UNIVERSITE MONTPELLIER II

SCIENCES ET TECHNIQUES DU LANGUEDOC

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline : *Bioinformatique*
Formation Doctorale : *Interface Chimie-Biologie*
Ecole Doctorale : *Sciences Chimiques et Biologiques pour la Santé*

Présentée et soutenue publiquement par

Jérôme LANE

Le 14 décembre 2009

Titre :

Analyse et conception au sein d'IMGT® d'une approche bioinformatique intégrée pour l'identification et la description des gènes d'immunoglobulines et de récepteurs T de locus génomiques de grandes tailles

JURY

M. Gérard Lefranc, Professeur, Université Montpellier II, Président du jury

Mme Marie-Paule Lefranc, Professeur, Université Montpellier II, Directeur de thèse

M. Pierre Pontarotti, Directeur de recherche au CNRS, Université de Provence, Rapporteur

M. Jean-Pol Frippiat, Maître de conférences, Université Henri Poincaré, Rapporteur

M. Antoine Blancher, Professeur, Université Paul-Sabatier, Examinateur

M. Patrice Duroux, Ingénieur de Recherche, UPR CNRS 1142, Examinateur

REMERCIEMENTS

Je tiens à remercier tout d'abord les rapporteurs de cette thèse, Pierre Pontarotti et Jean-Pol Frippiat. Je remercie également les autres membres du jury pour leur examen rigoureux de mon travail et le Professeur Gérard Lefranc pour la présidence de ce jury et ses conseils avisés.

Je remercie mon directeur de thèse, le Professeur Marie-Paule Lefranc, pour le temps qu'elle m'a accordé, son aide et son soutien tout au long de cette thèse.

Mille mercis à l'équipe IMGT pour m'avoir accueilli avec beaucoup de gentillesse: Eltaf Alamyar, Fatena Bellahcene, Xavier Brochet, Patrice Duroux, François Ehrenmann, Géraldine Folch, Elodie Gemrot, Chantal Ginestoux, Véronique Giudicelli, Joumana Jabado-Michaloud, Amandine Lacan, Christophe Le Roy, Odile Lucas, Zohra Ouaray, Vijay Phani Garapati, Laëtitia Regnier, Marie Schumeng, Emmanuel Servier et Yan Wu. Merci pour leur soutien, leur amitié et pour leur bonne humeur.

Je voudrais particulièrement remercier Patrice Duroux, qui m'a aidé et conseillé tout au long de cette thèse, Fatena Bellahcene, Géraldine Folch et Joumana Jabado-Michaloud pour leur évaluation d'IMGT/LIGMotif.

Je remercie également les institutions qui m'ont apporté leur soutien financier durant ces trois années: le Centre National de la Recherche Scientifique (CNRS), le Ministère de l'Enseignement Supérieur et de la Recherche MSER (Université Montpellier 2), l'Agence Nationale de la Recherche (ANR-06-BYOS-0005-01) et la Communauté Européenne (ImmunoGrid, FP6-2004-IST-4).

Mes derniers remerciements vont à ma famille. Merci à ma p'tit sœurette pour ses conseils sur les relations humaines et sa franchise. Merci à mes parents pour leur soutien et générosité. Un merci spécial à ma mère pour ses conseils d'ordre ménager; chaque visite à mon appartement a été l'occasion de le rendre plus propre et meublé. Un autre merci spécial à mon père pour ses coups de ciseau experts qui me rendent plus beau depuis que j'ai des cheveux et de ses blagues à la « papa » qui me font toujours autant rigoler.

PUBLICATIONS

Lane, J., Duroux, P. and Lefranc, M.-P. IMGT/LIGMotif: a tool for immunoglobulin and T cell receptor gene identification and description in large genomic sequences, (2009; article en révision à BMC Bioinformatics).

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., **Lane, J.,** Regnier, L., Ehrenmann, F., Lefranc, G. and Duroux, P. IMGT®, the international ImMunoGeneTics information system®, *Nucl. Acids Res*, 37, D1006-D1012 (2009); doi:10.1093/nar/gkn838. PMID: 18978023

Duroux, P., Kaas, Q., Brochet, X., **Lane, J.**, Ginestoux, C., Lefranc, M.-P. and Giudicelli, V IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm, *Biochimie*, 90, 570-583 (2008). Epub 2007 Sep11. PMID: 17949886

Duroux, P., Giudicelli, V., Kaas, Q., Jabado-Michaloud, J., Folch, G., Ginestoux, C., Brochet, X., **Lane, J.**, Regnier, L., Wu, Y., Garapati, V.P., Bellahcene, F., Servier, E.-J., Ehrenmann, F., Lefranc, G. and Lefranc, M.-P. IMGT®, the international ImMunoGeneTics information system®, the reference in immunogenetics and immunoinformatics, In: Proceedings First International Congress on Macromolecular Biochemistry and Genetics Gafsa, Tunisia, 12-16 April 2007, A Focus on Biochemistry and Genetics: from concepts to therapeutic advances (Ed. Fattoum A.), pp. 185-198 (2007).

CONFERENCE

IMGT/LIGMotif: a tool for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. EMBnet conference 2008. 20th Anniversary Celebration "Leading applications and technologies in Bioinformatics". Martina Franca, TA, Italy (17-20 September 2008).

COMMUNICATIONS ORALES ET POSTERS

Lane, J., Lefranc, M.-P. and Duroux, P. "IMGT/LIGMotif: a tool to annotate immunoglobulin and T cell receptor genes of vertebrates in genomic DNA" Communication orale et poster, Journées de l'Ecole Doctorale des Sciences Chimiques et Biologiques pour la Santé, Journées CBS2 2008 Montpellier, France (5-7 mai 2008).

Lane, J., Lefranc, M.-P. and Duroux, P. "IMGT/LIGMotif: a tool to identify and describe germline vertebrate genes of immunoglobulins and T cell receptors". Poster, 8èmes Journées Ouvertes Biologie, Informatique et Mathématiques JOBIM 2007, Marseille, France (10-12 juillet 2007) Publié dans les Actes des Journées Ouvertes Biologie, Informatique et Mathématiques JOBIM 2007, Brun C. and Didier G. (eds.), pp. 339-340.

Duroux, P., Ehrenmann, F., Régnier, L., Brochet, X., **Lane, J.**, Ginestoux, C., Lefranc, M.-P. and Giudicelli, V. "IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm" Communication, Workshop Towards Systems Biology Grenoble, France (8-10 October 2007).

Giudicelli, V., Wu, Y., Kaas, Q., Brochet, X., **Lane, J.**, Folch, G., Jabado-Michaloud, J., Régnier, L., Ehrenmann, F., Bellahcene, F., Lucas, O., Gemrot, E., Ginestoux, C., Lefranc, G., Duroux, P. and Lefranc, M.-P. "IMGT® resources for cancer research" Colloque Cancer Genome and epigenome: new technologies for new challenges Paris, France (13-14 December 2007).

Giudicelli, V., Regnier, L., Folch, G., Jabado-Michaloud, J., Bellahcene, F., Ginestoux, C., Gemrot, E., Wu, Y., Brochet, X., **Lane, J.**, Lefranc, G., Ehrenmann, F., Duroux, P. and Lefranc, M.-P. "IMGT-ONTOLOGY" Data integration in the Life Sciences 2008, DILS'08 Evry (near Paris), France (25-27 June 2008) Publié dans 'Poster and Poster/Demo. Abstracts Proceedings', pp. 37 (2008).

Regnier, L, Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., **Lane, J.**, Brochet, X., Ehrenmann, F., Duroux, P., Lefranc, G. and Lefranc, M.-P. "IMGT®, the international ImMunoGeneTics information system®: démarche

qualité au sein d’un système d’information" Sixième École Inter-Organismes "Qualité en Recherche et en Enseignement Supérieur" La Grande-Motte, France (15-17 September 2008).

Lane, J., Lefranc, M.-P. and Duroux, P. "IMGT/LIGMotif: a tool for immunoglobulin and T cell receptor gene identification and description in large genomic sequences" EMBnet conference 2008 – 20th Anniverary Celebration "Leading applications and technologies in Bioinformatics" Martina Franca, Italy (September 17-20, 2008) Publié dans EMBnet.news, 14 Nr.3, 40, Abstract 36, September 2008.

TABLE DES MATIERES

INTRODUCTION

La méthode de séquençage de l'ADN simple brin développé par Sanger a représenté une avancée technologique majeure en 1977 [1], reconnue par le Prix Nobel attribué en 1980. Lee Hood a mis au point les premiers séquenceurs automatiques en utilisant des systèmes capillaires et des fluorochromes. L'utilisation intensive des séquenceurs automatiques ont permis de réaliser le séquençage du génome humain en 2001 (Human Genome Project lancé en 1990) [2-3]. Récemment une nouvelle génération de séquençeurs permettant un séquençage à très haut débit (« Next Generation Sequencing ») [4] est en train de révolutionner le domaine. Ceci se traduit par la production accrue de nombreuses séquences de génomes. En juillet 2009, « Entrez Genome Project » du NCBI référençait 966 espèces différentes dont le génome est complètement séquencé (notamment 944 de Procaryotes, 22 d'Eucaryotes) et 2292 autres espèces dont le séquençage complet était en cours de réalisation. Cependant, tous ces génomes n'ont d'intérêt pour les biologistes que s'ils sont annotés. L'annotation des séquences d'ADN (codantes et non-codantes) est à la base de la génomique et consiste, entre autres finalités, à localiser au nucléotide près la position des motifs, à les labelliser et à leur attribuer une fonctionnalité. Les experts en charge de ce travail essentiel à la valorisation des génomes réalisent un véritable travail de fourmis sachant que la taille du génome humain atteint 3 gigabases, et que celle d'un exon peut ne faire que quelques dizaines de paires de bases et donc ne représente qu'une très faible portion du génome: l'annotation des séquences codantes revient à rechercher une aiguille dans une botte de foin.

La bioinformatique, dont le but est l'analyse des données de la biologie (génomique, transcriptomique, protéomique, métabolomique,…), bénéficie de l'évolution du matériel informatique (capacité de stockage, vitesse de traitement des données) et des réseaux informatiques (calcul parallèle et distribué et grilles de calculs). Les données biologiques sont stockées, organisées et triées dans des bases de données généralistes (the European Molecular Biology Laboratory (EMBL) Bank [5], the DNA Data Bank of Japan (DDBJ) [6], GenBank [7]) et spécialisées (IMGT/LIGM-DB [8], BRENDA [9], FlyBase [10]). L'accès à l'information est facilité par des interfaces Web afin de pouvoir gérer les connaissances d'un domaine dans des ontologies de domaines généraux et spécialisés (Gene Ontology (GO) [11], Sequence Ontology (SO) [12], Protein Ontology (PRO) [13], IMGT-ONTOLOGY [14-15] Les ontologies bénéficient du développement de méthodologies et d'outils comme Protégé

(http://protege.stanford.edu/). Le plus souvent, les développements informatiques et la construction d'ontologies permettent l'automatisation de la plupart des activités d'un expert. Les données de la biologie sont intégrées dans des modèles pour expliquer les mécanismes de fonctionnement des systèmes étudiés et prédire leur comportement, leur évolution. Ce domaine de la biologie systémique est en plein développement (CellML [16-17], ImmunoGrid [18]).

Parmi les systèmes biologiques, le système immunitaire est l'un des plus complexes et des plus importants en recherche médicale et thérapeutique. Sa fonction est de défendre les organismes multicellulaires contre les pathogènes (i.e. bactéries, parasites, virus et champignons) et cellules tumorales. Pour cela, il existe deux types de réponses immunitaires. La réponse immunitaire innée, ou naturelle, dans laquelle les défenses du système immunitaire sont immédiatement disponibles et, chez les vertébrés à mâchoires seulement, la réponse immunitaire dite spécifique (adaptative). Cette dernière fait intervenir, au cours de la vie d'un individu infecté, des cellules spécialisées dans la reconnaissance et la destruction du pathogène. La réponse adaptative confère une immunité contre la réinfection par le même pathogène spécifiquement reconnu par des récepteurs d'antigènes majeurs, immunoglobulines (IG) ou anticorps [19] et récepteurs T (TR) [20]. Ils présentent une énorme diversité des sites de reconnaissance antigénique (2.10^{12} pour les IG et 2.10^{12} pour les TR par individu). Les nombreuses protéines différentes sont codées par un nombre relativement limité de gènes organisés dans différents locus (7 chez l'homme). Les IG et TR sont synthétisés à partir de différents types de gènes: variable (V), diversité (D), jonction (J) et constant (C). La synthèse des IG et TR requiert des mécanismes complexes incluant, au niveau de l'ADN, d'abord les réarrangements des gènes V et J ou des gènes V, D et J créant la diversité combinatoire [21], puis la N-Diversité au cours de laquelle la jonction V-J ou V-D-J est produite [22-23], et enfin pour les IG, les hypermutations somatiques [24-25]. Ces réarrangements sont suivis au niveau de l'ARN de l'épissage des gènes réarrangés V-J et V-D-J suivi de leur transcription et de celle du gène C.

La complexité des séquences des IG et TR rend difficile leurs description et leur gestion par des bases généralistes. Dans le but de gérer les données des IG et TR, IMGT® (http://www.imgt.org/), « the international ImMunoGeneTics® information system » [26] a été créé en 1989, par le Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 et CNRS). L'un des premiers objectifs d'IMGT® était d'identifier et de décrire

tous les gènes d'IG et TR de l'homme, un prérequis indispensable avant l'analyse du répertoire. Face à la complexité des IG et TR, IMGT-ONTOLOGY [14-15, 27], la première ontologie dans le domaine de l'immunogénétique et immunoinformatique a été construite pour assurer l'efficacité et la consistance des données d'IMGT®, tout comme la cohérence entre les bases de données, les outils et les ressources Web d'IMGT®. Plusieurs années d'expertise et d'annotation manuelle ont été nécessaires à la nomenclature IMGT® des gènes d'IG et TR [19-20]. La nomenclature IMGT® a été approuvée par le « Human Genome Organisation (HUGO) Nomenclature Committee » (HGNC) en 1999 [28] et par le comité de nomenclature de l'Organisation Mondiale de la Santé - Union internationale des Sociétés Immunologiques (World Health Organization - International Union of Immunological Societies, WHO-IUIS) [29-30]. Les gènes humains d'IG et TR ont été entrés dans IMGT/GENE-DB [31], la base de données de gènes d'IMGT®, dans le « Human Genome Database » (GDB) [32], dans LocusLink [33], et dans Entrez Gene [34] au National Center for Biotechnology Information (NCBI), quand cette base de données a remplacé LocusLink. Ensembl [35] à l' « European Bioinformatics Institute » (EBI) et Vega [36] au « Wellcome Trust Sanger Institute » utilisent les gènes IMGT®.

Bien que tous les gènes d'IG et TR de l'homme soient connus, de grands contigs restent à annoter précisément. Il existe de nombreuses sources d'erreurs inhérentes à l'annotation des génomes. L'assemblage des génomes est le résultat de la jonction de plusieurs fragments d'ADN provenant de différents haplotypes de l'un ou l'autre des chromosomes. Ce mélange reflète seulement une image approximative d'un 'vrai' haplotype. Une analyse rigoureuse est nécessaire pour l'identification des gènes et des allèles d'IG et TR. Les gènes d'IG et TR appartiennent à des sous groupes multigéniques. Dans chacun des sous groupes, les gènes sont dupliqués voir tripliqués et, en conséquence, partagent un pourcentage d'identité élevé (>75%). Ainsi, il est souvent difficile de déterminer si des séquences pratiquement identiques provenant d'individus d'une même espèce sont des gènes ou des allèles différents d'autant plus que le polymorphisme d'insertions et délétions est fréquent dans les familles multigéniques. Les locus des IG et TR contiennent plusieurs centaines de gènes. Ce nombre varie en fonction de l'haplotype considéré soit de 608 à 665 gènes d'IG et TR par génome haploïde chez l'homme, et de 619 à 628 gènes chez la souris. Il faut noter le nombre important de pseudogènes, de 227 à 253 chez l'homme et de 212 à 240 chez la souris. Beaucoup de ces gènes sont dégénérés et partiels, ce qui les rend particulièrement difficiles à annoter. Les gènes orphons d'IG et TR sont des cas particuliers car leur emplacement est en

dehors des locus principaux. Les orphons sont non fonctionnels mais partagent un pourcentage d'identité élevé avec les gènes des locus majeurs. Il est donc important de replacer une séquence à annoter dans son contexte génomique pour identifier leur fonctionnalité. Finalement, les gènes D et J possèdent de très petites régions codantes, de 8 à 37 paires de bases (bp) et de 37 à 69 bp, respectivement. L'ensemble de ces sources d'erreurs et de confusion fait de l'identification des gènes d'IG et TR une tâche ardue. IMGT® a mis en place un système d'annotation semi-automatique des gènes d'IG et TR qui tient compte de ces difficultés. L'annotation des gènes dans de grandes séquences génomiques est encore réalisée en grande partie manuellement par les experts d'IMGT®. En effet, les logiciels de prédiction de gènes en ligne tels que GENEMARK [37], GENESCAN [38] et N-SCAN [39] ne sont pas adaptés à l'annotation des gènes d'IG et TR en raison des particularités de leur structure.

Les objectifs de ma thèse étaient d'intégrer au sein d'IMGT® une nouvelle composante dans le domaine génomique, avec la mise en place d'un outil capable, dans des séquences d'ADN génomique de grande taille, d'identifier les gènes V, D et J des IG et TR, de les décrire et d'identifier leur fonctionnalité, pour les locus de l'homme, de la souris et du rat dans une structure facilitant l'expertise des annotateurs.

Le premier objectif était d'analyser l'organisation des locus tels qu'ils sont gérés par IMGT, d'approfondir tous les aspects d'IMGT-ONTOLOGY et d'analyser l'existant nécessaire pour concevoir et développer un outil Java capable d'annoter les gènes V, D et J des IG et TR dans l'ADN génomique et, ainsi, d'accélérer le processus d'annotation d'IMGT® en regroupant des alignements locaux de motifs appartenant au même gène. Ensuite, les gènes identifiés sont décrits précisément. Pour cela, il est nécessaire de faire ressortir les motifs les plus difficiles à retrouver en les alignant avec une base de référence de motifs dans une zone restreinte à proximité ou/et dans le gène identifié. La fonctionnalité est attribuée à chaque gène à partir des critères de la charte scientifique d'IMGT®. Finalement, les UTRs sont délimités pour assembler les gènes entre eux et en identifier leur cluster. Le nombre de gènes V, D et J retrouvés dans le brin direct et complémentaire de la séquence analysée est aussi calculé.

Le deuxième objectif était l'implémentation et l'évaluation d'IMGT/LIGMotif, au niveau de séquences de l'homme, de la souris et du rat provenant de différents locus (IGL, TRB chez l'homme, TRG chez la souris et IGK chez le rat) par comparaison du fichier

contenant les annotations d'IMGT/LIGMotif avec le fichier contenant les annotations des experts.

Enfin le troisième et dernier objectif était de concevoir et de développer une interface web permettant à l'annotateur d'utiliser IMGT/LIGMotif, de sélectionner les bases de motifs de références ainsi que le format de la séquence à analyser et de visualiser simplement ses résultats. L'interface permet, d'une part, d'avoir accès à tous les labels des gènes sélectionnés dans une vue synthétique comprenant les statistiques des gènes identifiés, et d'autre part, d'avoir leur positionnement précis dans la séquence et une vue plus fine des résultats. Les labels peuvent être exportés sous le format 'comma-separated value' (CSV) ou XML (en anglais, eXtensible Markup Language).

Le **chapitre 1** expose les caractéristiques et particularités d'identification et de description des gènes V, D et J des IG et TR dans leur locus chez l'homme et la souris. Le **chapitre 2** présente les axiomes et concepts majeurs de l'IMGT-ONTOLOGY. Le **chapitre 3** décrit les principales bases de données et outils d'analyse de séquences d'ADN au sein d'IMGT®. Le **chapitre 4** présente l'analyse du système d'annotation des IG et TR existant au sein d'IMGT® et en particulier du logiciel LIGMotif, la version qui a servi de base au développement d'IMGT/LIGMotif. Le **chapitre 5** décrit le modèle d'IMGT/LIGMotif, son algorithme, son implémentation et son évaluation.

CHAPITRE 1
Récepteurs des antigènes et organisation de leurs locus

Le système immunitaire défend les organismes multicellulaires contre les cellules devenues tumorales et les infections bactériennes, parasitaires et virales. La stratégie de protection repose sur différents systèmes de reconnaissances et mécanismes de destructions des pathogènes. L'immunité innée ou naturelle est la première ligne de défense. La barrière physique (p.ex. la peau), les cellules phagocytaires, les interférons et le système du complément en sont les principaux acteurs. La barrière physique prévient de la pénétration des pathogènes dans l'organisme, les cellules phagocytaires ingèrent ou digèrent les pathogènes par endocytose, les interférons agissent contre les virus et le système du complément détruit la surface des cellules étrangères. Cette défense a l'avantage d'être immédiate, mais les pathogènes ne sont pas reconnus spécifiquement.

L'immunité adaptative vient compléter ce système. Dans ce cas, chaque pathogène est reconnu par une signature spécifique de protéine qui les identifie (antigène). Le mécanisme de la mémoire immunitaire 'mémorise' l'antigène et ainsi accélère et amplifie les réponses ultérieures vis-à-vis d'un pathogène spécifiquement et préalablement reconnu. Les IG et TR sont les principaux acteurs de l'immunité adaptative. Ces récepteurs détectent les pathogènes présentant l'antigène pour lequel ils sont spécifiques. Les IG et TR ont la capacité de reconnaître une multitude de pathogènes par le biais de plusieurs mécanismes à l'origine de leur diversité.

Dans ce chapitre, deux composantes du système immunitaire adaptatif seront abordées: l'immunité humorale et l'immunité cellulaire. Nous décrirons d'abord les mécanismes à l'origine de la diversité des récepteurs d'antigènes, en prenant comme exemple les IG, puis la répartition des gènes V, D, J et C dans leurs locus pour les génomes de l'homme et de la souris.

1.1 Immunité humorale et cellulaire

Le système immunitaire adaptatif est souvent décrit par ses deux composantes, l'immunité humorale (sécrétion d'anticorps) et l'immunité cellulaire (cytolyse des cellules infectées ou cancéreuses) qui font intervenir différentes cellules et molécules (Figure 1.1) et défendent l'organisme contre les pathogènes extracellulaires et intracellulaires.

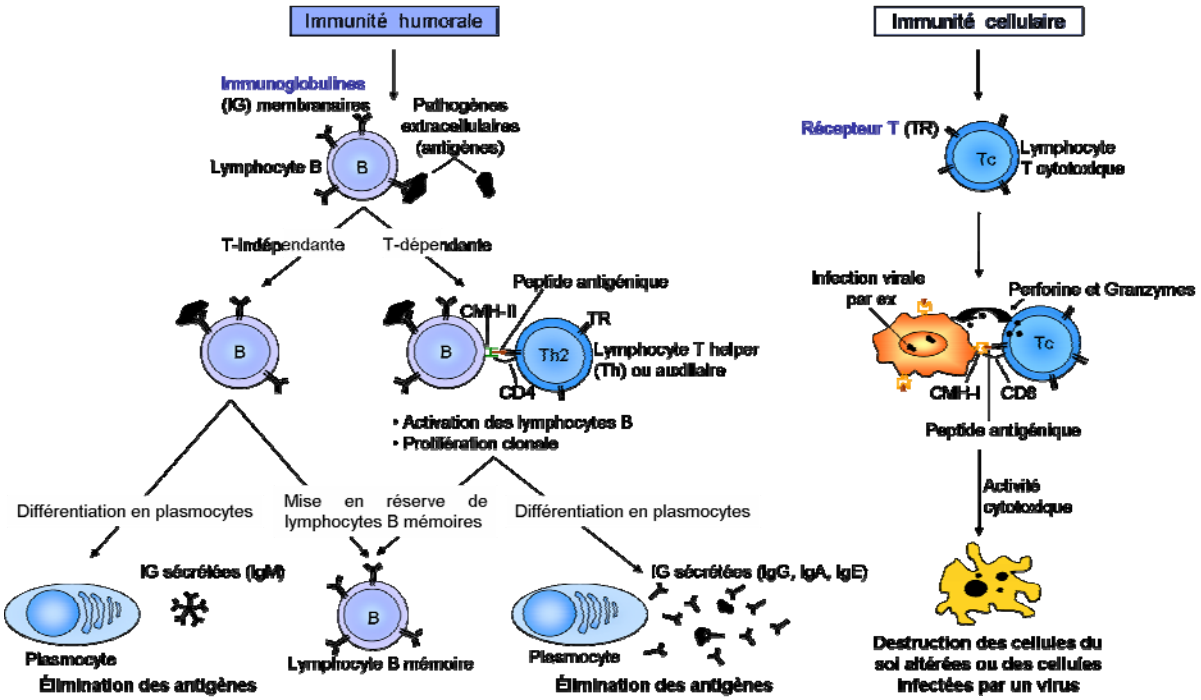


Figure 1.1. Les deux composantes du système immunitaire adaptatif: immunité humorale et cellulaire. Les cellules de la lignée B comprennent les lymphocytes B et les plasmocytes. Les cellules de la lignée T comprennent les lymphocytes T, qui peuvent être des lymphocytes T helper ou auxiliaire (Th), ou des lymphocytes T cytotoxiques (Tc). Les protéines caractéristiques de la réponse immunitaire adaptative comprennent les immunoglobulines (IG), les récepteurs T (TR), le complexe majeur d'histocompatibilité de classe I (CMH-I) et de classe II (CMH-II) (avec la permission d'IMGT®, http://www.imgt.org).

L'immunité humorale agit contre les pathogènes (bactéries et virus) circulant dans le sang et la lymphe et elle est basée sur la reconnaissance spécifique de déterminants antigéniques (ou épitopes) par les sites anticorps (ou paratopes) des domaines variables (V) des IG. Les IG existent en tant que protéines membranaires à la surface des lymphocytes B ou sont sécrétées par les plasmocytes, cellules qui représentent le stade de différenciation terminal des cellules B (Figure 1.1). Les lymphocytes B se développent à partir de cellules-souches dans la moelle osseuse. Durant la synthèse des IG, des mécanismes de réarrangements de l'ADN permettent la génération d'une énorme diversité de lymphocytes B (10^{12} chez l'homme), le facteur limitant étant le nombre de lymphocytes B génétiquement programmés pour une espèce donnée. Toutes les IG exprimées à la surface d'un lymphocyte

B sont identiques et ont la particularité d'avoir la même spécificité de reconnaissance d'un antigène. Les lymphocytes B matures circulent alors dans la lymphe et gagnent les organes lymphoïdes secondaires (ganglions lymphatiques, rate). Dans les organes lymphoïdes secondaires, un lymphocyte B qui reconnaît un antigène pour lequel il est spécifique, est activé et prolifère et après contact avec un lymphocyte T spécifique, se différencie soit en plasmocyte qui sécrète des IG, soit en lymphocyte B mémoire (Figure 1.1). Les anticorps sécrétés par les plasmocytes neutralisent le pouvoir infectieux des pathogènes en se liant à leurs antigènes de surface, qui interfère avec leur capacité à se fixer aux cellules de l'hôte (anticorps neutralisants). Les anticorps peuvent entraîner également une destruction de l'agent pathogène, par le complément (complement dependent cytotoxicity ou CDC) ou par une cellule cytotoxique (antibody dependent cellular cytotoxicity ou ADCC) (IMGT®, http://www.imgt.org). Enfin en recouvrant les pathogènes, les anticorps favorisent la phagocytose par les macrophages (opsonisation). Les lymphocytes B mémoires issus de lymphocytes B déjà sélectionnés et ayant subi l'expansion clonale et la commutation de classe possèdent à leur surface des IG membranaires. Ces lymphocytes B mémoires ont une durée de vie beaucoup plus longue que les plasmocytes, et pourront être activés et se différencier en plasmocytes lors d'une nouvelle rencontre avec le même antigène. L'immunité cellulaire est chargée de la défense de l'organisme vis-à-vis des cellules infectées par des agents pathogènes intracellulaires (virus) ou des cellules cancéreuses. Les lymphocytes T sont issus des cellules-souches de la moelle osseuse (comme les lymphocytes B) mais ils se différencient ensuite dans le thymus. Les mécanismes de synthèse des TR, semblables à ceux des IG sont basés sur des réarrangements de l'ADN, qui génèrent une grande diversité combinatoire de TR et de lymphocytes T (potentiellement 10^{12} chez l'homme). Chaque lymphocyte T exprime des TR d'une seule et même spécificité. Les lymphocytes T sont sélectionnés par une double sélection négative et positive qui permet premièrement d'éliminer les lymphocytes T fortement autoréactifs spécifiques des peptides du soi, et deuxièmement de sélectionner les lymphocytes T qui reconnaissent des peptides du non soi. L'interaction TR/peptide, celui-ci est présenté par une molécule du CMH, aboutit à l'expansion clonale du lymphocyte T impliqué dans la reconnaissance spécifique de l'agent pathogène et à la différenciation des clones en lymphocytes T effecteurs, cytotoxiques (Tc), ou auxiliaires ou helper (Th) et en lymphocytes T mémoire. Au sein de toute cellule, les protéines subissent une dégradation par le protéasome et les peptides de 8 à 10 acides aminés issus de cette protéolyse sont ensuite transportés à la surface de la cellule pour être présentés par l'intermédiaire du CMH-I (ou MHC-I, le complexe majeur d'histocompatibilité de classe I). Ainsi, les cellules

saines présentent à leur surface des peptides du soi qui ne déclenchent pas de réaction immunitaire. Au contraire, les cellules étrangères, les cellules tumorales ou infectées par un virus ou un autre agent pathogène présentent à leur surface des peptides du non soi, qui entraînent de manière spécifique une activation des lymphocytes T cytotoxiques (Tc) qui reconnaissent de manière spécifique ce complexe peptide-MHC-I (pMHC-I) et qui les détruisent. Les cellules T auxiliaires ou helper (Th) sécrètent des cytokines qui stimulent la réaction immunitaire auprès des autres cellules. Elles contribuent notamment à l'activation des cellules présentatrices d'antigènes (CPA professionnelles) lesquelles comprennent les cellules dendritiques, les macrophages et les lymphocytes B. Ceux-ci prolifèrent et se différencient en plasmocytes dans les organes lymphoïdes secondaires. On parle alors de réponse humorale T-dépendante (Figure 1.1), par opposition à la réponse humorale T-indépendante qui ne requiert pas l'aide des Th, et dans laquelle les lymphocytes B se différencient en plasmocytes sans contact au préalable avec un Th. Au sein d'une cellule CPA, les protéines exogènes (antigènes extracellulaires) sont dégradées dans les vésicules d'endocytose en peptides de 10 à 15 acides aminés, lesquels sont présentés à la surface de la cellule par l'intermédiaire du CMH-II (ou MHC-II), dont l'un des domaines (le G-DOMAIN) forme un sillon ou 'groove' où se loge le peptide. C'est le complexe pMHC-II qui est reconnu de manière spécifique par le TR des Th CD4+.

