Information Society Technologies

ImmunoGrid
The European Virtual Human Immune System Project

D1.2
Scientific chart rules and ontologies report

Date: *Aug 15, 2007*

Due Date: *Sept 15, 2007*

Principal Authors: Marie-Paule Lefranc, François Ehrenmann, Patrice Duroux and Véronique Giudicelli (**CNRS**); CINECA, CNRS, UQ, DTU, BIRKBECK, UNIBO, UNICT

Revision: 1.0

Content

Section 1. Introduction

The immune system has evolved to preserve the integrity of the organism (self) and to control infection by pathogens (parasites, bacteria, virus) and abnormal cell proliferation (cancer, allergy, autoimmune diseases). The immune system is characterized by a dynamic interplay and by a fine-tuned balance at the organism, cellular and molecular level. At the molecular level, the study of the immune responses includes the study of the genes and proteins involved in the innate immune responses of every living organism, and in the adaptive immune responses of the vertebrates. Molecular immunogenetics has blossomed considerably since 1979 when Tonegawa showed that the vast and extremely diverse repertoire of antigen receptors (10^{12} immunoglobulins (IG) or antibodies, and 10^{12} different T cell receptors (TR) per individual, in human) results from complex rearrangement mechanisms at the DNA level [1-3]. To the complexity of these mechanisms, must be also added the incredible diversity due to the somatic hypermutations of IG and a considerable polymorphism of major histocompatibility complex (MHC) - designated as HLA, in humans - which is particularly important for bone marrow and organ transplantation.

The focus of WP1 "Immune system standardized concepts" is the setting up of the standardized rules and concepts which are part of the identification, description and classification of the biological components and processes, in the modelling of the "Virtual Immune System" (VIS).

The first deliverable D1.1 provided the new and enhanced concepts and rules necessary for the VIS modelling. This deliverable D1.2 formalizes these new and enhanced concepts and rules for the identification, description and classification of the antigen receptors (IG, TR) and MHC, that are major molecular components of the "Virtual Immune System" modelling. The IG, TR and MHC proteins are 450 to 500 million years “old” and are characteristic of the adaptive immune responses in vertebrates. They allow a very fine specific recognition of the "non self" represented by infectious pathogens, viruses, bacteria, parasites and their products (toxins...), and by vaccine and tumoral antigens. These complex and heterogenous data are managed in IMGT® (http://imgt.cines.fr), the flagship of Europe in Immunogenetics and immunoinformatics (BIOMED, BIOTECH, 5th PCRDT) [4, 5] and a key component of the ImmunoGrid project (http://www.immunogrid.org/).

The D1.2 standardization is based on IMGT-ONTOLOGY [6], the first and so far unique ontology in immunogenetics and immunoinformatics. An ontology is a formal representation of a knowledge domain [7-12]. IMGT-ONTOLOGY provides a semantic specification of the terms to be used in immunogenetics and immunoinformatics and manages the related knowledge [13-15], thus allowing the standardization for immunogenetics data from genome, proteome, genetics and three-dimensional (3D) structures [16-27]. IMGT-ONTOLOGY results from a deep expertise in the domain and an extensive effort of conceptualization.

Novelty resides in the emergence, identification and characterization of new standards and concepts in IMGT-ONTOLOGY that are required for a systemic approach of the adaptive immune responses and that can represent the corresponding knowledge in other fields of biology. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope.

Three of these axioms, "IDENTIFICATION", "DESCRIPTION", "CLASSIFICATION" are presented in this deliverable, with the concepts that have been essential for the conceptualization of the molecular immunogenetics knowledge. As the same axioms can be used to generate concepts for multi-scale level approaches, the Formal IMGT-ONTOLOGY represents a paradigm for system biology ontologies, which need to identify, to describe and to classify objects, processes and relations at the molecule, cell, tissue, organ, organism or population levels.

The concepts are currently being formalized in the Ontology Web Language OWL (http://www.w3.org/2004/OWL/), using the Protégé editor (http://protege.stanford.edu/) [28]. This formalization will allow interoperability of the ImmunoGrid components with other major medical or biological ontologies (Unified Medical language System UMLS, Gene Ontology GO, Sequence Ontology SO, Ontology for Biomedical Investigations (OBI) (formerly Functional Genomics Investigation Ontology FuGO), MGED Network Ontology Working Group, Immunome Epitope database and Analysis Resource IEDB ontology, etc.) [29-32].

Section 2. The ImmunoGrid project and the IMGT-ONTOLOGY axioms and concepts

2.1. The Formal IMGT-ONTOLOGY or IMGT Kaleidoscope

IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets relying on seven axioms, "IDENTIFICATION", "CLASSIFICATION", "DESCRIPTION", "LOCALIZATION", "NUMEROTATION", "ORIENTATION" and "OBTENTION". These axioms 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 (Fig. 1). The axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope.

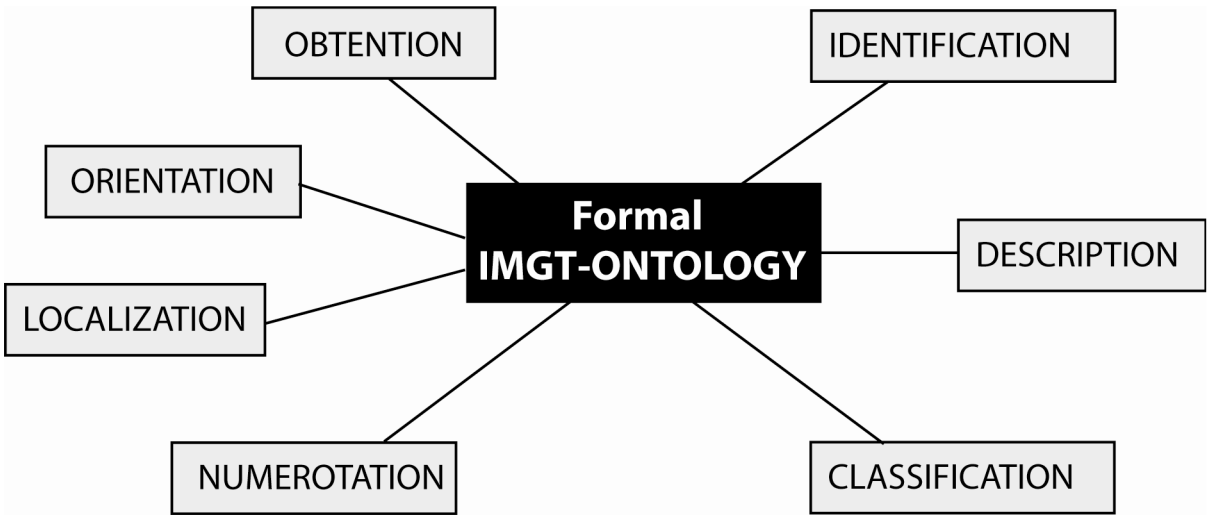

Fig. 1. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope.

Each axiom gives rise to a set of concepts. Concepts are general in the reality [7-15]. Concept instances correspond to all possible examples of representation of a concept at a given granularity. A concept may be exemplified by one or several concept instances. New concept instances may be defined with the advancement of science. Concepts are linked by relations, the simplest being "is_a" which represents the edge between concepts at different levels of granularity and organizes the main hierarchy of IMGT-ONTOLOGY. Properties are semantic characteristics of a concept or of a concept instance: they may be simple attributes as a name alias, or they may be specific relations between concepts and instances across the main hierarchy. These relations are fundamental since they reveal strong semantic constraints and dependencies on which relies the coherence within or between IMGT® components.

The same term can be used to define a concept in the ontology and a class in the databases, however ontologies (e.g., IMGT-ONTOLOGY) only contain concepts (and examples or concept instances) and their relations which are virtual representations, whereas databases (e.g. IMGT/LIGM-DB) contain classes (and class instances) which are data objects (Table 1). In the setting up of the Virtual Immune System this distinction is fundamental as the modelling is not dependent of the data description but of the concepts, and can therefore be applied to data from any database provided that they share the same ontology.

"general in reality" terms in ontologies (e.g. IMGT-ONTOLOGY)	"particular in reality" terms in databases (e.g. IMGT/LIGM-DB)
concepts concept instances[1]	Classes class instances[1]

1 When the context (ontology or database) is unambiguous, the word "instances" can be used without being preceded by the word "concept" or "class"

Table 1. Terms used in ontologies and databases.

IMGT-ONTOLOGY provides a semantic classification and standardization of the knowledge in the field of immunogenetics in order to identify data, to classify them, to describe them in detail, to number, to localize and to orientate them, and finally to define in which experimental, biological or medical context the sequences have been obtained.

2.2. Implementation

The main hierarchy of the IMGT-ONTOLOGY concepts has previously been described [6, 13-15]. We have analysed these concepts in the light of the ImmunoGrid project.

IMGT-ONTOLOGY concepts are available for the biologists and IMGT® users in natural language in the IMGT Scientific chart [4,20], and have been formalized for programming purpose in IMGT-ML [33, 34] which is an XML Schema (http://www.w3.org/TR/xmlschema-0/). In order to formalize the semantic relations between concepts and instances that are essential for high-quality data processing and coherence control, IMGT-ONTOLOGY is currently designed with Protégé [28] and OBO-Edit (http://oboedit.org/), that are frequently used ontology editors for biological ontologies. Protégé and OBO-Edit ontologies can be exported into RDF

(http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/) and OWL (http://www.w3.org/2004/OWL/) which allow interoperability with other ontologies.

The IMGT-ONTOLOGY axioms and their related IMGT Scientific chart rules at the molecular level defined and formalized in this D1.2 deliverable are listed in Table 2. This deliverable represents the state of the art for the IDENTIFICATION, DESCRIPTION and CLASSIFICATION axioms. These axioms are particularly important as they represent the crucial step of the ImmunoGrid approach, linked to the specificity of the immune response (antigen recognition, specificity antigen-receptor, B cell epitope and T cell epitope characterization, peptides used in vaccinology and immunotherapy, humanized antibodies used in cancerology, etc). Immune responses at the cellular level and organism level depend on this molecular level, whose component interactions trigger the whole cascade of events.

IMGT-ONTOLOGY axioms	**IMGT Scientific chart** rules[1]	Examples in ImmunoGrid[2]
IDENTIFICATION	Keywords	IG, TR, MHC and RPI nucleotide and amino acid sequence identification IG, TR, MHC and RPI 3D structure identification
DESCRIPTION	Labels	IG, TR, MHC and RPI nucleotide and amino acid sequence description IG, TR, MHC and RPI 3D structure description
CLASSIFICATION	Nomenclature	IG, TR, MHC and RPI gene and allele names

1 The corresponding controlled vocabulary and rules are available in the IMGT Scientific Chart at http://imgt.cines.fr.

2 Some examples are given in annexes. RPI: related proteins of the immune system.

Table 2. IDENTIFICATION, DESCRIPTION and CLASSIFICATION axioms of the Formal IMGT-ONTOLOGY. The main related IMGT Scientific chart rules correspond to the ImmunoGrid concepts required for the modelling of the immune system at the molecular level.

2.3. An example of knowledge at the molecular level: the immunoglobulin synthesis

The immunoglobulin synthesis, an example of knowledge at the molecular level, is used here to define the concepts generated by the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope in the D1.2 deliverable. The concepts of identification (IDENTIFICATION axiom) identify the nucleotide and protein sequences and the 3D structures according to a structured terminology, the concepts of description (DESCRIPTION axiom) describe the composition of the sequences and structures with standardized labels, and the concepts of classification (CLASSIFICATION axiom) classify the genes and alleles with a standardized nomenclature.

An IG or antibody is composed of two identical heavy chains associated with two identical light chains, kappa or lambda. In humans, heavy chain genes (locus IGH), light chain kappa genes (locus IGK) and light chain lambda genes (locus IGL) are located on the chromosomes 14 (14q32.3), 2 (2p11.2) and 22 (22q11.2), respectively. The synthesis of an immunoglobulin requires rearrangements of the IGH, IGK and IGL genes during the differentiation of the B lymphocytes.

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). The configuration of the V-gene, D-gene and J-gene is identified as "germline" (Fig. 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-Chain) and the light chain (IG-Light-Chain) of an IG (or antibody) (Fig.2).

