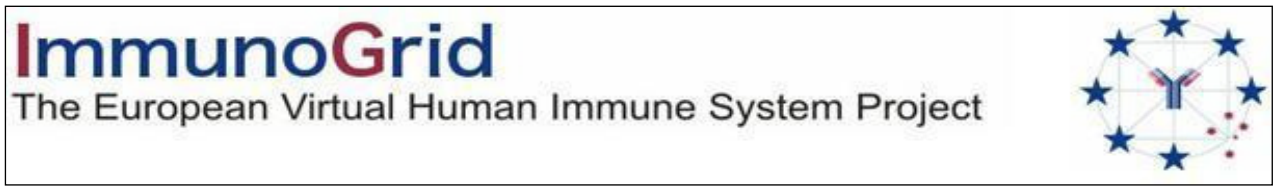

D1.3 IMGT unique numbering and controlled vocabularies report

Project Acronym: IMMUNOGRID

Contract no: IST-2004-028069

Date: March 14th, 2008

Due Date: Jan 31st, 2007

Principal Authors: Marie-Paule Lefranc, François Ehrenmann, Patrice Duroux and Véronique Giudicelli (CNRS)

Revision: 1.0

Dissemination Level: PU

Content

Section 1. Introduction

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 aim of this deliverable D1.3 *IMGT unique numbering and controlled vocabularies report* is to provide the IMGT unique numbering and controlled vocabularies for the antigen receptors - immunoglobulins (IG) and T cell receptors (TR) - and for the major histocompatibility complex (MHC). These proteins are key actors of the adaptive immune response and are characterized, for the antigen receptors, by an incredible diversity (10^{12} IG or antibodies and 10^{12} TR per individual) due to molecular mechanisms such as DNA rearrangements, N-diversity, and for the IG, somatic mutations and, for the MHC (designated as HLA in humans), by an extensive allelic polymorphism. The aim is also to extend the IMGT unique numbering to the domains of the related proteins of the immune system (RPI) that include proteins of the immunoglobulin superfamily (IgSF) other than IG and TR, and proteins of the MHC superfamily (MhcSF) other than MHC. These rules are crucial for a standardized analysis of the interactions between receptors and ligands and between proteins. They are delivered for WP2 (Molecular level modelling), WP3 (System level modelling) and WP4 (Simulator design).

Section 2 provides a brief overview of the previous deliverables in relation with the current deliverable D1.3. Section 3 specifies the controlled vocabularies for the chains and domains of the molecular components of the adaptive immune response (IG, TR, MHC) and for the proteins of the IgSF (other than IG and TR) and the proteins of the MhcSF (other than MHC). Section 4 provides the IMGT unique numbering for the three types of domains: variable (V), constant (C) and groove (G) and their two-dimensional (2D) representations or IMGT Colliers de Perles. Section 5 gives the implementation plan for the standardized IMGT unique numbering and IMGT Colliers de Perles in the VIS modelling. Section 6 provides the perspectives for ImmunoGrid and the modelling of the immune system.

Section 2. Overview with previous deliverables

2.1. D1.1– New and enhanced concepts and rules

The immunogenetics knowledge is particularly complex. IMGT-ONTOLOGY [1] is the first and so far unique ontology in immunogenetics and immunoinformatics. IMGT-ONTOLOGY provides a semantic specification of the terms to be used in immunogenetics and immunoinformatics and manages the related knowledge, thus allowing the standardization for immunogenetics data from genome, proteome, genetics and three-dimensional (3D) structures [2-5]. IMGT-ONTOLOGY results from a deep expertise in the

domain and an extensive effort of conceptualization and is crucial to make immunogenetics knowledge amenable to modelling.

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 [4].

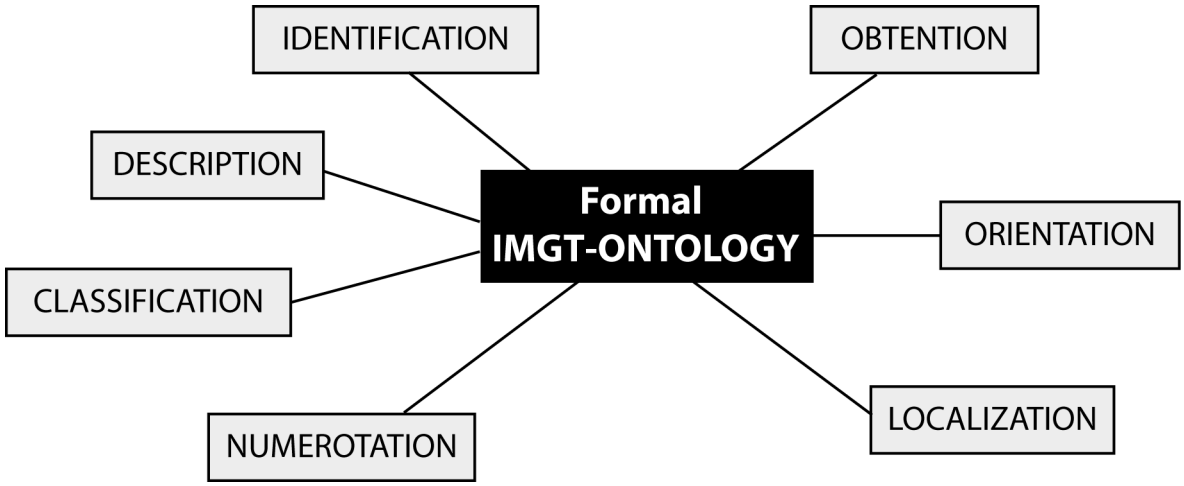

Figure 1. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope.

The deliverable D1.1 *New and enhanced concepts and rules* (internal ImmunoGrid report) provided to WP2 (Molecular level modelling), WP3 (System level modelling) and WP4 (Simulator design) the semantic classification and standardization of the knowledge necessary for the "Virtual Immune System" (VIS) modelling.

2.2. D1.2– Scientific chart rules and ontologies report

The deliverable D1.2 *Scientific chart rules and ontologies report* formalized the new and enhanced concepts and rules for the identification, description and classification of the antigen receptors (IG, TR) and the 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) [6, 7] and a key component of the ImmunoGrid project (http://www.immunogrid.org/).

The standardization in the deliverable D1.2 is based on IMGT-ONTOLOGY [1]. 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. Three axioms, "IDENTIFICATION", "DESCRIPTION" and "CLASSIFICATION", of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope, were defined in D1.2, with the concepts that have been essential for the conceptualization of the molecular immunogenetics knowledge. The IMGT-ONTOLOGY axioms, the related IMGT Scientific chart rules, ImmunoGrid deliverables and ImmunoGrid examples which were defined and formalized in D1.2 are listed in Table 1.

Table 1. IDENTIFICATION, DESCRIPTION, CLASSIFICATION and NUMEROTATION axioms of the Formal IMGT-ONTOLOGY.

