In Silico Biology 4 (2004) 17–29
IOS Press

IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics

Marie-Paule Lefranc*, Véronique Giudicelli, Chantal Ginestoux, Nathalie Bosc, Géraldine Folch, Delphine Guiraudou, Joumana Jabado-Michaloud, Séverine Magris, Dominique Scaviner, Valérie Thouvenin, Kora Combres, David Girod, Stéphanie Jeanjean, Céline Protat, Mehdi Yousfi Monod, Elodie Duprat, Quentin Kaas, Christelle Pommié, Denys Chaume and Gérard Lefranc

IMGT, the international ImMunoGeneTics information system® , Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France
Tel.: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01

Edited by E. Wingender; received 11 August 2003; accepted 16 November 2003; published 22 November 2003

ABSTRACT: IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr), is a high quality integrated knowledge resource specialized in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) and related proteins of the immune system (RPI) of human and other vertebrates, created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire LIGM. IMGT provides a common access to standardized data which include nucleotide and protein sequences, oligonucleotide primers, gene maps, genetic polymorphisms, specificities, 2D and 3D structures. IMGT consists of several sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB, IMGT/PRIMER-DB), one genome database (IMGT/GENE-DB) and one three-dimensional structure database (IMGT/3Dstructure-DB), interactive tools for sequence analysis (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/PhyloGene, IMGT/Allele-Align), for genome analysis (IMGT/GeneSearch, IMGT/GeneView, IMGT/LocusView) and for 3D structure analysis (IMGT/StructuralQuery), and Web resources ("IMGT Marie-Paule page") comprising 8000 HTML pages. IMGT other accesses include SRS, FTP, search by BLAST, etc. By its high quality and its easy data distribution, IMGT has important implications in medical research (repertoire in autoimmune diseases, AIDS, leukemias, lymphomas, myelomas), veterinary research, genome diversity and genome evolution studies of the adaptive immune responses, biotechnology related to antibody engineering (scFv, phage displays, combinatorial libraries) and therapeutical approaches (grafts, immunotherapy).
IMGT is freely available at http://imgt.cines.fr.

KEYWORDS: IMGT, ontology, database, information system, knowledge resource, immunoinformatics, immunogenetics, antibody, immunoglobulin, T cell receptor, immunoglobulin superfamily, leukemia, lymphoma, MHC, HLA, Collier de Perles, three-dimensional (3D) structure, primer, polymorphism.

*Corresponding author. E-mail: lefranc@ligm.igh.cnrs.fr.
Institut Universitaire de France.

Electronic publication can be found in *In Silico Biol.* **4**, 0004 <http://www.bioinfo.de/isb/2003/04/0004/>, 22 November 2003.

INTRODUCTION

The molecular synthesis and genetics of the immunoglobulin (IG) and T cell receptor (TR) chains is particularly complex and unique as it includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (for review, see Lefranc and Lefranc, 2001a; 2001b). The number of potential protein forms of IG and TR is almost unlimited. Owing to the complexity and high number of published sequences, data control and classification and detailed annotations are a very difficult task for the generalist databanks such as EMBL (UK) [Stoesser *et al*., 2003], GenBank (USA) [Benson *et al*., 2003] and DDBJ (Japan) [Miyazaki *et al*., 2003]. These observations were the starting point of IMGT, the international ImMunoGeneTics information system® [Lefranc, 2003a], created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) (Université Montpellier II and CNRS) at Montpellier, France. IMGT is a high quality integrated knowledge resource specialized in IG, TR, major histocompatibility complex (MHC) and related proteins of the immune systems (RPI) of human and other vertebrate species [Giudicelli *et al*., 1997; Lefranc *et al*., 1998; 1999; Ruiz *et al*., 2000; Lefranc, 2001a; 2002; 2003a; 2003b]. IMGT consists of several sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB, IMGT/PRIMER-DB, IMGT/PROTEIN-DB, this last one in development), one genome database (IMGT/GENE-DB) and one three-dimensional 3D structure database (IMGT/3Dstructure-DB), interactive tools for sequence analysis (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/PhyloGene, IMGT/Allele-Align), for genome analysis (IMGT/GeneSearch, IMGT/GeneView, IMGT/LocusView) and for 3D structure analysis (IMGT/StructuralQuery), and Web resources ("IMGT Marie-Paule page") comprising 8000 HTML pages which include IMGT Scientific chart, IMGT Repertoire (for IG and TR, MHC, RPI), IMGT Bloc-notes, IMGT Education and IMGT Index. IMGT other accesses include SRS, FTP, search by BLAST, etc. By its high quality and its easy data distribution, IMGT has important implications in medical research (repertoire in normal and pathological situations: autoimmune diseases, infectious diseases, AIDS, detection of residual diseases in leukemias, lymphomas, myelomas), veterinary research, genome diversity and genome evolution studies of the adaptive immune responses, biotechnology related to antibody engineering (single chain Fragment variable (scFv), phage displays, combinatorial libraries) and therapeutical approaches (grafts, immunotherapy). IMGT is freely available at http://imgt.cines.fr.

IMGT-ONTOLOGY

IMGT has developed a formal specification of the terms to be used in the domain of immunogenetics and immunoinformatics to ensure accuracy, consistency and coherence in IMGT. This has been the basis of IMGT-ONTOLOGY [Giudicelli and Lefranc, 1999], the first ontology in the domain, which allows the management of the immunogenetics knowledge for human and other vertebrate species. IMGT-ONTOLOGY comprises five main concepts: IDENTIFICATION, CLASSIFICATION, DESCRIPTION, NUMEROTATION and OBTENTION [Giudicelli and Lefranc, 1999]. Standardized keywords, standardized sequence annotation, standardized IG and TR gene nomenclature, the IMGT unique numbering, and standardized origin/methodology were defined, respectively, based on these five main concepts. The controlled vocabulary and the annotation rules for data and knowledge management of the IG, TR, MHC and RPI of human and other vertebrate species constitute the IMGT Scientific chart. All IMGT data are expertly annotated according to the IMGT Scientific chart. IMGT is the global internationally acknowledged reference in immunogenetics and immunoinformatics.