1.2 Synthèse des chaînes d'immunoglobulines

Les immunoglobulines (IG ou anticorps) sont exprimées en surface des cellules B matures et des cellules B mémoires ou sont sécrétées par les plasmocytes. Les différentes étapes de la différenciation des cellules souches hématopoïétiques en cellules B matures se produisent dans la moelle osseuse indépendamment de l'antigène (Figure 1.2). Les IG peuvent être classés en 5 classes ou isotypes (IgM, IgD, IgG, IgA ou IgE) selon la structure de leur région constante. En plus de la fonction de reconnaissance très fine exercée par les sites des anticorps, les immunoglobulines assurent d'autres fonctions extrêmement importantes par l'isotype; en théorie, n'importe quel domaine variable de chaîne lourde peut venir s'associer à la région constante de n'importe quel isotype. Chaque isotype contribue de façon différente à l'élimination des pathogènes. Les réarrangements de l'ADN qui sont à la base du changement d'isotype confèrent à la réponse humorale sa diversité fonctionnelle. Les cellules B matures produisent des IgM et des IgD. Les étapes finales de la différenciation des cellules B matures en cellules mémoires ou en plasmocytes, qui expriment ou sécrètent des immunoglobulines de

diverses classes ou sous-classes, se produisent dans le centre germinatif des organes lymphoïdes secondaires, et sont tributaires de l'antigène (Figure 1.2).

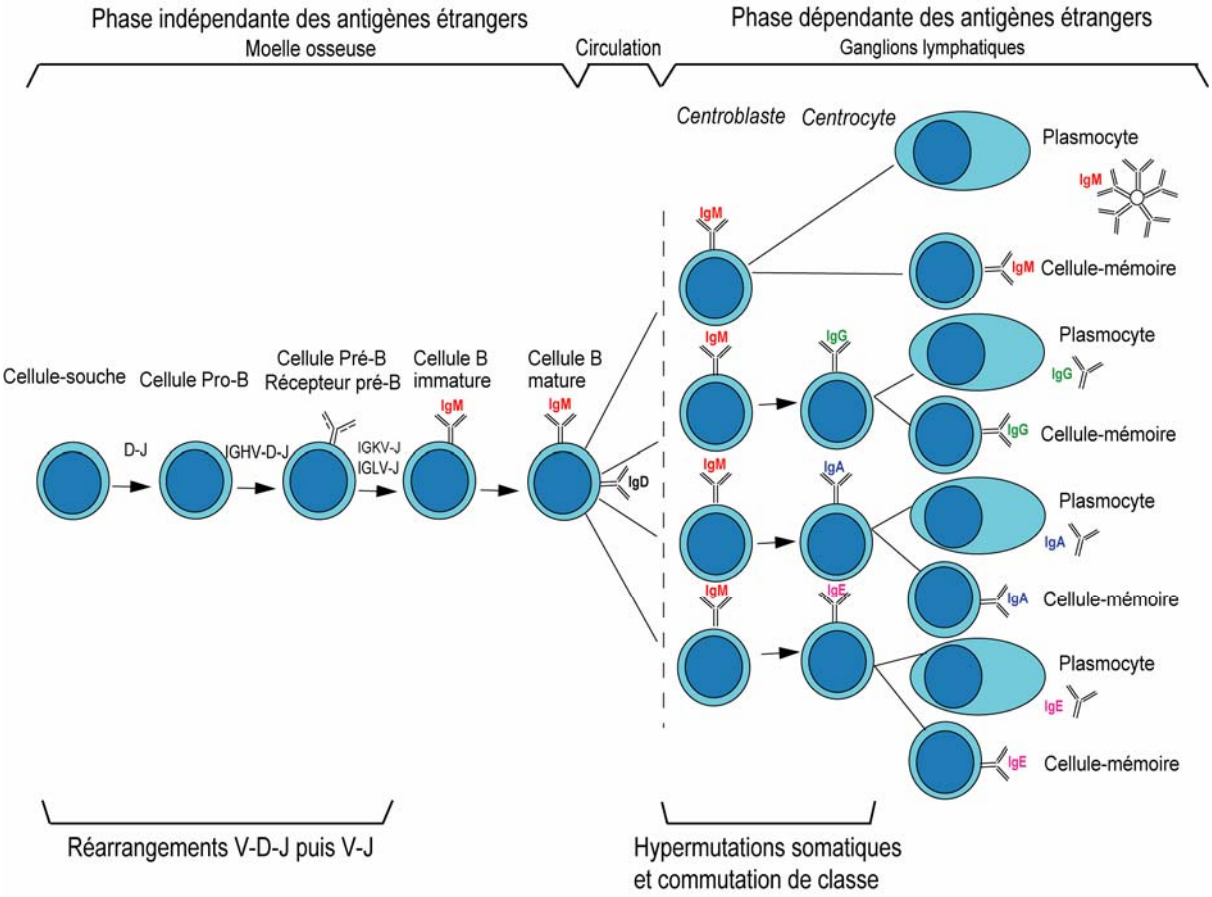


Figure 1.2. Différenciation des lymphocytes B. La différenciation des lymphocytes B comprend deux phases: une phase indépendante des antigènes étrangers, de la cellule souche hématopoïétique jusqu'au lymphocyte B mature, dans la moelle osseuse, et une phase dépendante des antigènes étrangers, du lymphocyte B mature au plasmocyte et au lymphocyte B mémoire, dans les centres germinatifs des organes lymphoïdes secondaires (rate, ganglions lymphatiques). Cette seconde phase requiert généralement une coopération entre les lymphocytes B et T [20] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Les immunoglobulines se composent de deux chaînes lourdes identiques, associées à deux chaînes légères identiques, kappa ou lambda (Figure 1.3). Chez l'homme, les gènes qui codent les chaînes lourdes, les chaînes légères kappa et les chaînes légères lambda, sont localisés dans les locus IGH, IGK et IGL, respectivement sur les chromosomes 14 (14q32.33), 2 (2p11.2) et 22 (22q11.2). La synthèse des chaînes lourdes et des chaînes légères des immunoglobulines requiert le réarrangement de trois types de gènes, variables (V), de diversité (D) et de jonction (J), au niveau de l'ADN, dans les locus IG durant la différenciation des cellules B [40-42]. Chronologiquement, la synthèse des chaînes lourdes mu précède celle des chaînes légères.

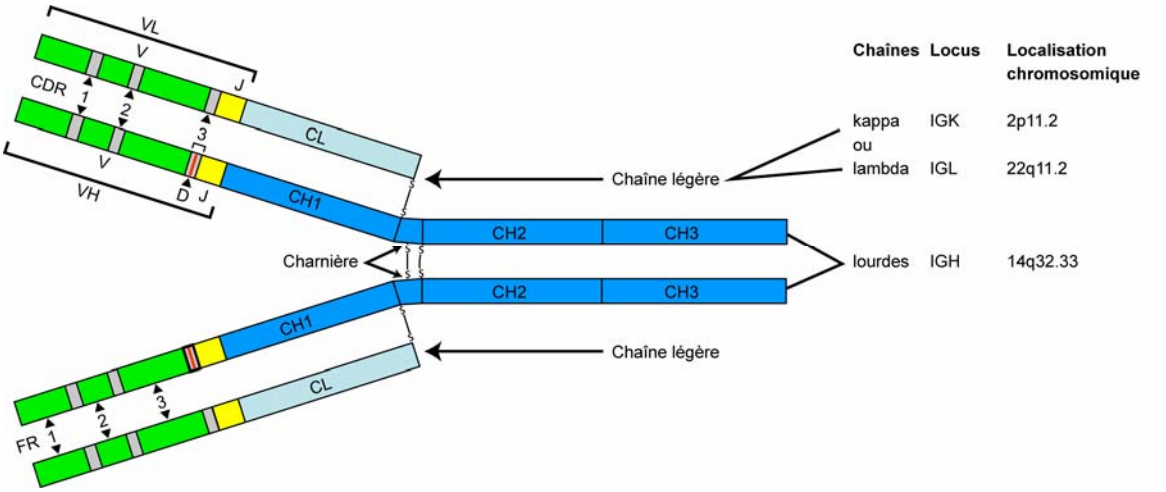


Figure 1.3. Représentation schématique d'une molécule d'IgG1 humaine sécrétée. Le domaine variable d'une chaîne lourde est codé par trois gènes réarrangés (un gène IGHV, un gène IGHD et un gène IGHJ). Le domaine variable d'une chaîne légère, ou V-J-REGION, est codé par deux gènes réarrangés (un gène IGKV réarrangé à un gène IGKJ pour une chaîne kappa, un gène IGLV réarrangé à un gène IGLJ pour une chaîne lambda). Les trois régions hypervariables ou complementarity determining regions (CDR) déterminent le site de reconnaissance et de liaison à l'antigène dans la structure tridimensionnelle. La région constante de la chaîne lourde, codée par des gènes IGHC, comprend 3 ou 4 domaines constants (domaines CH1, CH2, CH3 pour la région constante des IgG, des IgA et des IgD, 4 domaines CH1 à CH4 pour les IgE et les IgM). La région charnière située entre les domaines CH1 et CH2 des IgG est codée par 1 exon (cas des IgG1, IgG2 et IgG4) ou plusieurs exons, le plus souvent 4 (cas des IgG3). La région constante, ou C-REGION, de la chaîne légère est codée par le gène IGKC (cas des chaînes kappa) ou l'un des gènes IGLC (cas des chaînes lambda), et comprend un seul domaine constant (CL) [19] (avec la permission de M.-P. et G. Lefranc, IMGT® http://www.imgt.org).

1.2.1 Synthèse des chaînes lourdes mu: réarrangement D-J et V-D-J dans le locus IGH

Le locus des chaînes lourdes (IGH) comprend des gènes variables (V), de diversité (D), de jonction (J) et constants (C). Le domaine variable de la chaîne lourde, ou V-D-J-REGION, est généré par le réarrangement de l'ADN de trois gènes: un gène IGHV, un gène IGHD et un gène IGHJ. Il se fait en deux temps: le premier correspond au réarrangement d'un gène D à un gène J avec délétion de l'ADN intermédiaire (excision d'une boucle d'ADN) (Tableau 1.4), et le second correspond au réarrangement d'un gène V au D-J précédemment réarrangé pour générer la séquence réarrangée IGHV-D-J (Figure 1.4 et Figure 1.5). La séquence réarrangée IGHV-D-J est transcrite avec le gène IGHM en un pré-messager IGHV-D-J-M (ou IGHV-D-J-Cmu). Le gène IGHM code les quatre domaines CH1 à CH4 de la région constante de la chaîne lourde mu. Les séquences d'ARN correspondant aux introns et aux gènes J non utilisés sont alors excisées lors de l'épissage du pré-messager et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes.

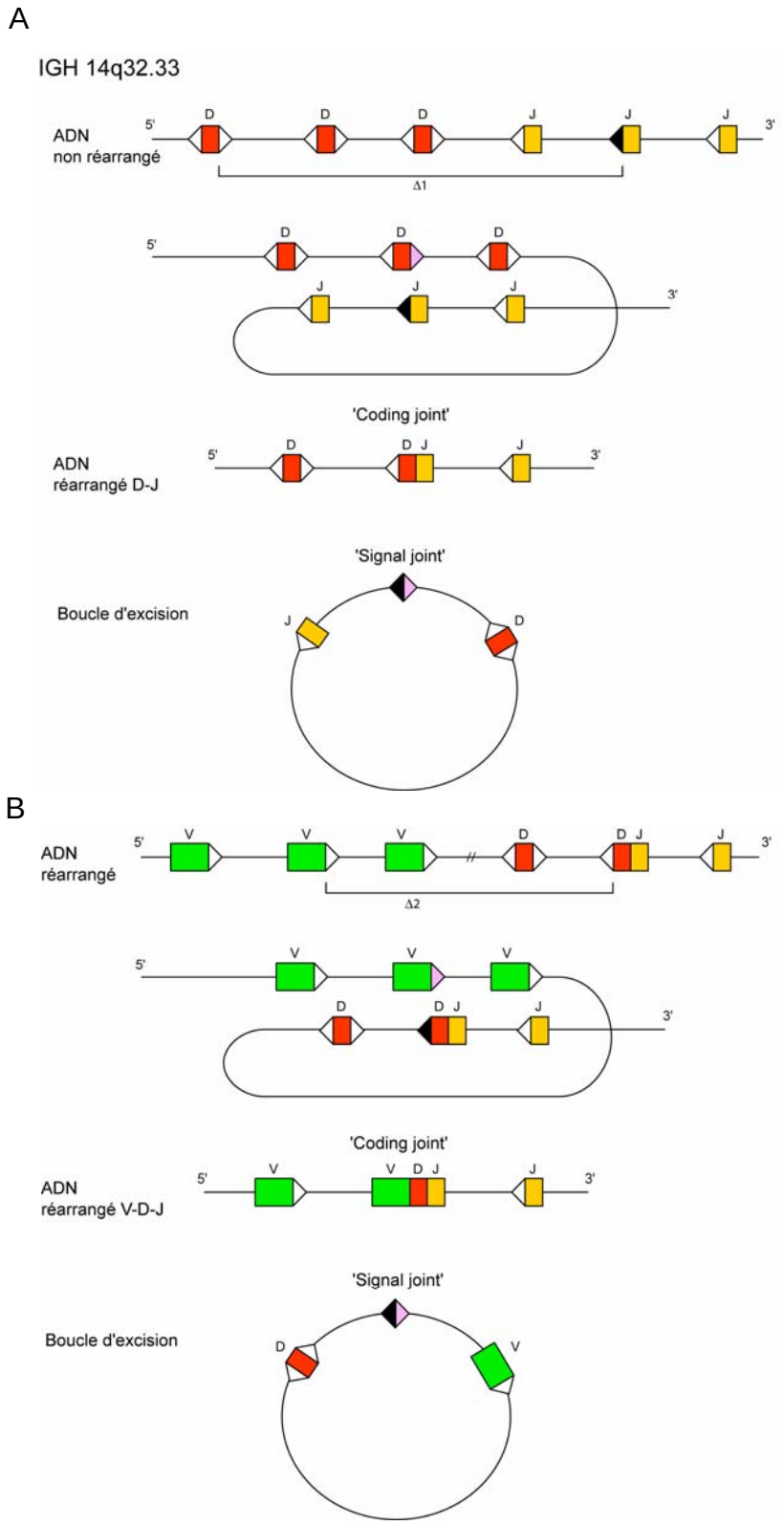


Figure 1.4. Réarrangement dans le locus IGH. (A) Réarrangement dans le locus IGH d'un gène D à un gène J avec délétion de l'ADN intermédiaire (excision d'une boucle d'ADN) (avec la permission d'IMGT®, http://www.imgt.org). (B) Réarrangement dans le locus IGH d'un gène V au D-J précédemment réarrangé pour générer la séquence réarrangée IGHV-D-J (avec la permission d'IMGT®, http://www.imgt.org).

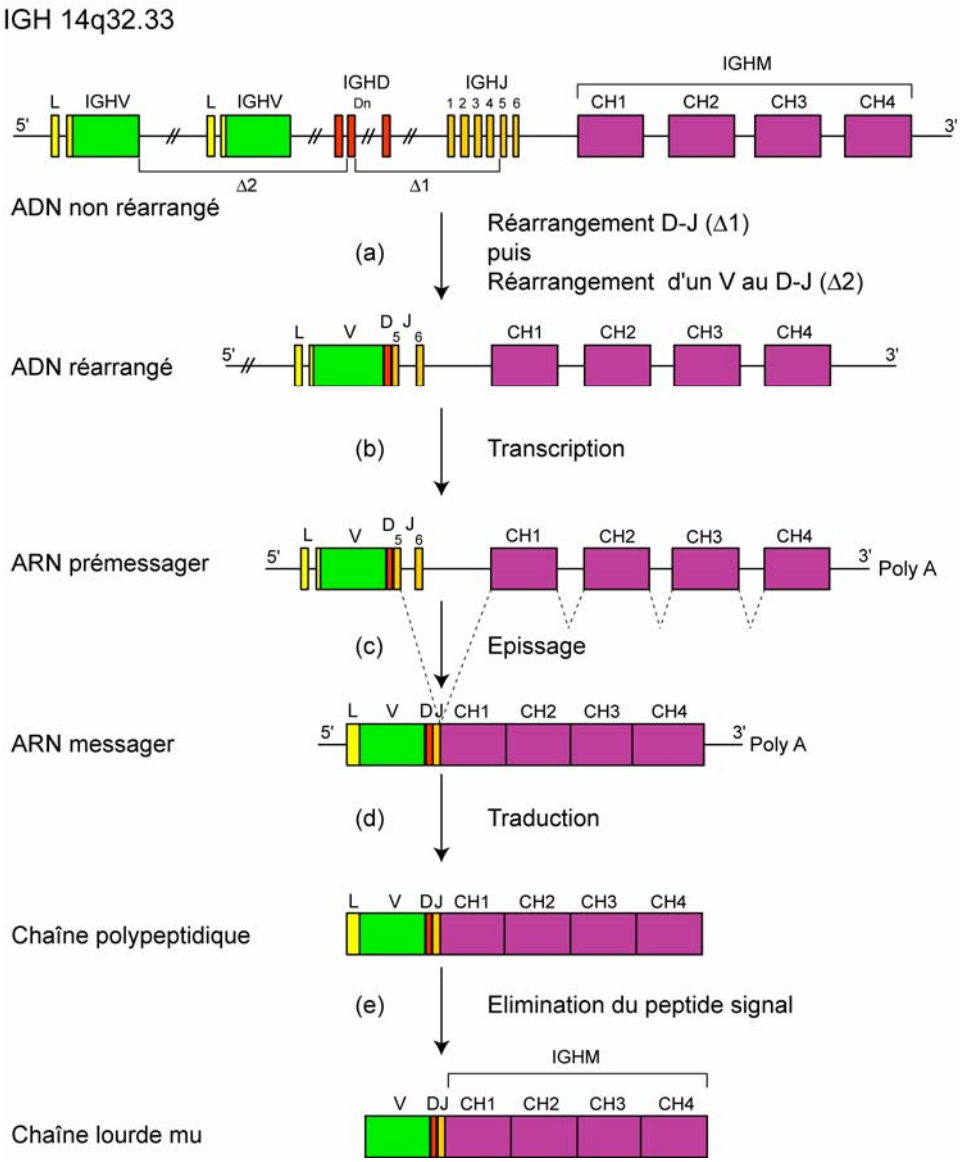


Figure 1.5. Synthèse d'une chaîne lourde mu d'immunoglobuline. (a) Au niveau de l'ADN, lors d'une première étape, l'un des gènes IGHD est joint à l'un des gènes IGHJ avec délétion de l'ADN intermédiaire, pour créer une séquence D-J partiellement réarrangée. Dans une deuxième étape, un des gènes IGHV est joint au D-J préalablement réarrangé, avec délétion de l'ADN intermédiaire pour générer un ensemble IGHV-D-J complètement réarrangé. (b) La séquence réarrangée IGHV-D-J est transcrite avec le gène IGHM en un ARN prémessager IGHV-D-J-M (ou IGHV-D-J-Cmu). (c) Les séquences d'ARN correspondant aux introns et aux gènes IGHJ non utilisés sont alors excisées lors de l'épissage de l'ARN prémessager et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. (d) L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes. (e) Le peptide signal est éliminé par une peptidase après pénétration de la chaîne polypeptidique dans la cavité du réticulum endoplasmique. Une chaîne lourde mu est produite. Dans l'ADN et l'ARN prémessager, L (pour leader) correspond à L-PART1 et L-PART2, dans l'ARN messager et la chaîne polypeptidique L correspond à la L-REGION [19] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Le peptide signal L est éliminé par une peptidase après pénétration de la chaîne polypeptidique dans la cavité du réticulum endoplasmique, et une chaîne lourde mu est alors produite.

1.2.2 Synthèse des chaînes légères lambda et kappa: réarrangement V-J dans les locus IGK et IGL

Le locus kappa (IGK) et le locus lambda (IGL) comprennent des gènes variables (V), des gènes de jonction (J) et des gènes constants (C). Le domaine variable de la chaîne légère (kappa ou lambda) ou V-J-REGION est généré par le réarrangement au niveau de l'ADN de deux gènes: un gène V et un gène J, avec délétion de l'ADN intermédiaire pour créer une séquence réarrangée IGKV-J dans le locus IGK (Figure 1.6) ou IGLV-J dans le locus IGL.

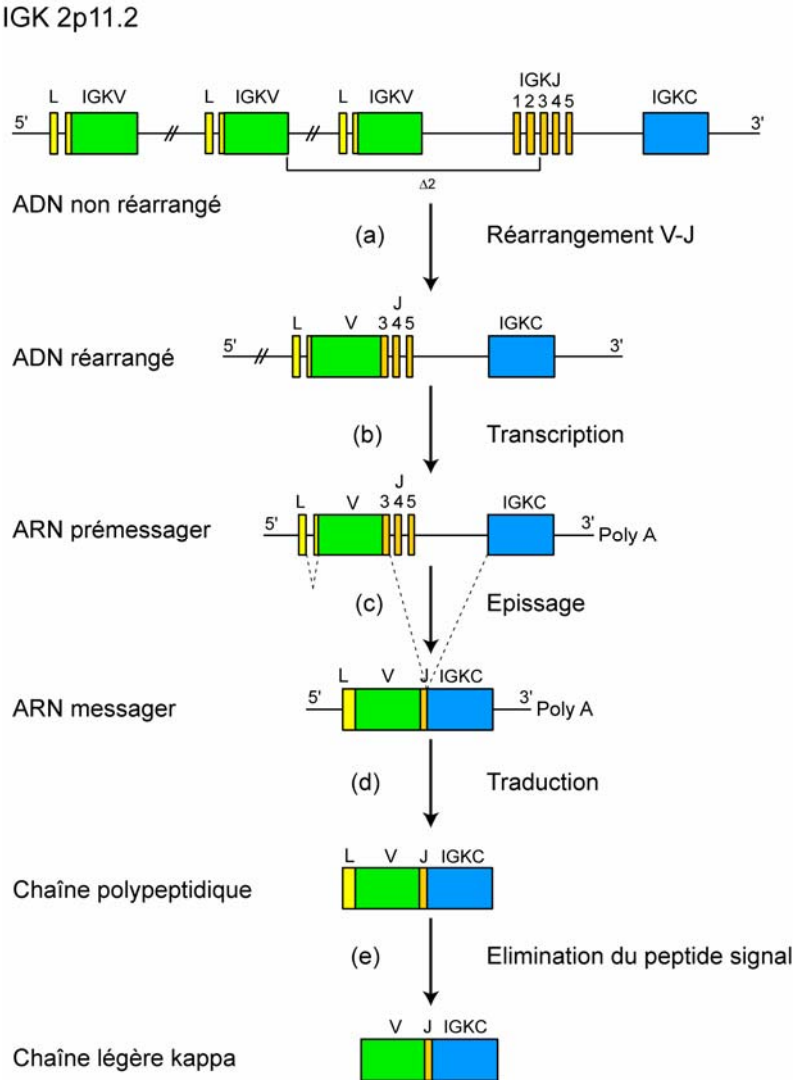


Figure 1.6. Synthèse d'une chaîne légère kappa d'immunoglobuline. (a) Au niveau de l'ADN, l'un des gènes IGKV est réarrangé à l'un des 5 gènes IGKJ avec délétion de l'ADN intermédiaire, pour créer un ensemble IGKV-J. (b) La séquence réarrangée IGKV-J est transcrite avec le gène IGKC en un ARN pré-messager IGKV-J-C. (c) Les séquences d'ARN correspondant aux introns et aux gènes IGKJ non utilisés sont alors excisées lors de l'épissage de l'ARN pré-messager, et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. (d) L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes. (e) Le peptide signal L est éliminé par une peptidase après pénétration de la chaîne polypeptidique dans la cavité du réticulum endoplasmique et une chaîne légère kappa mature est produite. Dans l'ADN et l'ARN pré-messager, L (pour leader) correspond à L-PART1 et L-PART2, dans l'ARN messager et la chaîne polypeptidique, L correspond à la L-REGION [19] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

La séquence réarrangée IGKV-J (ou IGLV-J) est transcrite avec le gène IGKC (ou un des gènes IGLC) en un ARN pré-messager IGKV-J-C (ou IGLV-J-C). L'unique gène IGKC, ou l'un des gènes fonctionnels IGLC, avec leur unique exon, code respectivement le seul domaine de la région constante des chaînes kappa ou lambda. Les séquences d'ARN correspondant aux introns (et pour le locus IGK aux gènes IGKJ non utilisés, pour le locus IGL aux gènes IGLJ et IGLC non utilisés) sont alors excisées par épissage de l'ARN prémessager, et l'on obtient un ARN messager mature qui comprend les régions codantes épissées et les régions 5' et 3' non traduites. L'ARN messager est ensuite traduit en une chaîne polypeptidique par les ribosomes. Le peptide signal L est éliminé par une peptidase après pénétration de la chaîne polypeptidique dans la cavité du réticulum endoplasmique et une chaîne légère mature (kappa ou lambda) est alors produite.

1.3 Locus des IG et TR

1.3.1 Locus des IG et TR humains

1.3.1.1 Locus IGH

Le locus IGH humain est localisé sur le chromosome 14 [43], à la bande 14q32.33 du bras long [44-45]. Le locus IGH (Figure 1.7) comprend 123 à 129 gènes IGHV [46-51] localisés sur une distance de plus de 900 kilobases (kb), dont 38 à 46 sont fonctionnels (Tableau 1.1 et Tableau 1.2) et sont répartis en 6 à 7 sous-groupes. Le locus IGH comprend également 27 gènes IGHD [52-55], dont 23 sont fonctionnels, disposés en tandem sur une distance de 9 kb, tandis que le gène IGHD7-27 est situé à 100 pb (paires de bases) en 5' des gènes IGHJ [55-56]. Il y a 9 gènes IGHJ localisés sur une distance de 8 kb, dont 6 gènes sont fonctionnels. Finalement, le locus IGH comprend 11 gènes IGHC [57-68] situés sur une distance de 300 kb.

