BIOINFORMATICS

Vol. 15 no. 12 1999
Pages 1047–1054

Ontology for immunogenetics: the IMGT-ONTOLOGY

Véronique Giudicelli and Marie-Paule Lefranc

Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Received on March 26, 1999; revised on June 4, 1999; accepted on June 11, 1999

Abstract

Motivation: *IMGT, the international ImMunoGeneTics database (http://imgt.cines.fr:8104), created by M.-P. Lefranc, is an integrated database specializing in antigen receptors (immunoglobulins and T-cell receptors) and major histocompatibility complex (MHC) of all vertebrate species. IMGT accurate immunogenetics data are based on the standardization of the biological knowledge provided by the 'ImMunoGeneTics' IMGT-ONTOLOGY.*

The IMGT-ONTOLOGY describes the classification and specification of terms needed for immunogenetics and bioinformatics. IMGT-ONTOLOGY covers four main concepts: 'IDENTIFICATION', 'DESCRIPTION', 'CLASSIFICATION' and 'OBTENTION'. These concepts allow an extensive and standardized description and characterization of immunoglobulin and T-cell receptor data. The controlled vocabulary and the annotation rules are indispensable to ensure accuracy, consistency and coherence in IMGT. IMGT-ONTOLOGY allows scientists and clinicians to use, for the first time, identical terms with the same meaning in immunogenetics. It provides a semantic repository that will improve interoperability between specialist and generalist databases.

Availability: *Controlled vocabulary and annotation rules are described in the IMGT Scientific chart from the IMGT Marie-Paule page at http://imgt.cines.fr:8104*

Contact: *lefranc@ligm.igh.cnrs.fr*

Introduction

The molecular synthesis of immunoglobulins (Ig) and T-cell receptors (TcR) (Lefranc, 1990; Honjo and Alt, 1995) is particularly complex as it includes biological mechanisms such as DNA molecular rearrangements, nucleotide deletions and insertions at rearrangement junctions and, for Ig, somatic hypermutations. This complex synthesis allows one individual to produce potentially $> 10^{12}$ different Ig or TcR (in humans). The rate of published data related to Ig and TcR sequences, structures, specificities and polymorphisms is currently growing exponentially. They are produced and used by various scientific fields, including fundamental research, clinical, veterinary and pharmaceutical studies which have different objectives and use of data. Thus, sequence data description and gene nomenclatures are very heterogeneous. Biological and immunogenetics terms are frequently used with different meanings. For example, there is no clear definition for a 'germline sequence' and no consensus agreement of what is a functional gene. As a consequence, there is poor interoperability between databases and inefficient links between data.

Moreover, the generalist databases (Benson *et al.*, 1999; Stoesser *et al.*, 1999; Sugawara *et al.*, 1999), which contain most of the published Ig and TcR sequences, propose a terminology of feature keys that is too limited (and then ambiguous) for immunogenetics sequence description. For example, there are no words to delimit specifically the coding variable region of Ig and TcR (without part of the leader peptide), although this is a very important component for the protein structure, functionality and specificity characterization.

IMGT, the international ImMunoGeneTics database (http://imgt.cines.fr:8104) (Lefranc *et al.*, 1998, 1999), is an integrated and specialized database for immunogenetics. IMGT comprises expertly annotated data and consists of three databases: LIGM-DB (for Ig and TcR), MHC/HLA-DB and PRIMER-DB [an Ig, TcR and major histocompatibility complex (MHC)-related primer database]. IMGT closely follows the content of Ig and TcR sequences from generalist databases and currently contains >35 000 entries. IMGT distributes high-quality data with an important increment value added by the LIGM expert annotations. Annotations are based on standardized and very detailed description of sequences in each step of Ig and TcR synthesis, whatever the species and whatever the chain type. Moreover, these annotations allow a description of two-dimensional (2D) and 3D structural data, and polymorphic data. This makes IMGT unique since it is the first, and so far the only, immunogenetics integrated database specializing in the genome, proteome, structure and polymorphism data of Ig and TcR of all vertebrate species, from fish to human. It was, therefore, crucial for the efficiency of IMGT to set up concise and standardized non-ambiguous classification of the immunogenetics knowledge, and to set up definitions of terms used in biology and immunogenetics. This standardization allows one to identify genes, describe sequences, structural and polymorphic data, perform a search, and improve data retrieval according to specific immunogenetics criteria. IMGT is a European project and its development policy has been agreed with the IMGT scientific committee (see IMGT information at http://imgt.cines.fr:8104).