IMGT-ONTOLOGY axioms	IMGT Scientific chart rules[1]	ImmunoGrid deliverables	Examples in ImmunoGrid
IDENTIFICATION	Keywords	D1.2 *Scientific chart rules and ontologies report (PU)*	IG, TR, MHC and RPI nucleotide and amino acid sequence, and 3D structure identification
DESCRIPTION	Labels		IG, TR, MHC and RPI nucleotide and amino acid sequence, and 3D structure description
CLASSIFICATION	Nomenclature		IG, TR, MHC and RPI gene and allele names
NUMEROTATION	Unique numbering	D1.3 *IMGT unique numbering and controlled vocabularies report (PU)*	IMGT Colliers de Perles for V-DOMAIN and V-LIKE-DOMAIN
			IMGT Colliers de Perles for C-DOMAIN and C-LIKE-DOMAIN
			IMGT Colliers de Perles for G-DOMAIN and G-LIKE-DOMAIN

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

The examples in ImmunoGrid include standardized IMGT keywords (IDENTIFICATION axiom), standardized IMGT labels[1] for the receptors, chains, domains and regions (DESCRIPTION axiom) and standardized IMGT gene and allele names (CLASSIFICATION axiom) [8-12]. The deliverable D1.2 represents the state of the art for the IDENTIFICATION, DESCRIPTION and CLASSIFICATION axioms and concepts. It was provided to WP2 (Molecular level modelling), WP3 (System level modelling) and WP4 (Simulator design) in September 2007 and made publicly available on the ImmunoGrid Web site at http://www.immunogrid.org/immunogrid/publications.

2.3. D1.3– IMGT unique numbering and controlled vocabularies report

The deliverable D1.3 (this report) defines the concepts which are necessary for a standardized description of the structural domains in the Virtual Immune System modelling. They are detailed for IG, TR and MHC, based on the NUMEROTATION axiom. The domains include the variable (V) domains and the constant (C) domains of the IG and TR which have an immunoglobulin fold, and the groove (G) domains of the MHC proteins. They have been extended to the V-like and C-like domains of the IgSF (other than IG and TR) and to the G-like domain of the MhcSF (other than MHC) (Table 1). These concepts are fundamental for the standardized description and characterization of the contact analysis and amino acid interactions between antigen and receptors, which trigger the signalling cascade of the cells involved in the immune responses. The deliverable D1.3 represents the state of the art for the NUMEROTATION axiom and for the standardization of both sequences and 3D structures by providing standardized amino acid positions according to the IMGT unique numbering (NUMEROTATION axiom). It was provided to WP2 (Molecular level modelling), WP3 (System level modelling) and WP4 (Simulator design) in February 2008.
The D1.3 deliverable is publicly available on the ImmunoGrid Web site at http://www.immunogrid.org/immunogrid/publications.

Section 3. Controlled vocabularies for molecular components

3.1. Immunoglobulins (IG) and T cell receptors (TR)

The immunoglobulin (IG) and T cell receptor (TR) proteins are antigen receptors formed, for the IG, by four chains (two identical heavy chains and two identical light chains) and, for the TR, by two chains of similar length (alpha and beta chains, or gamma and delta chains, depending on the receptor type) (Fig. 2).

[1] IMGT labels of the IMGT-ONTOLOGY DESCRIPTION concept are written in capital letters. Definitions of IMGT labels are available in the IMGT Scientific chart at http://imgt.cines.fr

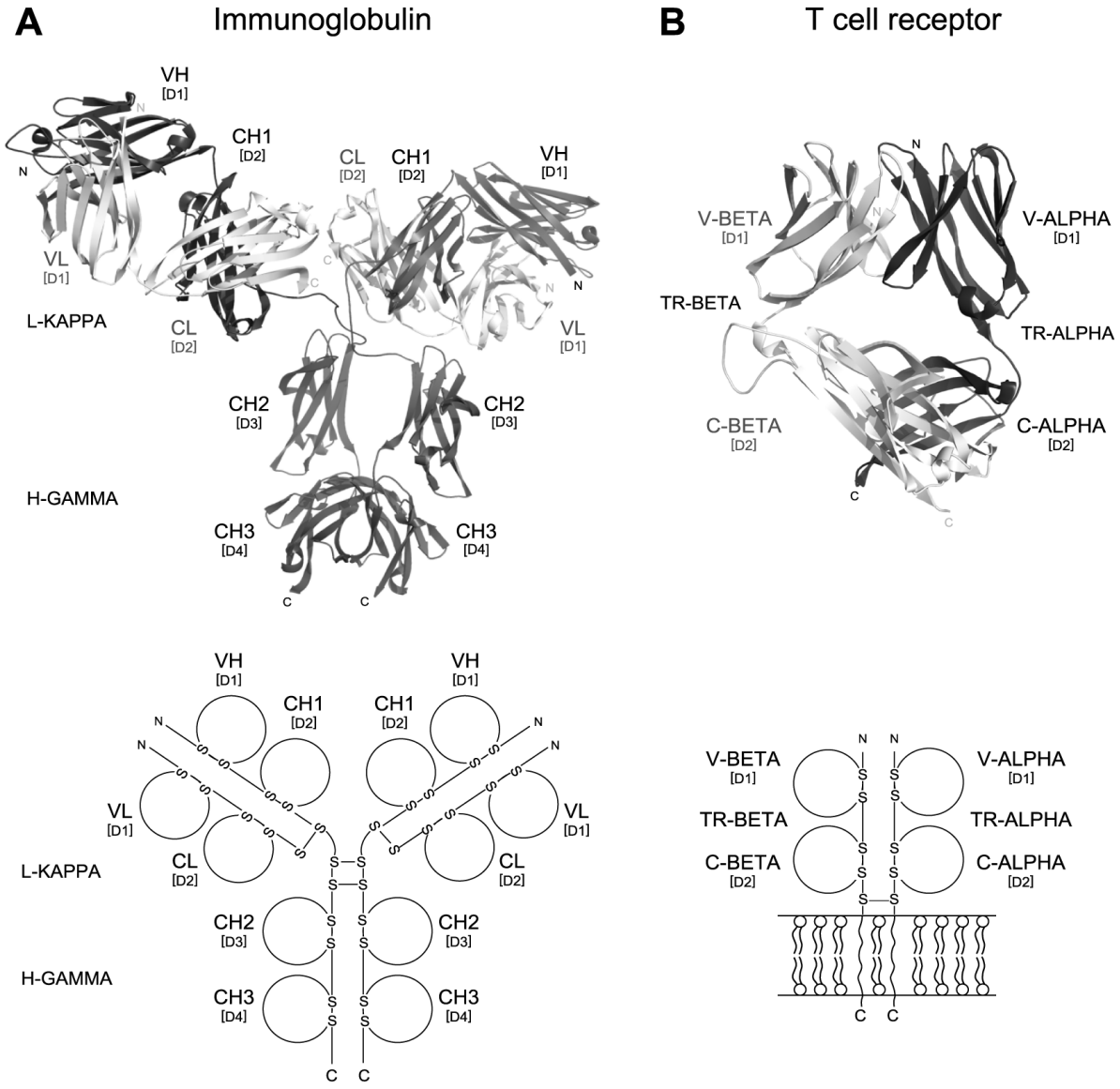

Figure 2: Three-dimensional (3D) structures and schematic representations of antigen receptors. (A) Immunoglobulin (IG). (B) T cell receptor (TR).