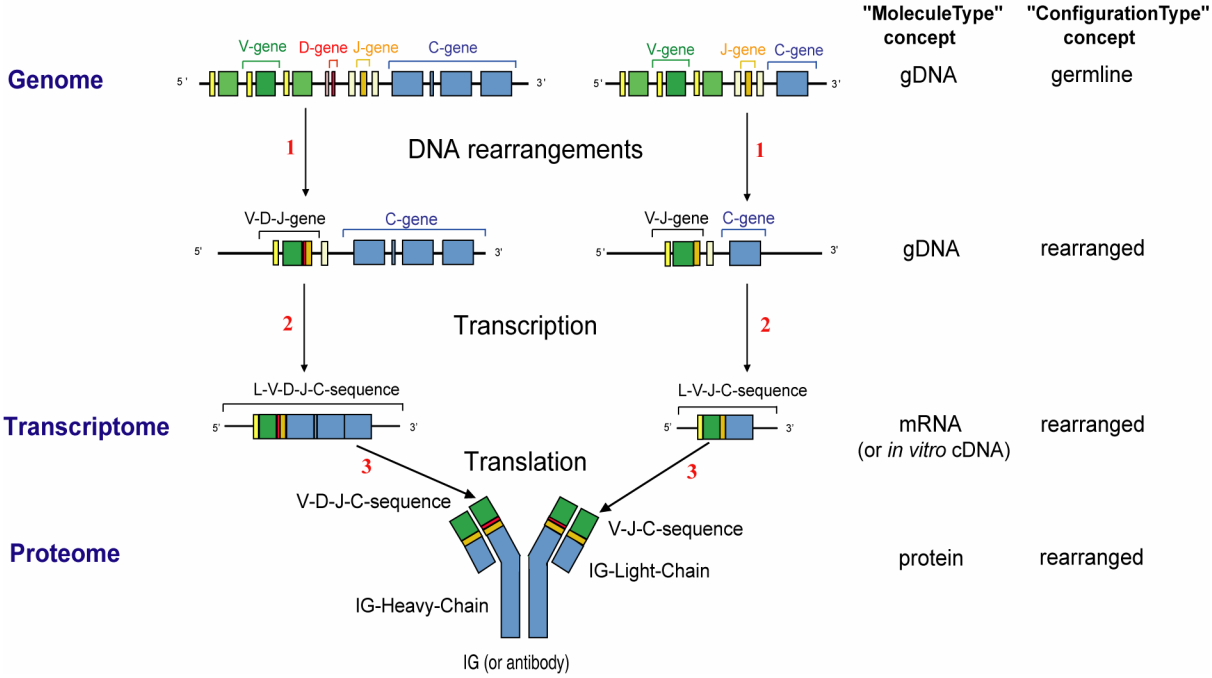

Fig. 2. An example of knowledge at the molecular level: the synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies [2]. 1: DNA rearrangements (is_rearranged_into), 2: Transcription (is_transcribed_into), 3: Translation (is_translated_into). The configuration of C-GENE is undefined.

The variable domains VH and VL are coded by the V-D-J-REGION and the V-J-REGION (Fig. 3). Each domain includes four framework regions (FR) (in pale blue in Fig.3) and three hypervariable loops or complementarity determining regions (CDR). The CDR, and more particularly the CDR3 that result from the junction of the V-D-J genes (in the VH domain) and V-J genes (in the VL domain), are involved in the antigen recognition. The VH and VL amino acids in contact with the antigen constitute the paratope. The part of the antigen recognized by the antibody is the epitope. The number of potential V-D-J and V-J

rearrangements depends on the number of functional V, D and J genes in the genome. Additional mechanisms (N diversity at the V-D-J and V-J junctions and somatic hypermutations) allow to reach 10^{12} different antibodies per individual [2] (IMGT®, http://imgt.cines.fr).

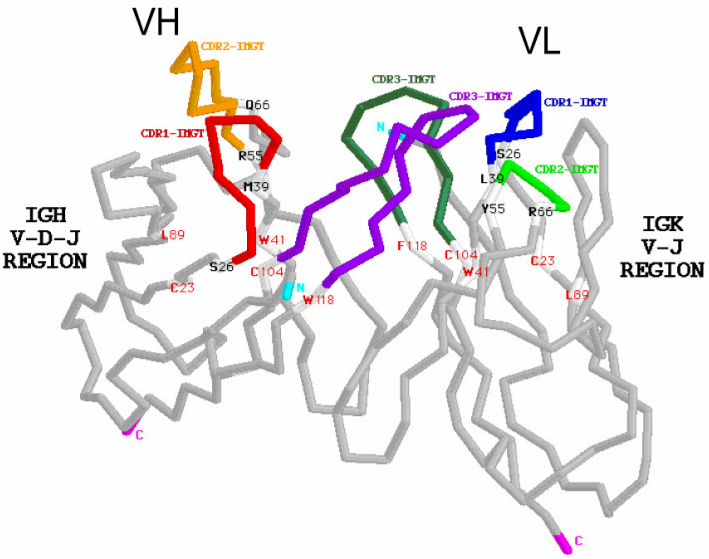

Fig. 3. The variable domains VH and VL of the heavy and light chains of an IG or antibody. VH CDR1-IMGT is in red, CDR2-IMGT in orange and CDR3-IMGT in purple. VL CDR1-IMGT is in blue, CDR2-IMGT in green and CDR3-IMGT in greenblue (IMGT color menu).

Section 3. Necessity of identification: the IDENTIFICATION axiom

The IDENTIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and relations, have to be identified. 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 molecular biology, the concepts of identification allow to identify the molecules, their processes and their relations at the genome, transcriptome and proteome levels.

3.1. Identification of an organism: the "Taxon" concept

The "Taxon" concept allows to identify the type of taxon in which an object, process or relation is found. The "Taxon" concept manages a hierarchy of concepts at various levels of granularity. The corresponding hierarchical taxonomy is that provided by the National Center for Biotechnology Information NCBI (http://www.ncbi.nlm.nih.gov/) up to the rank of species ("Species" concept) and subspecies ("Subspecies" concept) in order to establish complete interoperability with generalist databases. Since IG, TR and MHC genes are only

present in jawed vertebrates (gnathostoma), only vertebrate species were originally represented in IMGT-ONTOLOGY. However, with the extension of IMGT-ONTOLOGY to the IgSF and MhcSF, invertebrate species are incorporated whenever necessary. The "EthnicGroup", "Breed" and "Strain" concepts have been added to IMGT-ONTOLOGY to allow the identification of data specific to ethnic groups for humans (http://www.ebi.ac.uk/imgt/hla/help/ethnic_help.html), breeds for domestic animals or strains for laboratory and wild animals.

3.2. Identification of an entity: the "EntityType" concept

The "EntityType" concept identifies the type of entity. An entity can be a molecule, a cell, a tissue, an organ, an organism or a population. The molecules identified in the ImmunoGrid project are the IG, TR, MHC and RPI. Examples of concepts of identification required to identify them precisely are defined below as examples.

If the object is a molecule, the "EntityType" concept is designated as "Molecule_EntityType", which is defined by the "MoleculeType", "GeneType" and "ConfigurationType" concepts of identification and has properties identified in the "Functionality" and "StructureType" concepts (Fig. 4).

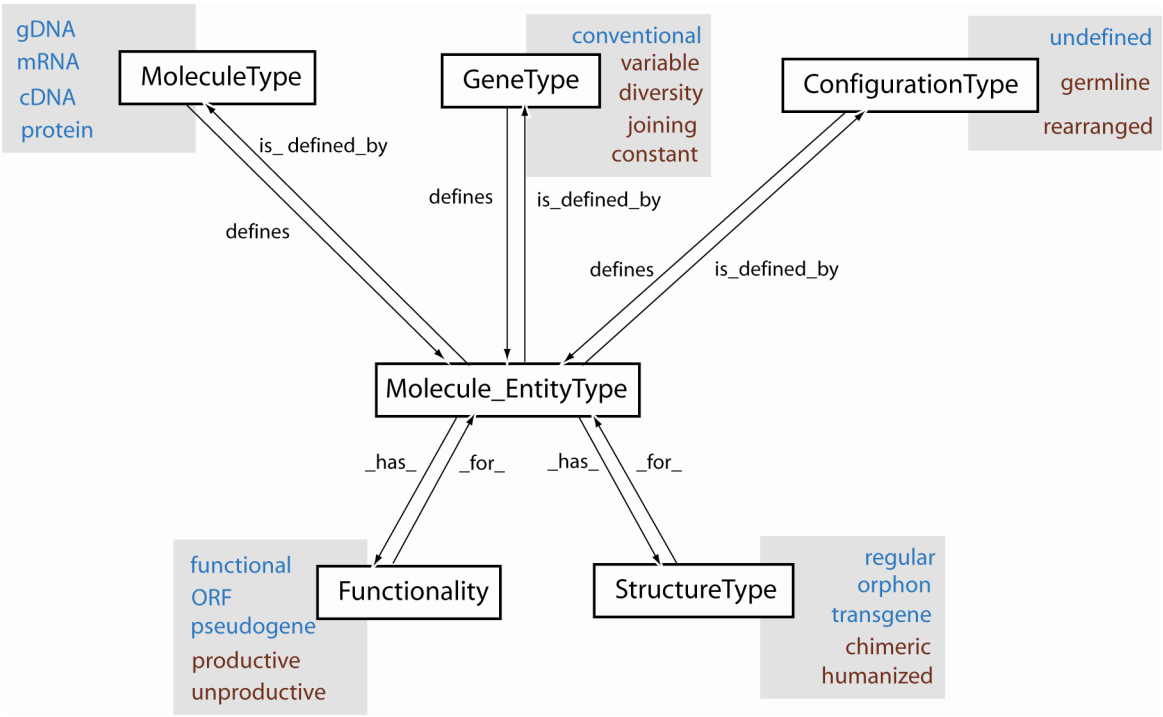

Fig. 4. The "Molecule_EntityType" concept. The "Molecule_EntityType" concept is defined by the "MoleculeType", "GeneType" and "ConfigurationType" concepts of identification and has properties identified in the "Functionality" and "StructureType" concepts (IDENTIFICATION axiom). Arrows indicate the reciprocal relations "is_defined_by" and "defines", "_has_" and "_for_". Concept instances which are general are in blue, those which are specific of the IG and TR are in red. The "Molecule_EntityType" concept instances comprise nineteen instances which are listed in Table 3. Only a few examples of the "StructureType" concept instances are shown (other instances are quoted in 2.2.6 and in Table 4).

3.2.1. The "MoleculeType" concept

The "MoleculeType" concept identifies the type of molecule based on the type of the constitutive elements and on the concepts of obtention (not detailed here). The four main instances of the "MoleculeType" concept are 'gDNA' (genomic DNA, a nucleotide sequence made of A, T, C, G, obtained from a genome), 'mRNA' (messenger RNA or transcript, a nucleotide sequence made of A, U, C, G, obtained by transcription of a genomic DNA), 'cDNA' (complementary DNA, a nucleotide sequence made of A, T, C, G, obtained *in vitro* by reverse transcription of the messenger RNA) and 'protein' (a sequence made of amino acids, obtained by translation of a transcript).

Thus, the instances of the "MoleculeType" concept allow to identify a sequence: nucleotide sequence that can be either genomic ('gDNA') or a transcrit ('mRNA', 'cDNA'), and amino acid sequence ('protein').

3.2.2. The "GeneType" concept

The "GeneType" concept identifies the type of gene and comprises five instances (Fig. 4). The first instance, 'conventional', refers to any (coding or not coding) gene other than IG or TR genes. The other four instances are specific to immunogenetics: 'variable' (V), 'diversity' (D) and 'joining' (J) gene types that rearrange at the DNA level and code the variable domains of IG and TR, and 'constant' (C) gene type that codes the constant region of IG and TR [2,3].

3.2.3. The "ConfigurationType" concept.

The "ConfigurationType" concept identifies the type of gene configuration and comprises three instances (Fig. 4). The instance 'undefined' identifies the configuration of the conventional and of the constant (C) genes. The instances 'germline' and 'rearranged' identify the status of the V, D and J genes, before and after DNA rearrangements, respectively [2,3].

3.2.4. The "Molecule_EntityType" concept.

The "Molecule_EntityType" concept, defined by the "MoleculeType", "GeneType" and "ConfigurationType" concepts, includes 19 instances (Table 3).

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.