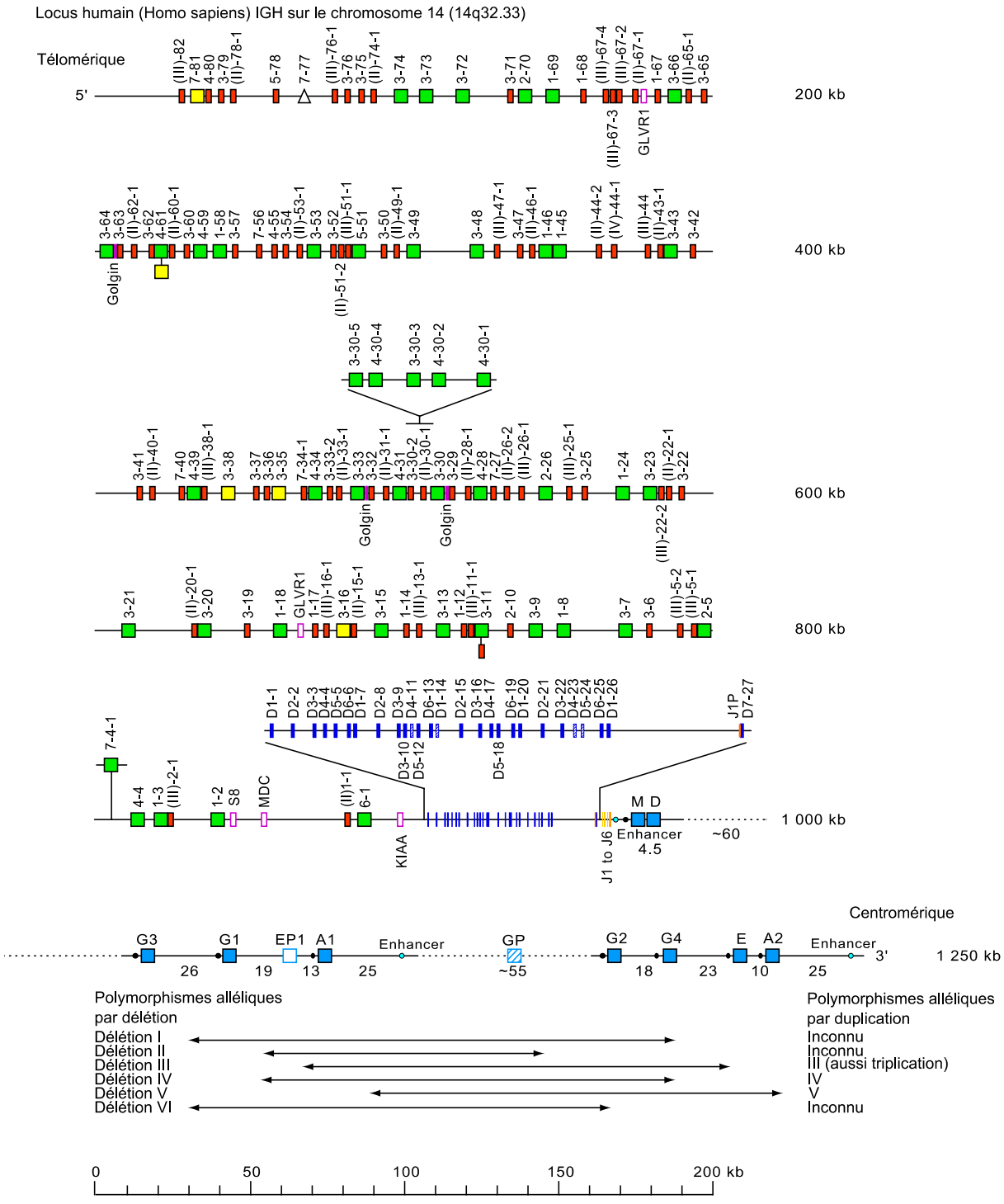


Figure 1.7. Représentation schématique du locus IGH chez l'homme. Le locus IGH comprend de 123 à 129 gènes IGHV dont 38 à 46 sont fonctionnels, 27 gènes IGHD, dont 23 sont fonctionnels et 9 gènes IGHJ, dont 6 sont fonctionnels. Finalement, le locus IGH comprend 11 gènes IGHC dont deux sont des pseudogènes. Les gènes V fonctionnels sont représentés en vert, les gènes V ORF en jaune clair, les gènes V pseudogènes en rouge, les gènes C fonctionnels en bleu (carrés), les gènes D fonctionnels en bleu (traits) et les gènes J fonctionnels en jaune foncé [19] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.1.2 Locus IGK

Le locus IGK humain est localisé sur le chromosome 2 [69], sur le bras court, à la bande 2p11.2 [70]. Le locus kappa (IGK) (Figure 1.8) comprend 76 gènes IGKV [71-77] dont 34 à 37 sont fonctionnels qui appartiennent à 5 sous-groupes. Il existe 5 gènes IGKJ [71, 74, 78] situés en 3' des gènes IGKV et à 2,5 kb en 5' de l'unique gène IGKC [79] qui code la région constante des chaînes légères kappa (Tableau 1.1 et Tableau 1.2).

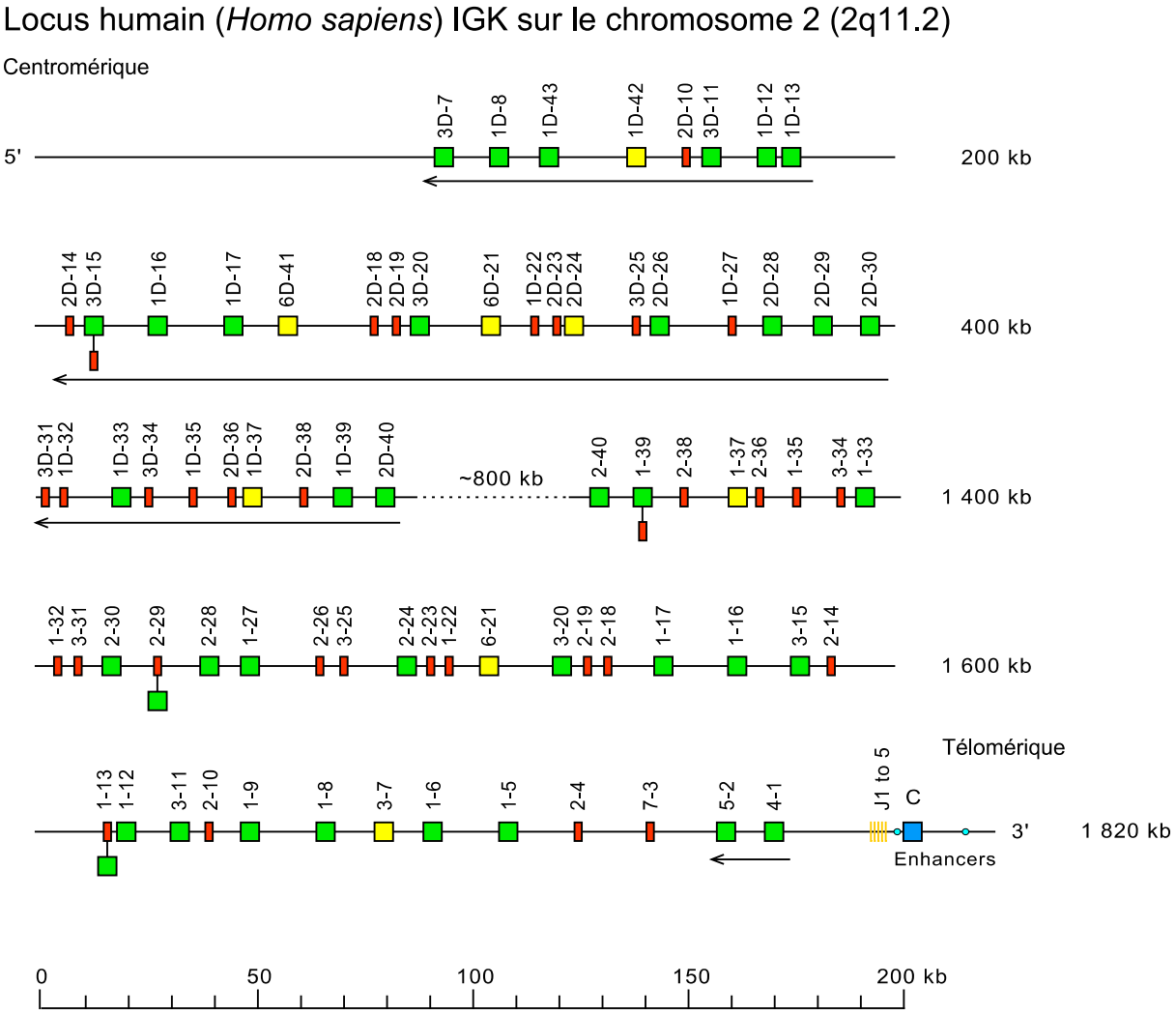


Figure 1.8. Représentation schématique du locus IGK chez l'homme. Le locus IGK comprend 76 gènes IGKV dont 34 à 37 sont fonctionnels, 5 gènes IGKJ et un unique gène IGKC [19] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.1.3 Locus IGL

Le locus IGL humain est localisé sur le chromosome 22 [80], sur le bras long, à la bande 22q11.2 [81]. Le locus lambda (IGL) (Figure 1.9) comprend de 73 ou 74 gènes IGLV [74, 82-86] dont 29 à 35 sont fonctionnels qui appartiennent à 10 sous-groupes (Tableau 1.1 et

Tableau 1.2). Le nombre de gènes IGLC varie chez l'homme de 7 à 11, dont 4 au moins sont fonctionnels et sont en tandem sur une distance de 50 à 70 kb. Chaque gène IGLC fonctionnel est précédé en 5' d'un gène IGLJ [87-90] situé à 1,5 kb.

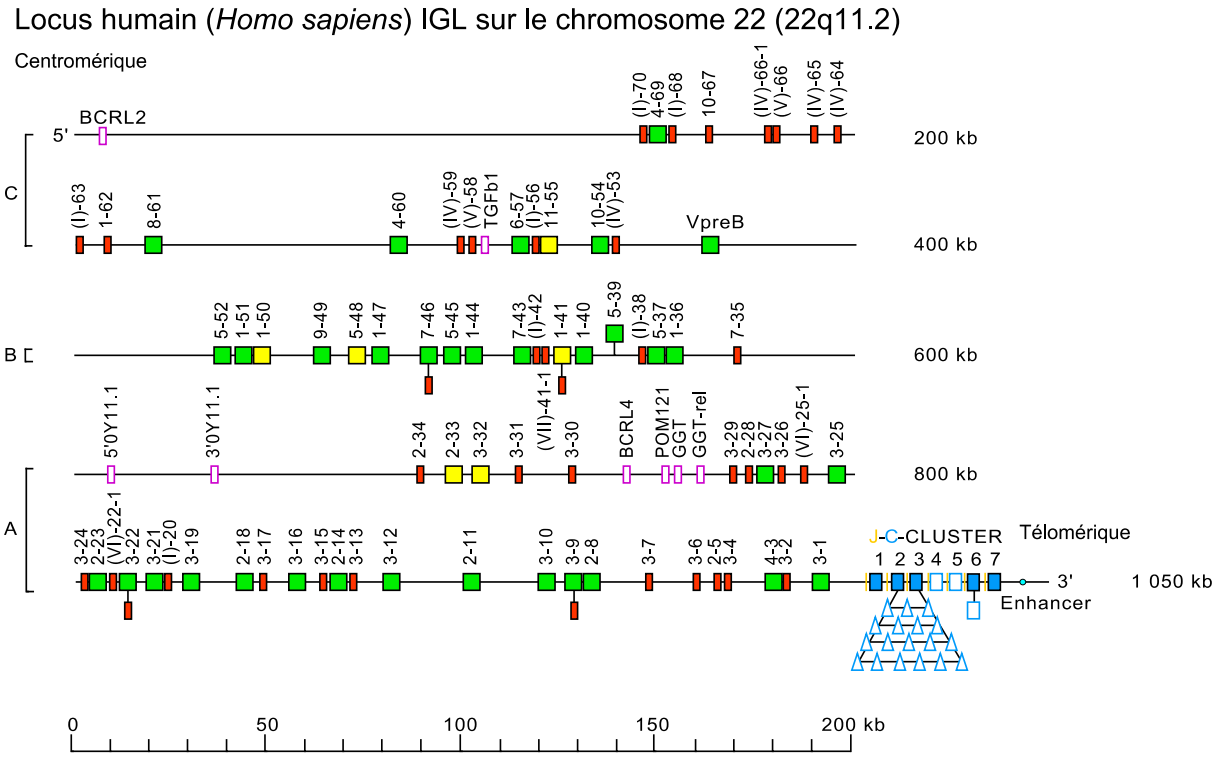


Figure 1.9. Représentation schématique du locus IGL chez l'homme. Le locus IGL comprend 73 ou 74 gènes IGLV (29 à 35 sont fonctionnels). Le nombre de gènes IGLC varie chez l'homme de 7 à 11, dont au moins 4 sont fonctionnels, et chaque gène IGLC fonctionnel est précédé en 5' d'un gène IGLJ situé à 1,5kb [19] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www/imgt.org).

1.3.1.4 Locus TRA/TRD

Les locus TRA et TRD ont la particularité d'être localisés sur le bras long du chromosome 14, à la bande 14q11.2 [91], le locus TRD étant niché dans le locus TRA. Le locus TRA humain (Figure 1.10) comprend 54 gènes TRAV [92-93] incluant les 5 TRAV/DV (Tableau 1.1 et Tableau 1.2). Les gènes TRAV/TRDV ont la particularité d'être retrouvés réarrangés aussi bien avec un TRAJ qu'avec un TRDD. Le locus comprend également 61 gènes TRAJ [93-94] et un unique gène constant TRAC [95-96]. Le locus TRA humain comprend 94 à 96 gènes fonctionnels par génome haploïde, dont 43 à 45 TRAV (si l'on inclut les 5 TRAV/DV), 50 TRAJ et 1 TRAC. L'ensemble des 116 gènes TRA, fonctionnels ou non, s'étend sur une distance de 1000 kb.

Le locus TRD humain comprend 16 gènes (fonctionnels ou non) et sur une distance de 530 kb, dont 3 TRDV, 5 TRAV/DV, 3 TRDD [97-98], 4 TRDJ [97, 99] et un unique TRDC [98]. Le nombre de gènes fonctionnels par génome haploïde du locus TRD est de 15 à 16 dont 5 TRAV/DV, 3 TRDV, 3 TRDD, 4 TRDJ et 1 TRDC. Le gène TRDV3, situé en 3' du gène TRDC, et est en orientation inverse de transcription par rapport aux gènes TRDJ et TRDC et réarrangé par un mécanisme d'inversion.

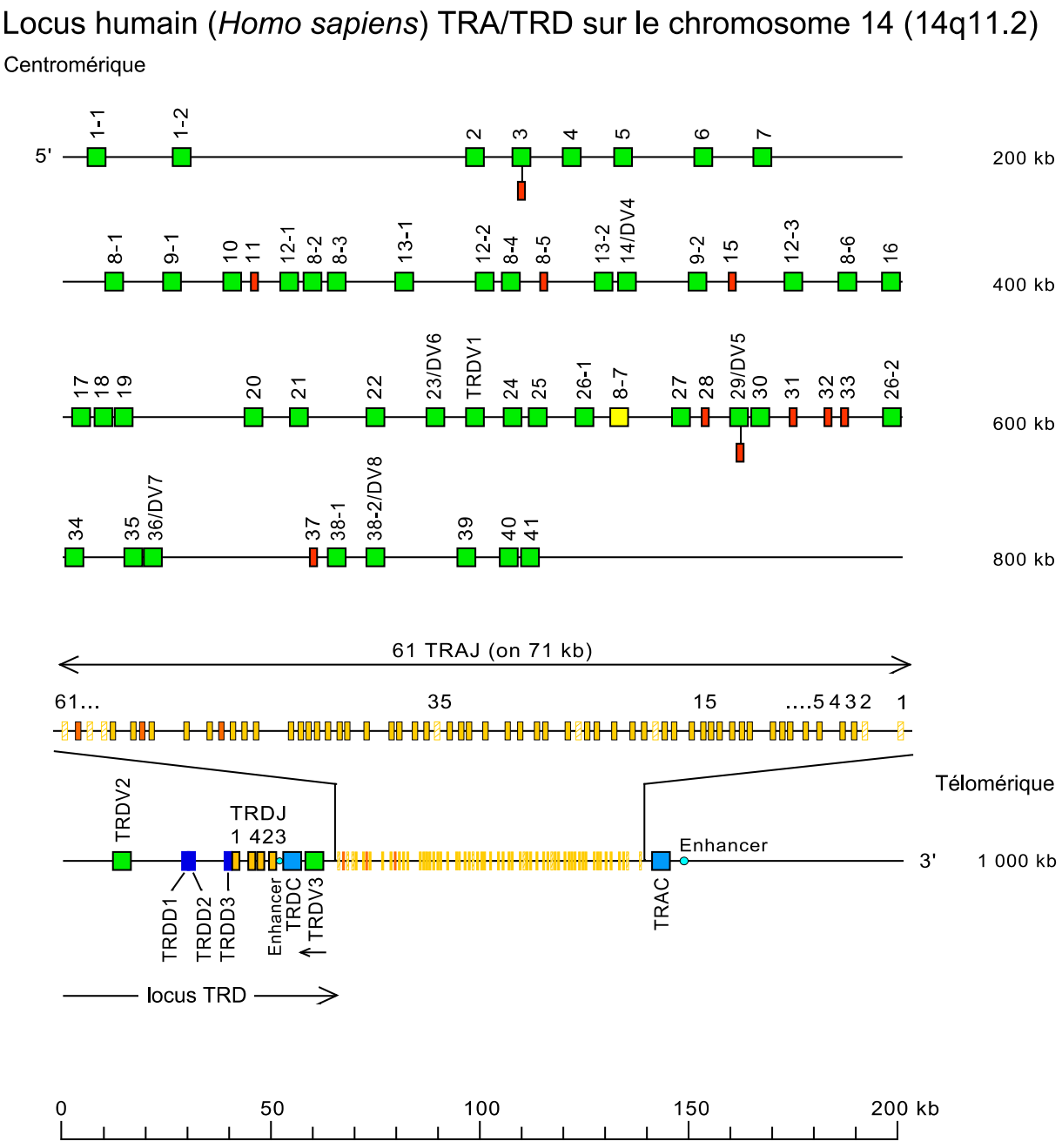


Figure 1.10. Représentation schématique du locus TRA/TRD chez l'homme. Le locus TRA comprend 53 gènes TRAV (43 à 45 sont fonctionnels), 61 gènes TRAJ (50 sont fonctionnels) et 1 gène TRAC (fonctionnel) [20]. Le locus TRD comprend 3 gènes TRDV, 4 gènes TRDJ, 3 gènes TRDD et 1 gène TRDC, tous fonctionnels [20] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.1.5 Locus TRB

Le locus TRB humain est localisé sur le bras long du chromosome 7 [100-102], à la bande 7q34 [102]. Le locus TRB (Figure 1.11) comprend de 64 à 67 gènes TRBV [92, 103-108] qui appartiennent à 32 sous groupes, dont 40 à 48 TRBV sont fonctionnels (Tableau 1.1 et Tableau 1.2). Il existe 2 gènes TRBD [109-110] et 2 TRBC [109] tous fonctionnels et 14 gènes TRBJ dont 12 ou 13 sont fonctionnels [109-110]. Le gène TRBV30 en 3' du gène TRBC2 et est en orientation inverse de transcription par rapport aux gènes TRBJ et TRBC et réarrangé par un mécanisme d'inversion.

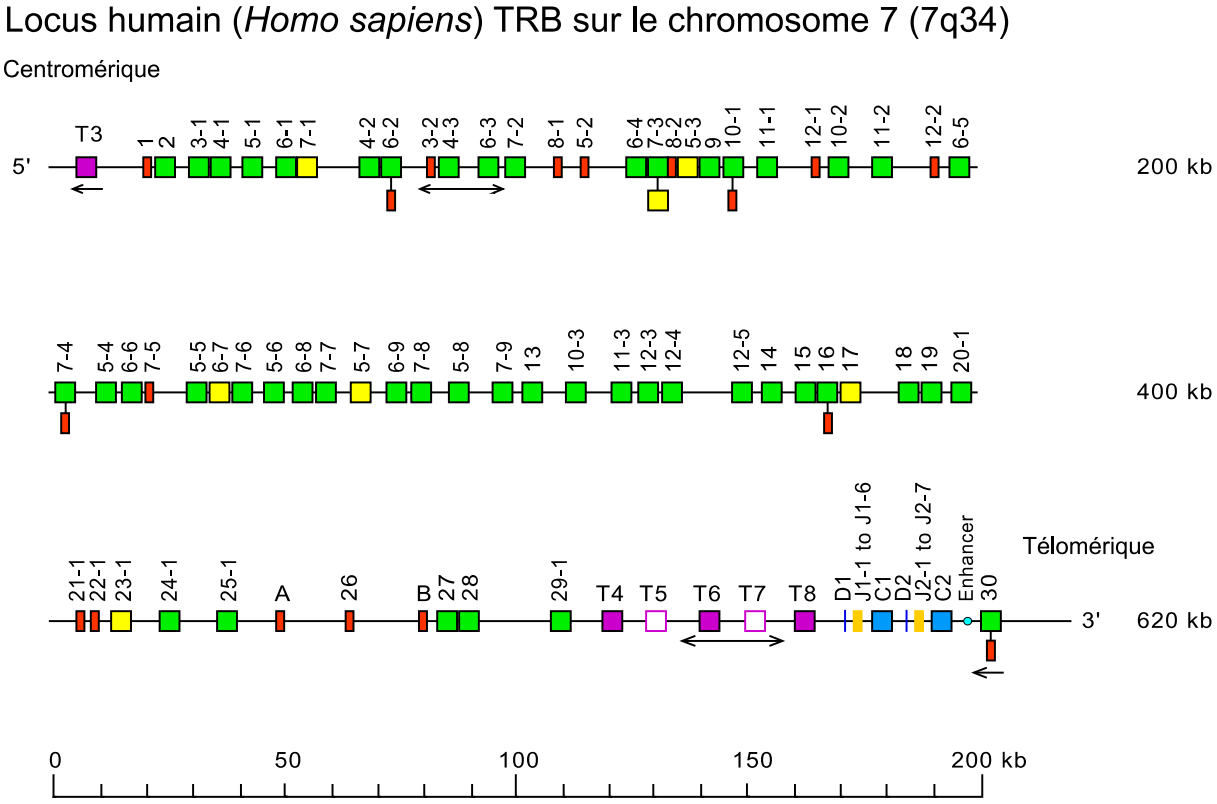


Figure 1.11. Représentation schématique du locus TRB chez l'homme. Le locus TRB comprend de 64 à 67 gènes V (40 à 48 sont fonctionnels), 2 TRBD (tous fonctionnels), 14 TRBJ (12 ou 13 sont fonctionnels) et 2 TRBC (tous fonctionnels) [20] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.1.6 Locus TRG

Le locus TRG humain est localisé sur le bras long du chromosome 7 [111], à la bande 7q14 [112]. Le locus TRG s'étend sur 160 kb (Figure 1.12) et comprend de 12 à 15 gènes TRGV [66, 113-115] appartenant à 6 sous-groupes, dont 4 à 6 gènes TRGV sont fonctionnels et appartiennent à 2 sous-groupes. Il existe 5 gènes TRGJ [116] et 2 TRGC [117-118] fonctionnels (Tableau 1.1 et Tableau 1.2).

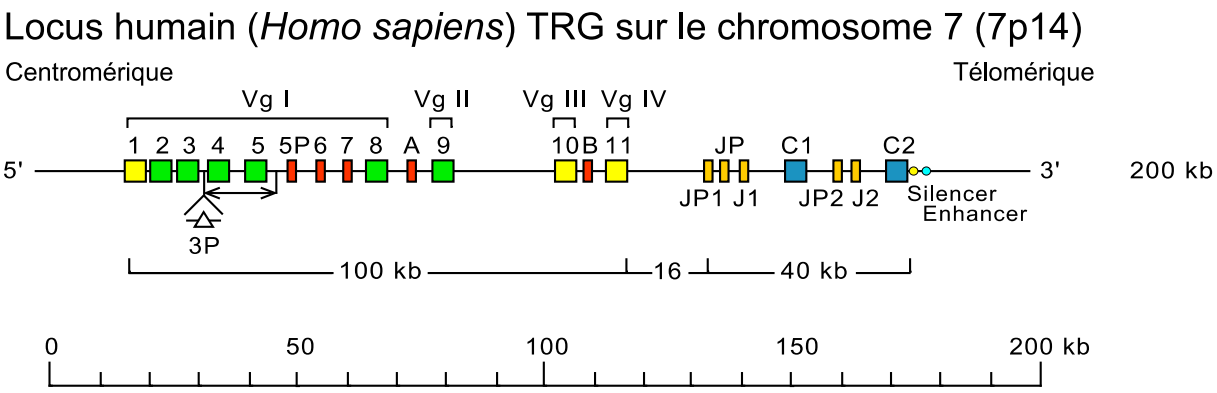


Figure 1.12. Représentation schématique du locus TRG chez l'homme. Le locus TRG comprend de 12 à 15 gènes TRGV (4 à 6 sont fonctionnels), 5 TRGJ (tous fonctionnels) et 2 TRGC (tous fonctionnels) [20] (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Tableau 1.1. Nombre total de gènes d'IG et TR par génome haploïde chez l'homme (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org/)

Locus	Localisation chromosomique	Type de gènes				Total (V+D+J+C)	Nombre d'orphons	Total (V+D+J+C+orphons)
		V	D	J	C			
IGH	14q32.33	123-129	27	9	11[a]	170-176[b]	36	206-212[b,c]
IGK	2p11.2	(40[d] ou) 76	0	5	1	(46[d] ou) 82	25	(71[d] ou)-107
IGL	22q11.2	73-74	0	7-11	7-11	87-96	7[e]	94-103
TRA	14q11.2	54[f]	0	61	1	116[f]	0	116[f]
TRD	14q11.2	3(8[f])	3	4	1	11 (16[f])	0	11 (16[f])
TRB	7q34	64-67	2	14	2	82-85	9	91-94
TRG	7p14	12-15	0	5	2	19-22	0	19-22

[a]des délétions multigéniques, des duplications et triplications alléliques des gènes IGHC ont été décrites chez les individus sains. Le nombre de gènes IGHC peut varier de 5 (délétion I, dans la
Figure 1.7) à 19 (triplication III, dans la Figure 1.7), par génome haploïde.
[b]comprend les 7 gènes IGHV non localisés.
[c]inclut les 'processed gene' IGHEP2 localisé sur le chromosome 9 (9p24.2-p24.1).
[d]nombre de gènes dans l'haplotype rare IGKV dépourvu du V-CLUSTER distal.
[e]inclut le processed gene IGLJ-C/OR18.
[f]inclut les 5 TRAV/DV.

Les locus comprennent le locus IGH (14q32.33), le locus IGK (2p11.2), et le locus IGL (22q11.2). Ces gènes sont impliqués dans la synthèse des chaînes d'immunoglobulines. Les orphons sont localisés en dehors des principaux locus, et ne contribuent pas à la synthèse des chaînes d'IG. 25 IGHV, 10 IGHD, 25 IGKV, 4 IGLV et 2 IGLC orphons ont été identifiés. Les deux 'processed genes' d'immunoglobulines décrits à ce jour, IGHEP2 et IGLJ-C/OR18, ont été inclus avec les orphons dans ce tableau.

Tableau 1.2. Nombre de gènes d'IG et TR fonctionnels par génome haploïde chez l'homme (avec la permission de M.-P. et G. Lefranc, IMGT ®, http://www.imgt.org).

Locus	Localisation chromosomique	Taille des locus (kb)	Type de gènes				Nombre de gènes fonctionnels	Diversité combinatoire	
			V	D	J	C		Minimum	Maximum
IGH	14q32.33	1250	38-46	23	6	9[a]	76-84	38 x 23 x 6 = 5244	46 x 23 x 6 = 6348
IGK	2p11.2	1820	34-38	0	5	1	40-44	34 x 5 = 170	38 x 5 = 190
		500	17-19[b]				23-25[b]	17 x 5 = 85[b]	19 x 5 = 95[b]
IGL	22q11.2	1050	29-33	0	4-5	4-5	37-43	29 x 4 = 116	33 x 5 = 165
TRA	14q11.2	1000	43-45[c]	0	50	1	94-96[c]	44 x 50 = 2200	46 x 50 = 2300
TRD	14q11.2	60[d] (530[e])	3 (7-8[c])	3	4	1	11 (15-16[c])	3 x 3 x 4 = 36	8 x 3 x 4 = 96[c]
									8 x 7 x 4 = 224[f]
TRB	7q34	620	40-48	2	12-13	2	56-65	40 x 2 x 12 = 960	48 x 2 x 13 = 2300
TRG	7p14	160	4-6	0	5	2	11-13	4 x 5 = 20	6 x 5 = 30

[a]dans les haplotypes avec des délétions multigéniques, le nombre de gènes IGHC fonctionnels est de 5 (délétions I, III et V), 6 (délétions IV et VI), ou 8 (délétion II) par génome haploïde. Dans les haplotypes contenant des duplications ou triplications de multigènes, le nombre exact de gènes fonctionnels IGHC par génome haploïde n'est pas connu.
[b]dans un haplotype rare, le nombre d'IGKV varie de 17 à 19.
[c]inclut les 5 TRAV/DV. Le nombre de TRAV et TRDV fonctionnels dans le locus varie de 46 à 48. Le nombre total de gènes fonctionnels contenu dans le locus TRA/TRD varie de 105 à 107.
[d]taille du cluster depuis le TRDV2 au TRDV3.
[e]distance comprise entre le gène TRA/DV le plus en 5' (TRAV14/DV4) et le gène le plus en 3' du locus TRD (TRDV3).
[f]le nombre prend en compte les réarrangements avec 2 ou 3 TRDD (4 combinaisons sont possibles: D1, D2; D1, D3; D2, D3; D1, D2, D3).