The IDENTIFICATION concept: standardized keywords

IMGT standardized keywords have been assigned to all IMGT/LIGM-DB entries. They include (i) *general keywords*: indispensable for the sequence assignments, they are described in an exhaustive and non redundant list, and are organized in a tree structure, and (ii) *specific keywords*: they are more specifically associated with particularities of the sequences (orphon, transgene, etc.) or with diseases (leukemia, lymphoma, myeloma, etc.) [Giudicelli *et al*., 1997]. The list is not definitive and new specific keywords can easily be added if needed.

The DESCRIPTION concept: standardized labels and annotations

Two hundred fifteen feature labels are necessary in IMGT/LIGM-DB to describe all structural and functional subregions that compose IG and TR sequences [Giudicelli *et al*., 1997], whereas only seven of them are available in EMBL, GenBank or DDBJ. Annotation of sequences with these labels constitutes the main part of the expertise. Levels of annotation have been defined, which allow the users to query sequences in IMGT/LIGM-DB even though they are not fully annotated [Giudicelli et *et al*., 1997]. An internal tool, IMGT/Automat, has been developed to automatically perform the annotation of the rearranged cDNA sequences in IMGT/LIGM-DB [Giudicelli *et al*., 2003]. One hundred seventy two additional labels were defined for IG, TR, MHC and RPI amino acid sequences and domain structures in IMGT/PROTEIN-DB and IMGT/3Dstructure-DB. Prototypes represent the organizational relationship between labels and give information on the order and expected length (in number of nucleotides) of the labels [Giudicelli *et al*., 1997; Lefranc *et al*., 1999]. Prototype can apply to general configuration of IG, TR or MHC, independently of the chain type, the species or any other parameters like functionality. However, prototypes may also be established for very precise cases when sequence characteristics are clearly established.

The CLASSIFICATION concept: standardized IG and TR gene nomenclature

The objective is to provide immunologists and geneticists with a standardized nomenclature per locus and per species which will allow extraction and comparison of data for the complex B and T cell antigen receptor molecules. The CLASSIFICATION concept has been used to set up a unique nomenclature of human IG and TR genes, which was approved by the Human Genome Organization (HUGO) Nomenclature Committee (HGNC) in 1999 [Wain *et al*., 2002] and has become the community standard. The complete list of the human IG and TR gene names [Lefranc and Lefranc, 2001a; 2001b; Lefranc, 2000a; 2000b; 2000c; 2000d; Lefranc, 2001b; 2001c; 2001d] has been entered by the IMGT Nomenclature Committee in the Genome DataBase GDB (Canada), LocusLink at NCBI (USA), and GeneCards. The complete list of the mouse IG and TR gene names was sent by IMGT, in July 2002, to the Mouse Genome Informatics MGI Mouse Genome Database MGD (USA), LocusLink at NCBI, and HGNC. Both lists are available from the IMGT site [Lefranc, 2003a] and queries on the human and mouse IG and TR gene classification and gene names can be made from IMGT/GENE-DB. IMGT reference sequences have been defined for each allele of each gene based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence, mapped sequence [Lefranc *et al*, 1999].

The NUMEROTATION concept: the IMGT unique numbering

A uniform numbering system for IG and TR sequences of all species has been established to facilitate

sequence comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor (IG or TR), the chain type, the domain (variable V or constant C), or the species [Lefranc, 1997; 1999; Lefranc *et al*., 2003]. This numbering results from the analysis of more than 5000 IG and TR variable region sequences of vertebrate species from fish to human. It takes into account and combines the definition of the framework (FR) and complementarity determining regions (CDR) [Kabat, 1991], structural data from X-ray diffraction studies [Satow *et al*., 1986] and the characterization of the hypervariable loops [Chothia and Lesk, 1987]. In the IMGT numbering, conserved amino acids from frameworks always have the same number whatever the IG or TR variable sequence, and whatever the species they come from (as examples: Cysteine 23 (in FR1), Tryptophan 41 (in FR2), Leucine 89 and Cysteine 104 (in FR3). Based on the IMGT unique numbering, standardized 2D graphical representations, designated as IMGT Colliers de Perles [Lefranc *et al*., 1999], are available in IMGT Repertoire. This IMGT unique numbering has several advantages:

- It has allowed the redefinition of the limits of the FR and CDR of the IG and TR variable regions [Lefranc and Lefranc, 2001a; 2001b] and domains [Ruiz and Lefranc, 2002]. The FR-IMGT and CDR-IMGT lengths become in themselves crucial information which characterize variable regions belonging to a group, a subgroup and/or a gene.
- Framework amino acids (and codons) located at the same position in different sequences can be compared without requiring sequence alignments. This also holds for amino acids belonging to CDR-IMGT of same length.
- The unique numbering is used as the output of the IMGT/V-QUEST alignment tool. The aligned sequences are displayed according to the IMGT numbering and with the FR-IMGT and CDR-IMGT delimitations.
- The unique numbering has allowed a standardization of the description of mutations and the description of IG and TR allele polymorphisms [Lefranc and Lefranc, 2001a; 2001b]. These mutations and allelic polymorphisms are described by comparison to the IMGT reference sequences of the alleles *01 [Lefranc, 1998].
- The unique numbering allows the description and comparison of somatic hypermutations of the IG IMGT variable domains.