IMGT accurate immunogenetics data are based on the semantic standardization of the biological knowledge provided by IMGT-ONTOLOGY. IMGT-ONTOLOGY is the first ontology to be developed for immunogenetics. ‘An ontology is a specification of conceptualization’ (Gruber, 1993). Ontologies have so far been developed to manage, to share and to represent knowledge in various scientific fields, such as the management of medical terminology (Rector *et al.*, 1998), the management of molecular biology knowledge (Schulze-Kremer, 1998), the management of the ribosome resources (Altman *et al.*, 1998), the management and the representation of the *Escherichia coli* genes and metabolism (Karp *et al.*, 1999), and the management of *Drosophila melanogaster* genomic and genetic resources (The FlyBase Consortium, 1999). The different interpretations of the word ontology have been discussed elsewhere (Guarino and Giaretta, 1995). IMGT-ONTOLOGY is considered as a vocabulary of well-defined terms commonly used in the field of biology and immunogenetics. IMGT-ONTOLOGY is composed by a hierarchy of concepts. A concept is a relevant and fundamental criterion which is needed to characterize Ig and TcR data. Instances of concepts represent the standardized terms that are associated with the concept. Concepts and instances are linked by relations (‘is a subset of’, ‘is a member of’, ‘is a part of’, etc.).

Material and software

Implementation

IMGT is a relational database managed by the Sybase RDBMS, like many other biological databases. It has been developed according to a systematized approach (Giudicelli *et al.*, 1998a). This approach allows an easy coordination of core data entries (sequence-related data coming from EMBL), knowledge management (IMGT-ONTOLOGY) and annotation integration (expertise realized according to the rules defined in IMGT-ONTOLOGY by the IMGT annotator team of 10 curators). Concepts, instances and relations are managed in the ‘knowledge subsystem’ which comprises mainly two sets of relational tables: (i) in the first set are recorded the concepts and the instances with their biological characteristics; (ii) in the second set are recorded the relationships between concepts and instances. The knowledge subsystem allows one to organize and control the knowledge, as illustrated by the management of the prototypes (described below). An object-oriented model of IMGT has been designed and implemented with Java and IMGT-ONTOLOGY related data are recorded in the knowledge package. This architecture, correlated with the use of Sybase procedures and Java programs, allows the control of the biological coherence of the entries before their distribution and an automatic update of old annotations with the evolution of IMGT-ONTOLOGY. The evolution of IMGT-ONTOLOGY results from new knowledge coming from experimental research, and from the IMGT expertise (Figure 1).

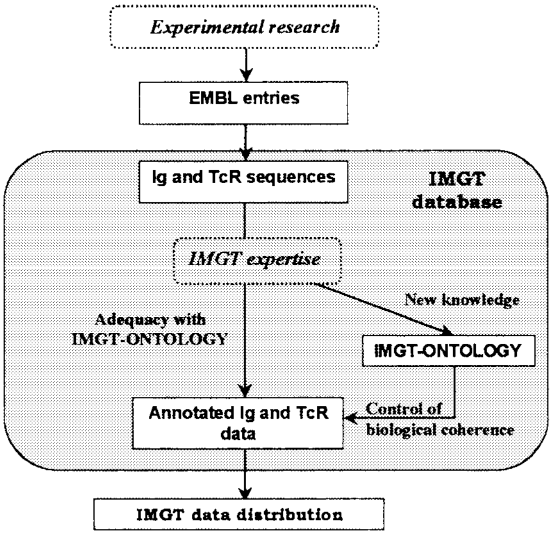

Fig. 1. Role and evolution of the IMGT-ONTOLOGY. IMGT-ONTOLOGY is used for the control of biological coherence. New concepts or instances are added to IMGT-ONTOLOGY following new knowledge which comes from experimental research and from the IMGT expertise.