Sixteen instances allow to identify the IG and TR. Ten of them are represented in Fig. 2:

- six for the gDNA ('V-gene', 'D-gene', 'J-gene', 'C-gene', 'V-D-J-gene' and 'V-J-gene'),
- two for the mRNA, 'L-V-D-J-C-sequence' and 'L-V-J-C-sequence', also valid for cDNA,
- and two for the protein, 'V-D-J-C-sequence' and 'V-J-C-sequence'.
- For examples:
- the instance 'V-gene' identifies a gDNA ("MoleculeType") containing a gene V ("GeneType"), in germline configuration ("ConfigurationType") (Table 3).
- the instance 'L-V-J-C-sequence' identifies a sequence of mRNA or cDNA ("MoleculeType") corresponding to V, J and C genes ("GeneType"), in rearranged configuration ("ConfigurationType") (Table 3).

The last six instances correspond to partial rearrangement ('D-J-gene') or to sterile transcripts ('L-V-sequence', 'D-sequence', 'J-sequence', 'J-C-sequence' and 'C-sequence').

	"Molecule_EntityType" concept instances	"GeneType" concept instances[1]	"MoleculeType" concept instances	"ConfigurationType" concept instances
Genomic sequence	V-gene	V	gDNA	germline
	D-gene	D		
	J-gene	J		
	C-gene	C		undefined
	gene	conventional		
	V-J-gene	V, J		rearranged
	V-D-J-gene	V, D, J		
	D-J-gene	D, J		
Transcript	L-V-J-C-sequence	V, J, C	mRNA (or cDNA)	
	L-V-D-J-C-sequence	V, D, J, C		
	L-V-sequence	V		germline
	D-sequence	D		
	J-sequence	J		
	J-C-sequence	J, C		
	C-sequence	C		undefined
	nt-sequence	conventional		
Amino acid sequence	AA-sequence	conventional	protein	undefined
	V-J-C-sequence	V, J, C		rearranged
	V-D-J-C-sequence	V, D, J, C		

1 V: variable, D: diversity, J: joining, C: constant

Table 3. "Molecule_EntityType" concept instances. Nineteen "Molecule_EntityType" concept instances are defined based on the "GeneType", "MoleculeType" and "ConfigurationType" concept instances (IDENTIFICATION).

3.2.5. The "Functionality" concept

The "Functionality" concept identifies the type of functionality for the "Molecule_EntityType" concept (Fig. 4). It includes five instances, divided into two categories, according to the configuration type.

Three instances, 'functional', 'ORF' (open reading frame) and 'pseudogene' identify the functionality of a "Molecule_EntityType" instance in undefined or germline configuration. They allow to identify the functionality of conventional genes, that of C genes, and that of V, D and J genes before their rearrangement in the genome, and by extension the functionality of their transcripts and proteins.

The two instances 'productive' and 'unproductive' identify the functionality of "Molecule_EntityType" instances in rearranged configuration. They allow to identify the functionality of IG and TR entities after their rearrangement in the genome, that of fusion genes resulting from translocations and that of hybrid genes obtained by biotechnology molecular engineering, and by extension the functionality of their transcripts and proteins.

3.2.6. The "StructureType" concept

The "StructureType" concept identifies the structure for the "Molecule_EntityType" concept. This concept allows to identify structures with a classical organization (‘regular’), from those which have been modified either naturally *in vivo* (‘orphon’, ‘processed orphon’, ‘unprocessed orphon’, ‘unspliced’, ‘partially spliced’, etc.), or artificially *in vitro* (‘chimeric’, ‘humanized’, transgene, etc.).

Table 4 shows the main instances of the "StructureType" concept which are relevant to the nineteen instances of the "Molecule_EntityType" concept. Other "StructureType" concept instances (‘engineered’, ‘fusion’, ‘transgene’, ‘translocated’, ‘unusual’, etc.) which are more general, and may apply if needed to several "Molecule_EntityType" concept instances, are available in IMGT Scientific chart http://imgt.cines.fr.

Definitions of the "Molecule_StructureType" concept instances are available from IMGT® http://imgt.cines.fr/cgi-bin/IMGTlect.jv?query=19. Examples of definitions for instances frequently used in antibody engineering are given below:

chimeric	defines a natural or synthetic IG or TR transcript or chain, (L)-V-J-C-sequence or (L)-V-D-J-C-sequence, obtained *in vitro* or *in vivo* from 2 sources. [1 source (murine, rat, ...) for V-J or V-D-J + 1 source (human) for C region].
humanized	defines a natural or synthetic human IG or TR gene, V-J-gene or V-D-J-gene, transcript or chain, (L)-V-J-C-sequence or (L)-V-D-J-C-sequence, modified *in vitro* with non-human recognition site sequences (CDR1,2,3). [1 source (murine, rat...) for recognition site sequences (CDR1,2,3) + 1 source (human)]
engineered	defines a gene or chain modified by deliberate mutagenesis *in vitro*. [1 source]
fusion	defines a gene, transcript or protein resulting from the *in vitro* fusion between two (or more) different genes. [2 (or more) sources]

"Molecule_EntityType" concept instances[1]	"Functionality" concept instances2	"StructureType" concept instances	
V-gene	- F/ORF/P	- regular/orphon	- orphon (processed/ unprocessed)
D-gene	- F/ORF/P	- regular/orphon	
J-gene	- F/ORF/P	- regular/orphon	
C-gene	- F/ORF/P	- regular/orphon	- orphon (processed/ unprocessed/ partially processed)
gene	- F/ORF/P	- regular/orphon	- orphon (processed/ unprocessed/ partially processed)
V-J-gene	- productive/ unproductive	- regular/humanized	
V-D-J-gene	- productive/ unproductive	- regular/humanized	
D-J-gene	- productive/ unproductive	- regular - partially rearranged	
L-V-J-C-sequence	- productive/ unproductive	- regular/chimeric/ humanized	- spliced/unspliced/ partially spliced
L-V-D-J-C-sequence	- productive/ unproductive	- regular/chimeric/ humanized	- spliced/unspliced/ partially spliced
L-V-sequence	- F/ORF/P	- sterile transcript	- spliced/unspliced/
D-sequence	- F/ORF/P	- sterile transcript	
J-sequence	- F/ORF/P	- sterile transcript	
J-C-sequence	- F/ORF/P	- sterile transcript	- spliced/unspliced/ partially spliced
C-sequence	- F/ORF/P	- sterile transcript	- spliced/unspliced/ partially spliced
nt-sequence	- F/ORF/P	- regular	- alternative splice/ spliced/unspliced/ partially spliced
AA-sequence	- F/ORF/P	- regular - protein fusion	
V-J-C-sequence	- productive/ unproductive	- regular/chimeric/ humanized	
V-D-J-C-sequence	- productive/ unproductive	- regular/chimeric/ humanized	

1 V: variable, D: diversity, J: joining, C: constant

2 F: functional, ORF: open reading frame, P: pseudogene

Table 4. "Molecule_StructureType" concept instances relevant to the nineteen "Molecule_EntityType" concept instances (IDENTIFICATION).

3.3. Identification of a receptor: the "ReceptorType" concept

The "ReceptorType" concept identifies the type of receptor. A receptor can be a molecule, a cell, a tissue, an organ, an organism or a population. If the object is a molecule, the "ReceptorType" concept is designated as "Molecule_ReceptorType" which is defined by the "ChainType" concept of identification and has properties identified in the "StructureType", "Specificity" and "Function" concepts (Fig. 5).

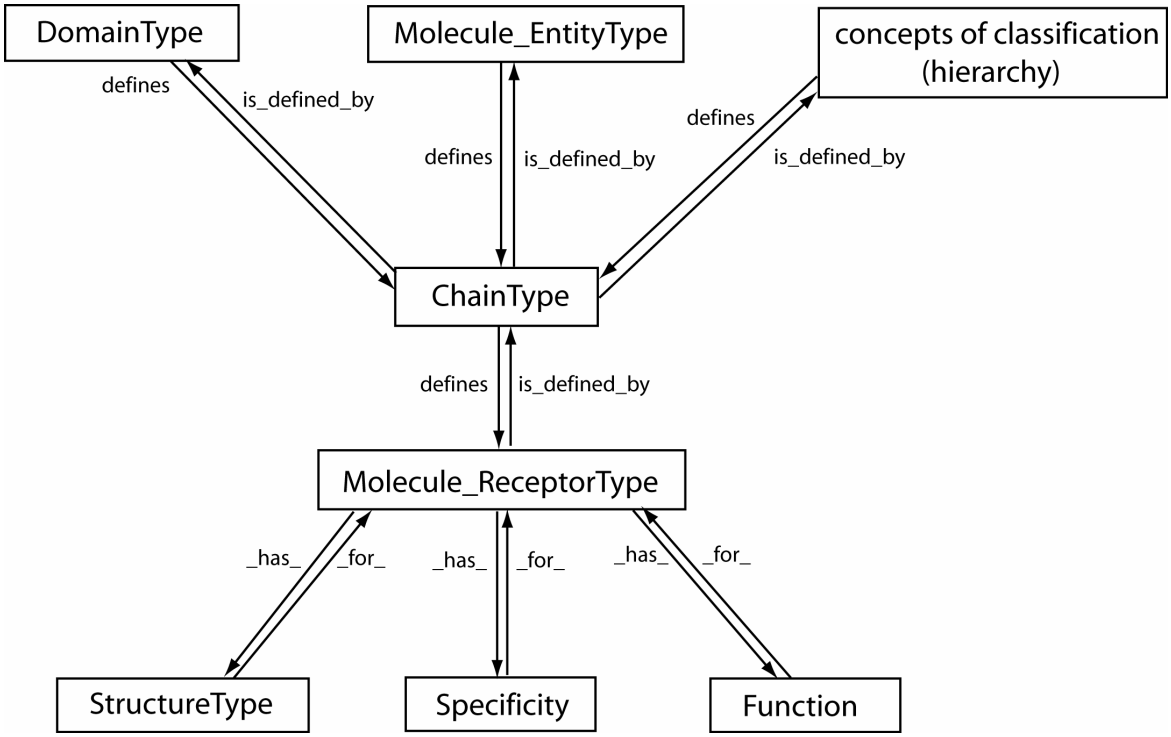

Fig. 5. The "Molecule_ReceptorType" concept. The "Molecule_ReceptorType" concept, defined by the "ChainType" concept of identification, has properties identified in the "StructureType", "Specificity" and "Function" concepts (IDENTIFICATION axiom). The "ChainType" concept is itself defined by the "Molecule_EntityType" and "DomainType" concepts and by concepts of classification (hierarchy). Arrows indicate reciprocal relations "is_defined_by" and "defines", "_has_" and "_for_". These concepts have different levels of granularity, up to six for "Molecule_ReceptorType" and "ChainType", as this is described in 2.3.1 and 2.3.2.

The "ChainType" concept is itself defined by the "Molecule_EntityType" and the "DomainType" concepts of identification and by concepts of classification (see CLASSIFICATION axiom, in section 4). These latter are organized in a hierarchy which confers different levels of granularity to the "Molecule_ReceptorType" and "ChainType" concepts, as this is described in 2.3.1 and 2.3.2.

3.3.1 The "Molecule_ReceptorType" concept.

The "Molecule_ReceptorType" concept identifies the type of protein receptor, defined by its chain composition. Thus, IG is an instance of the "Molecule_ReceptorType" concept, defined as comprising 4 chains, two heavy chains and two light chains, identical two by two and covalently linked (Fig. 6). A receptor may comprise one chain (monomer) or several associated chains (multimer).