1.3.2 Locus des IG et TR murins

1.3.2.1 Locus IGH

Le locus IGH murin (Mus musculus) est localisé sur le chromosome 6, à la bande 12F2. Le locus IGH (Figure 1.13) comprend de 181 à 185 gènes localisés sur une distance de plus de 2300 kb, dont 152 gènes IGHV [53, 119-129] qui appartiennent à 15 sous-groupes, 17 à 20 gènes IGHD [130-132], 4 gènes IGHJ [133-135] et 8 ou 9 gènes IGHC [136-138]. Ces variations dépendent de la souche de souris. Le nombre de gènes fonctionnels du locus IGH murin par génome haploïde est de 119 à 124, dont 97 gènes IGHV qui appartiennent à 15 sous-groupes, 10 à 14 gènes IGHD, 4 gènes IGHJ et 8 ou 9 gènes IGHC.

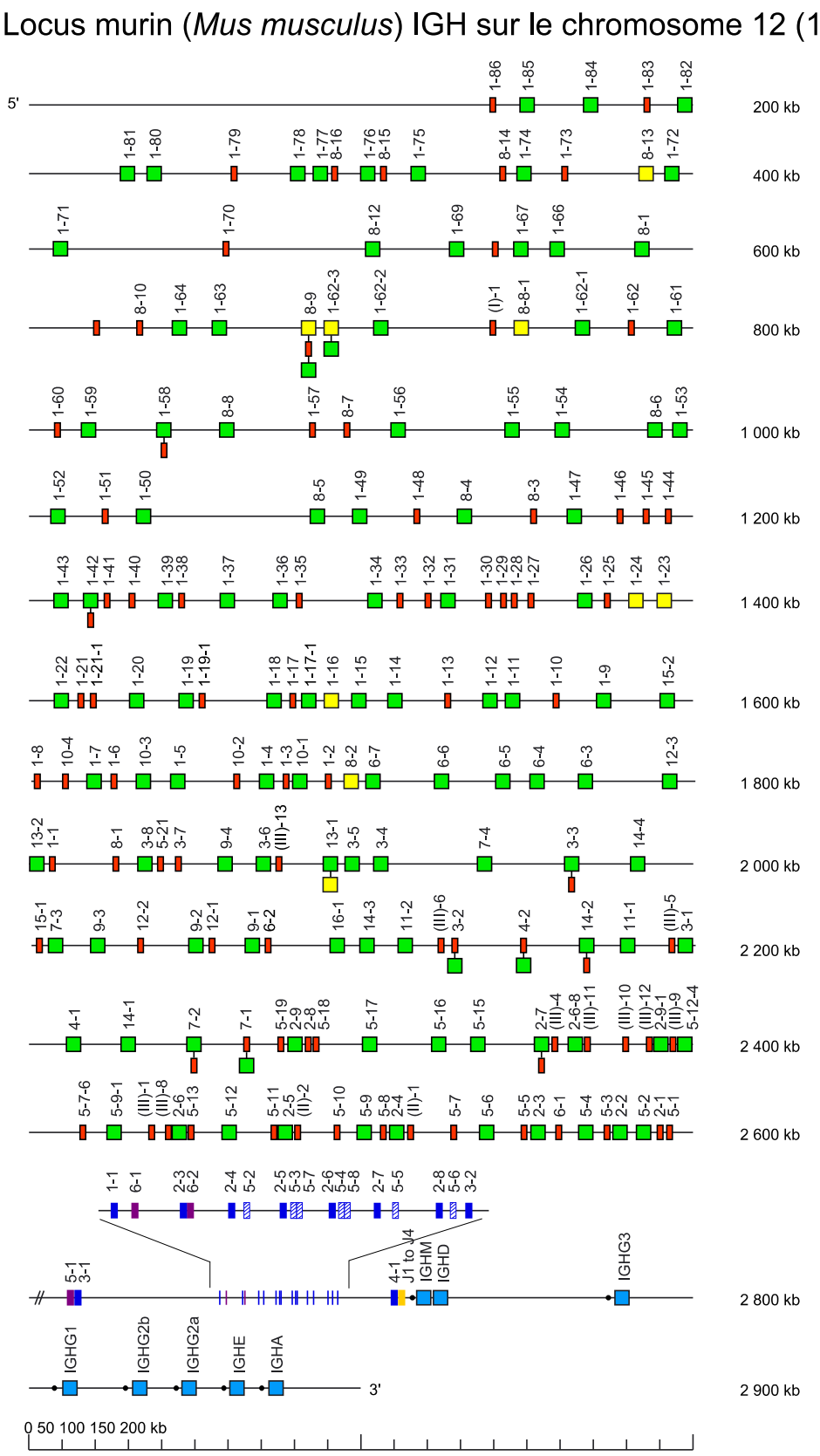


Figure 1.13. Représentation schématique du locus IGH chez la souris. Le locus IGH comprend de 181 à 185 gènes dont 152 sont des IGHV (97 fonctionnels), 17 à 20 IGHD (10 à 14 fonctionnels), de 4 IGHJ (4 fonctionnels) et de 8 ou 9 IGHC (tous fonctionnels) (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.2.2 Locus IGK

Le locus IGK murin est localisé sur le chromosome 6, à la bande 6C2. Le locus IGK (Figure 1.14) comprend 180 gènes localisés sur une distance de 3200 kb, dont 174 gènes IGKV [139-143] qui appartiennent à 19 sous-groupes et 3 clans, 5 gènes IGKJ [144-146] et un unique gène IGKC [145]. Le nombre de gènes fonctionnels du locus IGK murin par génome haploïde est de 99 à 101 dont 94 à 96 IGKV, 4 IGKJ et un unique gène IGKC.

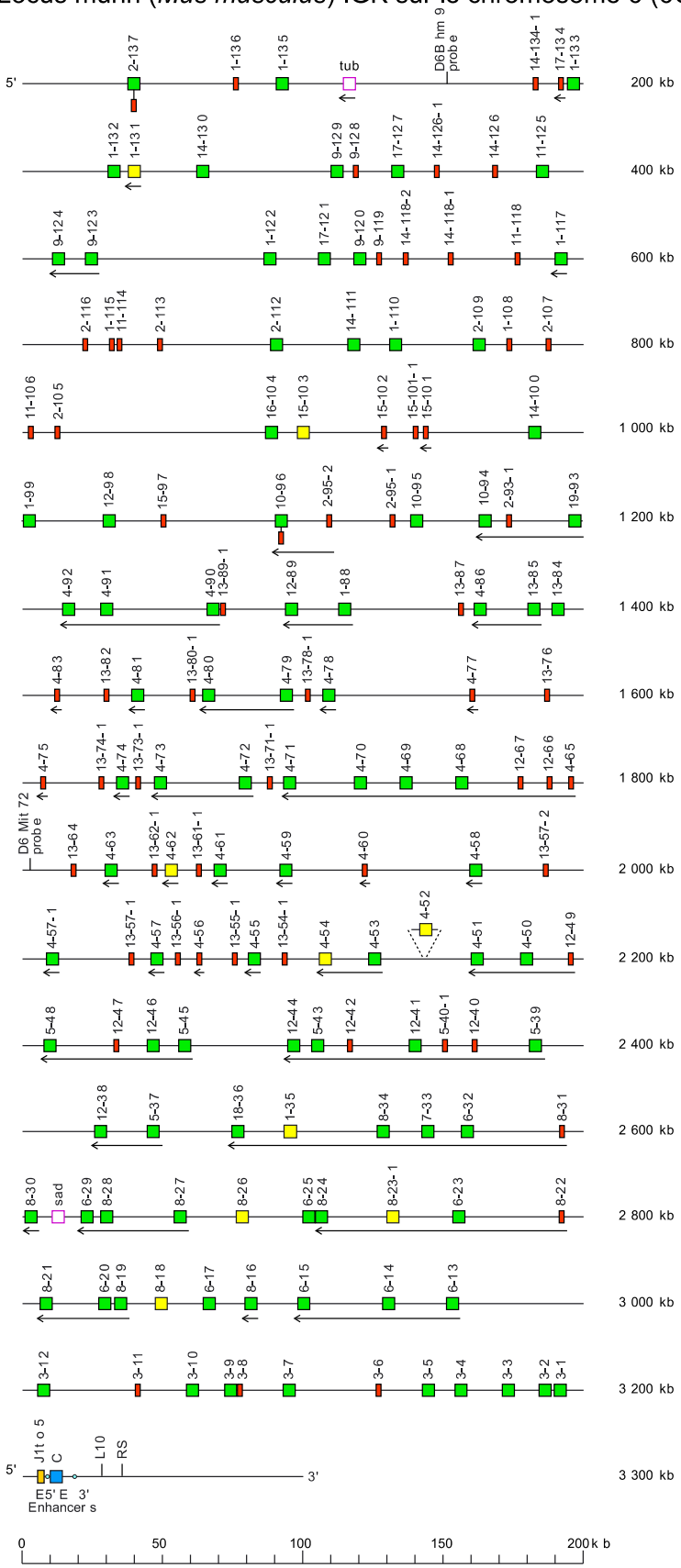


Figure 1.14. Représentation schématique du locus IGK chez la souris. Le locus IGK murin comprend 180 gènes dont 174 gènes IGKV, 5 IGKJ et un unique IGKC. Parmi ceux-ci 94 à 96 IGKV, 4 IGKJ et 1 IGKC sont fonctionnels (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.2.3 Locus IGL

Le locus IGL murin est localisé sur le chromosome 16, à la bande 16B1. Le locus IGL (Figure 1.15) comprend 12 gènes localisés sur une distance de 240 kb, 3 IGLV chez la souris de laboratoire et 8 au moins chez la souris sauvage (selon les espèces) [147-150] qui appartiennent à 2 sous-groupes, 5 IGLJ [151-152] et 4 IGLC [149, 152-154]. Chaque gène IGLC est précédé d'un ou deux gènes IGLJ. Les gènes IGLJ et IGLC sont organisés en deux cassettes: J2-C2-J4-C4 et J3-J3P-C3-J1-C1 précédés respectivement de 2 et d'un unique gène V. Le nombre de gènes fonctionnels du locus IGL par génome haploïde est de 8 ou 9 dont 3 IGLV, 3 IGLJ et 2 ou 3 IGLC.

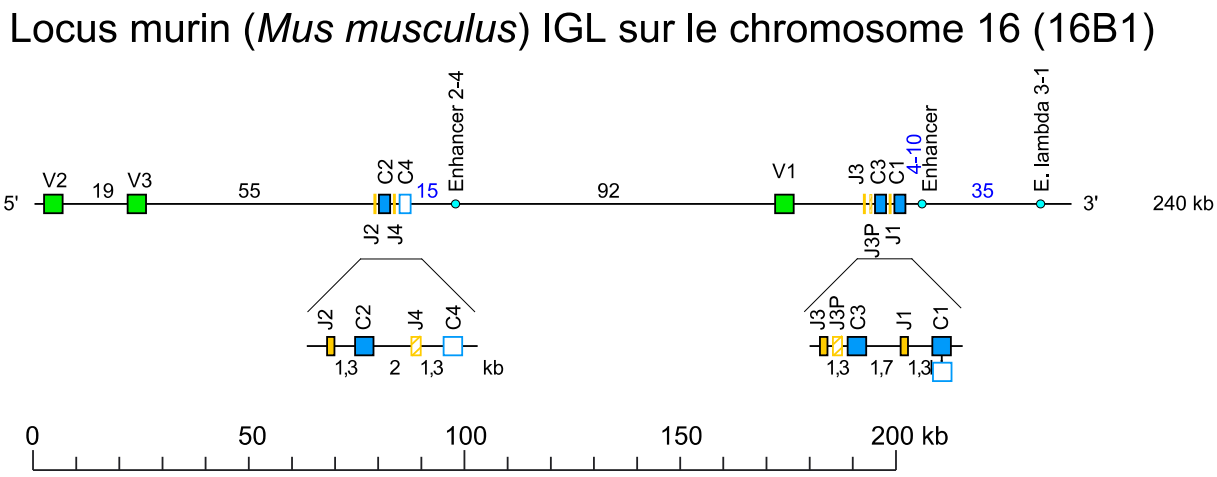


Figure 1.15. Représentation schématique du locus IGL chez la souris. Le locus IGL comprend, respectivement, 3 et 8 gènes (au moins) IGLV fonctionnels chez la souris de laboratoire et la souris sauvage, respectivement, 5 IGLJ et 4 IGLC (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.2.4 Locus TRA/TRD

Comme chez l'homme, les locus TRA et TRD ont une même localisation chromosomique. Ils sont localisés sur le chromosome 14, à la bande 14C1, le locus TRD est logé dans le locus TRA murin. Le locus TRA (Figure 1.16) comprend 159 gènes localisés sur une distance de 1650 kb, dont 98 gènes TRAV (si l'on inclut les 10 TRAV/DV) [155] Les gènes TRAV sont organisés en 2 clusters en amont des 60 gènes TRAJ [156-157] et de l'unique gène TRAC. Le nombre de gènes fonctionnels du locus TRA murin par génome haploïde est de 112 à 123, dont 73 à 84 TRAV (si l'on inclut les 10 TRAV/DV), 38 TRAJ et 1 gène TRAC.

Le locus TRD comprend 21 gènes localisés sur distance de 275 kb: 6 TRDV, 10 TRAV/DV [158-159], 2 TRDD [160], 2 TRDJ [160-161] et un unique TRDC [162]. Le gène TRDV5 est situé en 3' du gène TRDG et il est en orientation inverse de transcription par rapport au reste des gènes (TRDJ et TRDC). Le nombre de gènes fonctionnels du locus TRD par génome haploïde est de 20 gènes dont 10 gènes TRAV/DV, 5 TRDV, 2 TRDD, 2 TRDJ et un unique gène TRDC.

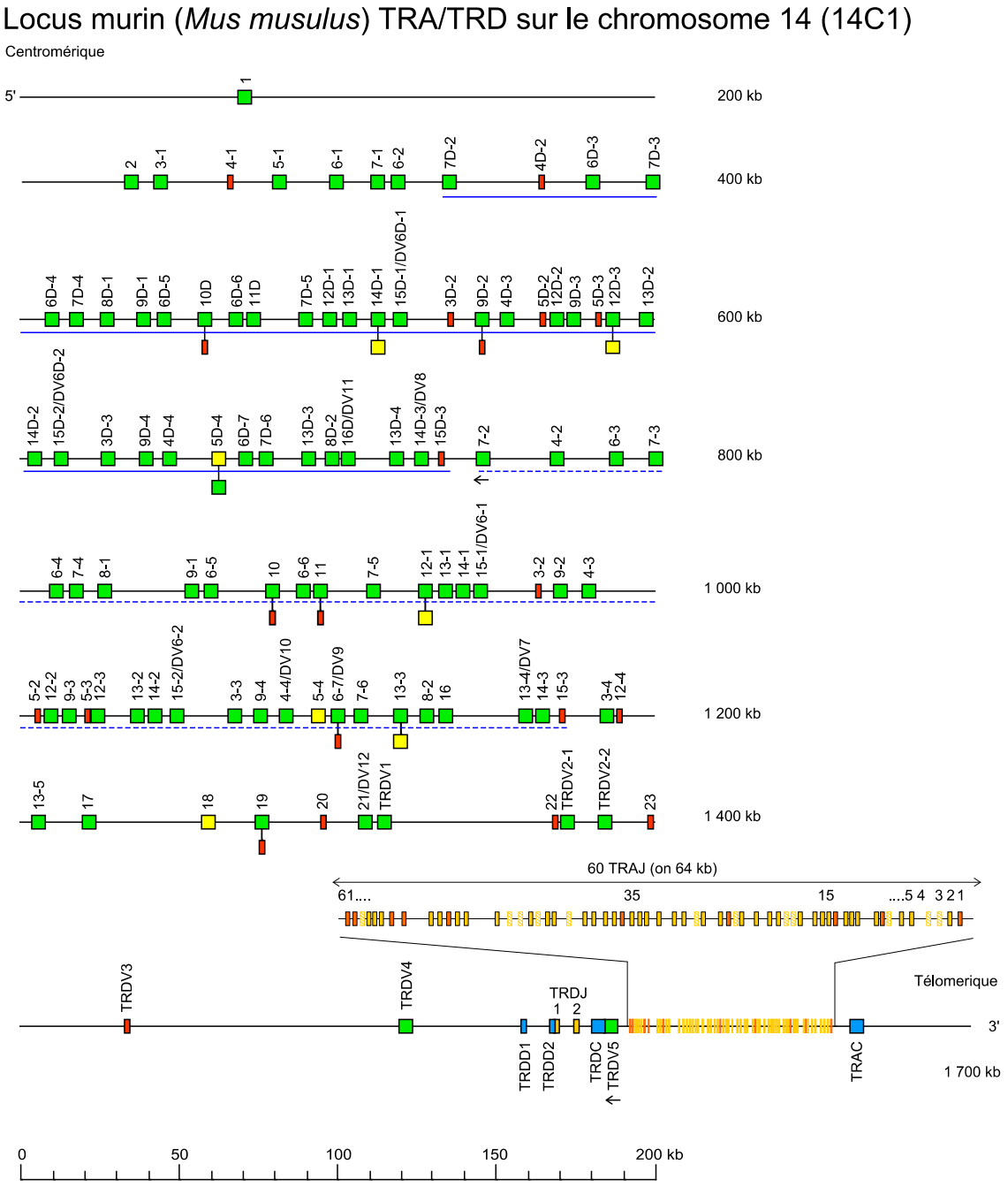


Figure 1.16. Représentation schématique du locus TRA/TRD chez la souris. Le locus TRA comprend 98 gènes TRAV, 60 gènes TRAJ et 1 gène TRAC. Parmi ceux-ci, 73 à 84 TRAV, 38 TRAJ et 1 TRAC sont fonctionnels. Le locus TRD comprend 6 gènes TRDV, 2 TRDD, 2 TRDJ et 1 TRDC. Parmi ceux-ci, 5 TRAV, tous les TRDD, TRDJ et TRDC sont fonctionnels (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.2.5 Locus TRB

Le locus TRB murin est localisé sur le chromosome 6, à la bande 6B2. Le locus TRB (Figure 1.17) comprend 53 gènes localisés sur une distance de 700 kb, dont 35 gènes TRBV [155, 163-165] qui appartiennent à 31 sous-groupes. Le gène TRBV31 en 3' du gène TRBC et est en orientation inverse de transcription par rapport aux gènes TRBJ et TRBC et réarrange par un mécanisme d'inversion. Les autres TRBV sont situés en amont d'un cluster dupliqué constitué d'un TRBD [163], de 7 TRBJ [163] et d'un unique TRBC. Le nombre de gènes fonctionnels du locus TRB par génome haploïde est de 36 ou 37 dont 21 ou 22 TRBV, 2 TRBD, 11 TRBJ et 2 TRBC.

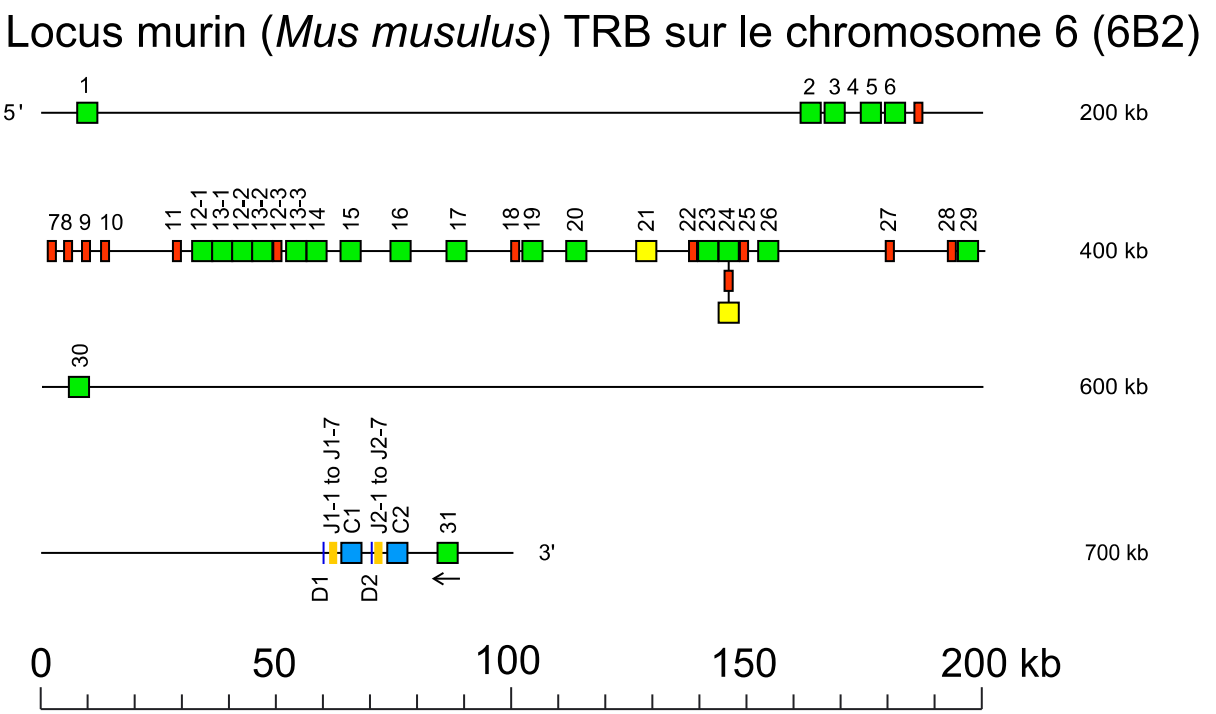


Figure 1.17. Représentation schématique du locus TRB chez la souris. Le locus TRB comprend 35 gènes TRBV (21 ou 22 fonctionnels), 2 TRBD (tous fonctionnels), 14 TRBJ (11 fonctionnels) et 2 TRBC (tous fonctionnels) (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

1.3.2.6 Locus TRG

Le locus TRG murin est localisé sur le chromosome 13, à la bande 13A3.1. Le locus TRG (Figure 1.18) comprend 15 gènes qui s'étendent sur 205 kb: y sont représentés 7 TRGV [166-170] appartenant à 5 sous groupes, 4 gènes TRGJ [166, 169-170] et 4 TRGC [166, 169-170]. Le locus TRG est organisé en 4 clusters. Le plus en amont est constitué de 4 TRGV, d'un TRGJ et d'un TRGC alors que les suivants sont constitués d'un unique TRGV, TRGJ et TRGC. L'avant dernier cluster est dans une orientation inversée. Le nombre de gènes

fonctionnels du locus TRG par génome haploïde est de 14 dont 7 TRGV, 4 TRGJ et 3 ou 4 TRGC.

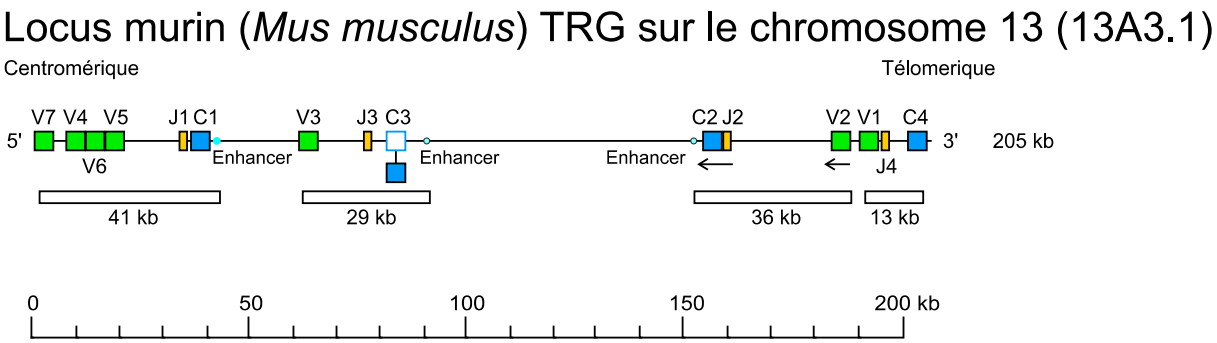


Figure 1.18. Représentation schématique du locus TRG chez la souris. Le locus TRG comprend 7 gènes TRGV (tous fonctionnels), 4 TRGJ (tous fonctionnels) et 4 TRGC (3 ou 4 fonctionnels) (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org).

Tableau 1.3. Nombre total de gènes d'IG et TR par génome haploïde chez la souris (avec la permission de M.-P. et G. Lefranc, IMGT®, http://www.imgt.org/)

Locus	Localisation chromosomique	Type de gènes				Total (V+D+J+C)	Nombre d'orphons	Total (V+D+J+C+orphons)
		V	D	J	C			
IGH	12F2	152	4	17-20	8-9	181-185	0	181-185
IGK	6C2	174	0	5	1	180	8	188
IGL	16B1	3	0	5	4	12	0	12
		8[a]				17[a]	0	17[a]
TRA	14C1	98[b]	0	60	1	159[b]	0	159[b]
TRD	14C1	6 (16[b])	2	2	1	11 (21[b])	0	11 (21[b])
TRB	6B2	35	2	14	2	53	0	53
TRG	13A3.1	7	0	4	4	15	0	15

[a]souches sauvages (*Mus musculus*, MBK, PWK, MAI).
[b]inclut les 10 TRAV/DV.

Tableau 1.4. Nombre de gènes d'IG et TR fonctionnels par génome haploïde chez la souris (avec la permission de M.-P. et G. Lefranc, IMGT ®, http://www.imgt.org).

Locus	Localisation chromosomique	Taille des locus (kb)	Type de gènes				Nombre de gènes fonctionnels	Diversité combinatoire	
			V	D	J	C		Minimum	Maximum
IGH	12F2	2300	97	10-14	4	8-9	119-124	97 x 10 x 4 = 3880	97 x 14 x 4 = 5432
IGK	6C2	3200	94-96	0	4	1	99-101	94 x 4 = 376	96 x 4 = 384
IGL	16B1	240	3	0	3	2-3	8-9	3 x 3 = 9	
			8[a]				13-14[a]	8 x 3 = 24	
TRA	14C1	1650	73-84[b]	0	38	1	112-123[b]	73 x 38 = 2774	84 x 38 = 3192
TRD	14C1	275	5 (15[b])	2	2	1	10 (20[b])	15 x 2 x 2 = 60	
TRB	6B2	700	21-22	2	11	2	36-37	21 x 2 x 11 = 462	22 x 2 x 11 = 484
TRG	13A3.1	205	7	0	4	3-4	14	7 x 4 = 28	

[a]souris sauvage.
[b]inclut les 10 TRAV/DV.

CHAPITRE 2
Une ontologie pour les récepteurs d'antigènes: IMGT-ONTOLOGY

Une ontologie est une description concise et non ambiguë de concepts les plus significatifs dans un domaine d'application. IMGT-ONTOLOGY est la première en immunogénétique. Elle permet la gestion des connaissances dans ce domaine pour toutes les espèces de vertébrés. IMGT-ONTOLOGY comprend un vocabulaire contrôlé et des règles d'annotation qui sont indispensables pour assurer la pertinence des annotations et la cohérence des composants dans IMGT®. De plus, cette ontologie permet aux scientifiques et aux cliniciens d'utiliser les mêmes termes avec les mêmes significations. Elle fournit un creuset sémantique qui peut être introduit dans les ontologies plus générales de la biologie moléculaire, et apparaît comme une aide fondamentale pour améliorer l'interopérabilité des bases de données spécialisées et généralistes. Sept axiomes principaux sont définis: IDENTIFICATION, CLASSIFICATION, DESCRIPTION, NUMEROTATION, LOCALIZATION, ORIENTATION et OBTENTION. Ces 7 axiomes postulent que les objets, les processus et les relations doivent être identifiés, classés, décrits, numérotés, localisés, orientés, et que la façon dont ils sont obtenus doit être déterminée. Ces 7 axiomes constituent l'IMGT-ONTOLOGY formalisée, aussi désignée par IMGT-Kaleïdoscope [15].

L'axiome OBTENTION regroupe les concepts qui permettent de sélectionner les séquences selon leur origine (méthodologie utilisée) ou les pathologies auxquelles elles sont associées. L'axiome de LOCALIZATION permet de caractériser les positions des gènes d'IG et TR. L'axiome ORIENTATION concerne les brins inversés ou directs de l'ADN. Dans ce chapitre, nous verrons de façon plus détaillée les axiomes d'IDENTIFICATION, de DESCRIPTION, de CLASSIFICATION et de NUMEROTATION.

2.1 Axiome IDENTIFICATION et concepts d'identification

L'axiome IDENTIFICATION postule que les molécules, cellules, tissus, organes, organismes ou populations, leurs processus et leurs relations, doivent être identifiés. L'axiome IDENTIFICATION contient les concepts d'identification qui fournissent les termes et règles pour identifier une entité, ses processus et ses relations. En biologie moléculaire, les concepts

d'identification permettent d'identifier les molécules, leurs processus et leurs relations au niveau du génome, transcriptome et protéome.