Distribution

IMGT-ONTOLOGY main concepts with controlled vocabularies and rules are distributed in the IMGT Scientific chart from the IMGT Marie-Paule Page at http://imgt.cines.fr:8104. The content of knowledge tables is either as direct links from the IMGT Informatics page or as descriptive HTML pages. A convivial WWW interface (http://imgt.cines.fr:8104) has been designed, in Java language, on a client-server basis to query the IMGT database. Most of the instances of the IMGT-ONTOLOGY can be used in the IMGT Web interface to perform a search through the five search menus.

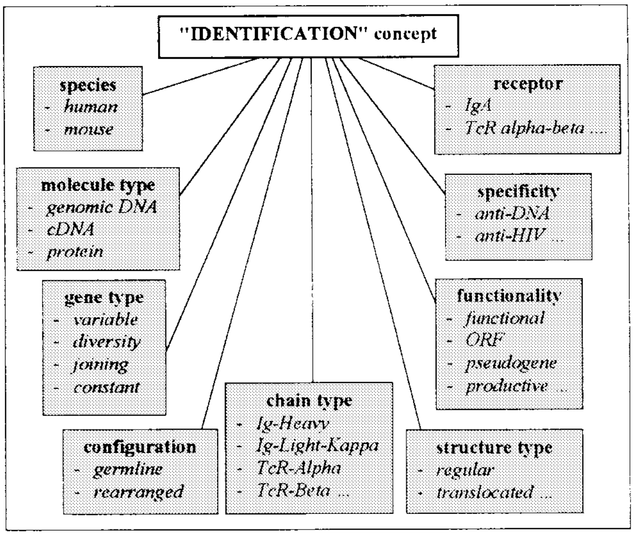

Fig. 2. The 'IDENTIFICATION' concept in the IMGT-ONTOLOGY. The names of the concepts of identification are shown in bold. Examples of instances for each concept are shown in italics.

IMGT-ONTOLOGY main concepts

IMGT-ONTOLOGY provides a semantic classification and standardization of the knowledge in the immunogenetics field used to identify the sequences, to describe their detailed composition, to classify Ig and TcR genes, and to define in which experimental, biological or medical context the sequences have been obtained. Four main concepts, 'IDENTIFICATION', 'DESCRIPTION', 'CLASSIFICATION' and 'OBTENTION', have been defined so far.

The 'IDENTIFICATION' concept

The 'IDENTIFICATION' concept (Figure 2) is a set of concepts that allow scientists to identify Ig or TcR sequences according to fundamental biological and immunogenetics characteristics. These are as follows.

The 'species' concept. Species names and the corresponding taxonomy are those provided by NCBI to keep complete interoperability with generalist databases. Since Ig and TcR proteins are synthesized by vertebrates, only vertebrate species are represented in IMGT-ONTOLOGY. However, invertebrate species can easily be added if the field of investigation is later extended (for instance, members of the immunoglobulin superfamily).

The 'molecule type' concept. Three instances are considered: genomic DNA, cDNA and protein.

The 'gene type' concept. Four types of genes are well known to be involved in Ig and TcR synthesis: the variable (V), diversity (D) and joining (J) genes which encode the antigen binding sites, and the constant (C) genes which encode the part of the protein which has effector properties.

The 'configuration' concept. The two instances of this concept are germline and rearranged. They define the status of the V, D and J genes before or after DNA rearrangements, respectively. This concept is particularly important because it is unique in the animal and plant genomes to the Ig and TcR V, D and J genes. Note that the C genes do not rearrange directly and therefore their configuration is not defined.

The 'chain type' concept. The chain type identifies the nature of the peptidic chain potentially encoded by Ig or TcR genes. There are seven main instances for the 'chain type' concept, which are defined by the C gene sequence characteristics: Ig-Heavy, Ig-Light-Kappa, Ig-Light-Lambda, TcR-Alpha, TcR-Beta, TcR-Gamma and TcR-Delta.

The 'structure type' concept. The structure type distinguishes sequences that show a classical organization (regular) from those which have been modified either naturally (processed, translocated, transposed, etc.) or artificially (transgene, humanized, engineered, etc.). Definitions of these terms are available from http://imgt.cines.fr:8104/textes/IMGTScientificChart.html.