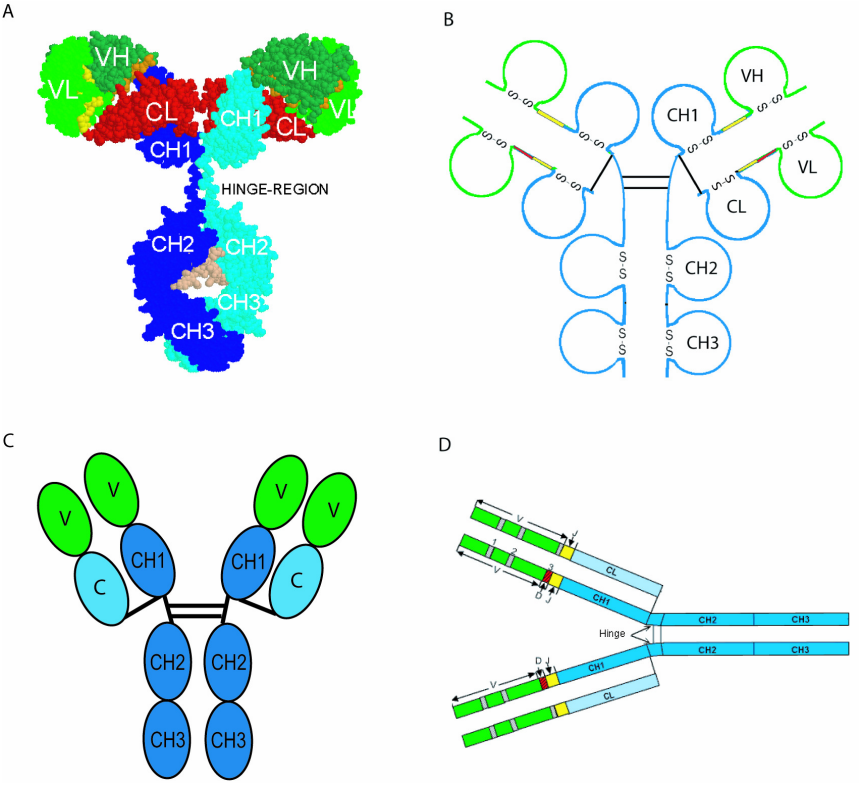

Fig. 6. Identification of an IG or antibody as an instance of the "Molecule_ReceptorType" concept made of four chains, two IG-Heavy-Chain and two IG-Light-Chain ("ChainType" concept). The four representations, although different, allow to identify an IG as a receptor of four chains, themselves organized in domains ("DomainType" concept). VH and VL are V type domains, coded by the V-D-J region and V-J region, respectively. CL, CH1, CH2 and CH3 are C type domains. (A) 3D structure, (B) organization in Ig-like domains, (C) organization in modules, (D) regions coded by the V, D, J and C gene types. The C gene type codes the constant region (CL for the IG-Light- Chain and CH1, hinge, CH2 and CH3 for the IG-Heavy-Chain). This representation, schematized as a Y shape, is frequently used to represent an IG.

The "Molecule_ReceptorType" concept contains a hierarchy of concept which identify the receptor type at different levels of granularity. Owing to the reciprocal relations "is_defined_by" and "defines" between the "Molecule_ReceptorType" and the "ChainType" concepts, a given level of granularity of the receptor type corresponds to a given level of granularity of the chain type. Three levels are general, whereas three additional intermediate levels are required for the MHC (two of them) and for the IG (one of them) (Fig. 7). The three general levels are the following:

- The "ReceptorLevelReceptorType" concept identifies a receptor type by reference to its constitutive chains defined at the level "ReceptorLevelChainType" (Ex: Integrin-Receptor).
- The "PartnerLevelReceptorType" concept identifies a receptor type by reference to its constitutive chains defined at the level "PartnerLevelChainType" (Ex: Integrin-Alpha_Beta-Receptor).
- The "GeneLevelReceptorType" concept identifies a receptor type by reference to its constitutive chains defined at the level "GeneLevelChainType" (Ex: Integrin-Alpha1_Beta1-Receptor).

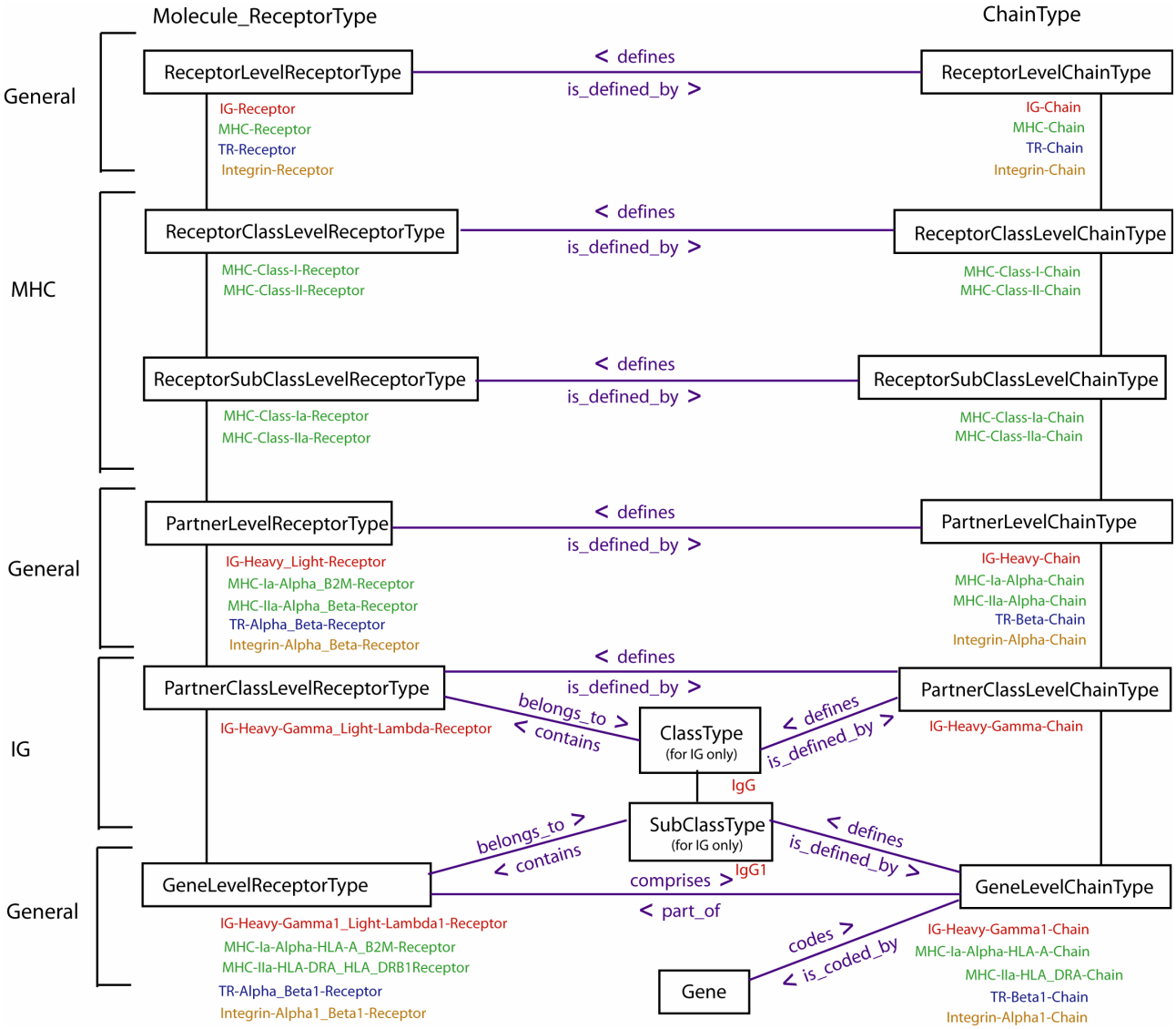

Fig. 7. "Molecule_ReceptorType" and "ChainType" concepts, their different levels of granularity and relations (IDENTIFICATION). For completeness, the relations with "ClassType" and "SubclassType", two concepts of identification specific of the IG, and the relation with the "Gene" concept (CLASSIFICATION) are shown. Examples of "Molecule_ReceptorType" and "ChainType" concept instances are shown for the IG (red), TR (blue), MHC (green) and Integrin chosen as RPI (orange).

The two intermediate levels required for the MHC are the following:

- The "ReceptorClassLevelReceptorType" concept identifies a MHC receptor type by reference to its constitutive chains defined at the level "ReceptorClassLevelChainType" (Ex: MHC-Class-I-Receptor).
- The "ReceptorSubClassLevelReceptorType" concept identifies a MHC receptor type by reference to its constitutive chains defined at the level "ReceptorSubClassLevelChainType" (Ex: MHC-Class-Ia-Receptor).

The intermediate level required for the IG is the following:

- The "PartnerClassLevelReceptorType" concept identifies an IG receptor type by reference to its constitutive chains defined at the level "PartnerClassChainType" (Ex: IG-Heavy-Gamma_Light-Lambda-Receptor).

3.3.2 The "ChainType" concept

The "ChainType" concept contains a hierarchy of concepts which identify the chain type at different levels of granularity. Three levels are general whereas three additional intermediate levels are required for the MHC (two of them) and one for the IG (one of them) as described above (Fig. 7). The three finest levels of granularity are defined by a concept of classification ('group', 'subgroup' and 'gene', respectively).

According to the hierarchy, the concepts are the following:

- The "ReceptorLevelChainType" concept identifies a chain type at the level of a receptor (Ex: Integrin-Chain).
- The "ReceptorClassLevelChainType" concept identifies a MHC chain type at the level of a MHC receptor class (Ex: MHC-Class-I-Chain).
- The "ReceptorSubClassLevelChainType" concept identifies a MHC chain type at the level of a MHC receptor subclass (Ex : MHC-Class-Ia-Chain).
- The "PartnerLevelChainType" concept identifies a chain type at the level of partner (Ex: Integrin-Alpha-Chain). This level is defined by the 'group' concept of classification.
- The "PartnerClassLevelChainType" concept identifies an IG chain type at the level of the partner class (Ex: IG-Heavy-Gamma-Chain). This level is defined by the 'subgroup' concept of classification.
- The "GeneLevelChainType" identifies a chain type at the level of the coding gene (Ex: Integrin-Alpha1-Chain). This level is defined by the 'gene' concept of classification.

The relation "codes" precises which gene codes for an instance of the "GeneLevelChainType" concept. The number of instances of the "GeneLevelChainType" concept depends on the number of functional and ORF genes per haploid genome in a given species (in the case of the IG and TR genes, 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 the concept correspond to the isotypes.

The "described_in_taxon" relation (not shown) allows to specify that some particular chain types have only been identified in some given species. This relation may be defined at any level of granularity of the "ChainType" concept.

Thanks to its different levels of granularity and relations with other concepts of identification and classification, the "ChainType" concept has been extremely useful to clarify the relations between biological knowledge. This concept is fundamental for modelling in ImmunoGrid and more generally for valid data interpretation in system biology.

3.3.3 The "DomainType" concept

A chain can be defined by its constitutive structural units ("DomainType" concept) (Fig. 5). A domain is a chain subunit characterized by its three-dimensional (3D) structure, and by extension its amino acid sequence and the nucleotide sequence which encodes it. This concept may theoretically comprise many instances, but so far only the instances which 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) [35-40].

3.3.4 The "Specificity" and "Function" concepts

The "Specificity" concept identifies the specificity of the "Molecule_ReceptorType" (Fig. 5), and by extension the specificity of the chains and domains and of the corresponding transcripts. Instances of the "Specificity" concept identify the antigen recognized by an antigen receptor (IG or TR). The "Specificity" concept is particularly important because of the unlimited number of antigens and of the complexity of the antigen/antigen receptor interactions. The conceptualization of knowledge associated with this concept is in the course of modelling. The instances of the "Specificity" concept (several hundreds at the present time) will be connected on the one hand, with the "Epitope" concept which identifies the part of the antigen recognized by the antigen receptor and on the other hand, with the "Paratope" concept which identifies the part of the antigen receptor (IG or TR) which recognizes and binds to the antigen.

The "Function" concept identifies the function of the "Molecule_ReceptorType" (Fig. 5), and by extension the function of the chains and domains and of the corresponding transcripts. Instances of the "Function" concept identify the dual function of the antigen receptors [2]. Their identification and definition are still in development.

3.3.5. The "ClassType" and "SubClassType" concepts

The "ClassType" and "SubClassType" concepts are specific to the IG. The "ClassType" concept identifies the class of an IG-Receptor by reference only to its heavy chains. It is defined by the "PartnerClasslevelChainType". For example, the IgG class is defined by the IG-Heavy-Gamma-Chain. The "SubClassType" is defined by the "GeneLevelChainType". For example, the IgG1 subclass is defined by the Heavy-Gamma1-Chain (Fig. 7).

Section 4. The necessity of description: the DESCRIPTION axiom

The DESCRIPTION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be described.

4.1. Description of an entity: the "EntityPrototype" concept

The "EntityPrototype" concept, generated from the DESCRIPTION axiom, provides the description of the "EntityType" concept (IDENTIFICATION axiom). Each instance of the "EntityPrototype" concept is linked to an instance of the "EntityType" concept by the reciprocal relations "describes" and "is_described_by". The "EntityPrototype" concept allows the description of the entity organization and of its constitutive motifs. The "Core" concept allows to describe the parts of the entities which need to be described in all instances of the "EntityPrototype" concept. The concepts of description, "EntityPrototype" and "Core" have been particularly highlighted by IMGT®.