An IG (for example an human IgG1κ as shown in Fig. 2A) comprises four chains, two identical disulfide-linked heavy chains and two identical light chains, each one disulfide-linked to an heavy chain. The heavy chain comprises a N-terminal V-DOMAIN (VH) and several C-terminal C-DOMAINs (number depending on the heavy chain type, here, CH1, CH2 and CH3 for H-GAMMA). The VH domain results from the junction of three genes (variable V, diversity D and junction J) and corresponds at the sequence level to the V-D-J-REGION [8]. The CH domains are part of the C-REGION that is coded by a constant C gene. The light chain comprises a N-terminal V-DOMAIN (VL, that can be V-KAPPA or V-LAMBDA depending on the light chain type) and one C-terminal C-DOMAIN (CL, that can be C-KAPPA or C-LAMBDA). The VL domain results from the junction of two genes (V and J) and corresponds to the V-J-REGION. The CL domain corresponds to the C-REGION and is coded by a C gene.

A TR (Fig. 2B) comprises two chains, TR-ALPHA and TR-BETA (disulfide-linked) for the TR-ALPHA_BETA, TR-GAMMA and TR-DELTA (disulfide-linked or not) for the TR-

GAMMA_DELTA. The TR chains comprise a N-terminal V-DOMAIN and one C-terminal C-DOMAIN. The V-ALPHA and V-GAMMA result from the junction of two genes (V and J) and correspond to the V-J-REGION, whereas the V-BETA and V-DELTA result from the junction of three genes (V, D and J) and correspond to the V-D-J-REGION [9]. The C-ALPHA, C-BETA, C-GAMMA and C-DELTA are part of the C-REGION, that also comprises the connecting region (CO), transmembrane region (TM), cytoplasmic region (CY) (not present in the 3D structures), and which is coded par a C gene [9]. [D1], [D2], ...indicate the positions of the domains from the N-terminal end of the chains.

Thus, an IG or TR chain comprises two types of structural units: one V-DOMAIN and one (for the IG light chains and TR chains) or several (for the IG heavy chains) C-DOMAINs (CH1, CH2,...). The unique V-DOMAIN (encoded by a rearranged V-J or V-D-J gene) of a IG or TR chain corresponds to the V-J-REGION or V-D-J-REGION, and is associated to a C-REGION encoded by the C-GENE (for review, see [8, 9] and IMGT Scientific chart, http://imgt.cines.fr "Correspondence between labels for IG and TR domains").

3.2. Immunoglobulin superfamily (IgSF)

The immunoglobulin superfamily IgSF comprises the IG and TR proteins described in 3.1 (each chain has one V-DOMAIN and one or several C-DOMAINs) and the proteins other than IG and TR defined as having at least one V-LIKE-DOMAIN or one C-LIKE-DOMAIN. The common feature of these IgSF proteins is indeed to have at least one immunoglobulin-like (Ig-like) domain [13, 14]. Despite a large divergence in the amino acid sequences, the Ig-like domains share the structural Ig-fold which typically consists of about one hundred amino acids in antiparallel beta strands, linked by beta turns or loops, and located on two layers maintained by a disulfide bridge [13, 14]. Whereas the IG and TR proteins are involved in antigen recognition, the other IgSF proteins are involved in many different functions (in ligand-receptor interactions in development, differentiation, activation, adhesion, regulation, etc.) [15-25].

The general organization of the IgSF other than IG and TR is more diverse and follows the modular shuffling between domains ranging from a unique V-LIKE-DOMAIN or a unique C-LIKE-DOMAIN or to any combination of those domains [15-25]. As examples, the MOG and MPZ (or P0) proteins have a unique N-terminal V-LIKE-DOMAIN, the CEA family proteins have a single N-terminal V-LIKE-DOMAIN followed by a variable number of C-LIKE-DOMAINs and the VCAM1 protein is composed of seven C-LIKE-DOMAINs [15, 23-25]. IgSF proteins with diverse V type domain and C type domain combinations, interspersed or not with domains belonging to other types, are continuously described [22].

3.3. Major histocompatibility complex (MHC)

The MHC proteins belong to two classes: MHC-I and MHC-II (Fig. 3). The MHC-I proteins, expressed on the cell surface of most cells, are formed by the association of a transmembrane heavy chain (I-ALPHA chain) and a non-covalently linked light chain beta-2-microglobulin (B2M) [26, 27]. The MHC-II proteins, expressed on the cell surface of professional antigen presenting cells (APC), are heterodimers formed by the association of two transmembrane chains, an alpha chain (II-ALPHA chain) and a beta chain (II-BETA chain) [26, 27].