The 'functionality' concept. The definition of functionality is based on the sequence analysis. As examples, the instances functional (for germline V, D, J, and for C sequences) and productive (for rearranged V-J and V-D-J sequences) mean that the coding regions have an open reading frame without a stop codon, and that there is no described defect in the splicing sites, and/or recombination signals, and/or regulatory elements. According to the gravity of the identified defects, the functionality can be defined as ORF, pseudogene or vestigial (for germline V, D, J, and for C sequences), or unproductive (for rearranged V-J and V-D-J sequences). Complete definitions are available at http://imgt.cines.fr:8104.

The 'specificity' concept. The specificity identifies the nature of the antigen recognized by the Ig or the TcR. The specificity is defined for rearranged Ig and TcR sequences, and standardization of instances is still in development.

The 'receptor' concept. A sequence is identified by the name of a receptor (e.g. IgA, TcR alpha-beta) when the 'chain type' of both polypeptides of the receptor has been unambiguously identified.

The 'DESCRIPTION' concept

The 'DESCRIPTION' concept (Figure 3) corresponds to the classification of terms and rules which are necessary to describe the organization and the components of the Ig and TcR sequences, and to characterize their specific and conserved motifs (Giudicelli, 1998). Three concepts of description are fundamental.

The 'entity' concept. Instances of that concept describe the conformation of an Ig or a TcR sequence. For example, the instance V-GENE represents a genomic V gene in germline configuration, whereas the instance V-J-GENE represents genomic V and J genes in rearranged configuration. Fourteen instances of the 'entity' concept have been defined so far to cover all cases of Ig and TcR sequence conformation. The different instances are linked to each other by relations that reflect the steps of the Ig and TcR chain synthesis: 'is rearranged into' and 'is transcribed into' (Giudicelli, 1998). Each instance is described by constitutive motifs which belong to the 'core' concept (described below), and to the 'other coding' and 'noncoding' subsets (Figure 2). Most of the concepts of description and their instances are features of Ig and TcR sequences (for a complete list, see Giudicelli, 1998). Others are common to biological sequences (splicing, regulation, etc.). The main relation that links an entity and its constitutive motifs is the relation 'is part of', related to the Component/Integral-Object part-whole relationship (Artale *et al.*, 1996a,b). Constitutive motifs are functional and separable parts of the entity. Relationships between motifs (dependency among parts) are detailed in prototypes (see below).

The 'core' concept. This concept contains four instances, V-REGION, D-REGION, J-REGION and C-REGION, which describe the coding sequences of the V, D, J and C genes, respectively. These instances are particularly important since they can be identified in all related germline or rearranged entities as well as in protein sequences and in effective antibodies or TcR.

The 'cluster' concept. This allows the description of sequences containing several entity instances. The instances can be of the same type as in a V-CLUSTER, which contains several V-GENEs, or of different types as in a V-(VJ)-CLUSTER, which contains at least one germline V-GENE and one rearranged V-J-GENE.

Prototypes of immunogenetics sequences

Prototypes represent the organizational relationships between the constitutive motifs of an entity instance, i.e. the 'horizontal' relationships between parts (Artale *et al.*, 1996b), and give information on the order and expected length (in number of nucleotides) of these motifs. Very precise prototypes can be established according to other concepts such as the 'species', the 'functionality', the 'chain type', etc. The relative organization of the constitutive motifs is based on the description of the relations that order two motifs. Six relations are necessary: 'is separated from', 'is adjacent downstream of', 'is included in', 'is included in and shares the same 5′ end', 'is included in and shares the same 3′ end', 'is overlapping' (Giudicelli, 1998).