2.1.1 Identification de l'organisme: le concept « Taxon »

Le concept « Taxon » permet d'identifier le type de taxon, dans lequel un objet, un processus ou une relation est trouvé(e). Le concept « Taxon » gère une hiérarchie de concepts à différents niveaux de granularité. La taxinomie hiérarchique correspondante est celle fournie par le National Center for Biotechnology Information (NCBI http://www.ncbi.nlm.nih.gov/) au rang d'espèces (concept « Species ») et sous-espèces (concept « Subspecies ») afin d'établir une interopérabilité complète avec les bases de données généralistes. Comme les gènes des IG, TR et du CMH ne sont présents que chez les vertébrés à mâchoire (Gnathostomes), seules des espèces de vertébrés ont été initialement représentées dans IMGT-ONTOLOGY. Toutefois, avec l'extension de IMGT-ONTOLOGY au IgSF et MhcSF, les espèces d'invertébrés sont incorporées chaque fois que nécessaire. Les concepts « EthnicGroup », « Breed » et « Strain » ont été ajoutés à IMGT-ONTOLOGY pour permettre l'identification des données spécifiques à des groupes ethniques pour les humains (http://www.ebi.ac.uk/imgt/hla/help/ethnic_help.html), à des races pour les animaux domestiques, à des souches de laboratoire [171] et à des animaux sauvages.

2.1.2 Identification d'une entité: le concept « EntityType »

Le concept « EntityType » identifie le type d'entité. Une entité peut être une molécule, une cellule, un tissu, un organe, un organisme ou une population. Si l'objet est une molécule, le concept « EntityType » est désigné comme « Molecule_EntityType », qui est défini par les concepts d'identification « MoleculeType », « GeneType » et « ConfigurationType » et possède des propriétés identifiées dans les concepts « Functionality » et « StructureType » Figure 2.1.

2.1.2.1 Le concept « MoleculeType »

Le concept « MoleculeType » identifie le type de molécule basé sur le type des éléments constitutifs et sur les concepts d'obtention (non détaillés ici). Les quatre principales instances du concept « MoleculeType » sont 'gDNA' (un ADN génomique ou ADNg est une séquence de nucléotides A, T, C, G, obtenue à partir d'un génome), 'mRNA' (un ARN

messager, ARNm ou transcript, est une séquence de nucléotides A, U, C, G, obtenus par la transcription d'un ADN génomique), 'cDNA' (un ADN complémentaire ou ADNc, est une séquence de nucléotides A, T, C, G, obtenue in vitro par transcription inverse de l'ARN messager) et 'protein' (une séquence d'acides aminés obtenue par traduction du transcript). Ainsi, les instances du concept « MoleculeType » permettent d'identifier une séquence de nucléotides qui peut être soit génomique ('gDNA'), soit un transcript ('mRNA', 'cDNA'), et une séquence d'acides aminés ('protein').

2.1.2.2 Le concept « GeneType »

Le concept « GeneType » identifie le type de gène et comprend cinq instances (Figure 2.1). La première instance, 'conventional', se réfère à toute séquence nucléotidique autre que celles des IG ou TR. Les quatre autres instances sont spécifiques à l'immunogénétique: les types de gènes 'variable' (V), 'diversity' (D) et 'joining' (J) qui réarrangent au niveau de l'ADN et codent les domaines variables des IG et TR, et le type de gène 'constant' (C) qui code la région constante des IG et TR [19-20].

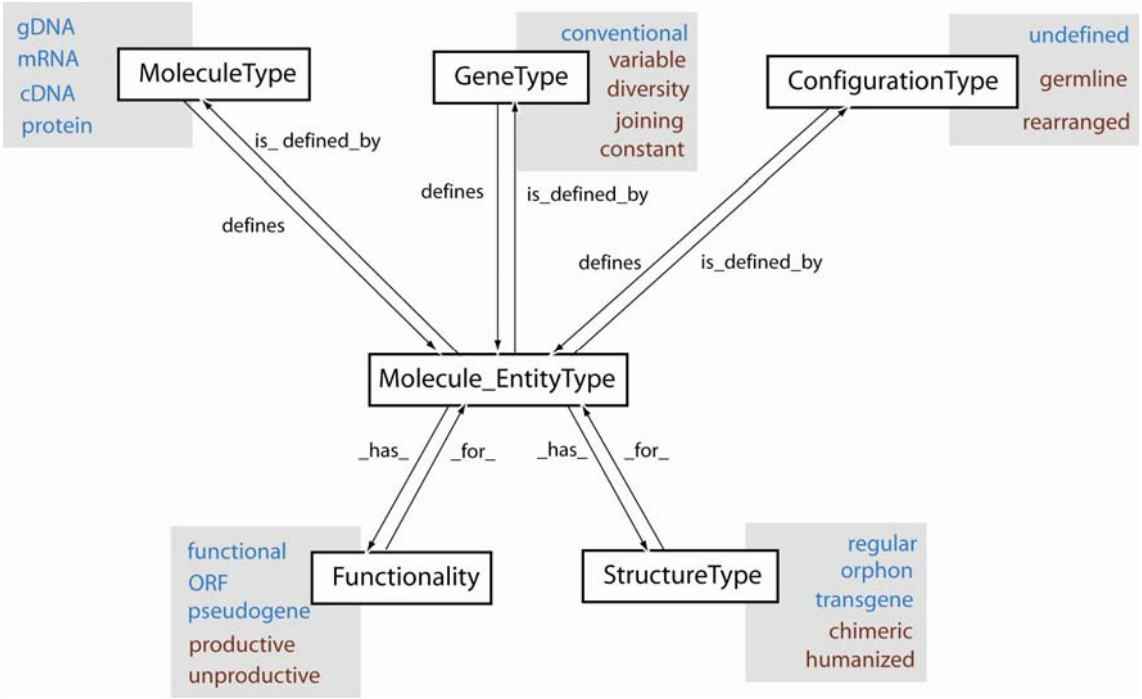


Figure 2.1. Le concept « Molecule_EntityType ». Le concept « Molecule_EntityType » est défini par les concepts d'identification « MoleculeType », « GeneType » et « ConfigurationType » et possède des propriétés définies dans les concepts « Functionality » et « StructureType » (axiome IDENTIFICATION) [15]. Les flèches indiquent les relations réciproques 'is_defined_by' et 'defines', '_has_' et '_for_'. Les instances des concepts qui sont d'ordres généraux sont en bleu, celles qui sont spécifiques des IG et TR sont en rouge. Le concept « Molecule_EntityType » a 19 instances. Seuls quelques exemples des instances du concept « StructureType » sont présentés.

2.1.2.3 Le concept « ConfigurationType »

Le concept « ConfigurationType » identifie le type de configuration du gène et comprend trois instances (Figure 2.1). L'instance 'undefined' identifie la configuration des gènes conventionnels et constants (C). Les instances 'germline' et 'rearranged' identifient le statut des gènes V, D et J, avant et après les réarrangements d'ADN, respectivement [19-20].

2.1.2.4 Le concept « Molecule_EntityType »

Le concept « Molecule_EntityType », défini par les concepts « MoleculeType », « GeneType » et « ConfigurationType », comprend 19 instances. Trois instances, 'gene', 'nt-sequence' et 'AA-sequence', respectivement, identifient les ADNg, ARNm et protéines (« MoleculeType ») d'un gène conventionnel (« GeneType ») en configuration 'undefined' (« ConfigurationType »). L'instance nt-sequence est également valable pour l'ADNc. Seize instances permettent d'identifier les IG et TR. Dix d'entre elles sont représentées dans la Figure 2.1: six pour l'ADNg ('V-gene', 'D-gene', 'J-gene', 'C-gene', 'V-D-J-gene' et 'V-J-gene'), deux pour l'ARNm, 'L-V-D-J-C-sequence' et 'L-V-J-C-sequence', valables également pour l'ADNc, et deux pour la protéine, 'V-D-J-C-sequence' et 'V-J-C-sequence'. Par exemple, l'instance 'V-gene' désigne un 'gDNA' (« MoleculeType ») contenant un gène V (« GeneType »), en configuration 'germline' (« ConfigurationType »). L'instance 'L-V-J-C-sequence' désigne une séquence d'ARNm ou d'ADNc (« MoleculeType ») correspondant aux gènes V, J et C (« GeneType »), en configuration réarrangée (« ConfigurationType ») (Figure 2.1). Les six dernières instances correspondent à des cas de réarrangements partiels ('D-J-gene') ou à des transcriptions stériles ('L-V-sequence', 'D-sequence', 'J-sequence', 'J-C-sequence' et 'C-sequence').

2.1.2.5 Le concept « Functionality »

Le concept « Functionality » identifie le type de fonctionnalité pour le concept « Molecule_EntityType » (Figure 2.1). Il comprend cinq instances divisées en deux catégories selon le type de configuration. Trois instances, 'functional', 'ORF' (cadre de lecture ouvert) et 'pseudogene' identifient la fonctionnalité d'une instance du concept « Molecule_EntityType » dans la configuration 'undefined' ou 'germline'. Ils permettent d'identifier la fonctionnalité des gènes conventionnels, tels que les gènes C, et les gènes V, D et J, avant leur réarrangement dans le génome, et, par extension, la fonctionnalité de leurs transcripts et protéines. Les deux instances 'productive' et 'unproductive' identifient la fonctionnalité des instances du concept « Molecule_EntityType » en configuration réarrangée. Ils permettent

d'identifier la fonctionnalité des entités d'IG et TR après leur réarrangement dans le génome, celle de gènes fusionnés par translocations, et encore celle de gènes hybrides obtenus par des techniques de biotechnologie moléculaire, et, par extension, la fonctionnalité de leurs transcripts et de leurs protéines.

2.1.3 Identification d'un récepteur: le concept « ReceptorType »

Le concept « ReceptorType » identifie le type de récepteur. Un récepteur peut être une molécule, une cellule, un tissu, un organe, un organisme ou une population. Si l'objet est une molécule, le concept « ReceptorType » est désigné comme « Molecule_ReceptorType » qui est défini par le concept d'identification « ChainType » et possède des propriétés recensées dans les concepts « StructureType », « Specificity » et « Function » (Figure 2.2). Le concept « ChainType » est lui-même défini par les concepts d'identification « Molecule_EntityType » et le « DomainType » et par les concepts de classification (voir axiome CLASSIFICATION). Ces derniers sont organisés en une hiérarchie qui confère différents niveaux de granularité pour les concepts « Molecule_ReceptorType » et « ChainType ».

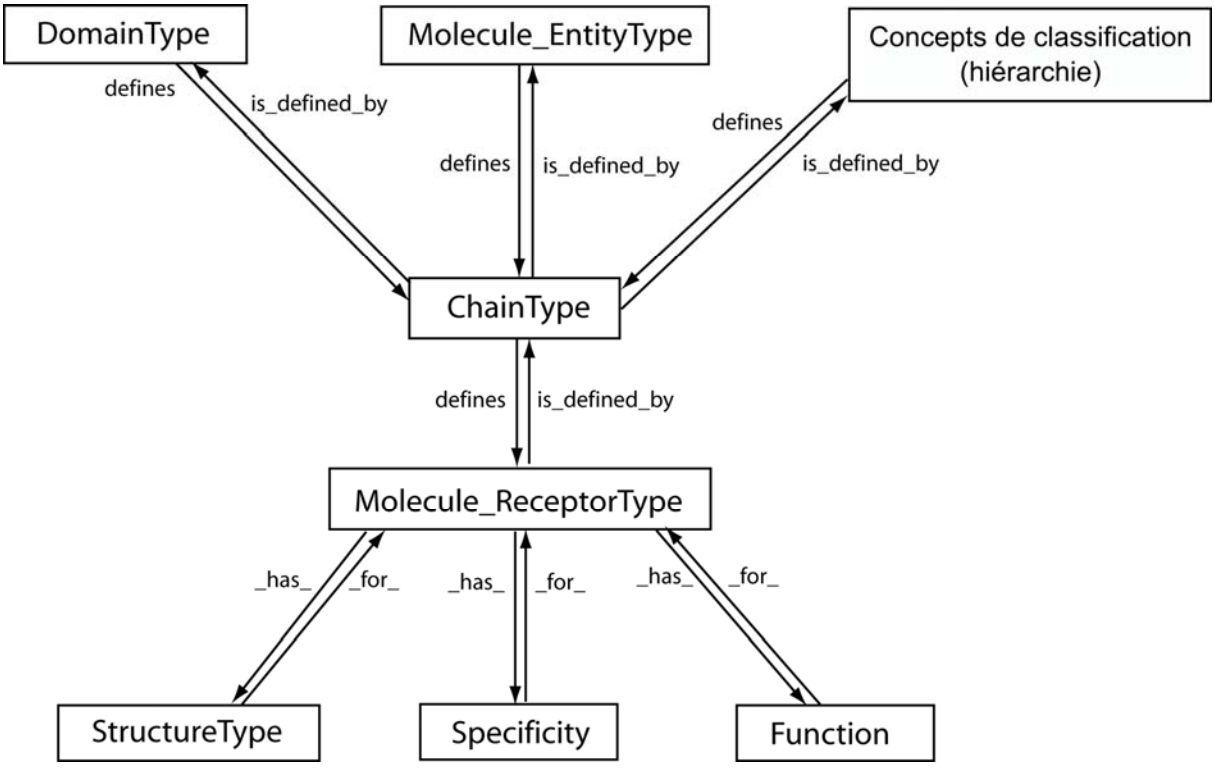


Figure 2.2. Le concept « Molecule_ReceptorType ». Le concept « Molecule_ReceptorType », défini par le concept « ChainType » de l'identification, a des propriétés recensées dans les concepts « StructureType », « Specificity » et « Function » (axiome IDENTIFICATION). Le concept « ChainType » est lui-même défini par les concepts « Molecule_EntityType » et « DomainType » et par les concepts de classification (hiérarchie). Les flèches indiquent les relations réciproques 'is_defined_by' et 'defines', '_has_' et '_for_'. Ces concepts ont différents niveaux de granularité, jusqu'à six pour « Molecule_ReceptorType » et « ChainType ».

2.1.3.1 Le concept « Molecule_ReceptorType »

Le concept « Molecule_ReceptorType » identifie le type de récepteur de protéine, défini par sa composition en chaîne. Ainsi, 'IG' est une instance du concept « Molecule_ReceptorType », définie comme comprenant 4 chaînes, deux chaînes lourdes et deux chaînes légères, identiques deux par deux et liées de façon covalente (

Figure 2.3). Un récepteur peut comprendre une chaîne (monomère) ou plusieurs chaînes associées (polymères).

2.1.3.2 Le concept « ChainType »

Le concept « ChainType » identifie le type de chaîne (Figure 2.2). Il est l'un des concepts les plus importants de l'identification pour la normalisation des données du génome, transcriptome et protéome en biologie des systèmes. En effet, être en mesure d'identifier un type de chaîne signifie qu'il est possible d'identifier la transcription et le gène codant. Le concept « ChainType » contient une hiérarchie de concepts qui permettent d'identifier le type de chaîne à différents niveaux de granularité. Le plus fin niveau de granularité, le concept « GeneLevelChainType », identifie le type de chaîne en fonction du (des) gène(s) qui code(nt) la chaîne. Il représente le principal concept d'une identification très précise car il établit une relation avec le concept « Gene » qui appartient aux concepts de classification (relations réciproques 'is_coded_by' et 'codes'). Le nombre d'instances du concept « GeneLevelChainType » dépend du nombre de gènes fonctionnels et ORF par génome haploïde étant donnée une espèce (dans le cas d'une IG et TR, c'est le nombre de gènes constants fonctionnels et ORF qui est pris en compte). Si seulement les gènes fonctionnels sont considérés, les instances de ce concept correspondent aux isotypes.

2.1.3.3 Le concept « DomainType »

Une chaîne peut être définie par ses unités structurales constitutives (le concept « DomainType ») (Figure 2.2). Un domaine est une sous-unité de la chaîne caractérisée par sa structure en trois dimensions (3D), et par extension sa séquence en acides aminés et la séquence nucléotidique qui code pour elle. Ce concept peut théoriquement comporter de nombreuses instances, soigneusement caractérisées par LIGM, qui ont été inscrites dans IMGT-ONTOLOGY. Le concept « DomainType » a actuellement trois instances: le type de domaine V (domaines variables des IG et TR et domaines V-like d'autres protéines IgSF); le

type de domaine C (domaines constants de l'IG et TR et domaines C-like d'autres protéines IgSF) et le type de domaine G (domaines groove des chaînes des molécules du CMH et les domaines G-like d'autres protéines MhcSF) [14-16].

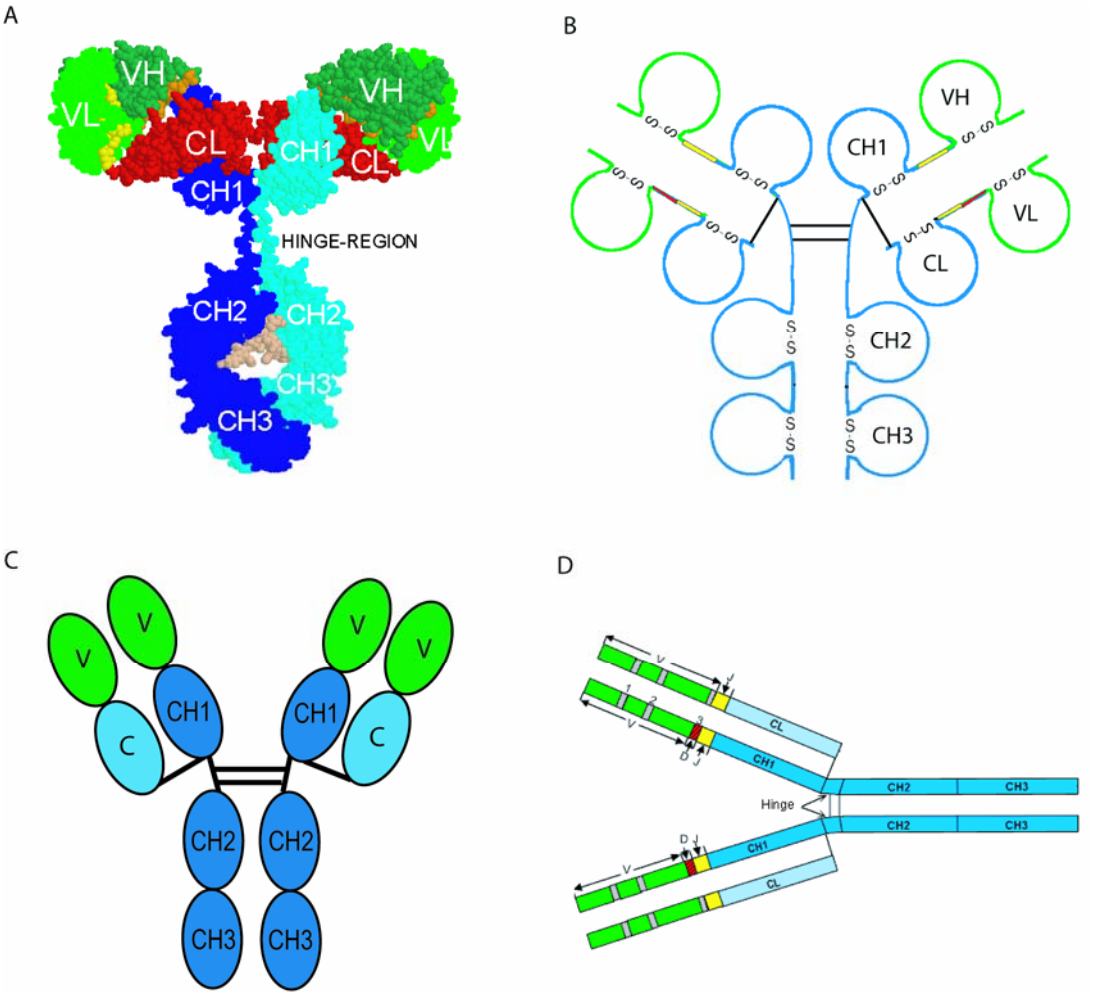


Figure 2.3. Identification d'une IG ou d'une instance d'anticorps comme du concept « Molecule_ReceptorType » composé de quatre chaînes, deux IG-Heavy-Chain et deux IG-Light-Chain (concept « ChainType »). Les quatre représentations, bien que différentes, permettent d'identifier une IG en tant que récepteur formé de quatre chaînes, qui sont elles-mêmes organisées en domaines (concept « DomainType »). VH et VL sont des domaines de type V, codées pour la région V-D-J et V-J, respectivement. CL, CH1, CH2 et CH3 sont des domaines de type C. (A) structure 3D, (B) organisation en domaines Ig-like, (C) organisation en modules, (D) les régions codées par les types de gènes V, D, J et C. Le gène de type C code pour la région constante (CL pour l'IG-Light-Chain et CH1, CH2 et CH3 de l'IG-Heavy-Chain). Cette représentation, schématisée sous forme de Y, est fréquemment utilisée pour représenter une IG.

2.1.3.4 Les concepts « Specificity » et « Function »

Le concept « Specificity » identifie la spécificité de « Molecule_ReceptorType » (Figure 2.2), et, par extension, la spécificité des chaînes et des domaines, et des transcripts correspondants. Les instances du concept « Specificity » identifient l'antigène reconnu par un récepteur d'antigène (IG ou TR). Le concept « Specificity » est particulièrement important en raison du nombre illimité d'antigènes et de la complexité des interactions antigène/récepteur

d'antigène. Les instances du concept « Specificity » (plusieurs centaines à l'heure actuelle) seront connectées, d'une part, avec le concept « Epitope » qui identifie la partie de l'antigène reconnue par le récepteur de l'antigène et, d'autre part, avec le concept « Paratope » qui identifie la partie du récepteur de l'antigène (IG ou TR), qui reconnaît et se lie à l'antigène. Le concept « Function » identifie la fonction de « Molecule_ReceptorType » (Figure 2.2), et par extension la fonction des chaînes et des domaines, et des transcripts correspondants. Les instances du concept « Function » identifient les fonctions de reconnaissance et d'élimination des pathogènes par les récepteurs d'antigènes [2].

2.2 Axiome DESCRIPTION et concepts de description

L'axiome DESCRIPTION d'IMGT-Kaleïdoscope postule que les molécules, cellules, tissus, organes, organismes ou populations, leurs processus et relations, doivent être décrits.

2.2.1 Le concept « Molecule_EntityPrototype »

En biologie moléculaire, l'axiome DESCRIPTION a généré les concepts de description qui définissent les termes et les règles pour décrire les motifs dans les séquences de nucléotides et d'acides aminés et dans les structures 3D. Ces concepts ont donné lieu à une terminologie standardisée et à une définition précise des règles d'annotation. Les instances de concepts de description sont référencées par des labels IMGT® écrits en lettre capitales. Plus de 550 labels ont été définis (270 pour les séquences de nucléotides (http://imgt.org/cgi-bin/IMGTlect.jv?query=7) [8] et 285 pour les structures 3D [172] (http://imgt.org/textes/IMGTScientificChart/SequenceDescription/IMGT3Dlabeldef.html). Il faut noter que 64 labels IMGT® définis pour des séquences de nucléotides sont utilisés, et référencés dans Sequence Ontology (SO) (http://song.sourceforge.net/) [12, 173] pour décrire l'organisation spécifique des gènes d'IG et TR (http://www.imgt.org/textes/IMGTindex/ontology.html).

Le concept « Molecule_EntityPrototype » permet la description de l'organisation de l'entité (gène, transcript et protéine) et de ses motifs constitutifs. Ce concept est fondamental dans IMGT-ONTOLOGY, car il permet la représentation des connaissances liées à la complexité des mécanismes de réarrangements des gènes d'IG et TR (Figure 2.4). La relation 'is_rearranged_into' est spécifique à la synthèse des IG et TR. Les relations 'is_transcribed_into' et 'is_translated_into' concernent de manière générale la biologie

moléculaire. Ces trois relations permettent l'organisation des différentes instances du concept « Molecule_EntityPrototype » lors de la synthèse des IG et TR, et d'une manière plus générale l'expression d'une protéine. Elles permettent en plus, par des relations plus spécifiques, de prendre en compte les transcrits alternatifs, les isoformes de protéines et les modifications post-traductionnelles.

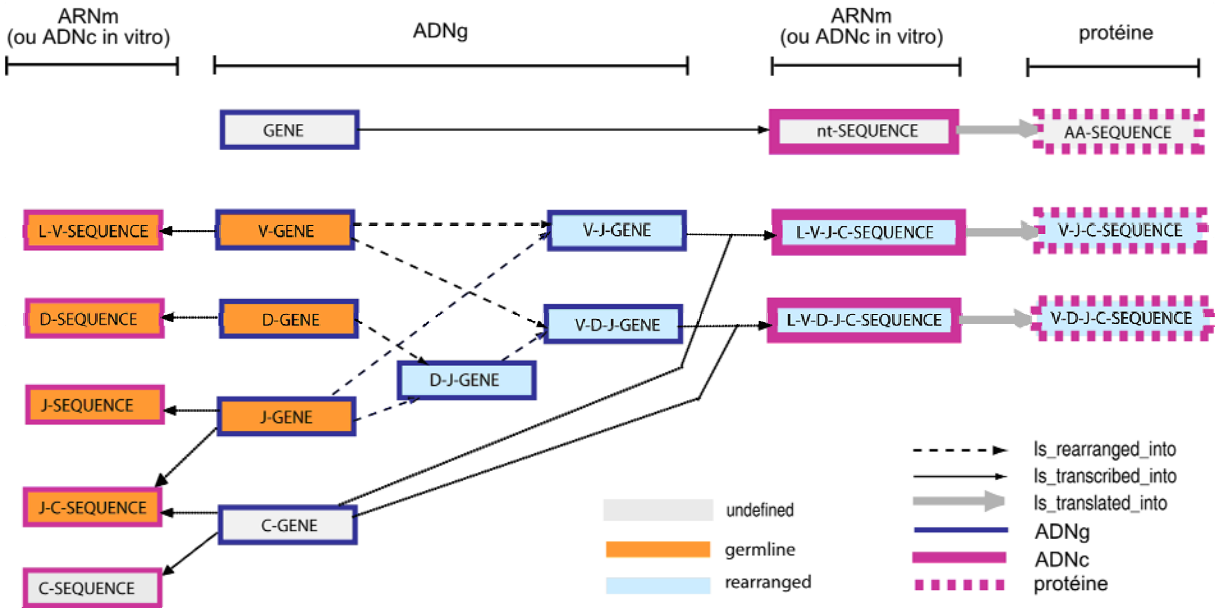


Figure 2.4. Les instances du concept « Molecule_EntityPrototype » (axiome DESCRIPTION). Les trois instances 'GENE', 'nt-SEQUENCE' et 'AA-SEQUENCE' correspondent à des gènes conventionnels tandis que les 16 autres instances sont spécifiques des IG et TR. Les instances du concept de l'ARNm sont également valables pour l'ADNc in vitro. La première colonne correspond aux instances 'sterile transcript'.

Chacune des 19 instances du concept « Molecule_EntityPrototype » peut être décrite avec ses motifs constitutifs qui appartiennent aux autres concepts de description. Ainsi, à chaque instance du concept de description « Molecule_EntityPrototype » correspond une instance du concept « GraphicalPrototype » (Figure 2.5).

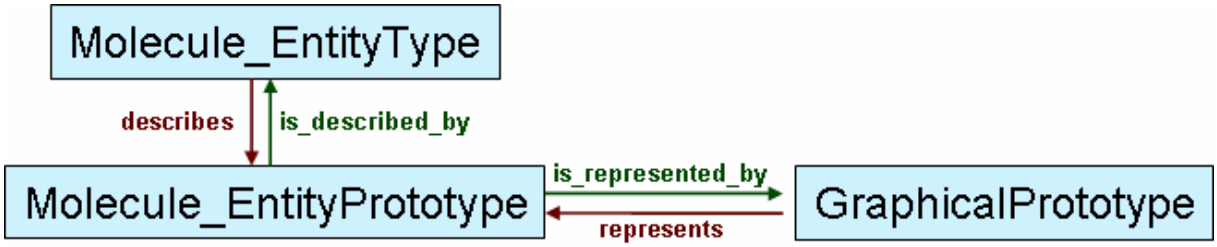


Figure 2.5. Relations entre le concept « Molecule_EntityType » et les concepts « Molecule_EntityPrototype » et « GraphicalPrototype » (axiome DESCRIPTION).

La Figure 2.6 montre la représentation graphique des instances ‘V-GENE’, ‘D-GENE’ et ‘J-GENE’ et leurs motifs constitutifs. L’instance réarrangée ‘V-D-J-GENE’ est montrée pour comparaison avec les instances ‘germline’ ‘V-GENE’, ‘D-GENE’ et ‘J-GENE’.