By facilitating the comparison between sequences and by allowing the description of alleles and mutations, the IMGT unique numbering represents a big step forward in the analysis of the IG and TR variable region (V-REGION) sequences of all vertebrate species [Pommié *et al*., 2003] (IMGT Repertoire (IG and TR)). Moreover, it gives insight into the structural configuration of the variable domain (V-DOMAIN encoded by the V-J- or V-D-J-REGION) [Ruiz and Lefranc, 2002; Lefranc *et al*., 2003]. The IMGT unique numbering opens interesting views on the evolution of the proteins belonging to the "immunoglobulin superfamily" (IgSF) [Williams and Barclay, 1988]. It has been applied with success to all the sequences of domains belonging to the IgSF V-set which comprises the V-DOMAINs if the IG and TR and the V-LIKE-DOMAINs of proteins other than IG or TR, these include non rearranging sequences in vertebrates (human CD4 [D1,D3], Xenopus CTXg1, etc.) and in invertebrates (drosophila amalgam, drosophila fasciclin II, etc.) [Lefranc, 1997; 1999; Lefranc *et al*., 2003]. This standardized approach has been applied to the IgSF C-set which comprises the constant domains (C-DOMAINs) of the IG and TR, and to the C-LIKE-DOMAINs of proteins other than IG and TR. An IMGT unique numbering has also been implemented for the groove domain (G-DOMAIN) [Duprat *et al*., 2003] of the MHC class I and II chains (IMGT Repertoire (MHC)), and for the G-LIKE-DOMAINs of proteins other than MHC (IMGT Repertoire (RPI)).

The OBTENTION concept: standardized origin/methodology

The OBTENTION concept is a set of standardized terms that precise the origins of the sequence (the

'origin' concept) and the conditions in which the sequences have been obtained (the 'methodology' concept). The 'origin' concept comprises the subsets of 'cell, tissue or organ', 'auto-immune diseases', 'clonal expansion diseases' (such as leukemia, lymphoma, myeloma), whereas the 'methodology' concept comprises the subsets related to the 'hybridoma', to the experimental conditions (sequences amplified by 'PCR'), to the obtention from 'libraries' (genomic, cDNA, combinatorial, etc.) or from 'transgenic' organisms (animal, plant). At this stage of development, the exhaustive definition of the concepts of obtention and of their instances is still in progress.

IMGT DATABASES

IMGT sequence databases: IMGT/LIGM-DB, IMGT/MHC-DB, IMGT/PRIMER-DB

IMGT/LIGM-DB is the comprehensive IMGT database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, based on the IDENTIFICATION and DESCRIPTION concepts, created in 1989 by LIGM, Montpellier, France, on the Web since July 1995 [Giudicelli *et al*., 1997; Lefranc *et al*., 1998; 1999; Ruiz *et al*., 2000; Lefranc, 2001a; 2002; 2003a; 2003b].

IMGT/LIGM-DB is the first and the largest database of IMGT, the international ImMunoGeneTics information system®. In March 2004, IMGT/LIGM-DB contained 81,300 nucleotide sequences of IG and TR from human and 150 other vertebrate species.

IMGT/LIGM-DB sequence data are identified by the EMBL/GenBank/DDBJ accession number. The unique source of data for IMGT/LIGM-DB is EMBL which shares data with the other two generalist databases GenBank and DDBJ (IMGT/LIGM-DB Sequence submission). Once the sequences are allowed by the authors to be made public, LIGM automatically receives IG and TR sequences by e-mail from EMBL. After control by LIGM curators, data are scanned to store sequences, bibliographical references and taxonomic data, and standardized IMGT/LIGM-DB keywords are assigned to all entries. Based on expert analysis, specific detail annotations are added to IMGT flat files in a second step [Giudicelli *et al*., 1997]. Since August 1996, the IMGT/LIGM-DB content has closely followed that of the EMBL for the IG and TR, with the following advantages: IMGT/LIGM-DB contains IG and TR entries which have disappeared from the generalist databases (as examples: the L36092 accession number which encompasses the complete human TRB locus is still present in IMGT/LIGM-DB, whereas it has been deleted from EMBL/GenBank/DDBJ due to its too large size (684973 bp); in 1999, IMGT detected the disappearance of 20 IG and TR sequences which inadvertently had been lost by GenBank, and allowed the recuperation of these sequences in the generalist databases); conversely, IMGT/LIGM-DB does not contain sequences which have previously been wrongly assigned to IG and TR.

The IMGT/LIGM-DB data, based on the DESCRIPTION concepts, are provided with a user friendly interface. The Web interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge in a computing language. The interface allows the users to get easily connected from any type of platform (PC, Macintosh, workstation) using freeware such as Explorer, Netscape. All IMGT/LIGM-DB information is available through five modules of search: Catalogue, Taxonomy and Characteristics, Keywords, Annotation labels and References. Selection is displayed at the top of the "results of your search" page, so the users can check their own queries [Lefranc *et al*., 1999]. Users have the possibility to modify their request or to consult the results. They can (i) add new conditions to increase or decrease the number of resulting sequences, (ii) view details: selecting this "View" option provides a list of resulting sequences; selection of one sequence in the list offers nine possibilities: annotations, IMGT flat file, coding regions with protein translation, catalogue and external references, sequence in dump format, sequence in FASTA format, sequence with three reading frames, EMBL flat file,

IMGT/V-QUEST, or (iii) search for sequence fragments: selecting this "Subsequences" option allows to search for sequence fragments (subsequences) corresponding to a particular label for the resulting sequences (available for fully annotated sequences) [Lefranc *et al*., 1999].

IMGT/LIGM-DB data are also distributed by anonymous FTP servers at the Centre Informatique National de l'Enseignement Supérieur (CINES), Montpellier, France (ftp://ftp.cines.fr/IMGT/), at the European Bioinformatics Institute (EBI), Hinxton, UK (ftp://ftp.ebi.ac.uk/pub/databases/imgt/) and at the Institut de Génétique Humaine (IGH), Montpellier, France (ftp://ftp.igh.cnrs.fr/pub/IMGT), and from many SRS (Sequence Retrieval System) sites (IMGT other accesses>SRS). IMGT/LIGM-DB releases are produced weekly. Users can compare their own sequences against IMGT/LIGM-DB data using BLAST or FASTA on different servers (EBI, IGH, INFOBIOGEN, Institut Pasteur Paris, etc.) (IMGT other accesses>Compare your sequence against IMGT (BLAST, FASTA)).