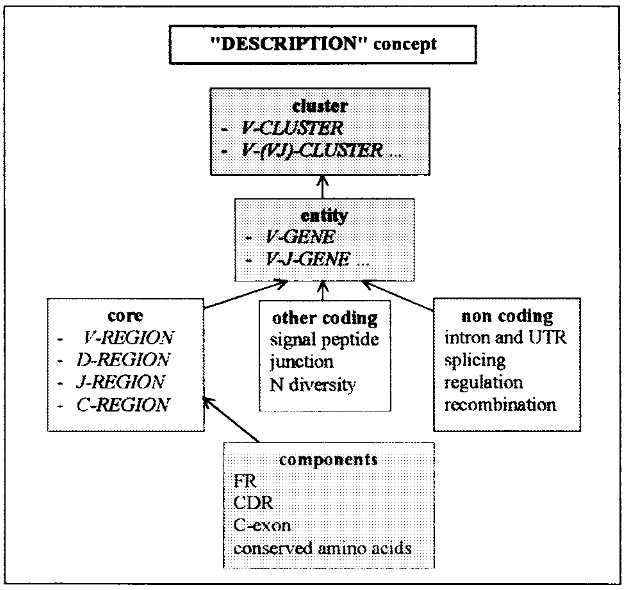

Fig. 3. The 'DESCRIPTION' concept in the IMGT-ONTOLOGY. The three fundamental concepts of description, 'entity', 'core', and 'cluster', are shown with examples of instances in italics. A number of 'other coding', 'non-coding' and 'component' concepts are necessary for a complete description of the sequences. Examples of these subsets are indicated in the figure. 'Coding' and 'non-coding' concepts are shown in yellow and green, respectively. Complete lists of concepts, instances and definitions are reported elsewhere (Giudicelli, 1998). Arrows indicate the relation 'is part of'.

The 'CLASSIFICATION' concept

The 'CLASSIFICATION' concept (Figure 4) organizes the immunogenetics knowledge useful to name and classify Ig and TcR genes in IMGT. Five distinct concepts of classification have been set up.

The 'locus' concept. A locus is a group of Ig or TcR genes that are ordered and are localized in the same chromosomal location, in a given species. The human genome includes seven main loci for Ig and TcR (three for Ig and four for TcR). Ig and TcR genes have also been identified in other chromosomal locations outside the main loci which represent new instances of the concept locus. However, the genes they contain, designated as orphons, are not functional.

The 'group' concept. A group is a set of genes which

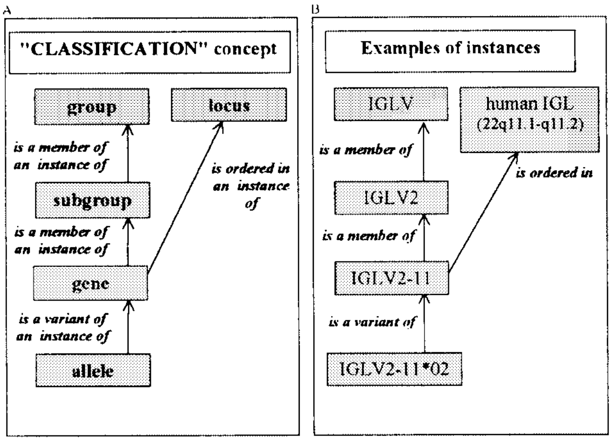

Fig. 4. The 'CLASSIFICATION' concept in the IMGT-ONTOLOGY. (A) Concepts of classification and their relations. (B) Examples of instances. For each concept of classification is shown the instance which allows the classification of the human IGLV2-11*02 allele.

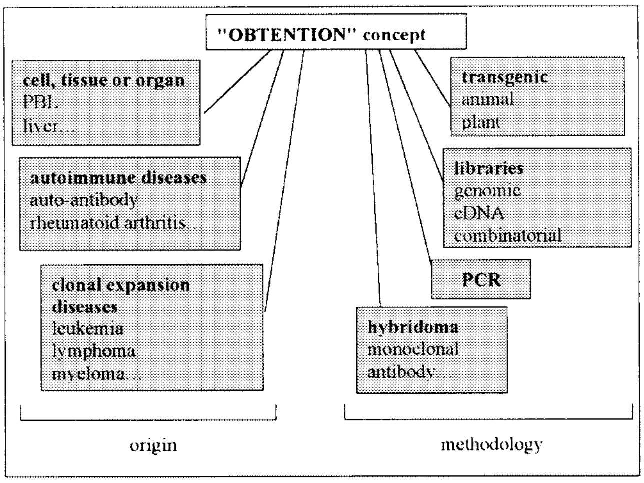

Fig. 5. The 'OBTENTION' concept in the IMGT-ONTOLOGY. Examples of subsets for the 'origin' and 'methodology' concepts are indicated in the figure. PBL, peripheral blood lymphocytes; PCR, polymerase chain reaction.

share the same 'gene type' (V, D, J or C) and participate potentially in the synthesis of a polypeptide of the same 'chain type'. By extension, a group includes the related pseudogenes and orphons.