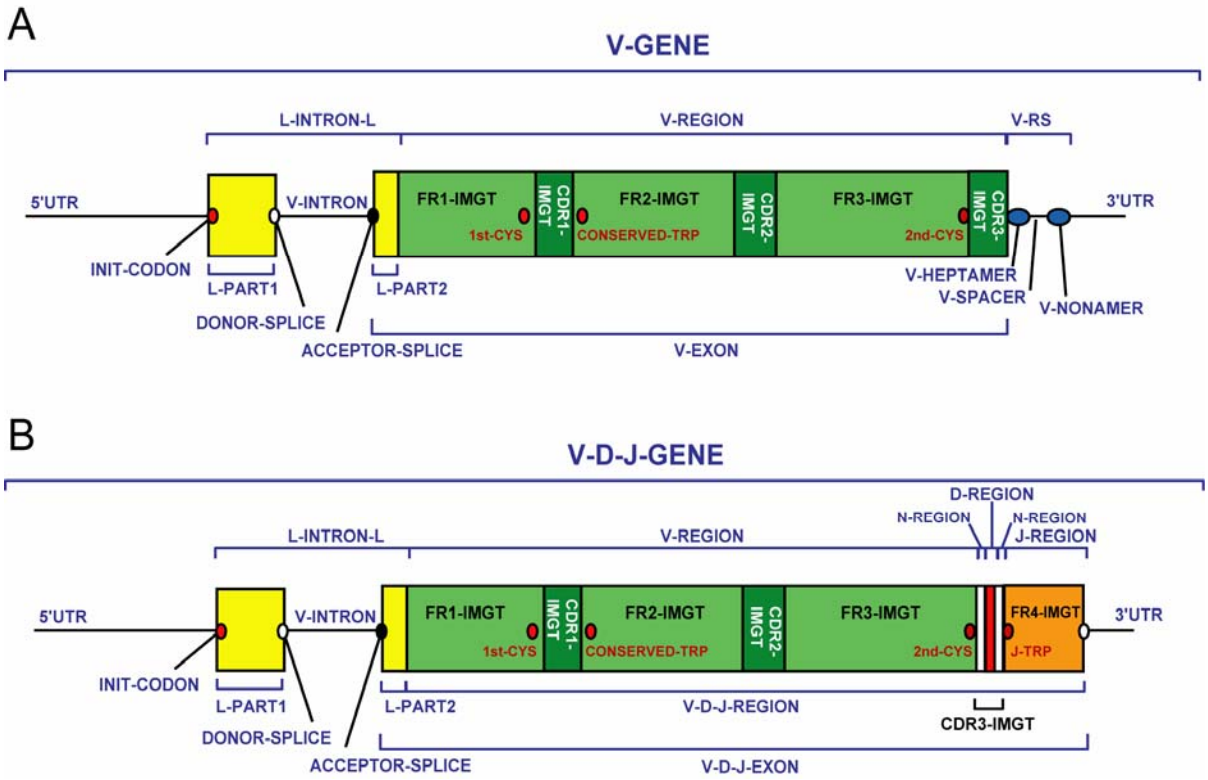


Figure 2.6. Représentation graphique des deux instances du concept ‘Molecule_EntityPrototype’ (DESCRIPTION axiome). (A) V-GENE. (B) V-D-J-GENE. Vingt-cinq labels et dix relations sont nécessaires et suffisantes pour une description complète de ces instances.

Une série de dix relations sont nécessaires et suffisantes pour une description complète des motifs d’une instance du concept « Molecule_EntityPrototype » (Figure 2.4A). Ces relations font partie des concepts de localisation (Tableau 2.1) (axiome LOCALIZATION) (IMGT Index, http://www.imgt.org). Deux relations additionnelles ‘is_in_5_prime_of’ et ‘is_in_3_prime_of’ (Figure 2.4B).

Tableau 2.1. Relations entre les labels d’IMGT-ONTOLOGY.

	Relation	Relation réciproque
A	‘adjacent_at_its_5_prime_to’ ‘included_with_same_5_prime_in’ ‘included_with_same_3_prime_in’ ‘overlaps_at_its_5_prime_with’ ‘included_in’	‘adjacent_at_its_3_prime_to’ ‘includes_with_same_5_prime’ ‘includes_with_same_3_prime’ ‘overlaps_at_its_3_prime_with’ ‘includes’
B	‘is_in_5_prime_of’	‘is_in_3_prime_of’

A. Relations entre les labels d’IMGT-ONTOLOGY utilisées pour la description des prototypes [14-15, 27].
B. Relations entre les labels d’IMGT-ONTOLOGY utilisées pour localiser respectivement deux labels en l’absence d’information sur la contigüité, inclusion ou chevauchement. Cette relation et la réciproque ont été ajoutées lors de la considération d’IMGT/LIGMotif.

2.2.2 Le concept « Core »

Le concept « Core » permet de décrire la région codante des gènes et contient les cinq instances suivantes: 'REGION' (pour les gènes de type conventionnels), 'V-REGION', 'D-REGION', 'J-REGION' et 'C-REGION' (pour les types de gènes V, D, J et C, respectivement). Ces instances sont particulièrement importantes, car elles peuvent être décrites dans toutes les instances du concept « Molecule_EntityPrototype ». Elles permettent de décrire les chaînes des récepteurs de l'antigène, en dépit de la complexité de leur structure et ainsi de relier séquences, structures et fonctions. De plus, ce sont les instances du concept « Core » qui ont permis la définition et la description standardisées des allèles d'IG et TR (concepts de classification), maintenant approuvées au niveau international [19-20].

2.2.3 Le concept « GeneCluster »

Le concept « GeneCluster » d'IMGT-ONTOLOGY permet de décrire les séquences génomiques contenant plusieurs gènes. Les gènes rassemblés dans un cluster peuvent être d'un même prototype (par exemple, un V-CLUSTER contient seulement des gènes V), ou de prototypes différents (par exemple, un V-D-J-CLUSTER contient au moins un gène V, un gène D et un gène J).

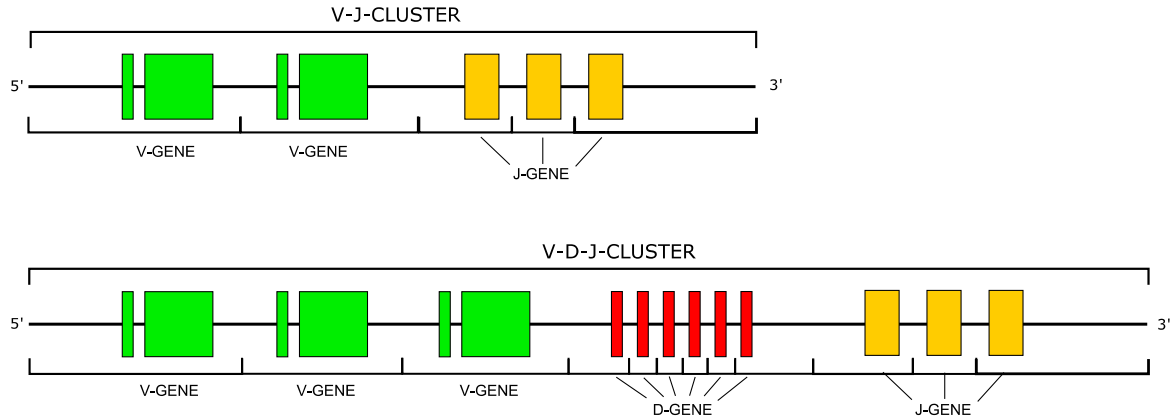


Figure 2.7. Représentation graphique des instances 'V-J-CLUSTER' et 'V-D-J-CLUSTER' du concept « GeneCluster » de l'axiome DESCRIPTION.

Sept instances du concept « GeneCluster » sont utilisées pour les séquences génomiques non réarrangées. De manière générale, ces instances sont particulièrement utiles à l'annotation à grande échelle génomique des locus d'IG et TR et sont utilisées par Sequence Ontology (SO) (Tableau 2.2).

Tableau 2.2. Instances du concept « GeneCluster ».

		Instance du concept 'Molecule_EntityPrototype'	
Instance du concept 'GeneCluster' d'IMGT-ONTOLOGY	Sequence Ontology	Nombre minimum des différentes instances	Nom des différentes instances
V-CLUSTER	SO:0000526	1	V-GENE
J-CLUSTER	SO:0000513	1	J-GENE
D-CLUSTER	SO:0000559	1	D-GENE
D-J-CLUSTER	SO:0000560	2	D-GENE
			J-GENE
V-D-CLUSTER		2	V-GENE
			D-GENE
V-J-CLUSTER	SO:0000534	2	V-GENE
			J-GENE
V-D-J-CLUSTER	SO:0000532	3	V-GENE
			D-GENE
			J-GENE

Sept instances du concept « GeneCluster » d'IMGT-ONTOLOGY sont présentées. Six d'entre elles sont aussi utilisées par Sequence Ontology (SO) [12, 173]. Relations avec les instances du concept « Molecule_EntityPrototype » comprenant le nombre minimum d'instances différentes pour chaque instance du concept « GeneCluster » et nom des différentes instances, (définies dans la liste des labels IMGT/LIGM-DB [8]).

2.3 Axiome CLASSIFICATION et concepts de classification

L'axiome CLASSIFICATION postule que les molécules, cellules, tissus, organes, organismes ou populations, leurs processus et leurs relations, doivent être classés. En biologie moléculaire, les concepts de la classification établis par l'axiome CLASSIFICATION permettent de classer et nommer les gènes et leurs allèles. Les gènes qui codent pour les IG et TR appartiennent à des familles multigéniques très polymorphes. Une contribution majeure de l'IMGT-ONTOLOGY a été de fixer les principes de leur classification et de proposer une nomenclature standardisée [1,2] (Figure 2.8). La nomenclature IMGT® des gènes a été approuvée au niveau international par le HGNC, en 1999 [28]. Les noms de gènes IMGT® des IG et TR sont la référence officielle pour les projets sur le génome et, comme tels, ont été intégrés dans la base de données du génome (GDB), dans LocusLink et dans Entrez Gene au NCBI [30]. Les gènes IG et TR [19-20] sont gérés dans la base de données IMGT/GENE-DB [174].

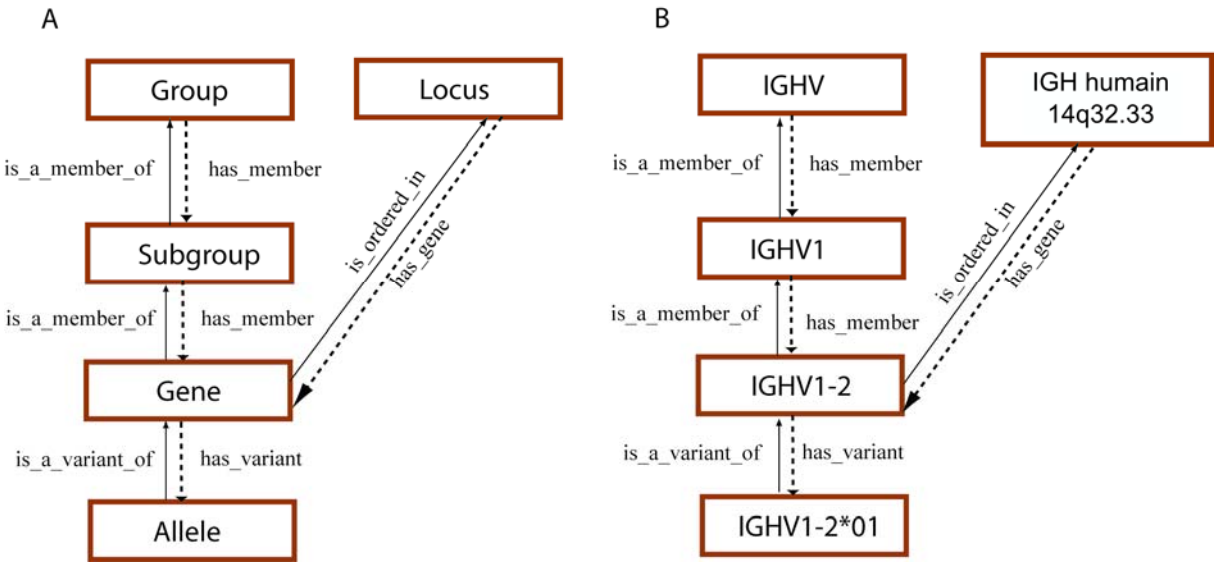


Figure 2.8. Concepts de classification et nomenclature des gènes et allèles (axiome de CLASSIFICATION). (A) Hiérarchie des concepts de classification et de leurs relations. (B) Exemples d'instances de concept de classification. Les instances sont associées à une instance du concept « Taxon », et plus précisément les instances des concepts « Gene » et « Allele » à une instance du concept « Species » (ici, *Homo sapiens*). Le concept « Locus » est un concept de localisation (axiome de LOCALIZATION).

2.3.1 Les concepts « Group » et « Subgroup »

Le concept « Group » classifie un ensemble de gènes qui appartiennent à la même famille multigénique, au sein d'une même espèce ou entre espèces différentes. Pour les IG et TR, l'ensemble des gènes est identifié par une instance du concept « GeneType » (V, D, J ou C). Le concept « Subgroup » classifie un sous-ensemble de gènes qui appartiennent au même groupe, et qui, dans une espèce donnée, partagent au moins 75% d'identité au niveau de la séquence nucléotidique (et dans la configuration 'germline' pour les gènes V, D, et J).

2.3.2 Les concepts « Gene » et « Allele »

Le concept « Gene » classe une unité de la séquence d'ADN qui peut être potentiellement transcrite et/ou traduite (cette définition comprend les éléments de régulation en 5' et 3', et les introns, s'ils sont présents). Les instances du concept « Gene » sont des noms de gènes. Dans IMGT-ONTOLOGY, un nom de gène est composé du nom de l'espèce (par exemple de l'instance du concept « Species » du concept parent « Taxon ») et du symbole international de gène HGNC/IMGT, par exemple, *Homo sapiens* IGLV1-2. Par extension, les orphons et les pseudogènes sont également des instances du concept « Gene ». Le concept « Allele » classe une variante polymorphique d'un gène. Les instances du concept « Allele » sont des noms d'allèles. Les allèles identifiés par les mutations de la séquence nucléotidique sont classés par référence à l'allèle *01. La description complète des mutations

et la désignation des noms d'allèles sont actuellement enregistrées pour les séquences 'Core' (V-REGION, D-REGION, J-REGION, C-REGION). Elles sont reportées dans les tableaux de l'alignement, dans IMGT Répertoire http://www.imgt.org et IMGT/GENE-DB [31].

2.4 Axiome NUMEROTATION

L'axiome NUMEROTATION postule que les molécules, cellules, tissus, organes, organismes ou populations, leurs processus et leurs relations, doivent être numérotés. Jusqu'à présent, ces concepts ont été essentiellement définis au niveau moléculaire. L'axiome NUMEROTATION et les concepts de numérotation déterminent les principes de la numérotation unique pour un domaine (séquences et structures 3D) [175-177]. Le concept « IMGT_unique_numbering » a trois instances: « IMGT_unique_numbering_for_V_type_ domain », « IMGT_unique_numbering_for_C_type_domain », « IMGT_unique_numbering _for_G_type_domain » [175-177].

Le concept « IMGT_unique_numbering » détermine les concepts « FR-IMGT_length », « CDR-IMGT_length », « Strand_length » et « Helix_length » [175-177]. Le concept « IMGT_unique_numbering » est illustré par le concept « IMGT_Collier_de_Perles » qui permet la représentation graphique en deux dimensions (2D) des séquences d'acides aminés du type de domaine V, C ou G [174] et comprend trois instances (Figure 2.9 et Figure 2.10). Ce concept est largement reconnu au niveau international et l'expression « IMGT Collier de Perles » est maintenant utilisée, en français, dans les publications scientifiques en anglais.

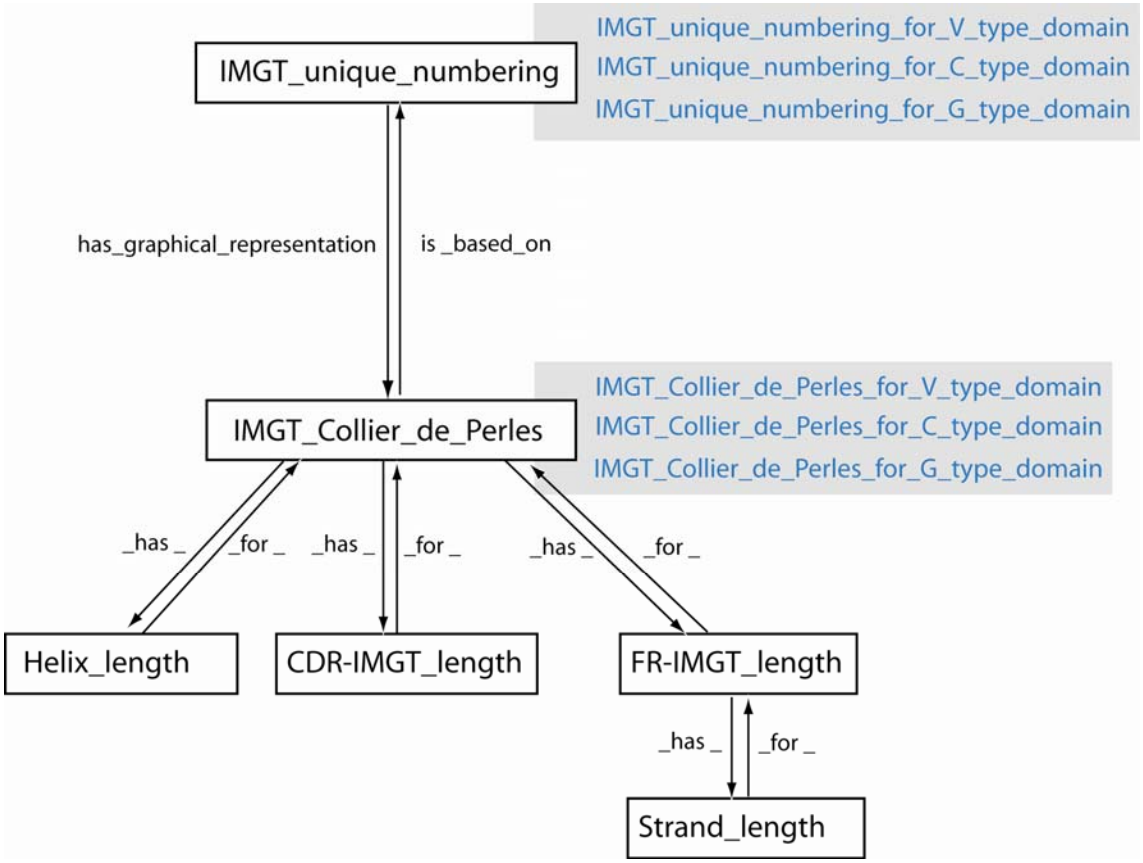


Figure 2.9. Les concepts « IMGT_unique_numbering » et « IMGT_Collier_de_Perles » et leurs relations avec d'autres concepts de numérotation (axiome de NUMEROTATION). Les instances de concepts sont écrites en bleu. Les flèches indiquent les relations réciproques 'has_graphical_representation' et 'is_based_on', '_has_' et '_for_'.

Le concept « IMGT_Collier_de_Perles » est particulièrement utilisé dans l'ingénierie des anticorps pour l'humanisation des anticorps murins dans lesquels il est nécessaire de délimiter précisément les CDR-IMGT murins à greffer, afin de préserver la spécificité des anticorps. Les concepts de numérotation sont également à l'origine de la standardisation de la description de l'allèle, plus généralement, de la description des mutations (IMGT Scientific chart, http://www.imgt.org).

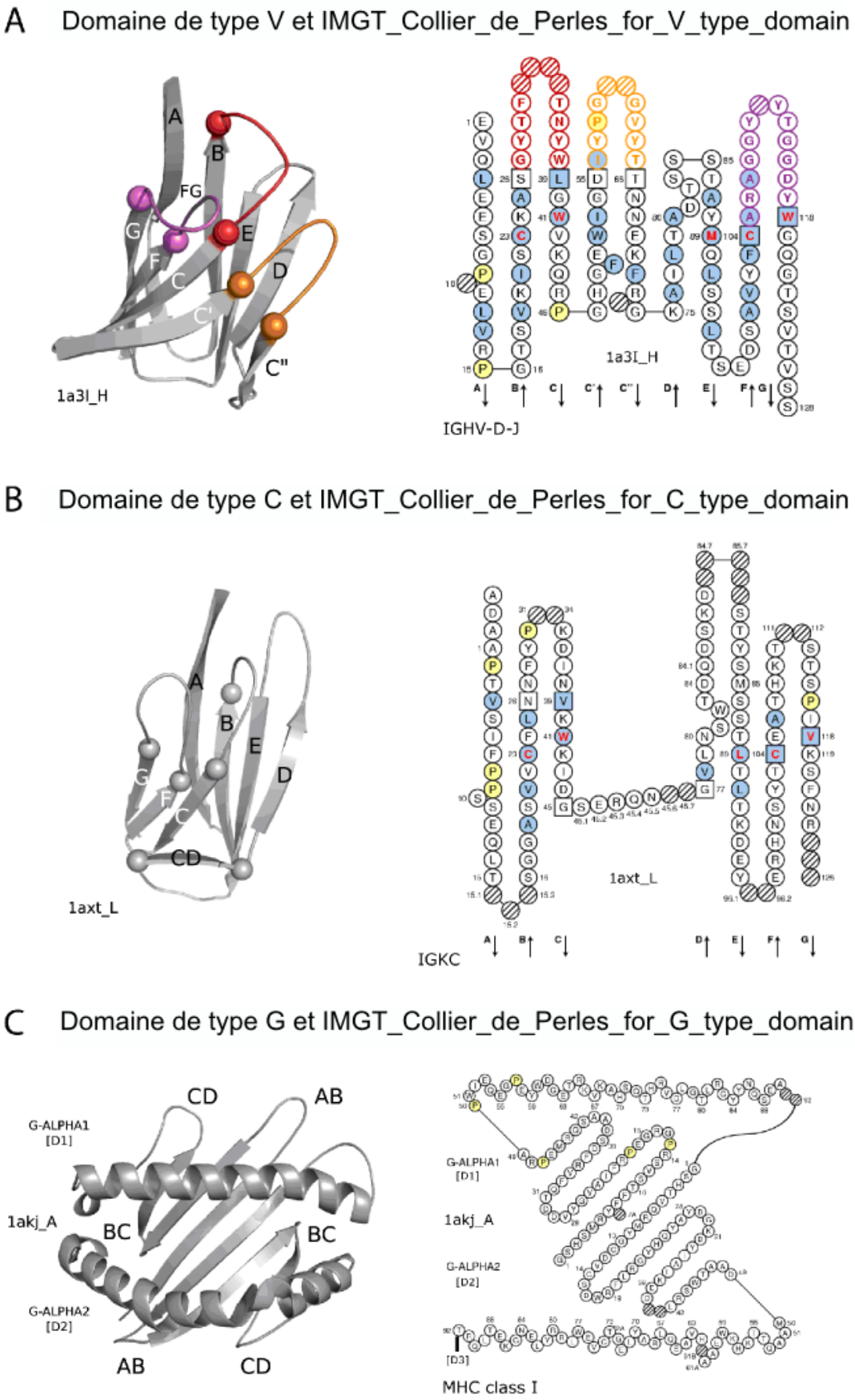


Figure 2.10. Instances des concepts « DomainType » et « IMGT_Collier_de_Perles ».

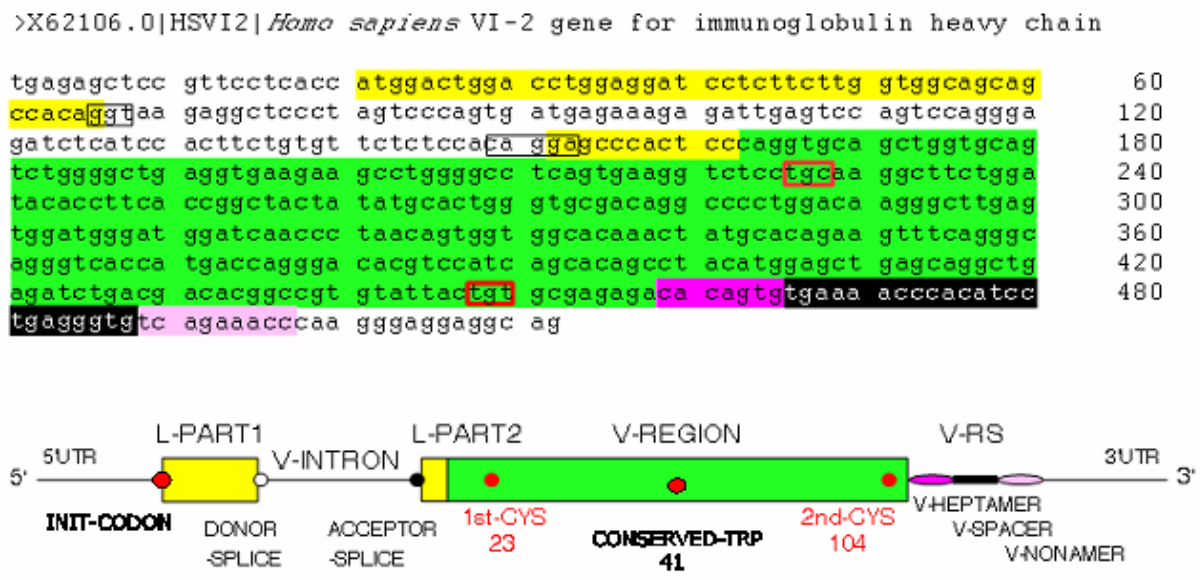


Figure 2.11. Organisation d'un V-GENE d'IG avec ses labels.

L'axiome NUMEROTATION attribue une position définie aux acides aminés conservés. La numérotation permet de repérer ainsi les acides aminés 1st-CYS à la position 23 et 2nd-CYS à la position 104 et CONSERVED-TRP à la position 41 (Figure 2.11) d'un gène V. Ces positions d'acides aminés proviennent d'un alignement de plusieurs milliers de séquences de domaines variables et constants d'IG et TR. Ce sont des positions de référence qui sont très utiles pour comparer les séquences d'IG et TR entre elles.

CHAPITRE 3
Le système d'information international en immunogénétique IMGT®

La complexité des molécules d'IG et TR nécessite une grande expertise. Lors du 10ième Workshop du « Human Genome Mapping » en 1989 à New Haven (Etats-Unis) un système d'information spécialisé pour les IG et TR s'est révélé nécessaire. IMGT® a été créé par le Professeur Marie-Paule Lefranc à Montpellier en collaboration avec EMBL-EBI.

Le système d'information IMGT® assure le recueil, le stockage, le traitement, la transmission, l'archivage et la traçabilité des informations produites pour répondre aux objectifs des chercheurs et des cliniciens. Il contient des informations expertisées sur les IG et TR, et d'autres protéines apparentées au Système Immunitaire de l'homme et d'autres vertébrés. Les données d'IMGT sont particulièrement utilisées dans la recherche fondamentale et médicale pour l'étude de pathologies (maladies auto-immunes et infectieuses, Sida, leucémies, lymphomes, myélomes), en génomique (étude de la diversité et de l'évolution des gènes des réponses immunitaires adaptatives), en biologie structurale, en recherche vétérinaire (répertoire des IG et TR dans les espèces domestiques et sauvages), en biotechnologie et ingénierie des anticorps, pour les diagnostics des leucémies, lymphomes et myélomes et pour les approches thérapeutiques (greffes, immunothérapie, vaccinologie).

IMGT comprend plusieurs bases de données accessibles sur le Web (IMGT/LIGM-DB [8], IMGT/GENE-DB [31], IMGT/3Dstructure-DB [172], IMGT/PRIMER-DB) des outils en ligne permettant l'analyse des séquences (IMGT/V-QUEST [178-179], IMGT/JunctionAnalysis [180], IMGT/PhyloGene [181], IMGT/Allele-Align (http://www.imgt.org/Allele-Align/), IMGT/Automat , IMGT/LIGMotif) des génomes (IMGT/GeneInfo [182-183], IMGT/LocusView, IMGT/GeneSearch, IMGT/GeneView) et des structures (IMGT/StructureQuery [172]) et plus de 10 000 pages de ressources Web (IMGT Répertoire, IMGT Scientific chart…). Ces outils ont été développés principalement pour faciliter le travail des annotateurs d'IMGT. L'annotation consiste à décrire les séquences d'ADN en recherchant les zones d'intérêts biologiques (gènes, exons…) et à leur attribuer l'expertise et les connaissances correspondantes. Dans ce chapitre nous nous intéresserons

aux bases de données et outils principaux d'IMGT® que sont IMGT/LIGM-DB, IMGT/Automat et IMGT/V-QUEST.