4.1.1 The "Molecule_EntityPrototype" concept

In molecular biology, the DESCRIPTION axiom has generated the concepts of description which provide the terms and the rules to describe motifs in the nucleotide and protein sequences and in 3D structures. These concepts gave rise to a standardized terminology and to a precise definition of the annotation rules. The instances of the concepts of description correspond to IMGT® labels. More than 550 labels were defined (270 for the nucleotide sequences (http://imgt.cines.fr/cgi-bin/IMGTlect.jv?query=7) [17] and 285 for the 3D structures [23]).

The ontology for sequences and 3D structures has been the focus of IMGT® for many years. Interestingly, 64 of the IMGT® labels defined for nucleotide sequences are used and cross-referenced in the recently created Sequence Ontology (SO) (http://song.sourceforge.net/) [31] to describe specific IG and TR gene organization (http://imgt.cines.fr/textes/IMGTindex/ontology.html).

The "Molecule_EntityPrototype" concept allows the description of the entity (gene, transcrit and protein) organization and of their constitutive motifs. This concept is fundamental in IMGT-ONTOLOGY because it allows the representation of the knowledge related to the complex mechanisms of IG and TR gene rearrangements (Fig. 8).

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.

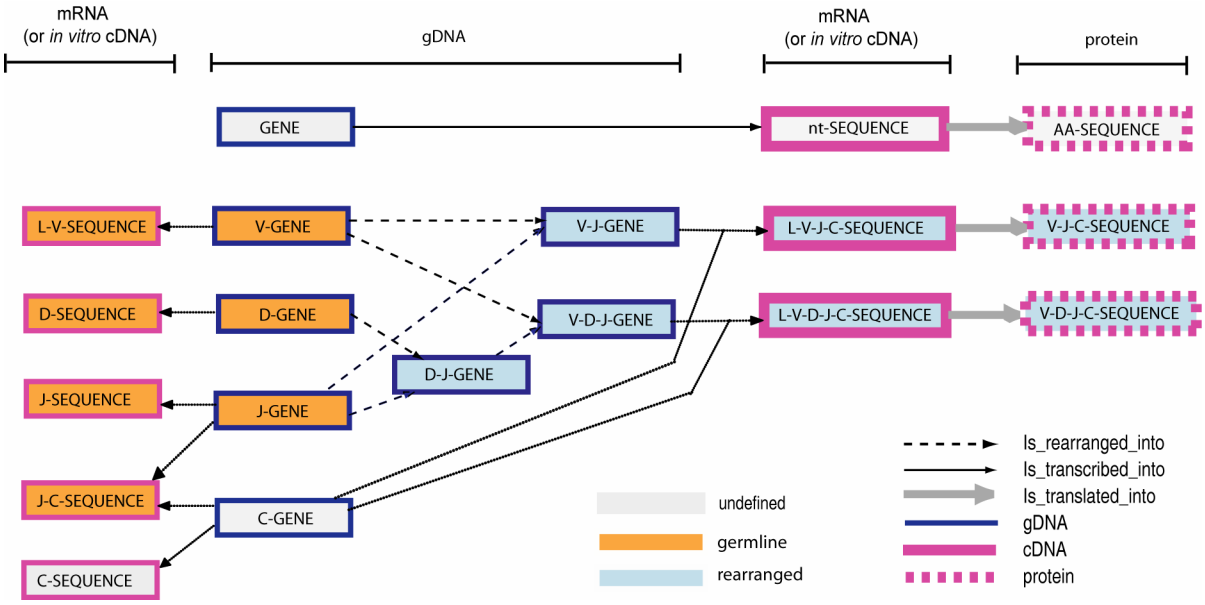

Fig. 8. Instances of the "Molecule_EntityPrototype" concept (DESCRIPTION axiom). The three instances "GENE", "nt-SEQUENCE" and "AA-SEQUENCE" correspond to conventional genes while the 16 other instances are specific of the IG and TR. The concept instances for mRNA are also valid for *in vitro* cDNA. The first column corresponds to 'sterile transcript' instances.

Each of the 19 instances of the concept "Molecule_EntityPrototype" can be described with its constitutive motifs which belong to the other concepts of description. Thus Fig. 9 shows as examples the graphic representation of the V-GENE and V-D-J-GENE instances with their constitutive motifs.

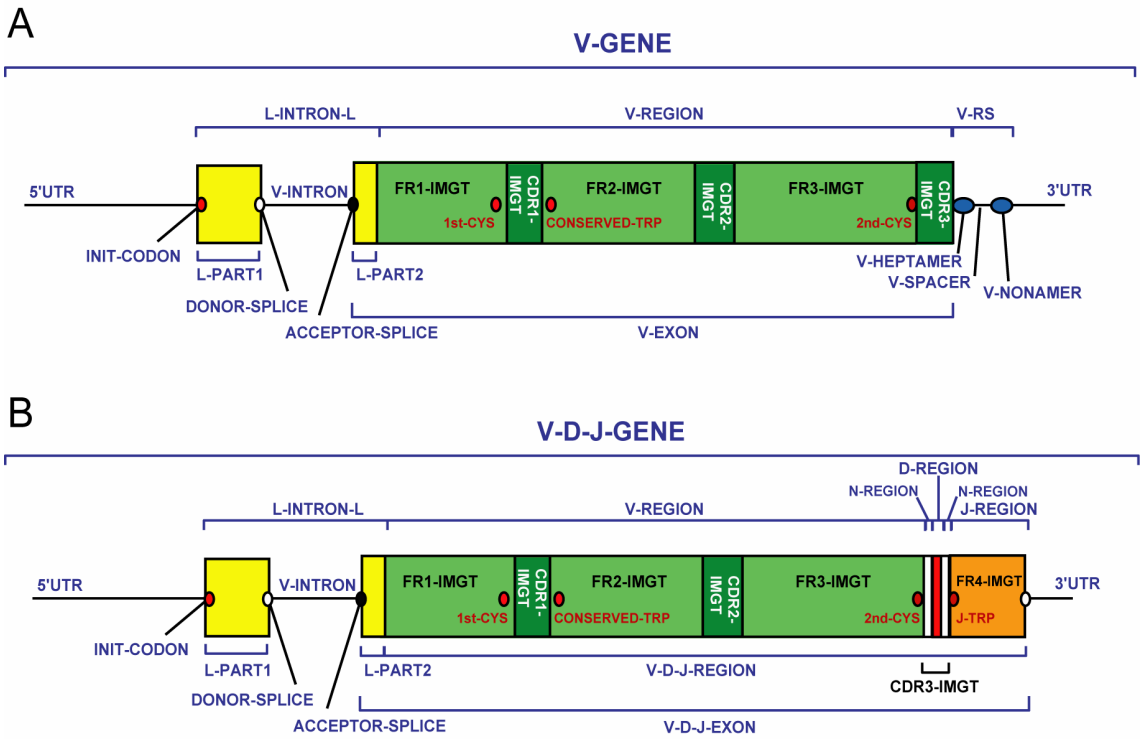

Fig. 9. Graphical representation of two instances of the "Molecule_EntityPrototype" concept (DESCRIPTION axiom). (A) V-GENE. (B) V-D-J-GENE. Twenty-five labels and ten relations are necessary and sufficient for a complete description of these instances.

A set of ten relations are necessary and sufficient to compare the localization of the motifs of an instance of the concept "Molecule_EntityPrototype" (Table 5). These relations are part of the concepts of localization (LOCALIZATION axiom) (IMGT Index, http://imgt.cines.fr).

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",
"overlap_at_its_5_prime_with"	"overlap_at_its_3_prime_with"
"included_in"	"includes"

Table 5. Relations between labels for sequence description (LOCALIZATION axiom).

The relations between the constitutive motifs of a V-GENE have been formalized in Protégé and are graphically displayed in Fig. 10.

Other prototypes can be found from IMGT Scientific chart and at http://imgt.cines.fr/textes/IMGTScientificChart/SequenceDescription/variable.html. Indeed, very precise "Molecule_EntityPrototype" instances can be established according to other concepts such as "Species", "Functionality", "ChainType"...

Cross-references

Four instances of the "Molecule_EntityPrototype" concepts ('V-GENE', 'D-GENE', 'J-GENE' and 'C-GENE') are currently cross-referenced in the Sequence Ontology Project (SO) [31].

4.1.2 The "Core" concept

The "Core" concept allows to describe the coding region of genes and contains five instances which are REGION (for conventional gene type), V-REGION, D-REGION, J-REGION and C-REGION (for V, D, J and C gene types, respectively). These instances are particularly important since they can be described in all the instances of the "Molecule_EntityPrototype" concept (Fig. 11). They allow to describe the chains of the antigen receptors in spite of the complexity of their structure and to link sequences, structures and functions. Moreover, these are the instances of the "Core" concept which allowed the definition and standardized description of the IG and TR alleles (concepts of classification), now approved at the international level [2,3].

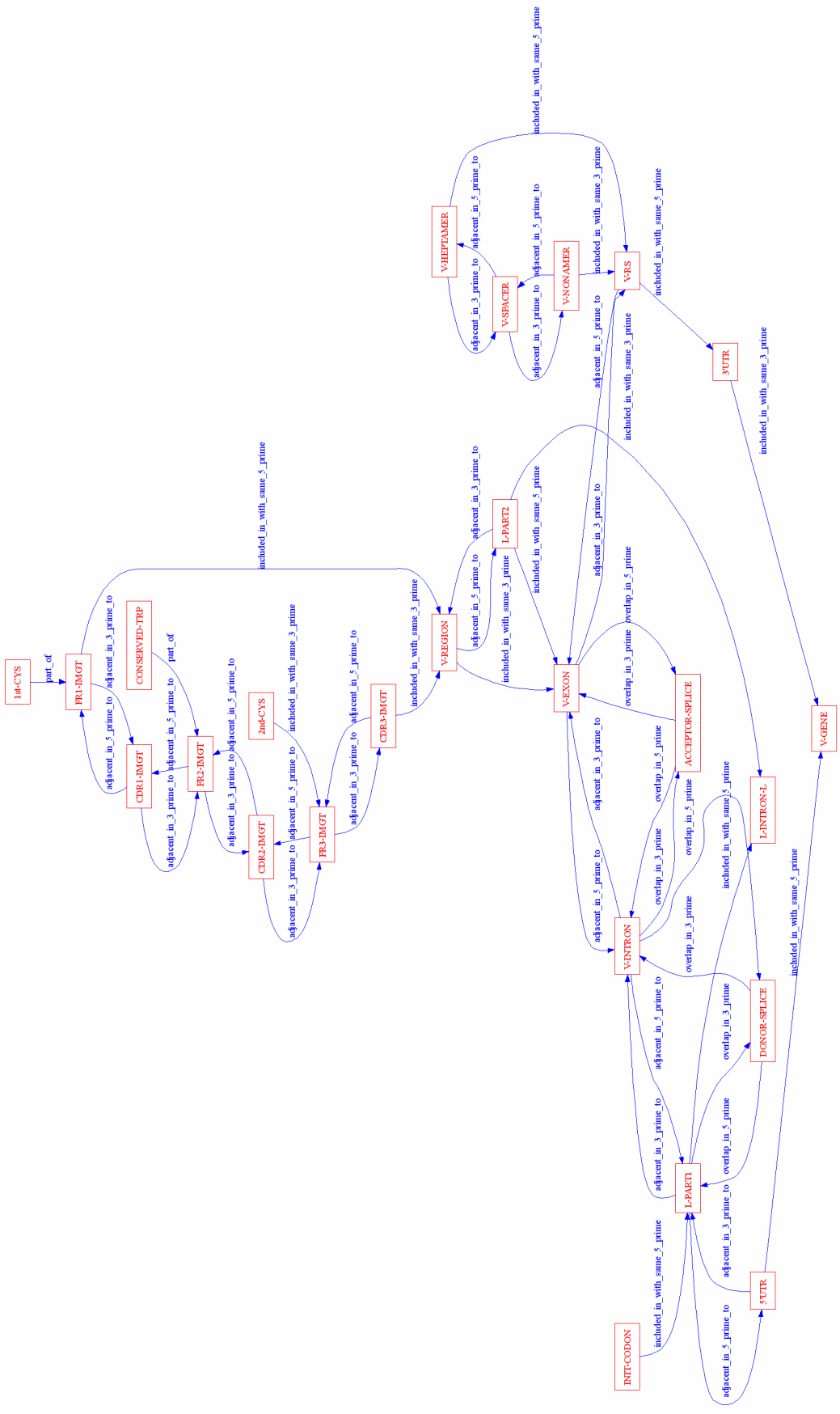

Fig. 10. Graphical representation of relations between the constitutive motifs of a V-GENE, concept instance of the “Molecule_EntityPrototype” concept (DESCRIPTION).