The 'subgroup' concept. A subgroup is a set of genes which belong to the same group, in a given species, and which share at least 75% identity at the nucleotide level (in the germline configuration for V, D and J).

The 'gene' concept. A gene is defined as a DNA sequence that can be potentially transcribed and/or translated (this definition includes the regulatory elements in 5′ and 3′, and the introns, if present). Instances of the 'gene' concept are gene names. By extension, orphons and pseudogenes are also instances of the 'gene' concept.

The 'allele' concept. An allele is a polymorphic variant of a gene. Currently, in IMGT, alleles are described at the sequence level. An allele is identified by the mutations of its sequence compared to the reference sequence designated as *01 (see IMGT Scientific chart at http://imgt.cines.fr:8104 for IMGT description of mutations and IMGT allele polymorphism description). Full descriptions of mutations and allele name designations are only recorded for the core sequences (V-REGION, D-REGION, J-REGION, C-REGION). They are reported in alignment tables.

IMGT unique nomenclature and reference sequences

The concepts of classification have been used to set up a unique nomenclature of Ig and TcR genes (Lefranc, 1998). A gene name comprise four letters for the group (IGHV, IGHD, TRAV, etc.). Queries on these four letters in IMGT allow the retrieval of all genes belonging to a given group, whatever the species. For each gene, IMGT has defined a reference sequence. For the V, D and J genes, the reference sequence corresponds to a germline entity (or rearranged, if the germline gene has not yet been isolated). The rules for the choice of the reference sequences are described at http://imgt.cines.fr:8104 in the IMGT Scientific Chart. The reference sequence of a gene is always the first allele (*01). The alleles of a gene are numbered in chronological order of their identification. Allele reference sequences constitute the IMGT reference directory which can be downloaded from the IMGT repertoire at http://imgt.cines.fr:8104. IMGT nomenclature for Ig and TcR of all species follows the Human Gene Mapping Nomenclature rules. This has been applied as early as 1988 for the human IGL and IGH loci (Bensmana *et al.*, 1988; Ghanem *et al.*, 1988), and 1989 for all the genes of the human TRG locus (Lefranc and Rabbitts, 1989, 1990). Using the IMGT nomenclature, the complete list of the Ig and TcR human gene has been established recently for the first time (Barbié and Lefranc, 1998; Lefranc, 1998; Pallarès *et al.*, 1998, 1999; Ruiz *et al.*, 1999).

The 'OBTENTION' concept

The 'OBTENTION' concept (Figure 5) 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.

Discussion and conclusion

IMGT-ONTOLOGY is the first ontology to be developed for immunogenetics knowledge management and distribution. The advantages provided by the use of an ontology in the development of a biological database are listed below.

Quality control

This knowledge organization (definition and classification of concepts) is indispensable to ensure the quality of the IMGT integrated database, whose goals are to provide a common access to immunogenetics data from all vertebrate species.

Quality criteria resulting from IMGT-ONTOLOGY are as follows.

- Consistency: Ig and TcR data are described with a standardized vocabulary and using the same rules whatever the species.
- Accuracy: sequences are expertly annotated according to the biological knowledge.
- Reliability: all entries (∼500 new or updated entries per week) are checked by the IMGT annotators according to the IMGT rules.
- Timeliness and coherence: with its systematized organization, IMGT-ONTOLOGY is easily updated with the appearance of new knowledge related to immunogenetics. Java programs developed for biological coherence control are used to check and update the annotations according to the new rules (Giudicelli *et al.*, 1998a,b).

Communication and interoperability

By dealing with the immunogenetics knowledge for all vertebrate species, IMGT-ONTOLOGY will improve communication and discussion between scientists from different fields in fundamental research, clinical, veterinary and pharmaceutical studies, in terms of immunogenetics knowledge and reuse of data. In particular, IMGT has important implications in medical research (repertoire in autoimmune diseases, AIDS, leukemia, lymphoma), therapeutic approaches (antibody engineering), genome diversity and genome evolution studies. It will be used to improve semantic signification of links and develop interoperability with the generalist EMBL and SWISS-PROT (Bairoch and Apweiler, 1999) databases, with the genome databases GDB (Fasman *et al.*, 1997), GENATLAS (Frezal, 1998) and OMIM (http://www.ncbi.nlm.nih.gov/omim/) for human, ArkDB (http://www.ri.bbsrc.ac.uk/arkdb/sites.html) for various species, and MGD (Blake *et al.*, 1999) for mouse, and with the structure PDB database (Sussman *et al.*, 1998).