IMGT/MHC-DB, hosted on the EBI server at Hinxton (UK), comprises a database of the human MHC allele sequences (IMGT/MHC-HLA, developed by Cancer Research, UK and Anthony Nolan Research Institute (ANRI), London, UK, on the Web since December 1998 (1,719 entries in March 2004) [Robinson *et al*., 2000; 2003], databases of the MHC class II sequences from non-human primates (IMGT/MHC-NHP, curated by the Biomedical Primate Research Centre (BPRC), Rijswijk, The Netherlands) and from felines and canines (IMGT/MHC-FLA and IMGT/MHC-DLA, curated by the Centre for Integrated Genomic Medical Research, Manchester, UK), on the Web since April 2002 [Robinson *et al*., 2003].

IMGT/PRIMER-DB is the IMGT oligonucleotide (primer) database for the IG and TR, created by LIGM (Montpellier, France) in collaboration with EUROGENTEC S.A. (Seraing, Belgium), on the Web since February 2002. The IG and TR primers are useful for the analysis of the IG and TR gene repertoire and expression, the detection of minimal residual diseases in B and T cell malignancies, the construction of antibody combinatorial libraries, scFv, phage display or microarray technologies. In March 2004, IMGT/PRIMER-DB contained 1,827 entries from *Homo sapiens* and other species.

IMGT/PRIMER-DB contains information on primers and combinations of primers described as "sets" [primers sharing identical properties (species, group and orientation)] and "couples" [sets of opposite orientation for which IMGT/LIGM-DB sequences are known (or expected)]. Primers, Sets and Couples are described in IMGT Primer cards, IMGT Set cards and IMGT Couple cards, respectively. An IMGT Primer is an oligonucleotide described by comparison to an IMGT/LIGM-DB reference sequence, according to the standardized rules of the IMGT Scientific chart, based on IMGT-ONTOLOGY [Giudicelli and Lefranc, 1999]. Taxonomy species and IMGT classification (group, subgroup, gene, allele) of a primer are those of the IMGT/LIGM-DB reference sequence, and not those of the PCR amplification products. This provides the following advantages for the data standardization: IMGT/PRIMER-DB primer definition, classification and description are independent from the experimental conditions, from DNA sources and from the different combinations (sets and couples) in which the primer can be used. That means that (i) the specificity of a primer (subgroup, gene or allele specific) which is either described experimentally or deduced from sequence comparison is not used for the primer classification, although these data are provided in the IMGT/PRIMER-DB Primer card in "Classification comments and specificity", (ii) the sequences resulting from the PCR amplifications are uniquely associated to the couples. The IMGT Primer cards are linked to IMGT/LIGM-DB flat files, IMGT Colliers de Perles and IMGT Repertoire>Alignments of alleles of the IMGT/LIGM-DB reference sequence used for the primer description.

IMGT genome database: IMGT/GENE-DB

IMGT/GENE-DB is the comprehensive IMGT genome database for IG and TR genes from human and mouse, and in development, from other vertebrates, created by LIGM, on the Web since January 2003.

IMGT/GENE-DB annotated data are extracted from IMGT Repertoire, the global ImMunoGeneTics Web Resource, and from the IMGT/LIGM-DB database. All the human IMGT gene names [Lefranc and Lefranc, 2001a; 2001b] were approved by HGNC in 1999 [Wain *et al*., 2002], and entered in GDB (Canada), LocusLink, NCBI (USA), and GeneCards. In March 2004, IMGT/GENE-DB contained 1,375 genes and 2,201 alleles (673 IG and TR genes and 1,024 alleles from *Homo sapiens*, and 702 IG and TR genes and 1,177 alleles from *Mus musculus*, *Mus cookii*, *Mus pahari*, *Mus spretus*, *Mus saxicola*, *Mus minutoides*).

IMGT/GENE-DB allows a search of IG and TR gene entries by locus, group, subgroup, based on the CLASSIFICATION concept of IMGT-ONTOLOGY [Giudicelli and Lefranc, 1999]. A short cut allows to search genes by a selection on gene name (according to the IMGT nomenclature) or on clone name(s) (data from the "Reference sequences" and "Sequences from the literature" columns in IMGT Repertoire> Gene tables). The selection is displayed at the top of the resulting genes page. The users can select the genes to view their detailed entries. Each IMGT/GENE-DB entry corresponds to one gene and provides, for each gene, the chromosomal localization, the gene name and definition, the number of alleles, links to IMGT Repertoire and to external sequence databases (EMBL, GenBank, DDBJ), genome databases (GDB, LocusLink, OMIM), and nomenclature database (HGNC Genew). Reciprocally, LocusLink, GDB, GeneCards and HGNC Genew have direct links to the IMGT/GENE-DB entries.

IMGT/GENE-DB provides for each allele, the functionality, the clone names, the IMGT/LIGM-DB reference sequence accession numbers (with link to the flat files), and the "IMGT/GENE-DB reference sequences in FASTA format" (nucleotide and amino acid sequences of the coding regions extracted from the IMGT/LIGM-DB reference sequence), with gaps according to the IMGT unique numbering [Lefranc *et al*., 2003] (based on the IMGT Scientific chart rules and on the NUMEROTATION concept of IMGT-ONTOLOGY).

IMGT 3D structure database: IMGT/3Dstructure-DB

IMGT/3Dstructure-DB is the IMGT 3D structure database for IG, TR, MHC and RPI of human and other vertebrate species, created by LIGM, on the Web since November 2001 [Ruiz and Lefranc, 2002; Kaas and Lefranc, 2002]. IMGT/3Dstructure-DB comprises IG, TR, MHC and RPI with known 3D structures. In March 2004, the IMGT/3Dstructure-DB database managed 725 coordinate files which correspond to 744 different proteins (540 IG, 33 TR and 171 MHC).