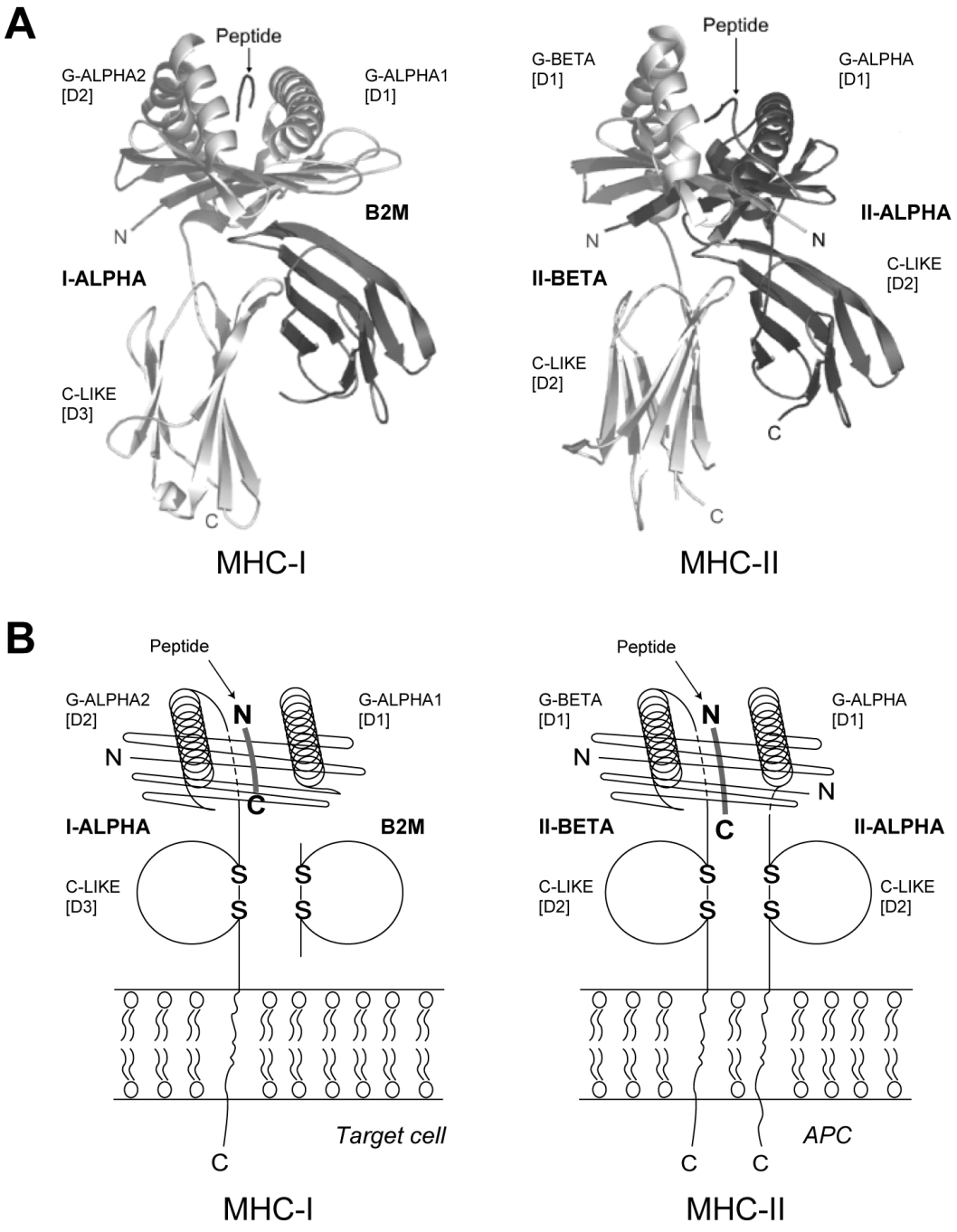

Figure 3: Three-dimensional (3D) structures and schematic representations of the MHC-I and MHC-II proteins. (A) 3D structures of MHC-I and MHC-II. The MHC-I comprises the I-ALPHA and the B2M chains. The I-ALPHA chain is shown with its extracellular domains (G-ALPHA1, G-ALPHA2 and C-LIKE) [26]. The MHC-II comprises the II-ALPHA and II-BETA chains that are shown with their extracellular domains (G-ALPHA and C-LIKE for the II-ALPHA chain, G-BETA and C-LIKE for the II-BETA chain). (B) Schematic representations of the MHC-I and MHC-II proteins. The MHC-I and MHC-II are shown as transmembrane proteins, at the surface of a target cell and of an antigen presenting cell (APC), respectively. Complete MHC-I and MHC-II chains comprise the extracellular domains (shown in A) and the connecting, transmembrane and cytoplamic regions (not present in 3D structures). [D1], [D2] and [D3] indicate the position of the domains from the N-terminal end of the chains. Arrows indicate the peptide localization in the MHC groove (the N-terminal end of the peptide is in the back).

The I-ALPHA chain of the MHC-I, and the II-ALPHA and II-BETA chains of the MHC-II proteins, comprise an extracellular region made of three domains for the MHC-I chain and of two domains for each MHC-II chain, a connecting region, a transmembrane region and an intracytoplamic region. The I-ALPHA chain comprises two groove domains (G-DOMAINs), the G-ALPHA1 [D1] and G-ALPHA2 [D2] domains, and one C-LIKE-DOMAIN [D3] [26]. The II-ALPHA chain and the II-BETA chain each comprises two domains, the G-ALPHA [D1] and one C-LIKE-DOMAIN [D2], and the G-BETA [D1] and one C-LIKE-DOMAIN [D2], respectively. The four G-DOMAINs, G-ALPHA1 and G-ALPHA2 of the MHC-I proteins, and G-ALPHA and G-BETA of the MHC-II proteins have a similar groove 3D structure [26]. Interestingly this groove is found in the "classical" MHC (MHC-Ia and MHC-IIa) proteins that present peptides to the T cells (the groove is part of the cleft that is the peptide binding site), and also in "nonclassical" MHC (MHC-Ib and MHC-IIb) proteins with more specific functions or which do not present peptides to the T cells [26].

3.4. Major histocompatibility complex superfamily (MhcSF)

The major histocompatibility complex (MHC) superfamily (MhcSF) comprises the MHC proteins which belong to two classes, MHC class I (one chain has two G-DOMAINs and one C-LIKE-DOMAIN, associated to the beta-2-microglobulin (B2M)) and MHC class II (each chain has one G-DOMAIN and one C-LIKE-DOMAIN) described in 3.3, and the proteins other than MHC defined as having a groove-like domain made up of two G-LIKE-DOMAINs, associated or not to one C-LIKE-DOMAIN [26]. The common feature of these MhcSF proteins is to have two Mhc-like domains which together contribute to a similar groove 3D structure that consists of one sheet of eight antiparallel beta strands ("floor" of the groove or platform) and two helical regions ("walls" of the groove) [26, 27]. Each domain made of four antiparallel beta strands and one helix belongs to the G type [which comprises the G-DOMAIN of the MHC, and the G-LIKE-DOMAIN of the MhcSF proteins other than the MHC] [26].

So far, only MHC-I-like chains have been identified in the MhcSF [28-32]. These chains either include a C-LIKE-DOMAIN and are bound to the B2M (e.g. CD1, FCGRT, HFE, MR1) or not (MIC, AZGP1), or do not even include a C-LIKE-DOMAIN (EPCR, RAE) [29]. The G-ALPHA1-LIKE [D1] and G-ALPHA2-LIKE [D2] domains of these proteins show a striking structural homology with the MHC G-ALPHA1 and G-ALPHA2 domains and this, despite a high sequence divergence [26]. Whereas the MHC proteins are involved in the antigen presentation to the T cells [27], the other MhcSF proteins are involved in different functions (display of phospholipid antigens for CD1, iron homeostasis for HFE, maternal IG transport for FCGRT, stress induced for MICA, etc.) [28-32].

Section 4. IMGT unique numbering

4.1. IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN

The V type domain which comprises the V-DOMAIN of IG and TR, and the V-LIKE-DOMAIN of the IgSF proteins other than the IG or TR [13] is defined by 9 antiparallel beta strands on two layers. The 3D structure of a V-LIKE-DOMAIN is very similar to that of an IG and TR V-DOMAIN (Table 2, Fig. 4A). Both domains are made of 9 antiparallel beta strands (A, B, C, C', C", D, E, F and G) linked by beta turns (AB, CC', C"D, DE and EF) or loops (BC, C'C" and FG), forming a sandwich of two sheets.