3.1 Bases de données

3.1.1 IMGT/LIGM-DB

La base de données IMGT/LIGM-DB contient l'ensemble des séquences nucléotidiques IG et TR de l'homme et de 104 autres espèces de vertébrés. Ces séquences ont été publiées dans les divisions « HUM », « MUS », « VRT », et « PRI » de la base de données généraliste EMBL Bank [11] (les séquences des autres divisions telles que les EST ne sont pas suffisamment fiables pour être intégrées dans IMGT). Ces séquences se répartissent en séquences d'ADN génomique [« germline » (non réarrangées) et réarrangées (résultant de la recombinaison des gènes V-D-J ou V-J)] et en séquences d'ADNc, réarrangées et épissées qui codent, lorsqu'elles sont productives et non partielles, une chaîne complète d'IG ou de TR [3,4]. Cette dernière catégorie représente plus de la moitié de la base de données IMGT/LIGM-DB [2] et peut être annotée automatiquement. Le traitement des séquences des IG et TR dans IMGT® consiste à produire un fichier IMGT/LIGM-DB contenant les annotations expertes d'IMGT® à partir d'un fichier EMBL contenant des annotations généralistes. Le même format est utilisé dans les deux fichiers afin de faciliter l'interopérabilité avec les bases généralistes. Les fichiers EMBL Bank et IMGT/LIGM-DB sont composés d'un simple texte dont chaque ligne fait au plus 80 caractères (Figure 3.1). L'information que renferme chaque ligne est désignée par un identifiant constitué des deux premières lettres de la ligne.

Tableau 3.1. Définition des identifiants et leurs occurrences dans un fichier IMGT/LIGM-DB.

identifiant	définition	occurrence
ID	IDentification	(commence chaque entrée, une par entrée)
AC	ACcession number	(=1 par entrée)
DT	DaTe	(2 par entrée)
DE	DEscription	(=1 par entrée)
KW	KeyWord	(=1 par entrée)
OS	Organism Species	(=1 par entrée)
OC	Organism Classification	(=1 par entrée)
RN	Reference Number	(=1 par entrée)
RC	Reference Comment	(=0 par entrée)
RP	Reference Positions	(=1 par entrée)
RX	reference cross-reference	(=0 par entrée)
RA	Reference Author(s)	(=1 par entrée)
RT	Reference Title	(=1 par entrée)
RL	Reference Location	(=1 par entrée)
DR	Database cross Reference	(=0 par entrée)
FH	Feature table Header	(=1 par entrée)
FT	Feature Table data	(=0 par entrée)
CC	Comments or notes	(=0 par entrée)
XX	ligne d'espace	(plusieurs par entrée)
SQ	Sequence Header	(=1 par entrée)
blancs	séquence	(=1 par entrée)
//	ligne de fin	(termine chaque entrée, 1 par entrée)

L'information elle-même commence au caractère qui est en position 6 de chaque ligne. Selon le type d'identifiant, la totalité de l'information peut être sur une seule ligne ou répartie sur plusieurs lignes consécutives. De plus, selon le type d'identifiant, il y a un seul ou plusieurs éléments d'informations. On y retrouve notamment le type de chaîne 'IG', 'IG-Heavy' le type de gènes 'variable', 'diversity', 'joining', les mots clés 'antigen receptor' qui s'appliquent à toutes les entrées de IMGT/LIGM-DB (Figure 3.1). Les concepts de description et de numérotation se retrouvent dans les lignes FT. IMGT/LIGM-DB possède plus de 200 labels alors qu'EMBL n'en possède que 60 dont seulement 7 ([5], http://www.ebi.ac.uk/embl/WebFeat/) correspondent à des annotations d'IG et TR.

IMGT FLAT-FILE ADNg

```
ID   X07448 IMGT/LIGM annotation : by annotators; genomic DNA; HUM; 618 BP.
XX
AC   X07448;
XX
DT   15-MAY-1995 (Rel. 2, arrived in LIGM-DB )
DT   20-OCT-2008 (Rel. 200843-1, Last updated, Version 10)
XX
DE   Human V35 gene for Ig heavy chain  ;
DE   genomic DNA; germline configuration; Ig-Heavy; regular; functionality
DE   functional; group IGHV; subgroup HV1.
XX
KW   antigen receptor; Immunoglobulin superfamily (IgSF);
KW   Immunoglobulin (IG); IG-Heavy; variable; IMGT reference sequence; gDNA;
KW   germline; functional; V-gene.
XX
OS   Homo sapiens (human)
OC   cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa;
OC   Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata;
OC   Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda;
OC   Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates;
OC   Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae;
OC   Homo/Pan/Gorilla group; Homo.
XX
RN   [1]
RP   1-618
RX   PUBMED; 2841108.
RA   Matsuda F., Lee K.H., Nakai S., Sato T., Kodaira M., Zong S.Q., Ohno H.,
RA   Fukuhara S., Honjo T.;
RT   "Dispersed localization of D segments in the human immunoglobulin
RT   heavy-chain locus";
RL   EMBO J. 7(4):1047-1051(1988).
XX
DR   EMBL-ALIGN; ALIGN_000425;
DR   GDB; 118731;
DR   GDB; 9931660;
DR   GOA; P20003;
DR   InterPro; IPR003596; Ig_V-set_sub
DR   InterPro; IPR007110; Ig-like
DR   InterPro; IPR013106; Ig_V-set
DR   InterPro; IPR013783; Ig-like_fold
DR   PDB; 1WT5; X-ray
DR   UniProtKB/Swiss-Prot; P23083; HV103_HUMAN
DR   EMBL; X07448.
XX
CC   Data kindly reviewed (09-JUN-1900) by Matsuda F.
XX
FH   Key             Location/Qualifiers
FH
FT   V-GENE          1..618
FT                   /db_xref="taxon:9606"
FT                   /cell_line="FLEB14-14"
FT                   /allele="IGHV1-2*01"
FT                   /gene="IGHV1-2"
FT                   /organism="Homo sapiens"
FT                   /mol_type="genomic DNA"
FT   L-INTRON-L      126..260
FT   L-PART1         126..171
FT                   /protein_id="CAB56703.1"
FT                   /translation="MDWTWRILFLVAAAT"
FT   INIT-CODON      126..128
FT   DONOR-SPLICE    171..173
FT   V-INTRON        172..257
FT   ACCEPTOR-SPLICE 255..259
FT   V-EXON          258..564
FT                   /protein_id="CAB56703.1"
FT                   /codon_start=3
FT                   /translation="AHSGVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAP
FT                   QAPGQGLEWMGRINPNSGGTNYAQKFQGRVTSTPDTSISTAYMELSRLRSDDTVVYYC
FT                   YTCAR"
FT   L-PART2         258..268
FT                   /codon_start=3
FT                   /translation="AHS"
FT   V-REGION        269..564
FT                   /allele="IGHV1-2*01"
FT                   /gene="IGHV1-2"
FT                   /CDR_length="[8.8.21"
FT                   /translation="QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAP
FT                   GQGLEWMGRINPNSGGTNYAQKFQGRVTSTPDTSISTAYMELSRLRSDDTVVYYC
FT                   AR"
FT   FR1-IMGT        269..343
FT                   /AA_IMGT="1 to 26, AA 10 is missing"
FT                   /translation="QVQLVQSGAEVKKPGASVKVSCKAS"
FT   1st-CYS         332..334
FT   CDR1-IMGT       344..367
FT                   /AA_IMGT="27 to 34"
FT                   /translation="GYTFTGYY"
FT   FR2-IMGT        368..418
FT                   /AA_IMGT="39 to 55"
FT                   /translation="MHWVRQAPGQGLEWMGR"
FT   CONSERVED-TRP   374..376
FT   CDR2-IMGT       419..442
FT                   /AA_IMGT="56 to 63"
FT                   /translation="INPNSGGT"
FT   FR3-IMGT        443..556
FT                   /AA_IMGT="66 to 104, AA 73 is missing"
FT                   /translation="NYAQKFQGRVTSTPDTSISTAYMELSRLRSDDTVVYYC"
FT   2nd-CYS         554..556
FT   CDR3-IMGT       557..564
FT                   /AA_IMGT="105 to 106"
FT                   /translation="AR"
FT   3'UTR           565..618
FT   V-RS            565..603
FT   V-HEPTAMER      565..571
FT   V-SPACER        572..594
FT   V-NONAMER       595..603
XX
SQ   Sequence 618 BP; 156 A; 165 C; 173 G; 124 T; 0 other;
     ctcatgaata tgcaaataac ctgagattta ctgaagtaaa tacagatctg tcctgtgccc      60
     tgagagcatc acccagcaac cacatctgtc ctctagagaa tcccctgaga gctccgttcc     120
     tcaccatgga ctggacctgg aggatcctct tctggtggc agcagccaca ggtaagaggc      180
     tccctagtcc cagtgatgag aaagaggatt gagtccagtc cagggagatc tcatccactt     240
     ctgtgtctcc tccacaggag cccactccca ggtgcagctg gtgcagtctg gggctgaggt     300
     gaagaagcct gggagcctca tgaaggtctc ctgcaaggct tctggatacа ccttcaccgg     360
     ctactatatg cactgggtgc gacaggcccc tggacaaggg cttgagtgga tgggacggat     420
     caaccctaac agtggtggca caaactatgc acagaagttt cagggcaggg tcaccagtac     480
     cagggacacg tccattagca cagcctacat ggagctgagc aggctgagat ctgacgacac     540
     ggtcgtgtat tactgtgcga gagacacagt gtgaaaaccc acatcctgag ggtgtcagaa     600
     gccccaggga ggaggcag                                                  618
//
```

IMGT FLAT-FILE ADNc

```
ID   L39956 IMGT/LIGM annotation : automatic; RNA; HUM; 375 BP.
XX
AC   L39956;
XX
DT   28-SEP-2001 (Rel. 200139-S, arrived in LIGM-DB )
DT   13-NOV-2001 (Rel. 200146-2, Last updated, Version 2)
XX
DE   Homo sapiens monospecific anti-ssDNA antibody heavy chain variable region
DE   mRNA, complementarity determining regions  1-3 and framework regions 1-4.
DE   ;
DE   RNA; rearranged configuration; Ig-Heavy; regular; functionality
DE   productive; group IGHV; subgroup HV3; specificity anti-DNA single-stranded
DE   (ss) [human].
XX
KW   antigen receptor; Immunoglobulin superfamily (IgSF);
KW   Immunoglobulin (IG); IG-Heavy; variable; diversity; joining; hybridoma;
KW   cDNA; undefined; rearranged; L-V-D-J-C-sequence; partner.
XX
OS   Homo sapiens (human)
OC   cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa;
OC   Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata;
OC   Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda;
OC   Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates;
OC   Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae;
OC   Homo/Pan/Gorilla group; Homo.
XX
RN   [1]
RP   1-375
RX   MEDLINE; 96195158.
RA   Mitamura K., Suenaga R., Wilson K.B., Abdou N.I.;
RT   "V gene sequences of human anti-ssDNA antibodies secreted by lupus-derived
RT   CD5-negative B cell hybridomas";
RL   Clin. Immunol. Immunopathol. 70(2):152-160(1996).
XX
DR   IMGT/LIGM:partner; L39957;
DR   EMBL; L39956.
XX
FH   Key             Location/Qualifiers
FH
FT   L-V-D-J-C-SEQUENCE  <1..375>
FT                   /partial
FT                   /IMGT_note="automatically annotated with IMGT tools"
FT                   /cell_type="B-cell hybridoma 2F7"
FT                   /db_xref="taxon:9606"
FT                   /organism="Homo sapiens"
FT   V-D-J-REGION    1..375
FT                   /translation="QVHLVESGGAVFHPGRSLRLSRAASGFTFSSYGMHWVRQAP
FT                   AKGLEWVAVIWYDGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYC
FT                   AKHVTIAAAGRRGAGMDVWGQGTTVTVSS"
FT   V-REGION        1..296
FT                   /CDR_length="[8.8.18]"
FT                   /allele="IGHV3-33*01, putative"
FT                   /gene="IGHV3-33"
FT                   /putative_limit="3' side"
FT                   /translation="QVHLVESGGAVFHPGRSLRLSRAASGFTFSSYGMHWVRQAP
FT                   AKGLEWVAVIWYDGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYC
FT                   AK"
FT   FR1-IMGT        1..75
FT                   /AA_IMGT="1 to 26, AA 10 is missing"
FT                   /translation="QVHLVESGGAVFHPGRSLRLSRAAS"
FT   CDR1-IMGT       76..99
FT                   /AA_IMGT="27 to 34"
FT                   /translation="GFTFSSYG"
FT   FR2-IMGT        100..150
FT                   /AA_IMGT="39 to 55"
FT                   /translation="MHWVRQAPAKGLEWVAV"
FT   CONSERVED-TRP   106..108
FT   CDR2-IMGT       151..174
FT                   /AA_IMGT="56 to 63"
FT                   /translation="IWYDGSNK"
FT   FR3-IMGT        175..288
FT                   /AA_IMGT="66 to 104, AA 73 is missing"
FT                   /translation="YYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYC"
FT   2nd-CYS         286..288
FT   CDR3-IMGT       289..342
FT                   /AA_IMGT="105 to 117, including AA 112.1, 111.1,
FT                   112.2, 111.3, 112.3"
FT                   /translation="AKHVTIAAAGRRGAGMDV"
FT   JUNCTION        286..345
FT                   /in_frame
FT                   /translation="CAKHVTIAAAGRRGAGMDVW"
FT   N1-REGION       297..299
FT                   /translation="R"
FT   D-REGION        300..319
FT                   /allele="IGHD6-13*01, putative"
FT                   /gene="IGHD6-13"
FT                   /codon_start=2
FT                   /translation="TIAAAG"
FT   N2-REGION       320..326
FT                   /codon_start=3
FT                   /translation="R"
FT   J-REGION        327..375
FT                   /allele="IGHJ6*02, putative"
FT                   /gene="IGHJ6"
FT                   /putative_limit="5' side"
FT                   /codon_start=2
FT                   /translation="AGMDVWGQGTTVTVSS"
FT   J-TRP           343..345
FT   FR4-IMGT        343..375
FT                   /AA_IMGT="118 to 128"
FT                   /translation="WGQGTTVTVSS"
XX
SQ   Sequence 375 BP; 86 A; 97 C; 117 G; 75 T; 0 other;
     caggtgcacc tggtggagtc tgggggagcc gtgttccacc ctgggaggtc cctgagactc      60
     tcccgtgcag cgtctggatt caccttcagt agctatggca tgcactgggt ccgccaggct     120
     ccaggcaaag ggctggagtg ggtggcagtt atatggtatg atggaagtaa taaatactat     180
     gcagactccg tgaagggccg attcaccatc tccagagaca attccaagaa cacgctgtat     240
     ctgcaaatga acagcctgag agccgaggac acggctgtgt attactgtgc gaaacacgtg     300
     acgatagcag cagctggtag acgaggcgcc ggtatggacg tctgggggca agggaccacg     360
     gtcaccgtct cctca                                                     375
//
```

Figure 3.1. Fichiers IMGT Flat-File d'ADNg et d'ADNc. Le fichier à plat d'ADNg contient un V-GENE fonctionnel complet. Le fichier à plat d'ADNc contient une L-V-D-J-C-SEQUENCE complète. Les carrés rouges encadrent la numérotation des labels.

Dans notre cas, un motif est une courte séquence nucléotidique commune et spécifique à tous les membres d'une famille de gènes. Ils forment la charpente sur laquelle la description de la séquence va s'édifier. Les motifs d'un gène variable sont représentés par les labels INIT-CODON, DONOR-SPLICE, ACCEPTOR-SPLICE, 1st-CYS, CONSERVED-TRP, 2nd-CYS, V-HEPTAMER, V-NONAMER. La numérotation IMGT est indiquée au niveau protéique sous forme AA_IMGT dans le fichier à plat (Figure 3.1). La classification des gènes (et allèles qui sont les différentes versions d'un gène pour une espèce donnée) est ajoutée au niveau des labels qui définissent les régions codantes 'core' des gènes. L'obtention comprend des informations telles que le 'cell_type' et est attribuée au label majeur.

3.1.2 IMGT/GENE-DB

Le 9 septembre 2009, IMGT/GENE-DB contenait 1999 gènes et 3026 allèles d'IG et TR de l'homme, la souris, du lapin, du rat et du chien (673 gènes et 1244 allèles de l'*Homo sapiens*, et 832 gènes et 1264 allèles de la souris la plupart des entrées proviennent de *Mus musculus*, quelques entrées de *Mus cookii*, *Mus minutoides*, *Mus Pahari*, *Mus saxicola* et *Mus spretus*), 19 gènes et 40 allèles du lapin *Oryctolagus cuniculus*, 398 gènes et 401 allèles du rat dont un seul gène et allèle proviennent du *Rattus rattus*, le reste provient du *Rattus Norvegicus*, 77 gènes et allèles proviennent du chien *Canis lupus familiaris*, Tableau 3.2). Ces données peuvent être observées en cliquant directement sur le lien 'IMGT/GENE-DB Statistics' de la page de requête d'IMGT/GENE-DB (Figure 3.2). Un tableau dynamique affiche les statistiques des données actuelles contenues dans la base.

Tableau 3.2. Statistiques globales par espèces d'IMGT/GENE-DB

Espèces	Nombre gènes d'IG et TR	Nombre d'allèles d'IG et TR
Canis lupus familiaris	77	77
Homo sapiens	673	1244
Mus cookii	1	1
Mus minutoides	2	2
Mus musculus	811	1240
Mus pahari	3	3
Mus saxicola	1	1
Mus spretus	14	17
Oryctolagus cuniculus	19	40
Rattus norvegicus	397	400
Rattus rattus	1	1
Total	1999	3026

L'ensemble des gènes d'IG et TR est complet pour l'homme incluant les sept locus et pour les ensembles orphons chromosomiques [19-20]. La page de requête d'IMGT/GENE-DB comprend trois types de recherches: (i) 'GENERAL CRITERIA' permet une recherche des

gènes d'IG et TR, pour une espèce donnée, par locus ou par ensemble d'orphons chromosomique, par type de gène, groupe ou sous-groupe, ou fonctionnalité. L'utilisateur peut sélectionner les gènes qui ont été trouvés réarrangés, transcripts ou traduits. (ii) 'SHORT CUT' permet une sélection sur le nom de gène ou de clone, pour une espèce donnée. (iii) 'IMGT/GENE-DB direct links' donne accès à un ensemble de liens, qui permettent de récupérer l'information liée à des gènes donnés, ou des gènes d'un groupe à l'aide d'URL configurables.

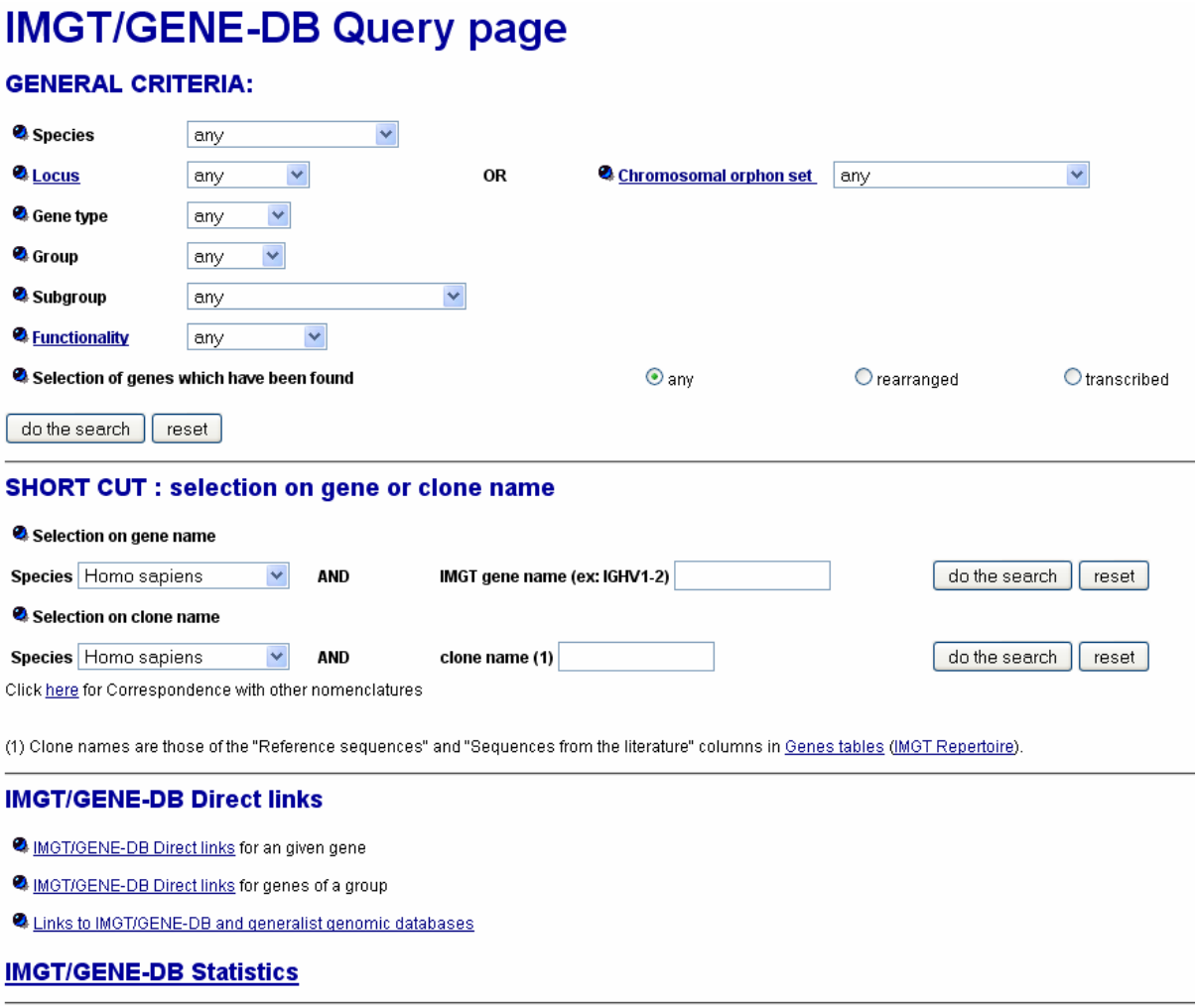


Figure 3.2. Page de requête d'IMGT/GENE-DB.

Suite à une sélection par 'GENERAL CRITERIA' ou par 'SHORT CUT', la page de résultat IMGT/GENE-DB affiche, en haut, la sélection de l'utilisateur, le nombre de gènes et allèles résultant, puis la liste des gènes avec, pour chacun des gènes, leur espèce, leur nom de gène IMGT®, leur fonctionnalité, leur définition de gène selon IMGT®, leur nombre

d’allèles, leur localisation chromosomique et leur séquence(s) de référence provenant d’IMGT/LIGM-DB pour l’allèle *01 (Figure 3.3).

Dans la section « Choose your display », l’utilisateur peut choisir entre trois types d’affichage: (i) les entrées complètes individuelles d’IMGT/GENE-DB pour les gènes sélectionnés dans la liste des gènes résultant ; (ii) les séquences de références des allèles d’IMGT/GENE-DB au format FASTA: les séquences de nucléotides ou d’acides aminés, que ce soit avec des gaps de la numérotation unique IMGT [176-177, 184], ou sans gaps; (iii) les labels de séquences IMGT au format FASTA, extraite d’annotations expertes de séquences référence provenant d’IMGT/LIGM-DB. Ainsi, toute les séquences labellisées (par exemple, V-EXON, V-HEPTAMER) peuvent être récupérées, les régions ‘core’ out-of-frame des pseudogènes, qui ne sont pas disponibles dans la base de données de séquences référence allèlique d’IMGT/GENE-DB, et les séquences artificiellement épissées L-PART1+L-PART2 et L-PART1+V-EXON. Pour les séquences de nucléotides, l’utilisateur a la possibilité d’étendre les limites en 5’ ou 3’ en tapant le nombre de nucléotides de son choix.

RESULTS OF YOUR SEARCH:

```
Your selection :
        species='Homo sapiens' ...
AND     type=locus...
AND     collection='TRB locus' ...
AND     gene type=diversity...
```

Number of resulting genes : 2

Number of resulting alleles : 3

List of resulting genes

Select, in the first column, the genes to view their detailed IMGT gene entry.

	Species	IMGT gene name	Gene functionality	IMGT gene definition	Number of alleles	Chromosome	Chromosomal localization	IMGT/LIGM-DB reference sequence(s) for allele *01
☐	*Homo sapiens*	TRBD1	F	T cell receptor beta diversity 1	1	7	7q34	K02545
☐	*Homo sapiens*	TRBD2	F	T cell receptor beta diversity 2	2	7	7q34	X02987

☐ Select all genes

Choose your display

Complete IMGT/GENE-DB entries	IMGT/GENE-DB allele reference sequences in FASTA format	IMGT label extraction from IMGT/LIGM-DB reference sequences (F+ORF+all P)
◉ IMGT/GENE-DB entries	○ F+ORF+all P nucleotide sequences ○ F+ORF+in-frame P nucleotide sequences ○ F+ORF+in-frame P nucleotide sequences with IMGT gaps (1)	○ Choose label(s) for extraction 3'D-HEPTAMER 3'D-NONAMER 3'D-RS 3'D-SPACER 3'UTR ◉ Nucleotide sequences (option: add 0 nucleotides in 5' and 0 in 3' *) * only for labels that are not combined ○ Amino acid sequences
	○ F+ORF+in-frame P amino acid sequences ○ F+ORF+in-frame P amino acid sequences with IMGT gaps (1)	

Figure 3.3. Page de résultats d'IMGT/GENE-DB et les trois types de choix dans 'Choose your display'.

3.2 Outils d'analyse de séquences

3.2.1 IMGT/V-QUEST

IMGT/V-QUEST ('V-QUEry and STandardization') est un outil web paramétrable [178], spécialisé, accessible depuis Juillet 1997 [179]. C'est un outil très utile, notamment pour la recherche médicale : il permet d'analyser des séquences nucléotidiques réarrangées en prenant en compte la structure particulière des domaines V des récepteurs d'antigènes (IG et TR), de localiser et de caractériser des mutations, de déterminer des insertions et des délétions et il fournit une évaluation la fonctionnalité.

IMGT/V-QUEST identifie les gènes et allèles V, D et J dans les séquences réarrangées V-J et V-D-J par alignement avec les gènes et allèles germline d'IG et TR des locus de référence d'IMGT®. Il délimite les régions charpentes (FR-IMGT) et les régions hypervariables (CDR-IMGT) de la séquence soumise par l'utilisateur, en accord avec les règles de la charte scientifique d'IMGT® basée sur les axiomes et les concepts de description, de classification et de numérotation de l'IMGT-ONTOLOGY [14-15, 27].

3.2.1.1 Principes de la recherche par IMGT/V-QUEST

IMGT/V-QUEST cherche les régions constitutives des séquences d'IG et TR pour identifier les gènes V, D et J impliqués dans le réarrangement en comparant la séquence utilisateur avec les séquences de référence des gènes et allèles d'IG et TR présents dans IMGT/GENE-DB [31]. Les séquences de référence proviennent de données expérimentales, annotées selon les règles standardisées d'IMGT®. Les locus des séquences de référence, utilisés par IMGT/V-QUEST, contiennent des séquences correspondant à la V-REGION pour les gènes et allèles V, à la J-REGION pour les gènes et allèles J et à la D-REGION pour les gènes et allèles D. IMGT/V-QUEST procède en une analyse séquentielle d'une séquence réarrangée d'IG ou de TR soumise par l'utilisateur, afin de déterminer les gènes et allèles de cette séquence et de délimiter les nucléotides appartenant à chaque région. L'ordre de recherche est le suivant: V-REGION, J-REGION, D-REGION. Lorsqu'une région est délimitée et lorsque les gènes et allèles 'germline' (les plus proches de la séquence utilisateur) sont identifiés, l'algorithme passe à la recherche de la région suivante, en excluant les régions précédemment déterminées. L'analyse précise et complète de la jonction est effectuée par un

outil dédié et intégré à IMGT/V-QUEST: IMGT/JunctionAnalysis [180]. La méthode d'alignement employée par l'outil IMGT/V-QUEST prend en compte les règles d'IMGT unique numbering [177]. Dans la numérotation IMGT, les acides aminés conservés de la V-REGION (et les codons correspondants) sont toujours localisés aux mêmes positions (cystéine 23, tryptophane 41 et cystéine 104) et des gaps sont introduits dans les CDR pour uniformiser les tailles variables des CDR-IMGT. Pour appliquer et préserver la numérotation, les alignements réalisés par IMGT/V-QUEST ne doivent pas autoriser l'insertion de gaps supplémentaires. Afin de respecter ces contraintes, nous avons mis en place un algorithme d'alignement par paire (alignement de séquences deux à deux) dérivé d'un algorithme d'alignement classique de type global qui n'autorise ni les insertions ni les délétions. Le résultat de ces alignements permet de délimiter chaque région, dans le cas des gènes variables, de déterminer la séquence V germline numérotée selon les règles du IMGT unique numbering [177] qui sert de modèle pour numéroter la V-REGION de la séquence utilisateur. L'outil peut ensuite comparer la région délimitée, V-REGION (numérotée), D-REGION, ou J-REGION, avec l'ensemble des V-REGION, D-REGION ou J-REGION, respectivement, des gènes et allèles 'germline' de référence afin de déterminer les gènes et allèles impliqués dans le réarrangement de la séquence de l'utilisateur.

