc M-P. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res* 2004;**32**:D208–10.
10. Lefranc M-P. Unique database numbering system for immunogenetic analysis. *Immunol Today* 1997;**18**:509.
11. Lefranc M-P. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist* 1999;**7**:132–6.
12. Lefranc M-P, Pommié C, Ruiz M, *et al*. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 2003;**27**:55–77.
13. Lefranc M-P, Pommié C, Kaas Q, *et al*. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev Comp Immunol* 2005;**29**:185–203.
14. Lefranc M-P, Duprat E, Kaas Q, *et al*. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* 2005;**29**:917–38.
15. Giudicelli V, Ginestoux C, Folch G, *et al*. IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res* 2006;**34**:D781–4.

16. Giudicelli V, Chaume D, Jabado-Michaloud J, *et al*. Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. *Stud Health Technol Inform* 2005;**116**:3–8.
17. Eilbeck K, Lewis SE. Sequence ontology annotation guide. *Comp Funct Genomics* 2004;**5**:642–7.
18. Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* 2005; **33**:D256–61.
19. Wain HM, Bruford EA, Lovering RC, *et al*. Guidelines for human gene nomenclature. *Genomics* 2002;**79**:464–70.
20. Maglott D, Ostell J, Pruitt KD, *et al*. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2007;**35**: D26–31.
21. Lefranc M-P. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev Comp Immunol* 2008;**32**:461–63.
22. Lefranc M-P. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. *Immunogenetics* 2007;**59**:899–902.
23. Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucl Acids Res* 2004;**32**:W435–40.
24. Yousfi Monod M, Giudicelli V, Chaume D, *et al*. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* 2004;**20**:i379–85.
25. Belessi CJ, Davi FB, Stamatopoulos KE, *et al*. IGHV gene insertions and deletions in chronic lymphocytic leukemia: 'CLL-biased' deletions in a subset of cases with stereotyped receptors. *Eur J Immunol* 2006;**36**:1963–74.
26. Ghia P, Stamatopoulos K, Belessi C, *et al*. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. *Leukemia* 2007; **21**:1–3.
27. Ruiz M, Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* 2002;**53**:857–83.
28. Kaas Q, Lefranc M-P. IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics* 2007;**2**:21–30.
29. Kaas Q, Ehrenmann F, Lefranc M-P. IG, TR and IgSf, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? *Brief Funct Genomics and Proteomics* 2007;**6**: 253–64.
30. Kaas Q, Lefranc M-P. T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol* 2005; **5**:505–28.
31. Kaas Q, Duprat E, Tourneur G, *et al*. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: Schoenbach C, Ranganathan S, Brusic V (eds). *Immunoinformatics, Immunomics Reviews, Series of Springer Science and Business Media LLC*. New York, USA: Springer, 2008;19–49.
32. Duprat E, Kaas Q, Garelle V, *et al*. IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. *Recent Res Devel Human Genet* 2004;**2**:111–36.
33. Bertrand G, Duprat E, Lefranc M-P, *et al*. Characterization of human FCGR3B*02 (HNA-1B, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. *Tissue Antigens* 2004;**64**:119–31.
34. Garapati VP, Lefranc M-P. IMGT Colliers de Perles and IgSF domain standardization for T cell costimulatory activatory (CD28, ICOS) and inhibitory (CTLA4, PDCD1 and BTLA) receptors. *Dev Comp Immunol* 2007;**31**:1050–72.
35. Frigoul A, Lefranc M-P. MICA: standardized IMGT allele nomenclature, polymorphisms and diseases. *Recent Res Devel Human Genet* 2005;**3**:95–145.
36. Duprat E, Lefranc M-P, Gascuel O. A simple method to predict protein binding from aligned sequences - application to MHC superfamily and beta2-microglobulin. *Bioinformatics* 2006;**22**:453–9.
37. Magdelaine-Beuzelin C, Kaas Q, Wehbi V, *et al*. Structure-function relationships of the variable domains of monoclonal antibodies approved for cancer treatment. *Crit Rev Oncol/Hematol* 2007;**64**:210–25.
38. Laffly E, Danjou L, Condemine F, *et al*. Selection of a macaque Fab with human-like framework regions, high affinity, and that neutralizes the protective antigen (PA) of *Bacillus anthracis*. *Antimicrob Agents Chemother* 2005;**49**: 3414–20.
39. Pommié C, Levadoux S, Sabatier R, *et al*. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* 2004; **17**:17–32.
40. Lefranc M-P. IMGT-ONTOLOGY, IMGT® databases, tools and Web resources for Immunoinformatics. In: Schoenbach C, Ranganathan S, Brusic V (eds). *Immunoinformatics, Immunomics Reviews, Series of Springer Science and Business Media LLC*. New York, USA: Springer, 2008;1–18.
41. Chaume D, Giudicelli V, Lefranc M-P. IMGT-ML a XML language for IMGT-ONTOLOGY and IMGT/LIGM-DB data. In: *Proceedings of NETTAB 2001*, pp. 71–5. Network Tools and Applications in Biology, Genoa, Italy, 17–18 May 2001.
42. The Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 2006;**34**:D322–6.
43. Peters B, Sidney J, Bourne P, *et al*. The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol* 2005;**3**:e91.