Table 2. V type domain (V-DOMAIN and V-LIKE-DOMAIN).

Strands and loops	IMGT positions[a]	Lengths[b]	Characteristic positions	FR-IMGT and CDR-IMGT in V-DOMAIN
A-STRAND	1-15	15 (14 if gap at 10)		FR1-IMGT
B-STRAND	16-26	11	1st-CYS 23	
BC-LOOP	27-38	12 (or less)		CDR1-IMGT
C-STRAND	39-46	8	CONSERVED-TRP 41	FR2-IMGT
C'-STRAND	47-55	9		
C'C"-LOOP	56-65	10 (or less)		CDR2-IMGT
C"-STRAND	66-74	9 (or 8 if gap at 73)		FR3-IMGT
D-STRAND	75-84	10 (or 8 if gaps at 81,82)		
E-STRAND	85-96	12	hydrophobic 89	
F-STRAND	97-104	8	2nd-CYS 104	
FG-LOOP	105-117	13 (or less, or more)		CDR3-IMGT
G-STRAND	118-128	11	(1)	FR4-IMGT

[a] based on the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [13].
[b] in number of amino acids (or codons).
(1) In the IG and TR V-DOMAINs, the G-STRAND is the C-terminal part of the J-REGION, with J-PHE or J-TRP 118 and the canonical motif F/W-G-X-G at positions 118-121.

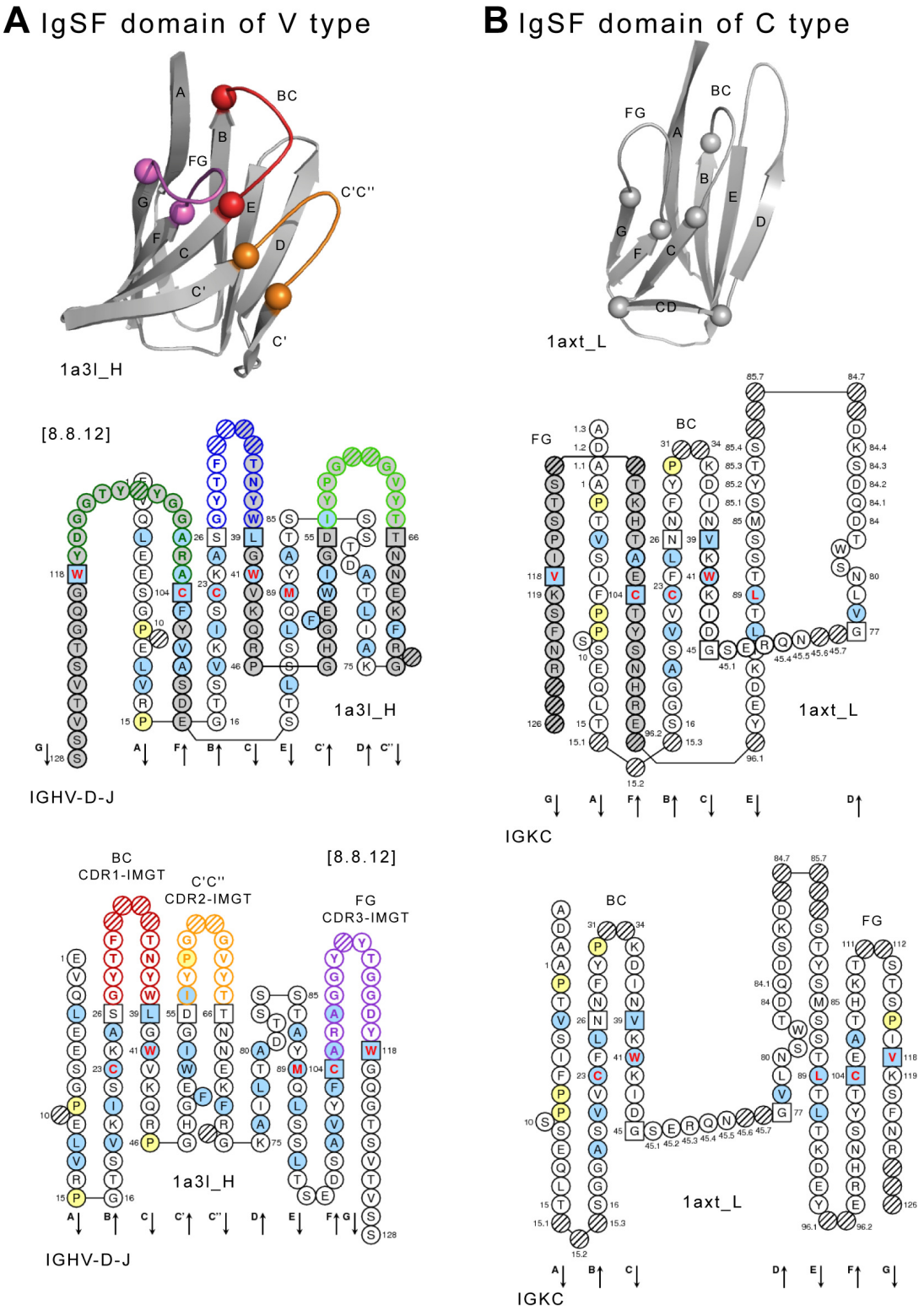

Figure 4: Ribbon representations and IMGT Colliers de Perles for V type and C type domains. (A) IgSF domain of V type, based on the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [13]. (B) IgSF domain of C type, based on the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN [14]. For the V type and C type domains, the 3D structure ribbon representation (upper part), IMGT Collier de Perles on two layers (middle part) and IMGT Collier de Perles on one layer (bottom part) are shown. Amino acids are shown in the one-letter abbreviation. Hatched circles or squares correspond to missing positions according to the IMGT unique numbering.

The sheets are closely packed against each other through hydrophobic interactions giving a hydrophobic core and joined together by a disulfide bridge between the B-STRAND in the first sheet and the F-STRAND in the second sheet [19, 33]. In the IMGT unique numbering [13, 34, 35], the conserved amino acids always have the same position, for instance cysteine 23 (1st-CYS), tryptophan 41 (CONSERVED-TRP), conserved hydrophobic amino acid 89, cysteine 104 (2nd-CYS). The IMGT Colliers de Perles are graphical two-dimensional representations [36]. IMGT Colliers de Perles for V-DOMAIN (IG and TR) and V-LIKE-DOMAIN (IgSF other than IG and TR) are based on the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [13].