Coordinate files are extracted from the Protein Data Bank PDB (Berman *et al* 2000), and IMGT annotations are added according to the IMGT Scientific chart rules, based on the IMGT-ONTOLOGY concepts [Giudicelli and Lefranc, 1999]. An IMGT/3Dstructure-DB card provides IMGT gene and allele identification (based on the CLASSIFICATION concept), domain delimitations (based on the DESCRIPTION concept), amino acid positions according to the IMGT unique numbering [Lefranc *et al*., 2003] (based on the NUMEROTATION concept). Domains that are analysed in IMGT/3Dstructure-DB include V-DOMAIN (variable) and C-DOMAIN (constant) found in IG and TR, V-LIKE and C-LIKE-DOMAINs found in proteins other than IG and TR, G-DOMAIN (groove) found in MHC, and G-LIKE-DOMAINs found in proteins other than MHC [Duprat *et al*., 2003]. Moreover, IMGT/3Dstructure-DB provides renumbered coordinate flat files, Collier de Perles (standard and two-layer 2D graphical representations) [Ruiz and Lefranc, 2002; Lefranc *et al*., 2003], and results of contact analysis. The IMGT unique numbering and gene standardization will provide a great help in large scale sequence-structure studies and more generally in protein engineering.

IMGT INTERACTIVE TOOLS

IMGT interactive immunoinformatics tools rely on the DESCRIPTION and NUMEROTATION concepts of IMGT-ONTOLOGY, and include sequence, genome and 3D structure analysis tools.

IMGT tools for sequence analysis

IMGT/V-QUEST (V-QUEry and STandardization) is an integrated software for IG and TR [Lefranc, 2001a; 2003b]. This tool, easy to use, analyses an input IG or TR germline or rearranged variable nucleotide sequence. IMGT/V-QUEST results comprise the identification of the V, D and J genes and alleles and the nucleotide alignments by comparison with sequences from the IMGT reference directory, the delimitations of the FR-IMGT and CDR-IMGT based on the IMGT unique numbering, the protein translation of the input sequence, the identification of the JUNCTION, and the two-dimensional Collier de Perles representation of the V-REGION. IMGT/V-QUEST is particularly useful for the analysis of the rearranged variable genes: it allows to identify the V-GENE and J-GENE and alleles involved in the IG and TR rearrangements, and to delimit the JUNCTION. Searches can be done related to IG and TR of human and mouse, and of other species (non-human primates, sheep, teleostei and chondrichthyes). IMGT/V-QUEST can also be used for the analysis of functional or ORF germline variable genes from other species, to delimit the FR-IMGT and CDR-IMGT, provided that the similarity with sequences of the IMGT/V-QUEST reference directory sets is sufficiently high. The sets of sequences from the IMGT reference directory, used for IMGT/V-QUEST, can be downloaded in FASTA format from the IMGT site.

IMGT/JunctionAnalysis is a tool, complementary to IMGT/V-QUEST, which provides a thorough analysis of the V-J and V-D-J junction of IG and TR rearranged genes. IMGT/JunctionAnalysis identifies the D-GENEs and alleles involved in the IGH, TRB and TRD V-D-J rearrangements by comparison with the IMGT reference directory, and delimits precisely the P, N and D regions (IMGT/JunctionAnalysis output results). Results from IMGT/JunctionAnalysis are more accurate than those given by IMGT/V-QUEST regarding the D-GENE identification. Indeed, IMGT/JunctionAnalysis works on shorter sequences (JUNCTION), and with a higher constraint since the identification of the V-GENE and J-GENE and alleles is a prerequisite to perform the analysis. Several hundreds of junction sequences can be analysed simultaneously.

IMGT/Phylogene is an easy to use tool for phylogenetic analysis of variable region (V-REGION) and constant domain (C-DOMAIN) sequences. This tool is particularly useful in developmental and comparative immunology. The users can analyse their own sequences by comparison with the IMGT standardized reference sequences for human and mouse IG and TR [Elemento and Lefranc, 2003].

IMGT/Allele-Align allows the comparison of two alleles highlighting the nucleotide and amino acid differences.

IMGT tools for genome analysis

IMGT/GeneSearch, IMGT/GeneView and IMGT/LocusView are tools which provide the display of physical maps for the human IG, TR and MHC loci. The mouse TRA/TRD locus is also available.

IMGT tool for 3D structure analysis

IMGT/StructuralQuery is a tool which allows to retrieve the IMGT/3Dstructure-DB entries, based on specific structural characteristics: phi and psi angles, accessible surface area ASA, amino acid type, distance in angstrom between amino acids, CDR-IMGT lengths [Kaas and Lefranc, 2003]. IMGT/StructuralQuery is currently available for the V-DOMAINs.

IMGT WEB RESOURCES ("IMGT MARIE-PAULE PAGE")

IMGT Web resources ("IMGT Marie-Paule page") [Lefranc, 2003a] consist of 8000 HTML pages and comprise the following sections: "IMGT Scientific chart", "IMGT Repertoire", "IMGT Bloc-notes", "IMGT Education" and "IMGT Index".

IMGT Scientific chart

IMGT Scientific chart provides the controlled vocabulary and the annotation rules and concepts defined by IMGT-ONTOLOGY [Giudicelli and Lefranc, 1999] for the identification, the description, the classification and the numerotation of the IG, TR, MHC and RPI data of human and other vertebrates. The IMGT Scientific chart rules are described in the corresponding sections: Sequence and 3D structure identification and description, Nomenclature, and Numbering.