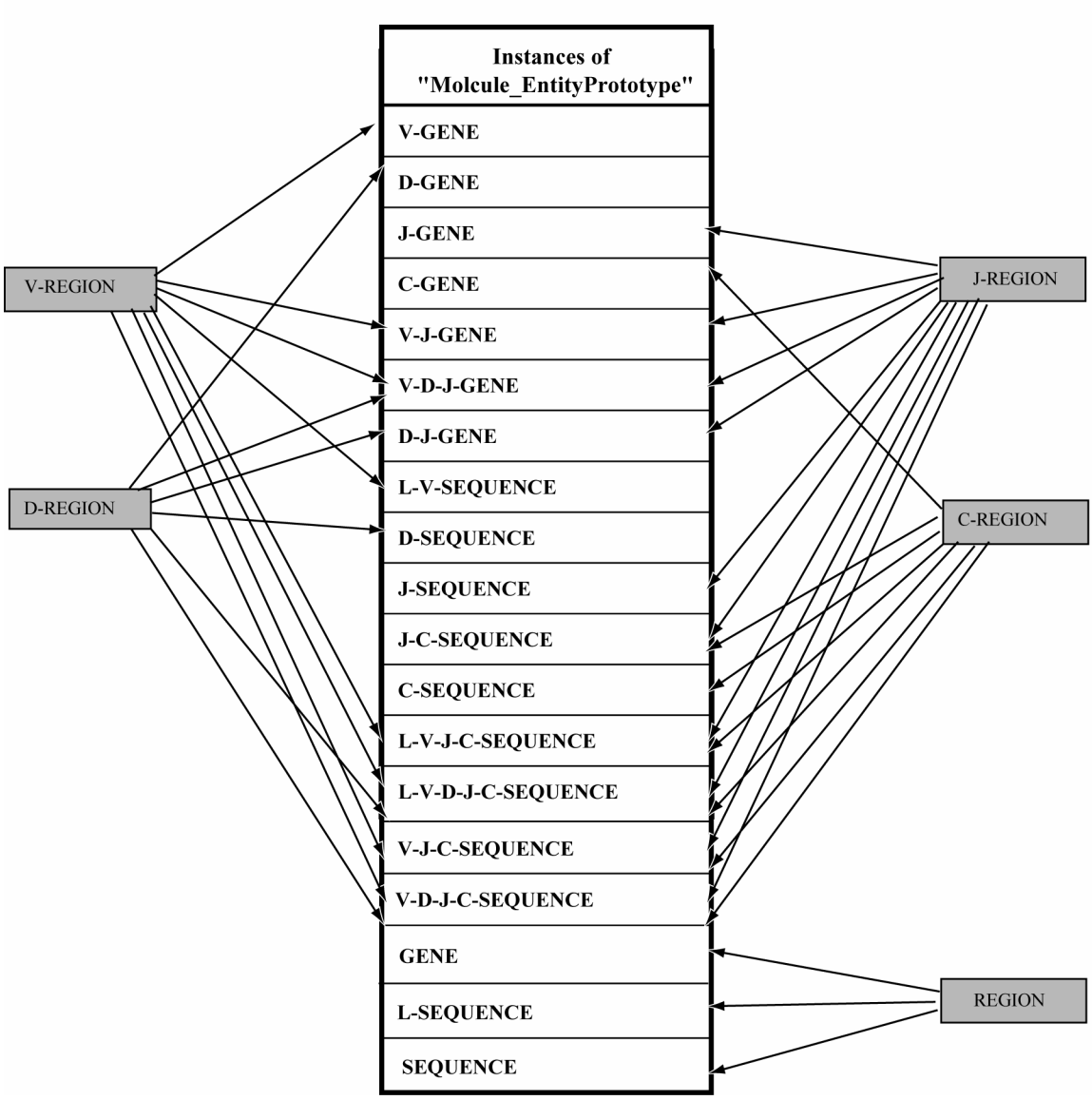

Fig. 11. The five "Core" concept instances in relation with the nineteen "Molecule_EntityPrototype" concept instances (DESCRIPTION). The relation "part_of" is represented with an arrow, the reciprocal relation being "includes".

4.1.3 The "RecombinationSignal" concept

The "RecombinationSignal" concept describes the non coding motifs specifically involved in the rearrangement of the germline V-GENE, D-GENE, J-GENE. There are four instances V-RS, J-RS, 5'D-RS and 3'D-RS. The organization of the 4 instances is graphically displayed in Fig. 12. Each instance of the "Recombination Signal" (RS) concept is composed by an heptamer (adjacent to the coding region), a spacer (12 or 23 nucleotides), and a nonamer.

The list of heptamer and nonamer motifs that have been found in human V, D and J are displayed at

http://imgt.cines.fr/textes/IMGTrepertoire/LocusGenes/RecombinationSignals/Hu_index.html

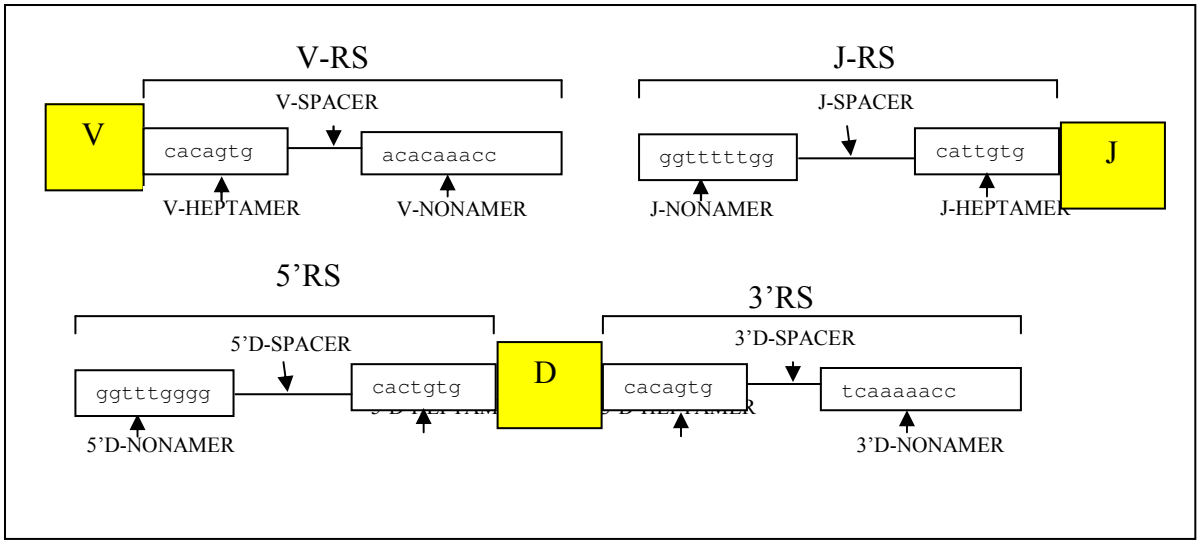

Fig. 12. Graphical representation of the instances V-RS, J-RS, 5'D-RS and 3'D-RS of the "RecombinationSignal" concept. Yellow boxes correspond to the 3'V-REGION (V), to the 5'J-REGION (J), and to the D-REGION (D) (DESCRIPTION) [40].

Relations

The "RecombinationSignal" concept instances are part of the "Molecule_EntityPrototype" concept instances shown in Table 6 (relation "is_part_of", the reciprocal relation being "includes").

"RecombinationSignal" concept instances		"Molecule_EntityPrototype" concept instances	
		Genomic instances	Sterile transcripts
V-RS	V-HEPTAMER V-SPACER V-NONAMER	V-GENE	V-SEQUENCE
J-RS	J-HEPTAMER J-SPACER J-NONAMER	J-GENE	J-C-SEQUENCE
	INTERNAL_HEPTAMER	V-GENE	V-SEQUENCE
5'D-RS	5'D-NONAMER 5'D-SPACER 5'D-HEPTAMER	D-GENE D-J-GENE	D-SEQUENCE
3'D-RS	3'D-HEPTAMER 3'D-SPACER 3'D-NONAMER	D-GENE	D-SEQUENCE

Table 6. Instances of the "Recombination Signal" concept instances (DESCRIPTION) in relation with the "ChainType" concept instances (IDENTIFICATION) and "Molecule_EntityPrototype" concept instances (DESCRIPTION).

Cross-references

Sixteen instances of the "RecombinationSignal" concepts are currently cross-referenced in the Sequence Ontology Project (SO) [31].

4.2. Description of a cluster for "EntityPrototype": the "Cluster" concept

The "Cluster" concept allows the description of a cluster for "EntityPrototype".

In molecular biology, the "Cluster" concept is designated as "Molecule_Cluster". So far only the instances corresponding to genomic sequences have been defined. They correspond to the "GeneCluster" concept.

The "GeneCluster" concept

The "GeneCluster" concept allows to describe genomic sequences containing several genes (instances of the "Molecule_EntityPrototype"concept). These instances can be of the same prototype as in a V-CLUSTER which contains several V-GENEs, or of different prototypes as in a V-(VJ)-CLUSTER which contains at least one germline V-GENE and one rearranged V-J-GENE. A graphical representation of the instances "V-J-CLUSTER" and "V-(VDJ)-CLUSTER" is shown, as example, in Fig. 13.

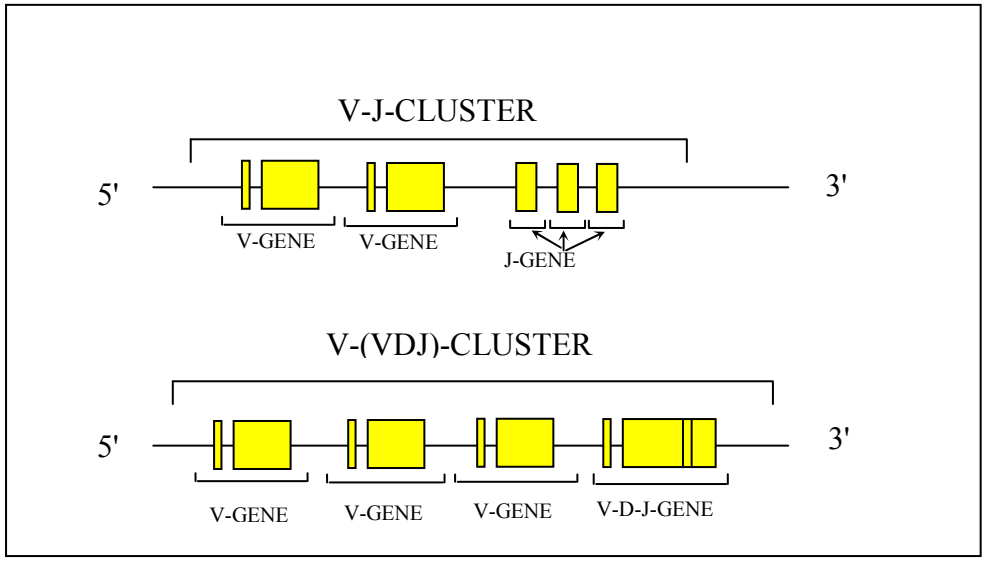

Fig. 13. Graphical representation of the instances "V-J-CLUSTER" and "V-(VDJ)-CLUSTER" of the "GeneCluster" concept (DESCRIPTION).

The instances of this concept are particularly useful for the annotation of large scale sequences as IG and TR whole loci, or chromosomes. They are listed in Table 7.

Cross-references

Forty-four instances of the "GeneCluster" concept are currently cross-referenced in the Sequence Ontology Project (SO) [31].