The hydrophobic amino acids of the framework regions are also found in conserved positions [13]. It is remarkable that the Ig-fold 3D structure has been conserved through evolution, despite the particularities of the IG and TR synthesis compared to the other proteins and the sequence divergence of the IgSF domains. Indeed, the V-LIKE-DOMAIN is usually encoded by a unique exon, whereas the IG and TR V-DOMAIN results from the rearrangement of two (V, J) or three (V, D, J) genes [8, 9]. The V-LIKE-DOMAIN is usually, as the IG and TR V-DOMAIN, the most N-terminal (and extracellular) domain of the protein. However, in contrast to the IG and TR V-DOMAIN which is always unique, the V-LIKE-DOMAIN may be present in several copies in the same protein and interspersed with C-LIKE-DOMAINs or with domains of other superfamilies.

The antiparallel beta strands of the V-LIKE-DOMAIN correspond to the conserved framework regions (FR-IMGT) described in the IG and TR V-DOMAIN, whereas the BC, C'C" and FG loops correspond to the complementarity determining regions (CDR-IMGT) [13] (Table 2). Squares in the IMGT Colliers de Perles (Fig. 4A) indicate positions that belong to strands and represent anchor positions for BC-LOOP, C'C"-LOOP and FG-LOOP of the V type. The loop length (number of amino acids or by extrapolation number of codons, that is number of occupied positions) is a crucial and original concept of IMGT-ONTOLOGY [1]. The lengths of the BC (CDR1-IMGT), C'C" (CDR2-IMGT) and FG (CDR3-IMGT) loops characterize the V-DOMAIN and V-LIKE-DOMAIN. Thus, the length of the three loops BC, C'C" and FG is shown, in number of amino acids (or codons), into brackets and separated by dots. In Figure 4A, the CDR-IMGT lengths are [8.8.12].

4.2. IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN

The IMGT Colliers de Perles for C-DOMAIN (IG and TR) and C-LIKE-DOMAIN (IgSF other than IG and TR) are based on the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN [14]. This numbering is itself derived from the IMGT unique numbering first described for the V-REGION [34, 35] and for the V-DOMAIN [13]. Indeed, the sandwich beta sheet of the C type (C-DOMAIN and C-LIKE-DOMAIN) has the same topology and similar 3D structure than the V type (V-DOMAIN and V-LIKE-DOMAIN), although it differs by the number of strands (Fig. 4B, Table 3). The C-DOMAIN and C-LIKE-DOMAIN are made of seven beta strands linked by beta turns or loops, and arranged so that four strands form one sheet and three strands form a second sheet [14]. A characteristic transversal CD-STRAND links the two sheets; depending on the CD-STRAND length, the D-STRAND is in the first or in the second sheet [14].

Table 3. C type domain (C-DOMAIN and C-LIKE-DOMAIN).

Strands, loops and turns	IMGT positions[a]	Lengths[b]	Characteristic positions
A-STRAND	1-15	15	
AB-TURN	15.1-15.3	0-3	
B-STRAND	16-26	11	1st-CYS 23
BC-LOOP	27-38	10 (or less)	no 32, 33[c]
C-STRAND	39-45	7	CONSERVED-TRP 41 no 46[c]
CD-STRAND	45.1-45.9	1-9	
D-STRAND	77-84	8	no 75, 76[c]
DE-TURN	84.1-84.7, 85.7-85.1	0-14	
E-STRAND	85-96	12	hydrophobic 89
EF-TURN	96.1-96.2	0-2	
F-STRAND	97-104	8	2nd-CYS 104
FG-LOOP	105-117	13 (or less, or more)	
G-STRAND	118-128	11 (or less)	

[a] based on the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN [14].
[b] in number of amino acids (or codons).
[c] compared to V type.

The C'-STRAND, C'C"-LOOP and C"-STRAND are missing in the C type and are replaced by the characteristic transversal CD-STRAND [14]. Squares in the IMGT Colliers de Perles (Fig. 4B) indicate positions that belong to strands and represent anchor positions for BC-LOOP, CD-STRAND and FG-LOOP of the C type domain. Additional positions in the C type define the AB-TURN, DE-TURN and EF-TURN [14].

4.3. IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN

The IMGT unique numbering for G-DOMAIN (MHC) and G-LIKE-DOMAIN (MhcSF other than MHC) [26] has allowed standardized representations or IMGT Colliers de Perles of the groove domain whatever the MhcSF protein (MHC-I, MHC-II or Mhc-I-like) whatever the G domain and whatever the species (Fig. 5, Table 4).