Data submission

IMGT-ONTOLOGY is indispensable to design and develop interfaces that will allow the authors to submit and annotate sequences directly in IMGT. To be efficient and useful, such an interface will need a powerful and user-friendly user guide, that will present the controlled vocabulary with semantics.

Data distribution and reuse

IMGT-ONTOLOGY enhances the reliability of data exchange and distribution. Since most of the terms of the controlled vocabulary are used as search criteria in the Web interface, this avoids multiple searches using synonyms in order to obtain complete and correct answers. This fits with user needs since IMGT records >14 000 queries a week. IMGT-ONTOLOGY will also facilitate the reuse of IMGT data for immunogenetics data mining studies, with a significant reduction of treatments commonly used to clean data.

IMGT concepts and rules have been used so far to develop an API which can be downloaded from http://imgt.cines.fr:8104/informatics/index.html. A new approach of data exchange using CORBA standard is currently in development with the collaboration of EMBL.

Although developed for immunogenetics purposes, the organization of the IMGT-ONTOLOGY main concepts can be reused by other biological databases. The 'IDENTIFICATION' concept ('species', 'molecule type', 'structure type') applies to all biological domains. The 'DESCRIPTION' concept shows a way semantically to set up and control sequence annotation in specialized databases. The 'CLASSIFICATION' concept allows the standardization of gene nomenclatures according to the HUGO nomenclature committee recommendations (http://www.gene.ucl.ac.uk/hugo). The 'OBTENTION' concept can be easily be applied by databases which use data coming from experimental research.

The elaboration of IMGT-ONTOLOGY was an imperative requirement for the management of IMGT, which is growing rapidly (3000 sequences in 1995, >35 000 in 1999), and whose aims are to distribute controlled and up-to-date genome, proteome, structure and polymorphism data according to the knowledge in immunogenetics of all vertebrate species, from fish to human.

Acknowledgements

We are deeply grateful to Johanne Abad, Sylvaine Artero, Valérie Barbié, Valérie Contet, Géraldine Folch, Chantal Ginestoux, Christèle Jean, Manuel Ruiz and Dominique Scaviner for the IMGT annotations and database development, Denys Chaume for expert computing, and Gérard Mennessier and Gérard Lefranc for helpful discussion. IMGT is funded by the European Union's BIOTECH programme (BIO4CT960037), the CNRS (Centre National de la Recherche Scientifique) and MENRT (Ministère de l'Education Nationale, de la Recherche et de la Technologie). Subventions have been received from ARC (Association de Recherche sur le Cancer), FRM (Fondation pour la Recherche Médicale) and the Région Languedoc-Roussillon.

References

Altman,R.B., Chen,R.O., Abernethy,N.F. and Bada,B. (1998) Using ontologies for a collaborative data resource in molecular biology: The RIBOWEB System. *ISMB-98 Tutorials—Ontologies for Molecular Biology* URL http://smi-web.stanford.edu/pubs/SMI_Abstracts/SMI-98-0729.html

Artale,A., Franconi,E. and Guarino,N. (1996a) Open problems for part-whole relations. *Proceedings of the 1996 International Workshop on Description Logics (DL'96)* AAAI Press, Boston, MA, pp. 70–73.

Artale,A., Franconi,E., Guarino,N. and Pazzi,L. (1996b) Part-whole relations in object-centered formalisms: an overview. *Data Knowl. Eng.*, **20**, 347–383.

Bairoch,A. and Apweiler,R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.

Barbié,V. and Lefranc,M.-P. (1998) IMGT locus on focus: The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Exp. Clin. Immunogenet.*, **15**, 171–183.

Bensmana,M., Huck,S., Lefranc,G. and Lefranc,M.-P. (1988) The human immunoglobulin pseudo-gamma IGHGP gene shows no major structural defect. *Nucleic Acids Res.*, **16**, 3108.