IMGT Repertoire

IMGT Repertoire is the global Web Resource in ImMunoGeneTics for the IG, TR, MHC and RPI of human and other vertebrates, based on the IMGT Scientific chart. IMGT Repertoire provides an easy-to-use interface to carefully and expertly annotated data on the genome, proteome, polymorphism and structural data, organized in three major sections: IMGT Repertoire (IG and TR), IMGT Repertoire (MHC) and IMGT Repertoire (RPI) [Lefranc, 2001a]. Only titles of this large resource are quoted here, with links as examples, to IMGT Repertoire (IG and TR). Genome data ("Locus and genes") include chromosomal localizations, locus representations, locus description, gene tables, potential germline repertoires, lists of IG and TR genes and links between IMGT, HUGO, GDB, LocusLink and OMIM, correspondence between nomenclatures [Lefranc and Lefranc, 2001a; 2001b], references sequences [Barbié and Lefranc, 1998; Pallarès *et al*., 1998; 1999; Ruiz *et al*., 1999; Folch and Lefranc, 2000a; 2000b; Scaviner and Lefranc, 2000a; 2000b; Martinez-Jean *et al*., 2001; Bosc and Lefranc, 2003]. Proteome and polymorphism data ("Proteins and alleles") are represented by protein displays which show translated sequences of the allele *01 of each functional or ORF gene [Lefranc and Lefranc, 2001a; 2001b; Scaviner *et al*., 1999; Folch et *al*., 2000], alignments of alleles, tables of alleles, allotypes, particularities in protein designations, IMGT reference directory in FASTA format, correspondence between IG and TR chain and receptor IMGT designations [Lefranc and Lefranc, 2001a; 2001b]. Structural data ("2D and 3D structures") comprise 2D graphical representations designated as Colliers de Perles [Lefranc *et al*., 1998; 1999], FR-IMGT and CDR-IMGT lengths, and 3D representations of IG and TR variable domains [Ruiz *et al*., 2002; Lefranc *et al*., 2003]. This visualization permits rapid correlation between protein sequences and 3D data retrieved from the Protein Data Bank (PDB). Other data comprise: Probes and RFLP with phages, probes used for the analysis of IG and TR gene rearrangements and expression, and Restriction Fragment Length Polymorphism (RFLP) studies, Taxonomy of vertebrate species present in IMGT/LIGM-DB, Gene regulation and expression with data on promoters, primers, cDNAs, reagent monoclonal antibodies, and Genes and clinical entities: translocations and inversions, humanized antibodies, monoclonal antibodies with clinical indications.

IMGT Bloc-notes

IMGT Bloc-notes is a selection of useful links for immunoinformatics, immunogenetics, immunology, genetics, molecular biology and bioinformatics. IMGT Bloc-notes is organized in several sections. The IMGT immunoinformatics page comprises links to databases, tools and resources on IG, TR, MHC and

RPI. Interesting links provide numerous hyperlinks towards the Web servers specializing in immunology, genetics, molecular biology and bioinformatics (associations, biopharmaceuticals, collections, companies, databases, immunology themes, journals, molecular biology servers, resources, societies, tools, etc.) [Lefranc, 2000e]. Other sections are meeting announcements, postdoctoral positions, etc.

IMGT Education

IMGT Education is a section which provides useful biological resources for students. It includes IMGT Aide-mémoire which provides an easy access to information such as genetic code, splicing sites, amino acid structures, restriction enzyme sites, etc., Questions and answers, Tutorials (in English and/or in French) on 3D structure, immunoglobulins and B cells, T cell receptors and T cells, NK receptors, pathologies of the immune system, cancer, AIDS, etc.

IMGT Index

IMGT Index is a referential index which provides a fast way to access data when information has to be retrieved from different parts of the IMGT site. For example, "allele" provides links to the IMGT Scientific chart rules for the allele description, and to the IMGT Repertoire Alignments of alleles and Tables of alleles.

CONCLUSION

Since July 1995, IMGT has been available on the Web at http://imgt.cines.fr. IMGT provides the biologists with an easy to use and friendly interface. Since January 2000, the IMGT www server at Montpellier has been accessed by more than 250,000 sites. IMGT has an exceptional response with more than 120,000 requests a month. Two thirds of the visitors are equally distributed between the European Union and the United States. To facilitate the integration of IMGT data into applications developed by other laboratories, we have built an Application Programming Interface (API) to access the database [Giudicelli *et al*., 1998a]. This API includes: a set of URL links to access biological knowledge data (keywords, labels, functionalities, list of gene names, etc.), a set of URL links to access all data related to one given sequence. To increase interoperability with other ontologies and information systems, IMGT-ONTOLOGY is currently being written using XML (Extensible Markup Language) approach, in IMGT-ML [Chaume *et al*., 2001; 2003]. By making data portable, XML is useful both internally for the integration of data and externally for sharing data with other information systems. Because of this data integration ability, XML has become the underpinning for Web-related computing. IMGT-ML defines XML schemas to encode data with XML tags respecting the IMGT-ONTOLOGY concepts. IMGT-ML schemas will be used for distributive data using the Web-services technology. IMGT distributes high quality data with an important incremental value added by the IMGT expert annotations, according to the rules described in the IMGT Scientific chart. Control of coherence in IMGT combines data integrity control and biological data evaluation [Giudicelli *et al*., 1998a; 1998b]. The information provided by IMGT is of much value to clinicians and biological scientists in general [Lefranc, 2002; 2003b; Chardes *et al*., 2002]. IMGT/PROTEIN-DB, a protein database for IG and TR, will contain translations of potentially functional and ORF sequences from IMGT/LIGM-DB, and protein data from Kabat [Kabat *et al*., 1991] and PDB. IMGT is designed to allow a common access to all immunogenetics data, and particular attention is given to the establishment of cross-referencing links to other databases pertinent to the users of IMGT.

CITING IMGT

If you use IMGT databases, tools and/or Web resources, please cite [Lefranc, 2003a], and this paper as references, and quote the IMGT Home page URL address, http://imgt.cines.fr.