"GeneCluster" concept instances	**"Molecule_EntityPrototype" concept instances**	
	Minimal number of different instances	Name of the different instances
C-CLUSTER	1	C-GENE
V-CLUSTER	1	V-GENE
J-CLUSTER	1	J-GENE
D-CLUSTER	1	D-GENE
V-J-CLUSTER	2	V-GENE J-GENE
J-C-CLUSTER	2	J-GENE C-GENE
V-D-CLUSTER	2	V-GENE D-GENE
D-J-CLUSTER	2	D-GENE J-GENE
V-J-C-CLUSTER	3	V-GENE J-GENE C-GENE
V-D-J-CLUSTER	3	V-GENE D-GENE J-GENE
D-J-C-CLUSTER	3	D-GENE J-GENE C-GENE
V-D-J-C-CLUSTER	4	V-GENE D-GENE J-GENE C-GENE
V-(VJ)-CLUSTER	2	V-GENE V-J-GENE
D-(DJ)-J-CLUSTER	2	D-GENE D-J-GENE J-GENE
(DJ)-C-CLUSTER	2	D-J-GENE C-GENE
(DJ)-J-CLUSTER	2	D-J-GENE J-GENE
(VDJ)-C-CLUSTER	2	V-D-J-GENE C-GENE
(VDJ)-J-CLUSTER	2	V-D-J-GENE J-GENE
(VJ)-J-CLUSTER	2	V-J-GENE J-GENE
D-(DJ) -CLUSTER	2	D-GENE D-J-GENE
V-(DJ) -CLUSTER	2	V-GENE D-J-GENE
(VJ)-C-CLUSTER	2	V-J-GENE C-GENE
V-(VDJ) -CLUSTER	2	V-GENE V-D-J-GENE
(DJ)-J-C-CLUSTER	3	D-J-GENE J-GENE C-GENE
(VDJ)-J-C-CLUSTER	3	V-D-J-GENE J-GENE C-GENE

(VJ)-J-C-CLUSTER	3	V-J-GENE J-GENE C-GENE
D-(DJ)-C-CLUSTER	3	D-GENE D-J-GENE C-GENE
V-(DJ)-C-CLUSTER	3	V-GENE D-J-GENE C-GENE
V-(DJ)-J-CLUSTER	2	V-GENE D-J-GENE J-GENE
V-(VDJ)-C-CLUSTER	3	V-GENE V-D-J-GENE C-GENE
V-(VDJ)-J-CLUSTER	3	V-GENE V-D-J-GENE J-GENE
V-(VJ)-C-CLUSTER	3	V-GENE V-J-GENE C-GENE
V-(VJ)-J-CLUSTER	3	V-GENE V-J-GENE J-GENE
V-D-(DJ) -CLUSTER	3	V-GENE D-GENE D-J-GENE
V-(DJ)-J-C-CLUSTER	4	V-GENE D-J-GENE J-GENE C-GENE
V-(VDJ)-J-C-CLUSTER	4	V-GENE V-D-J-GENE J-GENE C-GENE
V-D-(DJ)-C-CLUSTER	4	V-GENE D-GENE D-J-GENE C-GENE
V-(VJ)-J-C-CLUSTER	4	V-GENE V-J-GENE J-GENE C-GENE
V-D-(DJ)-J-CLUSTER	4	V-GENE D-GENE D-J-GENE J-GENE
V-D-(DJ)-J-C-CLUSTER	5	V-GENE D-GENE D-J-GENE J-GENE C-GENE

Table 7 : "GeneCluster" concept instances and relation with the "Molecule_EntityPrototype" concept instances (relation "includes", the reciprocal relation being "is_part_of"). The minimal number of different instances of "Molecule_EntityPrototype" and the name of these instances are indicated (DESCRIPTION).

Section 5. The necessity of classification: the CLASSIFICATION axiom

The CLASSIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be classified. In molecular biology, the concepts of classification generated from the CLASSIFICATION axiom 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 [2,3,41-45] (Fig. 14). The IMGT® gene nomenclature has been approved at the international level by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC), in 1999 [46]. The IMGT® IG and TR gene names are the official reference for the genome projects and, as such, have been integrated in the Genome Database (GDB), in LocusLink and in Entrez Gene at NCBI [47]. The IG and TR genes [2,3,41-45] are managed in the IMGT/GENE-DB database [16].

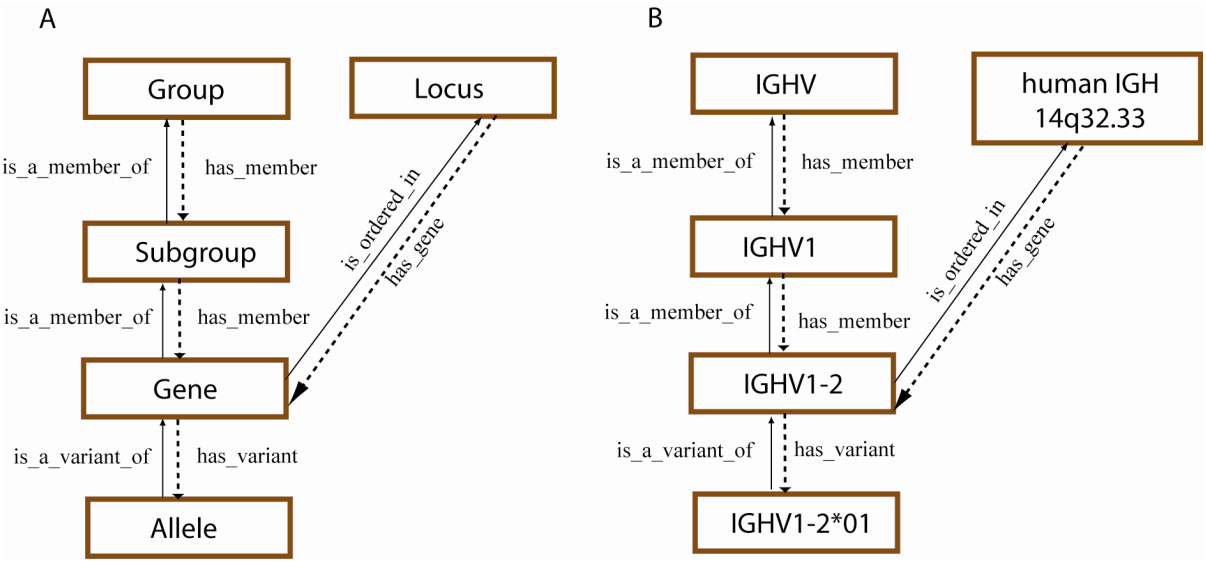

Fig. 14. Concepts of classification for gene and allele nomenclature (CLASSIFICATION axiom). (A) Hierarchy of the concepts of classification and their relations. (B) Examples of concept instances for each concept of classification. The concepts instances are associated to an instance of the "Taxon" concept, and more precisely for the "Gene" and "Allele" concepts to an instance of the "Species" concept (here, *Homo sapiens*). The "Locus" concept is a concept of localization (LOCALIZATION axiom).

5.1. The "Group" concept

The “Group” concept classifies a set of genes which 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" (V, D, J, or C). The list of the instances of the “Group” concept for the IG, TR and MHC is available from IMGT index at http://imgt.cines.fr/textes/IMGTindex/group.html.

Relations

The "Group" concept instances (CLASSIFICATION) in relation with the instances of "GeneType" and "ChainType" concept instances (IDENTIFICATION) are shown in Table 8.

"Group" concept instances	"GeneType" concept instances[1]	"ChainType" concept instances
IGHV	V	IG-Heavy-Chain
IGHD	D	
IGHJ	J	
IGHC	C	
IGKV	V	IG-Light-Kappa-Chain
IGKJ	J	
IGKC	C	
IGLV	V	IG-Light-Lambda-Chain
IGLJ	J	
IGLC	C	
TRAV	V	TR-Alpha-Chain
TRAJ	J	
TRAC	C	
TRBV	V	TR-Beta-Chain
TRBD	D	
TRBJ	J	
TRBC	C	
TRDV	V	TR-Delta-Chain
TRDD	D	
TRDJ	J	
TRDC	C	
TRGV	V	TR-Gamma-Chain
TRGJ	J	
TRGC	C	

1V: variable, D: diversity, J: junction, C: constant

Table 8. The "Group" concept instances (CLASSIFICATION) in relation with the "GeneType" and "ChainType" concept instances, in Taxon is "Mammalia" (mammals)" (IDENTIFICATION).

5.2. The "Subgroup" concept

The "Subgroup" concept classifies a set of genes which belong to the same group, and which, in a given species, share at least 75% identity at the nucleotide level (in the germline configuration for V, D, and J genes) as defined at

http://imgt.cines.fr/textes/IMGTindex/subgroup.html

The list of the instances of the "Subgroup" concept, for the human and mouse genes, is available in IMGT/GENE-DB [16].

5.3. The "Gene" concept

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). 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 international HGNC/IMGT gene symbol, for example, *Homo sapiens* IGHV1-2. By extension, orphons and pseudogenes are also instances of the "Gene" concept.

Gene names follow the rules of the IMGT gene name nomenclature for IG and TR of human and other vertebrates.

Relations

The relations of the instances of the "Gene" concept are multiple. They comprise, as shown in Table 9, relations with instances of "GeneType", "MoleculeType", "ConfigurationType" (IDENTIFICATION), with instances of the "Group", "Subgroup" (CLASSIFICATION) and with instances of "Position-In-Locus" (NUMEROTATION, not detailed in this deliverable).

Examples of "Gene" concept instances	"GeneType" concept instances	"MoleculeType" concept instances	"ConfigurationType" concept instances	"Group" concept instances	"Subgroup" concept instances	"Position-In-Locus" concept instances
IGLV6-57	V	gADN	germline	IGLV	IGLV6	57
IGLV7-35	V				IGLV7	35
IGLV7-43	V					43
IGLV7-46	V					46
IGLV8-61	V				IGLV8	61

Table 9. Examples of instances of the "Gene" concept (CLASSIFICATION) and relations with instances of "GeneType", "MoleculeType", "ConfigurationType" (IDENTIFICATION), "Group" and "Subgroup" concepts (CLASSIFICATION) and "Position-In-Locus" (NUMEROTATION).

5.4. The "Allele" concept

The "Allele" concept classifies a polymorphic variant of a gene. 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 [16].

Section 6. Implementation plan

This deliverable D1.2 “Scientific chart rules and ontologies report” formalizes the necessary concepts and rules for an exhaustive management of the molecular components as defined by the axioms IDENTIFICATION, DESCRIPTION and CLASSIFICATION of the key actors of the adaptative immune response, IG, TR and MHC, and essential parts of the Virtual Immune System modelling.

The concepts for the antigen receptors and MHC defined in this deliverable D1.2 are delivered for WP2 (Molecular level modelling), WP3 (System level modelling) and WP4 (Simulator design).

We have started to work on the next deliverable D1.3 for WP2, WP3 and WP4, and we are focusing on the peptides and T cell epitopes for which 3D structures of TR/peptide/MHC complexes are available, as a model. The concepts which are necessary for a standardized description of the structural domains in the Virtual Immune System modelling will be detailed for IG, TR and MHC, based on the NUMEROTATION axiom. This will include the variable domains and the constant domains which have an immunoglobulin fold and the groove domains of the MHC proteins. The priority will be given to the contact analysis and amino acid interactions between antigen and receptors, as they trigger the signalling cascade of the cells involved in the immune responses.

We are currently applying the concepts and instances of D1.2 to the building of the ontology of the Catania Mouse Model with UNICT.

We are currently formalizing the axioms of IDENTIFICATION, DESCRIPTION and CLASSIFICATION for the IG, TR and MHC nucleotide and amino acid sequence description and 3D structure identification with the Web Ontology Language OWL using the Protégé editor. This will allow interoperability of the ImmunoGrid components with other major medical or biological ontologies (The Immune Epitope Database and Analysis Resource IEDB ontology, Gene Ontology GO, Sequence Ontology SO, Ontology for Biomedical Investigations (OBI), MGED Network Ontology Working Group, UMLS,...) [29-32].

Section 7. Perspectives for ImmunoGrid and the modelling of the immune system

The inherent difficulties due to the complexity and diversity of immunogenetics knowledge gave rise to a conceptualization in IMGT-ONTOLOGY which has been developed on an original and unprecedented approach. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulate that the approach to manage biological data and to represent knowledge in biology comprises various facets. The IMGT-ONTOLOGY concepts generated from these axioms have allowed the representation, at the molecular level, of knowledge related to the genome, transcriptome, proteome, genetics and 3D structures. This multi-

faceted approach has great potential for multi-scale system biology. Indeed, the IDENTIFICATION, DESCRIPTION and CLASSIFICATION axioms defined in Deliverable D1.2 are valid, not only for molecules, but also for cells, tissues, organs, organisms or populations. In addition, the NUMEROTATION, LOCALIZATION, ORIENTATION and OBTENTION axioms (in development) will allow the integration of the time and space concepts and the follow-up of the components and their changes of states and properties, as well as the definition and characterization of processes, functions and activities. Thus, IMGT-ONTOLOGY represents, by its 7 axioms and the concepts generated from them, 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.

The concepts of IMGT-ONTOLOGY are available, for the users of the ImmunoGrid simulator and for the biologists in general, in natural language in IMGT Scientific chart (http://imgt.cines.fr), and have been formalized for programming purpose in IMGT-ML (XML Schema). 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) [30], and in immunology, such as the Immunome Epitope database and Analysis Resource (IEDB) [32] and other Open Biomedical Ontologies (OBO) (http://obo.sourceforge.net).