Benson,A.D., Bogusky,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F. F., Rapp,B.A. and Wheeler,D.L. (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.

Blake,J.A., Richardson,J.E., Davisson,M.T., Eppig,J.T. and the Mouse Genome Database Group (1999) The Mouse Genome Database (MGD): Genetic and genomic information about the laboratory mouse. *Nucleic Acids Res.*, **27**, 95–98.

Fasman,K.H., Letovsky,S.I., Cottingham,R.W. and Kingsbury,D.T. (1997) The GDB human genome database anno 1997. *Nucleic Acids Res.*, **25**, 72–80.

Frezal,J. (1998) Genatlas database, genes and development defects. *CR Acad. Sci.*, **321**, 805–817.

Ghanem,N., Dariavach,P., Bensmana,M., Chibani,J., Lefranc,G. and Lefranc,M.-P. (1988) Polymorphism of immunoglobulin lambda constant region genes in populations from France Lebanon and Tunisia. *Exp. Clin. Immunogenet.*, **5**, 186–195.

Giudicelli,V. (1998) Conception d'une ontologie en immunogénétique et développement d'un module de cohérence pour le contrôle de qualité de IMGT/LIGM-DB, PhD Thesis, Université Montpellier II, France.

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* AAAI Press, Menlo Park, CA, pp. 59–68.

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. *Proceedings of the Ninth World Congress on Medical Informatics* IOS Press, Amsterdam, pp. 351–355.

Gruber,T.R. (1993) A translation approach to portable ontology specifications. *Knowl. Acquis.*, **5**, 199–220. URL http://www.ksl-svc.stanford.edu:5915.

Guarino,N. and Giaretta,P. (1995) Ontologies and knowledge bases: towards a terminological clarification. In Mars,N.J. I. (ed.), *Towards Very Large Knowledge Bases* IOS Press, Amsterdam.

Honjo,T. and Alt,F.W. (1995) *Immunoglobulin Genes*. Academic Press, London, pp. 3–443.

Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. (1999) EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **26**, 50–53.

Lefranc,M.-P. (1990) Organization of the human T-cell receptor genes. *Eur. Cytokine Network*, **1**, 121–130.

Lefranc,M.-P. (1998) IMGT (ImMunoGeneTics) locus on focus. A new section of experimental and clinical immunogenetics. *Exp. Clin. Immunogenet.*, **15**, 1–7.

Lefranc,M.-P. and Rabbitts,T.H. (1989) The human T-cell receptor gamma (TRG) genes. *Trends Biochem. Sci.*, **14**, 214–218.

Lefranc,M.-P. and Rabbitts,T.H. (1990) A nomenclature to fit the organization of the human T cell receptor gamma and delta genes. *Res. Immunol.*, **141**, 615–618.

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.

Lefranc,M.-P.*et al.* (1999) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **27**, 209–212.

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.

Pallarès,N., Lefebvre,S., Contet,V., Matsuda,F. and Lefranc,M.-P. (1999) IMGT locus on focus: The human immunoglobulin heavy variable genes. *Exp. Clin. Immunogenet.*, **16**, 36–60.

Rector,A., Rossi,A., Consorti,M.F. and Zanstra,P. (1998) Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN Organisation. *Int. J. Med. Inf.*, **48**, 71–84.

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.

Schulze-Kremer,S. (1998) Ontologies for molecular biology. *Pacific Symposium on Biocomputing '98*, pp. 693–707. PSB98 On-line proceedings: http://smi-web.stanstead.edu/projects/helix/psb98.

Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **27**, 18–24.

Sugawara,H., Miyazaki,S., Gojobori,T. and Tateno,Y. (1999) DNA

Data Bank of Japan dealing with large-scale data submission. *Nucleic Acids Res.*, **27**, 25–28.

Sussman,J.L., Lin,D., Jiang,J., Manning,N.O., Prilusky,J., Ritter,O. and Abola,E.E. (1998) Protein Data Bank (PDB): database of three-dimensional structuralinformation of biological macromolecules. *Acta Crystallogr.*, **54**, 1078–1084.

The FlyBase Consortium (1999) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, **27**, 85–88.