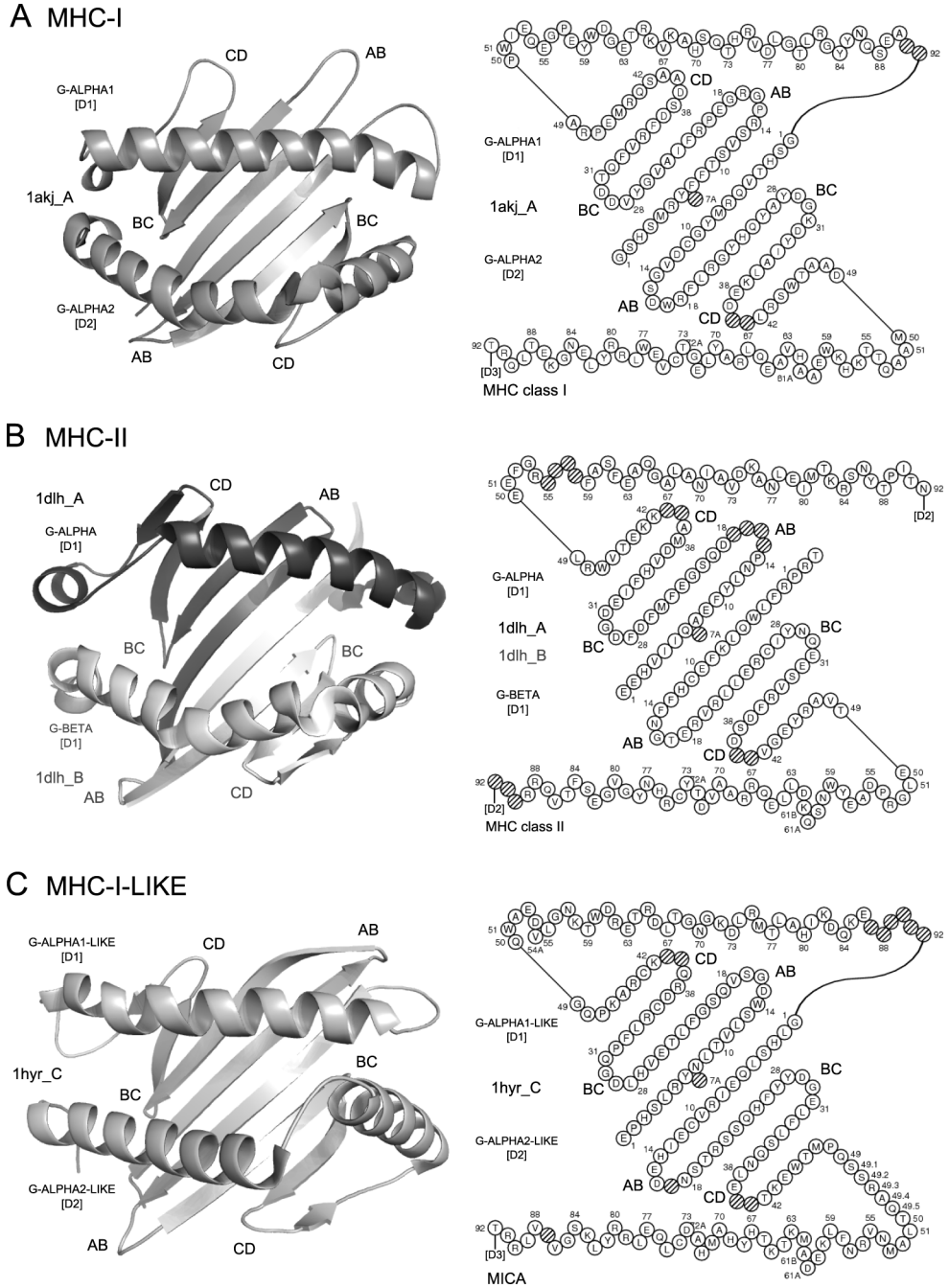

Figure 5: Ribbon representations and IMGT Colliers de Perles for G type domains. (A) MHC-I, (B) MHC-II, (C) MHC-I-like. The MhcSF domain of G type is based on the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN [26]. Note that the N-terminal end of a peptide in the cleft of MHC-I and MHC-II G-DOMAINs would be on the left hand side. Amino acids are shown in the one-letter abbreviation. Hatched circles correspond to missing positions according to the IMGT unique numbering.

Table 4. G type domain (G-DOMAIN and G-LIKE-DOMAIN).

Strands, turns and helix	IMGT positions[a]	Lengths[b]	Characteristic positions[c]
A-STRAND	1-14	14	7A, CYS-11
AB-TURN	15-17	3 (or 0)	
B-STRAND	18-28	11	
BC-TURN	29-30	2	
C-STRAND	31-38	8	
CD-TURN	39-41	3 (or 1)	
D-STRAND	42-49	8	
HELIX	50-92	43-48	54A, 61A, 61B, 72A, CYS-74, 92A

[a] based on the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN [26].
[b] in number of amino acids (or codons).
[c] for more details, see [26].

For each G type domain (G-DOMAIN and G-LIKE-DOMAIN), the positions that contribute to the groove floor comprise positions 1 to 49, with four strands linked by turns [26]. The numbering of the G type helix starts at position 50 and ends at position 92. Interestingly, we showed that, despite the high sequence divergence, only five additional positions (54A, 61A, 61B, 72A and 92A) are necessary to align any G-DOMAIN and G-LIKE-DOMAIN [26, 37]. It is worthwhile to note that position 54A in G-ALPHA1-LIKE is the only additional position needed to extend the IMGT numbering for G-DOMAIN to the G-LIKE-DOMAINs.

The helix (positions 50 to 92) seats on the beta sheet and its axis forms an angle of about 40 degrees with the beta strands. Two cysteines, CYS-11 (in strand A) and CYS-74 (in the helix) are well conserved in the G-ALPHA2 and G-BETA domains where they participate to a disulfide bridge that fastens the helix on the groove floor. The IMGT Colliers de Perles allow to describe specific features (detailed in [26, 37]). As an example, the G-ALPHA1 and G-ALPHA domains have a conserved N-glycosylation site at position 86 (N-X-S/T, where N is asparagine, X any amino acid except proline, S is serine and T is threonine).

Section 5. Implementation plan

Any domain represented by an IMGT Collier de Perles is characterized by the length of its strands, loops and turns and, for the G type, by the length of its helix [13, 14, 26]. The strand, loop, turn or helix lengths (the number of amino acids or codons, that is the number of occupied positions) become crucial information which characterizes the domains. This first feature of the IMGT standardization based on the IMGT unique numbering allowed, for instance, to show that the distinction between the C1, C2, I1 and I2 domain types found in

the literature and in the databases to describe the IgSF C type domains is unnecessary and moreover unapplicable when dealing with sequences for which no structural data are known (discussed in [14]).

A second feature of the IMGT standardization is the comparison of cDNA and/or amino acid sequences with genomic sequences, and the identification of the splicing sites, to delimit precisely the domains: a V-LIKE-DOMAIN, a C-DOMAIN, a C-LIKE-DOMAIN, a G-DOMAIN or a G-LIKE-DOMAIN is frequently encoded by a unique exon [13, 14, 26]. This IMGT standardization for the domain delimitation explains the discrepancies observed with the generalist UniProt/Swiss-Prot database which identifies domains based on amino acid sequences and does not take into account the genomic information. The IMGT Colliers de Perles also put the question of the leader region. Indeed, the N-terminal end of the first domain of an IgSF or MhcSF chain depends on the proteolytic cleavage site of the leader region (peptide signal) which is rarely determined experimentally. When this site is not known, the IMGT Colliers de Perles start with the first amino acid resulting from the splicing ("Splicing sites" in IMGT Aide-mémoire, http://imgt.cines.fr). For a IG and TR V-DOMAIN the leader proteolytic site is known (or is extrapolated) and the IMGT Colliers de Perles start with the first amino acid of the V-REGION [8, 9].

The IMGT Colliers de Perles allow a precise visualization of the inter-species differences for the IgSF V and C type domain strands and loops, and MhcSF G type domain strands and helix, even in the absence of 3D structures. This has been applied to the teleost CD28 family members and their B7 family ligands and to the BTLA protein, which belong to the IgSF by their V type and/or C type domains [21, 22]. The IMGT Colliers de Perles are particularly useful in molecular engineering and antibody humanization design based on CDR grafting. Indeed they allow to precisely define the CDR-IMGT and to easily compare the amino acid sequences of the four FR-IMGT (FR1-IMGT: positions 1 to 26, FR2-IMGT: 39 to 55, FR3-IMGT: 66 to 104, and FR4-IMGT: 118 to 128) between the murine and the closest human V-DOMAINs. A recent analysis performed on humanized antibodies used in oncology underlines the importance of a correct delimitation of the CDR regions to be grafted [38].

The IMGT Colliers de Perles also allow a comparison with the IMGT Collier de Perles statistical profiles for the human expressed IGHV, IGKV and IGLV repertoires [39]. These statistical profiles are based on the definition of eleven IMGT amino acid physicochemical characteristics classes which take into account the hydropathy, volume and chemical characteristics of the 20 common amino acids [39] ("Amino acids" in IMGT Aide-mémoire, http://imgt.cines.fr). The statistical profiles identified positions which are conserved for the physicochemical characteristics: 41 FR-IMGT positions for the human IGHV and 59 FR-IMGT positions for the human IGKV and IGLV at >80% threshold (see Plate 3 in [39]). After assignment of the IMGT Collier de Perles amino acids to the IMGT amino acid physicochemical classes, comparison can be made with the statistical profiles of the human expressed repertoires. This comparison is useful to identify potential immunogenic residues at given positions in chimeric or humanized antibodies [38] or to evaluate immunogenicity of primate antibodies [40].