The concepts of IMGT-ONTOLOGY are currently used for the exchange and the sharing of knowledge in very diverse fields of research at the molecular level: (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, leukemias, 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) biotechnology related to antibody engineering (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). 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.

Section 8. References

[1] Sakano H., Huppi K., Heinrich G. and Tonegawa S. (1979). Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature*, **280**, 288-294.

[2] Lefranc M.-P. and Lefranc G. (2001). The Immunoglobulin FactsBook. Academic Press, London, UK. 458 pages.

[3] Lefranc M.-P. and Lefranc G. (2001). The T cell receptor FactsBook. Academic Press, London, UK. 398 pages.

[4] Lefranc M.-P., Giudicelli V., Kaas Q., Duprat E., Jabado-Michaloud J., Scaviner D., Ginestoux C., Clément O., Chaume D. and Lefranc G. (2005). IMGT, the international ImMunoGeneTics information system®. *Nucl. Acids Res*., **33**, D593-D597.

[5] Lefranc M.-P., Giudicelli V., Ginestoux C. and Chaume D. (2003). Imgt, the international ImMunoGeneTics information system, http://imgt.cines.fr: the reference in immunoinformatics. *Studies in Health Technology and Informatics*, **95**, 74-79.

[6] Giudicelli V. and Lefranc M.-P. (1999). Ontology for Immunogenetics: IMGT-ONTOLOGY. *Bioinformatics*, **15**, 1047-1054.

[7] Gruber T.R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition,* **5**, 199-220.

[8] Guarino N. and Giaretta P. (1995). Ontologies and knowledge bases: towards a terminological clarification. In: Mars N. (Ed.), Towards very large knowledge bases, IOS Press, Amsterdam, pp. 29-45.

[9] Guarino N. (1997). Understanding, building and using ontologies. *International Journal of Human-Computer Studies,* **46**, 293-310.

[10] Noy N.F. and McGuinness D.L. (2001). Ontology development 101: A guide to creating your first ontology, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

[11] Smith B. (2003). Ontology. In: Floridi L. (Ed.), Blackwell Guide to the Philosophy of Computing and Information, Blackwell, Oxford, pp. 155–166.

[12] Soldatova L.-N., Clare A., Sparkes A. and King R.D. (2006). An ontology for a Robot Scientist. *Bioinformatics* **22**, e464-e471.

[13] Lefranc M.-P., Giudicelli V., Ginestoux C., Bosc N., Folch G., Guiraudou D., Jabado-Michaloud J., Magris S., Scaviner D., Thouvenin V., Combres K., Girod D., Jeanjean S., Protat C., Yousfi Monod M., Duprat E., Kaas Q., Pommié C., Chaume D. and Lefranc G. (2004). IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics, http://imgt.cines.fr. Epub 2003, 4, 0004, 22 Nov 2003. http://www.bioinfo.de/isb/2003/04/0004/. *In Silico Biology*, **4**, 17-29.

[14] Lefranc M.-P. (2004). IMGT-ONTOLOGY and IMGT databases, tools and web resources for immunogenetics and immunoinformatics. *Mol. Immunol*., **40**, 647-660.

[15] Lefranc M.-P., Clément O., Kaas Q., Duprat E., Chastellan P., Coelho I., Combres K., Ginestoux C., Giudicelli V., Chaume D. and Lefranc G. (2005). IMGT-Choreography for Immunogenetics and Immunoinformatics. Epub 2005, 5, 0006, 24 Dec 2004. http://www.bioinfo.de/isb/2004/05/0006/. *In Silico Biology*, **5**, 45-60.

[16] Giudicelli V., Chaume D. and Lefranc M.-P. (2005). IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucl. Acids Res*., **33**, D256-D261.

[17] Giudicelli V., Ginestoux C., Folch G., Jabado-Michaloud J., Chaume D. and Lefranc, M.-P. (2006). IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucl. Acids Res*., **34**, D781-D784.

[18] Giudicelli V., Chaume D. and Lefranc M.-P. (2004). IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucl. Acids Res*., **32**, W435-W440.

[19] Yousfi Monod M., Giudicelli V., Chaume D. and Lefranc M.-P. (2004). IMGT/JunctionAnalysis : the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics*, **20**, i379-i385.

[20] Lefranc M.-P. (2005). IMGT, the international ImMunoGeneTics information system®: a standardized approach for immunogenetics and immunoinformatics. 20 September 2005, http://www.immunome-research.com/content/1/1/3, doi:10.1186/1745-7580-1-3. *Immunome Res*., **1**, 3.

[21] Baum T.P., Hierle V., Pascal N., Bellahcene F., Chaume D., Lefranc M.-P., Jouvin-Marche E., Marche P.N. and Demongeot J. (2006). IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. *BMC Bioinformatics*, **7**:224.

[22] Elémento O. and Lefranc M.-P. (2003). IMGT/PhyloGene: an online software package for phylogenetic analysis of immunoglobulin and T cell receptor genes. *Dev. Comp. Immunol*., **27**, 763-779.

[23] Kaas Q., Ruiz M. and Lefranc M.-P. (2004). IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucl. Acids Res*., **32**, D208-D210.

[24] Kaas Q. and Lefranc M.-P. (2005). T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. Epub 2005, 5, 0046, 20 Oct 2005. http://www.bioinfo.de/isb/2005/05/0046/. *In Silico Biology*, **5**, 505-528.

[25] Duprat E., Kaas Q., Garelle V., Lefranc G. and Lefranc M.-P. (2004). IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. *Recent Research Developments in Human Genetics* (Pandalai S.G., ed.), Research Signpost, Trivandrum, Kerala, India, **2**, 111-136.

[26] Kaas Q. and Lefranc M.-P. (2007). IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. *Current Bioinformatics*, **2**, 21-30. http://www.ingentaconnect.com/content/ben/cbio/2007/00000002/00000001

[27] Pommié C., Levadoux S., Sabatier R., Lefranc G. and Lefranc M.-P. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.*, 17, 17-32.

[28] Noy N.F., Crubézy M., Fergerson R.W., Knublauch H., Tu S.W., Vendetti J. and Musen M.A. (2003). Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. In: *AMIA Annu. Symp. Proc.*, 953.

[29] McCray A.T., Nelson S.J. (1995). The representation of meaning in the UMLS. *Methods Inf. Med.* **34**, 193-201.

[30] The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genet*, **25**, 25–29.

[31] Eilbeck K. and Lewis S.E. (2004). Sequence Ontology annotation guide. *Comp. Funct. Genomics*, **5**, 642–647.

[32] Sathiamurthy M., Peters B., Bui H.H., Sidney J., Mokili J., Wilson S.S., Fleri W., McGuinness D.L., Bourne P.E. and Sette A. (2005). An ontology for immune epitopes:

application to the design of a broad scope database of immune reactivities. *Immunome Res*. 2005, **1**:2.

[33] Chaume D., Giudicelli V., Lefranc M.-P. (2001). IMGT-ML a XML language for IMGT-ONTOLOGY and IMGT/LIGM-DB data, In: *Proceedings of NETTAB 2001*, Network Tools and Applications in Biology, Genoa, Italy, May 17-18, 2001, pp 71-75.

[34] Chaume D., Combres K., Giudicelli V., Lefranc M.-P. (2005). Retrieving factual data and documents using IMGT-ML in the IMGT information system®. In: *Proceedings of NETTAB 2005*, Workflows management: new abilities for the biological information overflow, Naples, Italy, Oct 5-7, 2005, pp. 47-51.

[35] Lefranc M.-P. (1997). Unique database numbering system for immunogenetic analysis. *Immunology Today*, **18**, 509.

[36] Lefranc M.-P. (1999). The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist*, **7**, 132-136.

[37] Lefranc M.-P., Pommié C., Ruiz M., Giudicelli V., Foulquier E., Truong L., Thouvenin-Contet V. and Lefranc G. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol*., **27**, 55-77.

[38] Lefranc M.-P., Pommié C., Kaas Q., Duprat E., Bosc N., Guiraudou D., Jean C., Ruiz M., Da Piedade I., Rouard M., Foulquier E., Thouvenin V. and Lefranc G. (2005). IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol*., **29**, 185-203.

[39] Lefranc M.-P., Duprat E., Kaas Q., Tranne M., Thiriot A. and Lefranc G. (2005). IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev. Comp. Immunol*., **29**, 917-938.

[40] Bleakley K., Giudicelli V., Wu Y., Lefranc M.-P. and Biau G. (2006). IMGT standardization for statistical analyses of T cell receptor junctions: the TRAV-TRAJ example. Epub 2006, 6, 0051, http://www.bioinfo.de/isb/2006/06/0051/, *In Silico Biology*, **6**, 573-588.

[41] Lefranc M.-P. (2000). Nomenclature of the human immunoglobulin genes. In: *Current Protocols in Immunology* (Coligan J.E., Bierer B.E., Margulies D.E., Shevach E.M. and Strober W., eds.), John Wiley and Sons, Hoboken N.J., pp. A.1P.1-A.1P.37.

[42] Lefranc M.-P. (2000). Nomenclature of the human T cell receptor genes. In: *Current Protocols in Immunology* (Coligan J.E., Bierer B.E., Margulies D.E., Shevach E.M. and Strober W., eds.), John Wiley and Sons, Hoboken N.J., pp. A.1O.1-A.1O.23.

[43] Lefranc M.-P. (2001). Nomenclature of the human immunoglobulin heavy (IGH) genes. *Exp. Clin. Immunogenet*., **18**, 100-116.

[44] Lefranc M.-P. (2001). Nomenclature of the human immunoglobulin kappa (IGK) genes. *Exp. Clin. Immunogenet*., **18**, 161-174.

[45] Lefranc M.-P. (2001). Nomenclature of the human immunoglobulin lambda (IGL) genes. *Exp. Clin. Immunogenet*., **18**, 242-254.

[46] Wain H.M., Bruford E.A., Lovering R.C., Lush M.J., Wright M.W. and Povey S. (2002). Guidelines for human gene nomenclature**.** *Genomics*, **79,** 464-470.

[47] Maglott D., Ostell J., Pruitt K.D. and Tatusova T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucl. Acids Res.*, **35**, D26-D31.

Section 9. Web sites quoted in the deliverable D1.2

URL addresses of the IMGT resources are quoted in the text but are not reported below.

- FMA, Foundational Model of Anatomy.
  http://sig.biostr.washington.edu/projects/fm/AboutFM.html
- GALEN, General Architecture for Languages, Encyclopedias and Nomenclatures in Medicine.
  http://www.opengalen.org/index.html
- GO, Gene Ontology.
  http://www.geneontology.org/
- IEDB, Immune Epitope Database and Analysis Resource.
  http://www.immuneepitope.org/home.do
- IMGT-Choreography.
  http://www.bioinfo.de/isb/2004/05/0006/
- IMGT-ONTOLOGY.
  http://www.bioinfo.de/isb/2003/04/0004/
- IMGT®, the international ImMunoGeneTics information system®.
  http://imgt.cines.fr
- ImmunoGrid, the European Virtual Immune System Project.
  http://www.immunogrid.org/
- Logical Observation Identifiers Names and Codes (LOINC®).
  http://www.regenstrief.org/loinc/
- MGED Network Ontology Working Group
  http://mged.sourceforge.net/ontologies/index.php
- National Cancer Institute (NCI) Thesaurus
  http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do
- NCBI, National Center for Biotechnology Information
  http://www.ncbi.nlm.nih.gov/
- OBO-Edit.
  http://sourceforge.net/project/showfiles.php?group_id=36855
- Ontology for Biomedical Investigations (OBI), formerly FuGO
  http://obi.sourceforge.net/index.php
- OWL, Web Ontology Language.
  http://www.w3.org/2004/OWL/
- Protégé.
  http://protege.stanford.edu/
- Resource Description Framework (RDF) model and syntax specification. W3C recommendation 22 February 1999.
  http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
- SNOMED, The Systematized Nomenclature of Medicine.
  http://www.cap.org/apps/cap.portal?_nfpb=true&_pageLabel=snomed_page
- SO, The Sequence Ontology project.
  http://song.sourceforge.net/
- UMLS, Unified Medical language System.
  http://www.nlm.nih.gov/research/umls/
- W3C, Web Services Activity
  http://www.w3.org/2002/ws/

- XML Schema. W3C recommendation 28 October 2004
http://www.w3.org/TR/xmlschema-0/