ACKNOWLEDGEMENTS

We thank Marc Lemaitre, Marie-Claire Beckers, Géraldine Folch and Delphine Valette from EUROGENTEC S. A., Belgium, for their scientific contribution to IMGT/PRIMER-DB, Julien Bertrand, Christèle Jean-Martinez, Olivier Elemento and Manuel Ruiz for their previous work at LIGM, our "2003" students Naoufel Benabdelkrim el Filali, Elodie Boucomont, Patrick Chastellan, Laurent Douchy, Valérie Garelle, Valérie Torregrossa and Guillaume Tourneur for their motivation, and Ruth Henry for editorial work. We are deeply grateful to the IMGT team for its expertise and constant motivation and specially to our curators for their hard work and enthusiasm. IMGT is funded by the European Union's 5th PCRDT programme (QLG2-2000-01287), the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Education Nationale and the Ministère de la Recherche. Subventions have been received from Association pour la Recherche sur le Cancer (ARC) and the Région Languedoc-Roussillon.

REFERENCES

[1] Barbié, V. and Lefranc, M.-P. (1998). The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp. Clin. Immunogenet. **15**, 171-183.

[2] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2003). GenBank. Nucleic Acids Res. **31**, 23-27.

[3] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res. **28**, 235-242.

[4] Bosc, N. and Lefranc, M.-P. (2003). IMGT Locus in Focus: The mouse (*Mus musculus*) T cell receptor alpha (TRA) and delta (TRD) variable genes. Dev. Comp. Immunol. **27**, 465-497.

[5] Chardes, T., Chapal, N., Bresson, D., Bes, C., Giudicelli, V., Lefranc, M.-P. and Peraldi-Roux, S. (2002). The human anti-thyroid peroxidase autoantibody repertoire in Graves and Hashimoto's autoimmune thyroid diseases. Immunogenetics **54**, 141-157.

[6] Chaume, D., Giudicelli, V. and Lefranc, M.-P. (2001). IMGT-XML a language for IMGT-ONTOLOGY and IMGT/LIGM-DB data. *In:* CORBA and XML : Towards a bioinformatics integrated network environment. Proceedings of NETTAB 2001, Network tools and applications in biology, pp. 71-75.

[7] Chaume, D., Giudicelli, V., Combres, K. and Lefranc, M.-P. (2003). IMGT-ONTOLOGY and IMGT-ML for Immunogenetics and immunoinformatics. European Congress in Computational Biology ECCB'2003, Sequence databases and Ontologies satellite event, Paris.

[8] Chothia, C. and Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. **196**, 901-917.

[9] Duprat, E. and Lefranc, M.-P. (2003). IMGT standardization and analysis of V-LIKE, C-LIKE and G-LIKE-DOMAINS. European Congress in Computational Biology ECCB'2003, Paris.

[10] Elemento, 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.

[11] Folch, G. and Lefranc, M.-P. (2000a). The human T cell receptor beta variable (TRBV) genes. Exp. Clin. Immunogenet**. 17**, 42-54.

[12] Folch, G. and Lefranc M.-P. (2000b). The human T cell receptor beta diversity (TRBD) and beta joining (TRBJ) genes. Exp. Clin. Immunogenet. **17**, 107-114.

[13] Folch, G., Scaviner, D., Contet, V. and Lefranc, M.-P. (2000). Protein displays of the human T cell receptor

alpha, beta, gamma and delta variable and joining regions. Exp. Clin. Immunogenet. **17**, 205-215.
[14] Giudicelli, V. and Lefranc, M.-P. (1999). Ontology for Immunogenetics: IMGT-ONTOLOGY. Bioinformatics **12**, 1047-1054.
[15] Giudicelli, V., Chaume, D., Bodmer, J., Müller, W., Busin, C., Marsh, S., Bontrop, R., Lemaitre, M., Malik, A. and Lefranc, M.-P. (1997). IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. **25**, 206-211.
[16] Giudicelli, V., Chaume, D. and Lefranc, M.-P. (1998a). IMGT/LIGM-DB: A systematized approach for ImMunoGeneTics database coherence and data distribution improvement. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology. ISBM'98, pp. 59-68.
[17] Giudicelli, V., Chaume, D., Mennessier, G., Althaus, H. H., Müller, W., Bodmer, J., Malik, A. and Lefranc, M.-P. (1998b). IMGT, the international ImMunoGeneTics database: a new Design for Immunogenetics Data Access. *In:* B.Cesnik et al. (Eds.). Proceedings of the Ninth World Congress on Medical Informatics, MEDINFO' 98, IOS Press, Amsterdam. pp. 351-355.
[18] Giudicelli, V., Protat, C. and Lefranc, M.-P. (2003). The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. European Congress in Computational Biology ECCB'2003, Paris.
[19] Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. and Foeller, C. (1991). Sequences of proteins of immunological interest. National Institute of Health Publications, Washington D.C., USA. pp. 91-3242.
[20] Kaas, Q. and Lefranc, M.-P. (2002). IMGT/3DStructure-DB for immunoglobulin, T cell receptor and MHC structural data. European Congress in Computational Biology ECCB'2002, P72.
[21] Kaas, Q. and Lefranc, M.-P. (2003). IMGT/StructuralQuery: a tool for structural data analysis of immunoglobulin and T cell receptor variable domains (http://imgt.cines.fr). European Congress in Computational Biology ECCB'2003, Paris.
[22] Lefranc, M.-P. (1997). Unique database numbering system for immunogenetic analysis. Immunology Today **18**, 509.
[23] Lefranc, M.-P. (1998). IMGT (ImMunoGeneTics) Locus on Focus. A new section of Experimental and Clinical Immunogenetics. Exp. Clin. Immunogenet. **15**, 1-7.
[24] Lefranc, M.-P. (1999). The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. The Immunologist **7**, 132-136.
[25] Lefranc, M.-P. (2000a). Nomenclature of the human immunoglobulin genes. Current Protocols in Immunology. Wiley J. and Sons, New York, USA. Supplement **40**, A.1P.1-A.1P.37.
[26] Lefranc, M.-P. (2000b). Nomenclature of the human T cell Receptor genes. Current Protocols in Immunology. Wiley J. and Sons, New York, USA. Supplement **40**, A.1O.1-A.1O.23.
[27] Lefranc, M.-P. (2000c). Locus maps and genomic repertoire of the human Ig genes. The Immunologist **8**, 80-87.
[28] Lefranc, M.-P. (2000d). Locus maps and genomic repertoire of the human T-cell receptor genes. The Immunologist **8**, 72-79.
[29] Lefranc, M.-P. (2000e). Web sites of Interest to Immunologists. Current Protocols in Immunology. Wiley J. and Sons, New York, USA. A.1J.1-A.1J.33.
[30] Lefranc, M.-P. (2001a). IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. **29**, 207-209.
[31] Lefranc, M.-P. (2001b). Nomenclature of the human immunoglobulin heavy (IGH) genes. Exp. Clin. Immunogenet**. 18**, 100-116.
[32] Lefranc, M.-P. (2001c). Nomenclature of the human immunoglobulin kappa (IGK) genes. Exp. Clin. Immunogenet. **18**, 161-174.
[33] Lefranc, M.-P. (2001d). Nomenclature of the human immunoglobulin lambda (IGL) genes. Exp. Clin. Immunogenet. **18**, 242-254.
[34] Lefranc, M.-P. (2002). IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev. Comp. Immunol. **26**, 697-705.
[35] Lefranc, M.-P. (2003a). IMGT, the international ImMunoGeneTics database® (http://imgt.cines.fr). Nucleic Acids Res. **31**, 307-310.
[36] Lefranc, M.-P. (2003b). IMGT® databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis (http://imgt.cines.fr). Leukemia. **17**, 260-266.
[37] Lefranc, M.-P. and Lefranc G. (2001a). The Immunoglobulin FactsBook. Academic Press, London, UK, 458