IMGT Colliers de Perles are also of interest when 3D structures are available. In IMGT/3Dstructure-DB [33], "IMGT Collier de Perles on 2 layers" are displayed with hydrogen bonds for V type and C type domains. Clicking on a residue in ‘IMGT Collier de Perles on one layer’ gives access to the corresponding IMGT Residue@Position card which provides the atom contact types and atom contact categories for that amino acid. IMGT Colliers de Perles display the IMGT pMHC contact sites for 3D structures with peptide/MHC

(pMHC) complexes [27], which can be compared with the "IMGT reference pMHC contact sites" available in IMGT/3Dstructure-DB [41].

The IMGT Colliers de Perles for the V type, C type and G type, based on the IMGT unique numbering, represent therefore a major step forward for the comparative analysis of the sequences and structures of the IgSF and MhcSF domains, for the study of their evolution and for the applications in antibody engineering [38], IG and TR repertoires in autoimmune diseases and leukemias [42], pMHC contact analysis [41], and more generally ligand-receptor interactions involving V type, C type and/or G type domains.

Section 6. 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 the deliverable D1.2, and the NUMEROTATION axiom defined in this deliverable D1.3, are valid, not only for molecules, but also for cells, tissues, organs, organisms or populations. In addition, the 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. IMGT-ONTOLOGY will allow interactions with IMGT®, the global reference in immunogenetics and immunoinformatics. This will further strengthen the importance of standardization in pharmaceutical and clinical research, as demonstrated by companies which, in Europe (SANOFI-AVENTIS, Institut Pierre Fabre...), and in the USA (CENTOCOR Johnson and Johnson, MERCK, AMGEN...), have IMGT® licences and contracts.

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.

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.

Section 7. References

1. Giudicelli, V, Lefranc, M-P. Ontology for Immunogenetics: IMGT-ONTOLOGY. Bioinformatics 1999;15:1047-1054.
2. 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.
3. Lefranc M-P, Clément O, Kaas Q et al. IMGT-Choreography for Immunogenetics and Immunoinformatics (http://imgt.cines.fr). *In Silico* Biol 2005;29:185-203.
4. Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V. IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie 2007 Sep 11; [Epub ahead of print].
5. Lefranc M-P. IMGT-ONTOLOGY, IMGT® databases, tools and Web resources for Immunoinformatics. In: Immunoinformatics (Schoenbach C., Ranganathan S. and Brusic V. eds.), Immunomics Reviews, Springer, New York, USA, 2008, pp.1-18.
6. Lefranc M-P, Giudicelli V, Kaas Q et al. IMGT, the international ImMunoGeneTics information system®. Nucl Acids Res 2005;33:D593-D597.

7. Lefranc M-P. IMGT, the international ImMunoGeneTics information system®: a standardized approach for immunogenetics and immunoinformatics. Immunome Res 2005 Sep 20;1:3.

8. Lefranc M-P, Lefranc G. The Immunoglobulin FactsBook. London: Academic Press, 2001,1-458.

9. Lefranc M-P, Lefranc G. The T cell receptor FactsBook. London: Academic Press, 2001,1-398.

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

11. 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;32:461-463.

12. Lefranc M-P. WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report. Immunogenetics 2007;59:899-902.

13. 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 2002;27:55-77.

14. 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.

15. 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;2:111-136.

16. Williams AF, Barclay AN. The immunoglobulin superfamily-domains for cell surface recognition. Annu Rev Immunol 1988;6:381-405.

17. Hunkapiller T, Hood L. Diversity of the immunoglobulin gene superfamily. Adv Immunol 1989;44:1-63.

18. Jones EY. The immunoglobulin superfamily. Curr Opin Struct Biol 1993;3:846-852.

19. Bork P, Holm L, Sander C. The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol 1994;242:309-320.

20. 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;64:119-131.

21. Bernard D, Hansen JD, du Pasquier L et al. Costimulatory receptors in non mammalian vertebrates: conserved CD28, odd CTLA4 and multiple BTLAs. Dev Comp Immunol 2007;31:255-271.

22. Garapati VP and 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;31:1050-1072.

23. Gardinier MV, Amiguet P, Linington C et al. Myelin/Oligodendrocyte glycoprotein is a unique member of the immunoglobulin superfamily. J Neurosci Res 1992;33:177-187.

24. Schrewe H, Thompson J, Bona M et al. Cloning of the complete gene for carcinoembryonic antigen: analysis of its promoter indicates a region conveying cell type-specific expression. Mol Cell Biol 1990;10:2738-2748.

25. Cybulsky MI, Fries JWU, Williams AJ et al. Gene structure, chromosomal location, and basis for alternative mRNA splicing of the human VCAM1 gene. Proc Natl Acad Sci USA 1991;88:7859-7863.

26. 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;29:917-938.

27. Kaas Q, Lefranc M.-P. T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico* Biology 2005;5:505-528.

28. Maenaka K, Jones EY. MHC superfamily structure and the immune system. Curr Opin Struct Biol 1999;9:745-753.

29. 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;22:453-459.

30. Frigoul A, Lefranc M-P. MICA: standardized IMGT allele nomenclature, polymorphisms and diseases. Recent Res Devel Human Genet 2005;3:95-105.

31. Wilson IA, Bjorkman PJ. Unusual MHC-like molecules: CD1, Fc receptor, the hemochromatosis gene product, and viral homologs. Curr Opin Immunol 1998;10:67-73.

32. Braud VM, Allan DS, McMichael AJ. Functions of nonclassical MHC and non-MHC-encoded class I molecules. Curr Opin Immunol 1999;11:100-108.

33. 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. Nucl Acids Res 2004;32:D208-D210.

34. Lefranc M-P. Unique database numbering system for immunogenetic analysis. Immunol Today 1997;18:509.

35. Lefranc M-P. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 1999;7:132-136.

36. Ruiz M, Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulin with known 3D structures. Immunogenetics 2002;53:857-883.

37. Kaas Q, Lefranc M-P. IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. Curr Bioinformatics 2007;2:21-30.

38. 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;64:210-225.

39. Pommié C, Levadoux S, Sabatier R, et al. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognition 2004;17:17-32.

40. 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;49:3414-3420.

41. Kaas Q., Duprat E., Tourneur G. et al. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: Immunoinformatics (Schoenbach C., Ranganathan S. and Brusic V. eds.), Immunomics Reviews, Springer, New York, USA 2008, chap. 2, pp.19-49.

42. 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;36:1963-1974.