pages, ISBN:012441351X.
[38] Lefranc, M.-P., Lefranc, G. (2001b). The T cell receptor FactsBook. Academic Press, London, UK, 398 pages, ISBN:0124413528.
[39] Lefranc, M.-P., Giudicelli, V., Busin, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A. and Chaume, D. (1998). IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. **26**, 297-303.
[40] Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A., Barbié, V. and Chaume, D. (1999). IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. **27**, 209-212.
[41] Lefranc, M.-P., Lefranc, G., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L. and Thouvenin-Contet, V. (2003). IMGT, unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-Like domains. Dev. Comp. Immunol. **27**, 55-77.
[42] Martinez-Jean, C., Folch, G. and Lefranc, M.-P. (2001). Nomenclature and Overview of the Mouse (*Mus musculus* and *Mus* sp.) Immunoglobulin Kappa (IGK) Genes. Exp. Clin. Immunogenet. **18**, 255-279.
[43] Miyazaki, S., Sugawara, H., Gojobori, T. and Tateno, Y. (2003). DNA Data Bank of Japan (DDBJ) in XML. Nucleic Acids Res. **31**, 13-16.
[44] Pallarès, N., Frippiat, J.P., Giudicelli, V. and Lefranc, M.-P. (1998). The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp. Clin. Immunogenet. **15**, 8-18.
[45] Pallarès, N., Lefebvre, S., Contet, V., Matsuda, F. and Lefranc, M.-P. (1999). The human immunoglobulin heavy variable (IGHV) genes. Exp. Clin. Immunogenet. **16**, 36-60.
[46] Pommié, C, Levadoux, S., Sabatier, R., Lefranc, G. and Lefranc, M.-P. (2003). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. European Congress in Computational Biology ECCB'2003, Paris.
[47] Robinson, J., Malik, A., Parham, P., Bodmer, J. G. and Marsh, S. G. (2000). IMGT/HLA Database - a sequence database for the human major histocompatibility complex. Tissue Antigens **55**, 280-287.
[48] Robinson, J., Waller, M. J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L. J., Stoehr, P. and Marsh, S. G. (2003). IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. **31**, 311-314.
[49] Ruiz, M., Pallarès, N., Contet, V., Barbié, V. and Lefranc, M.-P. (1999). The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. Exp. Clin. Immunogenet. **16**, 173-184.
[50] Ruiz, M., Giudicelli, V., Ginestoux, C., Stoehr, P., Robinson, J., Bodmer, J., Marsh, S. G., Bontrop, R., Lemaitre, M., Lefranc, G., Chaume, D. and Lefranc, M.-P. (2000). IMGT, the international ImmunoGeneTics database. Nucleic Acids Res. **28**, 219-221.
[51] Ruiz, M and Lefranc, M.-P. (2002). IMGT gene identification and Colliers de Perles of human immunoglobulin with known 3D structures. Immunogenet. **53**, 857-883.
[52] Satow, Y., Cohen, G.H., Padlan, E.A., Davies, D.R. (1986). Phosphocholine binding immunoglobulin Fab. McPC603. J. Mol. Biol. **190**, 593-604.
[53] Scaviner, D., Barbié, V., Ruiz, M. and Lefranc, M.-P. (1999). Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp. Clin. Immunogenet. **16**, 234-240.
[54] Scaviner, D. and Lefranc, M.-P. (2000a). The human T cell receptor alpha variable (TRAV) genes. Exp. Clin. Immunogenet. **17**, 83-96.
[55] Scaviner, D. and Lefranc, M.-P. (2000b). The human T cell receptor alpha joining (TRAJ) genes. Exp. Clin. Immunogenet. **17**, 97-106.
[56] Stoesser, G., Baker, W., Van den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., Nardone, F., Stoehr, P., Tuli, M. A., Tzouvara, K. and Vaughan, R. (2003). The EMBL nucleotide sequence database: major new developments. Nucleic Acids Res. **31**, 17-22.
[57] 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.
[58] Williams, A. F. and Barclay, A. N. (1988). The immunoglobulin superfamily-domains for cell surface recognition. Annu. Rev. Immunol. **6**, 381-405.