3.2.1.2 Principes d'alignement global sans insertion ni délétions

Les programmes d'alignement de séquences sont basés sur le principe que les séquences de protéines et d'ADN de fonctions apparentée se ressemblent généralement il est donc probable que des séquences similaires aient la même fonction. L'alignement entre deux séquences semble être une solution pour identifier leur fonctionnalité. La recherche de similarité s'effectue par le calcul de la somme des scores élémentaires entre deux résidus (nucléotides ou acides aminés). L'identité de deux résidus et la ressemblance de deux résidus non identiques sont qualifiées de similitude. Le calcul du score peut prendre en compte les fonctions chimiques de certains acides aminés ou les structures très proches et que les nucléotides des séquences d'ADN (en Annexe 2 pour l'alphabet dégénéré de l'ADN) sont soumis à des biais mutationnels favorisant les transitions (C<=>T et A<=>G) aux transversions (A<=>C, A<=>T, G<=>C, G<=>T). C'est dans ce but que des tables de substitutions ont été développées. Ces tables attribuent un score pour évaluer la similitude entre deux acides aminés (par exemple PAM et BLOSUM) ou nucléotides (par exemple NUC4.4). Un alignement de deux séquences est considéré significatif lorsque son score est supérieur ou égal à un score seuil préalablement fixé. L'algorithme d'alignement par paire

employé dans IMGT/V-QUEST (Figure 3.1) est une heuristique car l'algorithme n'explore pas l'ensemble des alignements possibles entre deux séquences dans le but d'optimiser le temps de calcul. Pour une V-REGION, l'algorithme considère uniquement les alignements d'une longueur de plus d'un tiers de la plus petite des deux séquences. Les autres alignements sont jugés trop petits pour en déduire une relation entre une séquence réarrangée et la séquence d'un gène 'germline'. Cette longueur minimale appelée overlap a été déterminée empiriquement au cours du développement de la première version d'IMGT/V-QUEST (Figure 3.4). Pour évaluer le score des alignements (dont la matrice de substitution est présentée en Annexe 3), la séquence 'utilisateur' est décalée d'une position dans le sens 3' à partir de la position SP (StartPosition; correspond à la taille de la séquence utilisateur diminuée de l'overlap). La recherche du meilleur alignement prend fin quand le nucléotide en position EP (EndPosition; correspond à la taille de l'overlap augmentée de 1) de la séquence 'utilisateur' est aligné avec le dernier nucléotide de la séquence de référence. Pour chaque alignement un score est obtenu et seuls les alignements dont le score est jugé significatif seront mémorisés. Un score significatif doit être supérieur ou égal à un seuil préfixé (le seuil pour la V-REGION étant fixé à 600) et être supérieur au meilleur score précédemment calculé. Le score le plus grand détermine le meilleur alignement. Les valeurs des seuils utilisées pour chaque type d'alignement réalisé par IMGT/V-QUEST sont indiquées en Annexe 4.

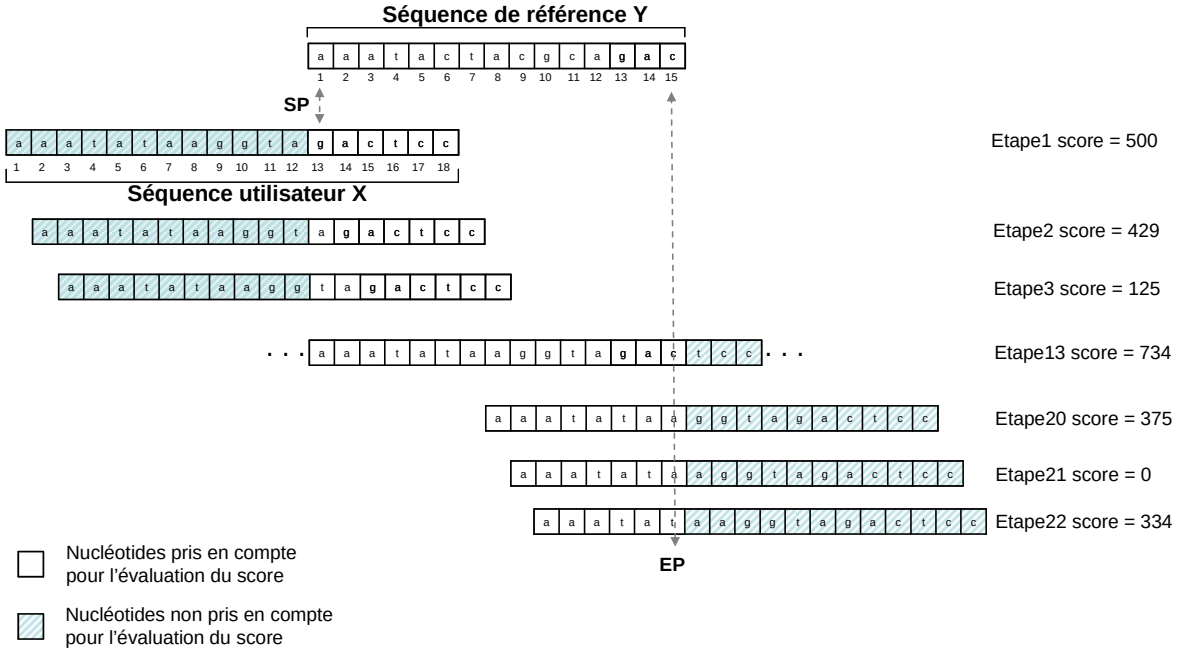


Figure 3.4. Méthode d'alignement entre la séquence utilisateur et une séquence de référence d'IMGT et évaluation du score normalisé. Exemple avec une séquence utilisateur (X) de longueur Tsu=18 nucléotides, et une séquence de référence de longueur Tsr=15 nucléotides; SP=Tsu-(Tsr*1/3) (SP=18-(15/3)=13), EP=1/3*Tsr+1 (EP=(15/3)+1=6). Le meilleur score est déterminé à l'étape 13 (source de l'image: [185]).

3.2.1.3 Etapes principales de l'analyse

3.2.1.3.1 Identification du type de chaîne

L'espèce, le type de récepteur (IG ou TR) et le type de chaîne de la séquence 'utilisateur' sont indispensables à l'analyse des séquences réarrangées d'IG et TR: ces paramètres conditionnent le choix des répertoires de référence comparés à la séquence utilisateur. Les séquences du répertoire de référence contiennent, pour une espèce et un type de récepteur donné, des séquences de référence représentatives de chaque groupe de gène variable (IGHV, IGKV, IGLV pour les IG, TRAV, TRBV, TRGV, TRDV pour les TR). Pour chaque groupe, les séquences sélectionnées ont pour caractéristique de présenter toutes les longueurs des CDR-IMGT possibles. Le type de chaîne de la séquence 'utilisateur' est alors celui de la séquence de référence donnant le meilleur score d'alignement.

3.2.1.3.2 Identification et description du gène V

Le gène et l'allèle V de la séquence 'utilisateur' peut être identifié une fois le type de chaîne connu. La recherche du gène et allèle V localise la V-REGION dans la séquence 'utilisateur' par alignement, elle numérote les codons de la V-REGION selon les règles du IMGT unique numbering (insertion des gaps IMGT) et la compare à l'ensemble des V-REGION des gènes et allèles V germline afin de déterminer le gène et allèle V le plus proche de la séquence 'utilisateur'.

3.2.1.3.3 Identification et description du gène J

IMGT/V-QUEST recherche la J-REGION sur un segment d'une longueur de 200 nucléotides, à partir de la fin 3' de la V-REGION avec l'ensemble des J-REGION des gènes et des allèles J pour un type de chaîne d'une espèce donnée. IMGT/V-QUEST compare ensuite chaque J-REGION des gènes et allèles J germline avec la J-REGION de la séquence utilisateur.

3.2.1.3.4 Identification et description du gène D

IMGT/V-QUEST recherche la D-REGION et le gène et allèle D le plus proche de la séquence 'utilisateur', uniquement pour les chaînes lourdes (IGH) des IG et les chaînes beta (TRB) et delta (TRD) des TR. IMGT/V-QUEST aligne la région de la séquence utilisateur située entre la fin de la V-REGION en 3' et le début de la J-REGION en 5' avec l'ensemble

des D-REGION des gènes et allèles D du répertoire de référence IMGT (pour un type de chaîne et une espèce donnés).

3.2.1.3.5 Analyse détaillée de la JUNCTION

Une fois la jonction délimitée et les gènes et allèles V et J déterminés, l'outil IMGT/JunctionAnalysis [180], intégré à IMGT/V-QUEST, identifie précisément le gène et allèle D germline le plus proche de la séquence utilisateur. Cette recherche est faite pour les chaînes lourdes des (IGH) IG, et les chaînes bêta (TRB) et delta (TRD) des TR. Il délimite les régions palindromiques ou P-REGION et les N-REGION résultant de la N-diversité. De plus, IMGT/JunctionAnalysis [180] ajuste la fin de la V-REGION en 3' et le début de la J-REGION en 5'. L'annotation de la séquence est mise à jour en prenant en compte les résultats de l'analyse de IMGT/JunctionAnalysis. Si l'identification et la description d'un gène D réalisé par IMGT/JunctionAnalysis s'avèrent plus précises qu'IMGT/V-QUEST, ce sont les informations d'IMGT/JunctionAnalysis qui sont affichées dans la page de résultats.

3.2.2 IMGT/Automat: annotation des séquences d'ADNc

L'annotation des séquences d'ADN complémentaires (ADNc) est automatisée par le programme IMGT/Automat développé en Java [186] et, ce, malgré l'origine génétique complexe de ces séquences. IMGT/Automat annote automatiquement les séquences d'ADNc des IG et des TR. A l'issue de l'analyse, le programme vérifie la cohérence globale de l'annotation et élimine les séquences qui pourraient nécessiter un complément d'expertise manuelle.

Lorsque les séquences d'ADNc sont issues de gènes et allèles connus et caractérisés dans IMGT, l'annotation peut être réalisée automatiquement par IMGT/Automat [186]. IMGT/Automat analyse chaque séquence d'ADNc et en déduit une annotation complète. IMGT/Automat s'appuie sur les principaux concepts d'IMGT-ONTOLOGY [14, 27]. IMGT/Automat utilise dans un premier temps le logiciel IMGT/V-QUEST [178-179] pour comparer et aligner la séquence d'ADNc avec les séquences référence du répertoire IMGT de la même espèce. Il en déduit l'IDENTIFICATION du type de chaîne, la CLASSIFICATION des gènes et des allèles V, D, J impliqués et la NUMEROTATION des codons (acides aminés). IMGT/Automat réalise ensuite la DESCRIPTION des motifs constitutifs et spécifiques aux IG et aux TR. Il délimite les «framework» (FR-IMGT) et « complementarity

determining region » (CDR-IMGT) [19-20]. La description précise de la zone de jonction des gènes V-D-J ou V-J est réalisée à l'aide d'IMGT/JunctionAnalysis [180]. Le programme permet par la comparaison de motifs, de délimiter le peptide signal (localisation du codon d'initiation) et la région constante (localisation du codon stop), les séquences non traduites en 5' et en 3', et les régions codantes composées (par exemple: L-V-D-J-C-REGION ou L-V-J-C-REGION). Dans une troisième étape, la fonctionnalité de la séquence est définie d'après les règles énoncées dans la charte scientifique. Les critères d'obtention de la séquence (origine biologique, méthodologie du concept d'OBTENTION), indiqués par les auteurs, sont ensuite reportés dans l'annotation. L'annotation complète est enfin intégrée dans IMGT/LIGM-DB.

A chaque étape, IMGT/Automat contrôle la signification et la cohérence des résultats. L'annotation est validée, à l'issue de la première étape, si le score d'alignement du gène V dépasse un seuil fixé (un gène V, dans la numérotation IMGT, a une longueur standard, voisine de 330 nucléotides). L'annotation est validée, à l'issue de la deuxième étape (description), si la délimitation des motifs est conforme au prototype des ADNc. Le module de cohérence de la troisième étape vérifie et valide les séquences définies comme 'productive' (les séquences 'unproductive' nécessitent une expertise complémentaire manuelle).

L'annotation des séquences d'ADNc par IMGT/Automat permet de traiter rapidement un grand nombre de ces séquences qui représentent plus de 50% d'IMGT/LIGM-DB. A l'heure actuelle IMGT/Automat, utilisé comme outil interne, annote principalement les ADNc d'homme et de souris. Ainsi sur 18000 séquences d'ADNc annotées dans IMGT/LIGM-DB, 8.000 ont été traitées et validées par IMGT/Automat. Nous estimons à 4.000 le nombre de séquences rejetées par IMGT/Automat pour cause d'incohérence ou besoin d'expertise manuelle complémentaire. IMGT/Automat permet aussi la vérification et la mise à jour des séquences annotées d'ADNc lorsqu'un nouveau gène ou un nouvel allèle a été identifié. IMGT® a adopté pour l'annotation des séquences nucléotidiques de la base de données IMGT/LIGM-DB, une stratégie qui comporte deux approches tenant compte des exigences spécifiques liées à la complexité de la génétique des IG et des TR, et de la nécessité d'une automatisation de plus en plus importante.

De manière remarquable, la qualité de l'annotation automatique des ADNc est équivalente à la qualité d'une annotation manuelle. Cette automatisation pourra être appliquée

à l’annotation des séquences génomiques réarrangées, moyennant la mise en place de contrôles et règles complémentaires.

CHAPITRE 4
LIGMotif

Le système d'annotation des séquences nucléotidiques d'IG et TR intègre les différents logiciels et bases de données du chapitre précédent. Cependant, IMGT® développe des prototypes pour s'améliorer sans cesse. LIGMotif est un logiciel dont l'objectif est d'annoter les séquences génomiques des gènes d'IG et TR. Nous présenterons dans ce chapitre comment les logiciels opérationnels et prototypes sont intégrés dans le système d'annotation des séquences de nucléotides des IG et TR. Nous décrirons ensuite plus en détails le prototype LIGMotif.

4.1 Traitement des séquences d'IG et TR

La première étape du traitement est la récupération des fichiers EMBL via le protocole FTP pour leur intégration dans la base IMGT/LIGM-DB. L'annotateur attribue ensuite à chaque séquence des mots clés qui l'identifient (Figure 4.1). Deux cas se présentent selon qu'il s'agit d'un ADNc ou d'un ADN génomique. Si la séquence est un ADNc, le fichier EMBL est traité par le programme IMGT/Automat qui, comme son nom l'indique, réalise une annotation automatique. L'efficacité d'IMGT/Automat repose sur une annotation préalable complète des séquences génomiques. Les ADNc des espèces autres que l'homme et la souris ne pourront être annotés par IMGT/Automat qu'après annotation des séquences génomiques des espèces correspondantes. Si la séquence est un ADN génomique, les annotateurs réalisent eux-mêmes l'annotation. Dans ce cas, les annotateurs utilisent des logiciels d'aide à l'annotation. Cette activité consiste, entre autres, à prédire les gènes, à les localiser dans la séquence, mais aussi à les décrire. Ainsi, la stratégie d'annotation mise en place par IMGT® comporte deux approches différentes dépendant de la nature des séquences (ADN génomique ou ADNc). Les deux branches du processus produisent au final un fichier IMGT/LIGM-DB [8] qui contient les données provenant de l'annotation automatique ou de l'annotation manuelle. Ce fichier est ensuite traité pour intégrer les informations qu'il contient dans la base de données IMGT/LIGM-DB [8]. L'activité la plus difficile à gérer par les annotateurs est l'annotation manuelle des ADN génomiques.

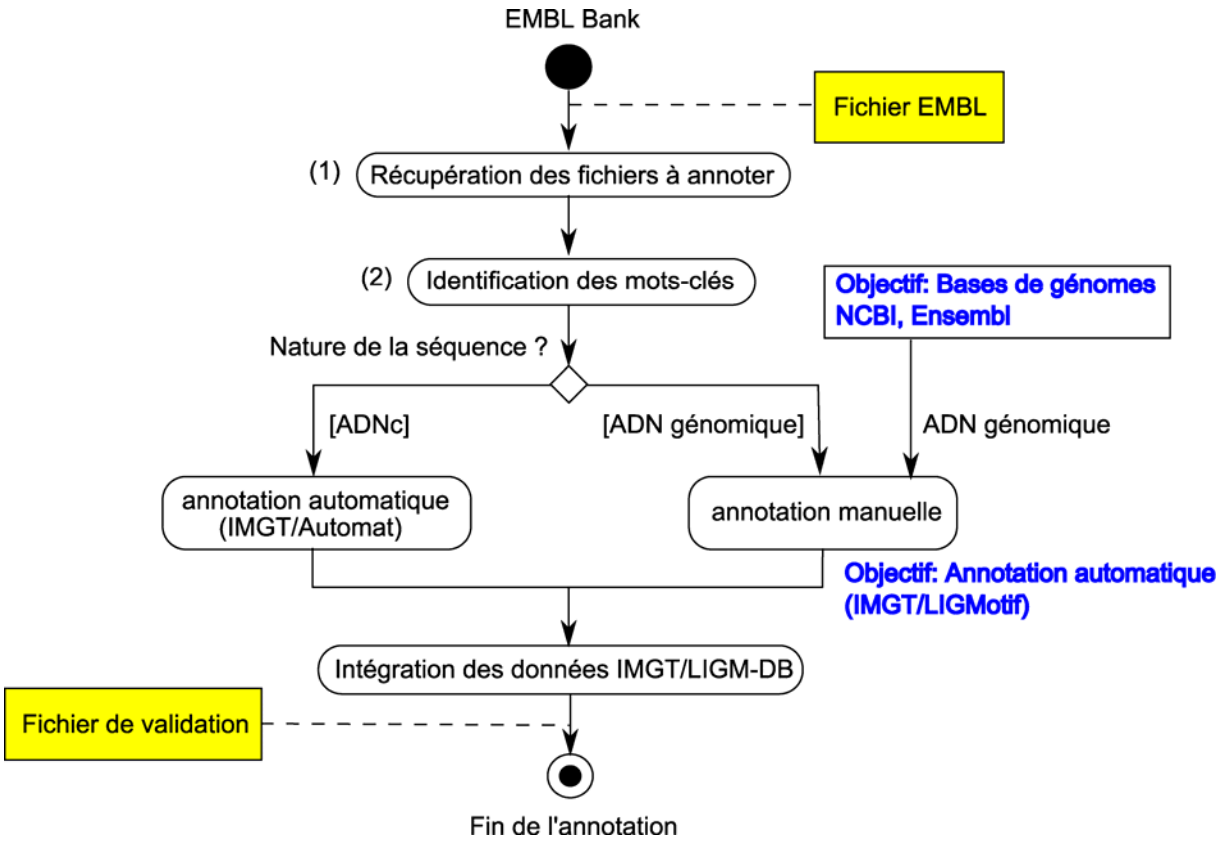


Figure 4.1. Diagramme d'activité du traitement des séquences d'IG et TR. Le système d'annotation des séquences nucléotidiques des IG et TR peut être compartimenté en deux parties. La première concerne l'annotation des séquences d'ADNc qui est complètement automatisée par IMGT/Automat. La deuxième partie concerne l'annotation manuelle des séquences d'ADN génomique et intègre LIGMotif.

La première étape de l'annotation repose sur l'utilisation d'un logiciel d'annotations spécifiques des IG et des TR, dont la dernière version n'est pas stabilisée, LIGMotif, (développé en 1995 par Gérard Ménessier pour venir en aide aux annotateurs d'IMGT®), d'IMGT/V-QUEST [178-179] et de BLASTN [187]. Le développement d'un nouveau programme basé sur LIGMotif, IMGT/LIGMotif, tout en s'acheminant vers une annotation automatique permettrait également d'envisager l'annotation automatique de sources telles que le NCBI (Figure 4.1).

4.2 Processus d'annotation manuelle des séquences génomiques d'IG et TR

L'analyse des activités de l'annotation des séquences génomiques et des fichiers manipulés, a permis de modéliser l'ensemble par le diagramme présenté Figure 4.2.

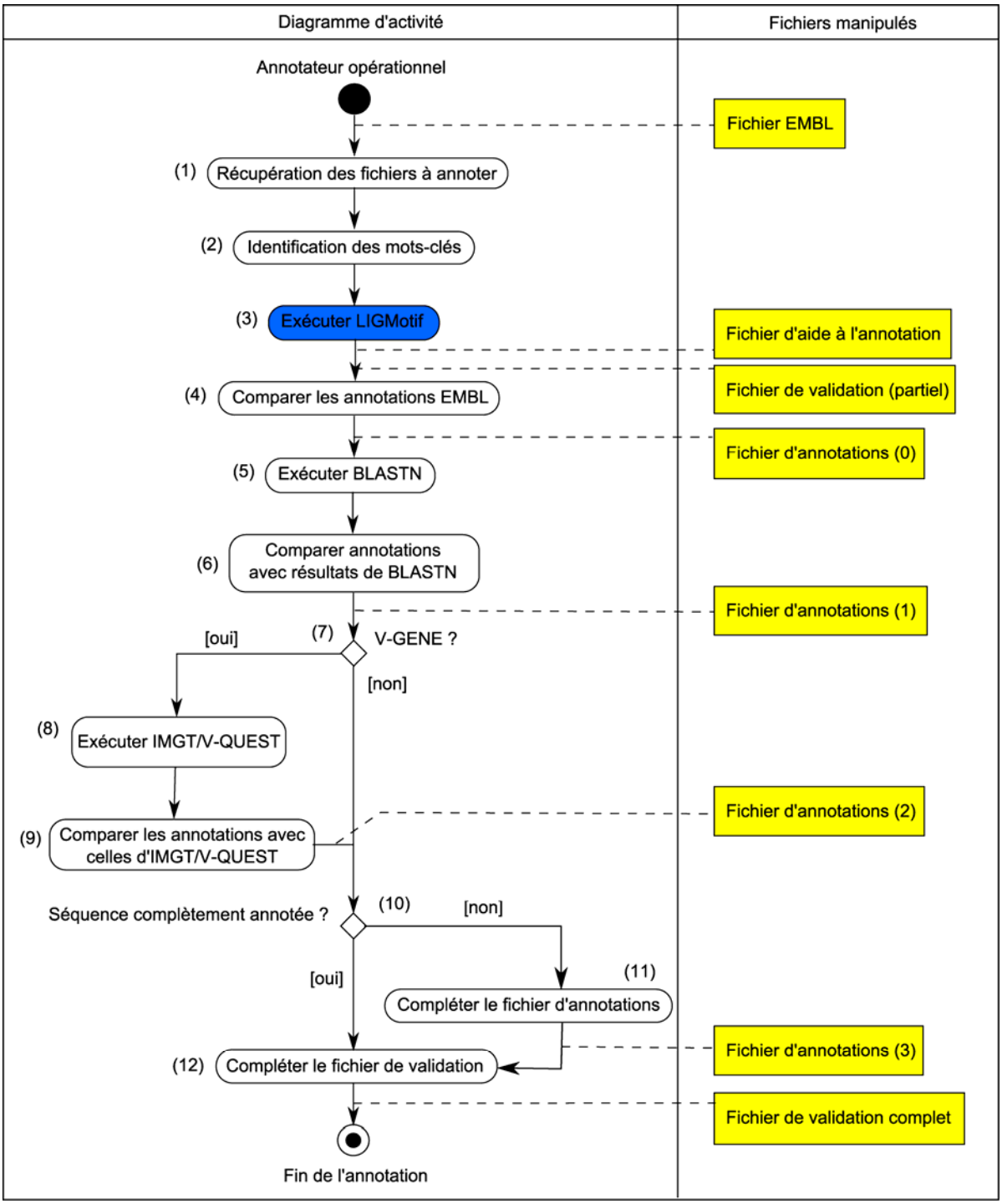


Figure 4.2. Parallèle entre le diagramme d'activité de l'annotation des séquences génomiques (à gauche) et les fichiers résultant du travail des annotateurs (à droite).

Chaque activité du diagramme est détaillée dans le Tableau 4.1.

Tableau 4.1. Détail du diagramme d'activité du processus d'annotation manuelle des séquences génomiques d'IG et TR.

Code	Etape	Commentaire
1	Récupérer le fichier EMBL	L'annotateur récupère les fichiers EMBL via le protocole FTP pour leur intégration dans la base de données IMGT/LIGM-DB.
2	Identifier les mots clés	L'annotateur attribue à la séquence des mots clés qui l'identifient.
3	Exécuter LIGMotif	L'annotateur exécute le logiciel LIGMotif qui identifie et décrit les gènes d'IG et TR dans la séquence analysée. Le programme produit eux fichiers. Un fichier de validation partiel et un fichier « d'aide à l'annotation » contenant les annotations des gènes qu'il prédit.
4	Comparer les annotations EMBL	L'annotateur compare les annotations qui sont contenues dans le fichier EMBL avec les annotations des gènes prédits par LIGMotif. Ensuite l'annotateur inscrit les annotations non redondantes dans un fichier d'annotations (0).
5	Exécuter BLASTN	L'annotateur exécute BLASTN à partir du Web Service d'IMGT en sélectionnant les paramètres requis pour décrire la séquence analysée et classifier les gènes qu'elle contient.
6	Comparer les annotations avec le résultat de BLASTN	L'annotateur compare les informations provenant des résultats de BLASTN avec celles contenues dans le fichier d'annotations. Il modifie les informations si nécessaire. (fichier d'annotation (1))
7	Contrôler le type de gènes	L'annotateur identifie les types des gènes contenus dans le fichier EMBL. Si le gène est du type V-GENE, V-D-J-GENE ou V-J-GENE, l'annotateur réalise l'activité 8. Sinon il passe à l'étape 10.
8	Exécuter IMGT/V-QUEST	L'annotateur exécute IMGT/V-QUEST à partir du Web Service d'IMGT pour décrire, identifier, numéroter et classifier les V-GENE.
9	Comparer les annotations avec celles de IMGT/V-QUEST	L'annotateur compare les annotations obtenues grâce à IMGT/V-QUEST avec celles qui sont contenues dans le fichier d'annotations. Il modifie les informations si nécessaire. (fichier d'annotations (2)).
10	Contrôler les annotations	L'annotateur vérifie que toute la séquence est complètement annotée en analysant le fichier d'annotations. Si la séquence n'est pas complètement annotée, l'annotateur réalise l'activité 11. Sinon il réalise l'activité 12.
11	Compléter le fichier d'annotations	L'annotateur termine lui-même les annotations en précisant la méthode utilisée pour obtenir la séquence (fichier d'annotations (3)).
12	Compléter le fichier validation	L'annotateur complète le fichier de validation provenant de LIGMotif en y répertoriant les informations récupérées lors de tous les processus c'est-à-dire celles qui sont répertoriées dans le fichier d'annotations (fichier de validation complet).

Les numéros de code correspondent aux étapes du diagramme d'activités.

4.3 Analyse de LIGMotif: déroulement général de LIGMotif

L'utilisation de LIGMotif est simple et s'effectue en ligne de commande. Cependant, la présentation de ses résultats est peu conviviale. En effet, les résultats sont contenus dans de simples fichiers texte et certaines données fournies par LIGMotif n'ont a priori aucune signification pour les annotateurs. Il n'y a pas de documentation sur ces données et sur les 143 fichiers sources de programme en C qui composent LIGMotif.

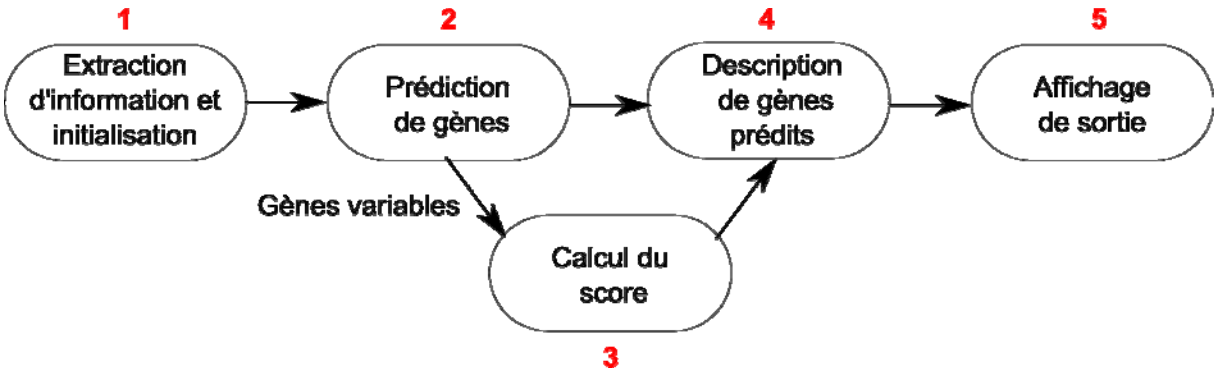


Figure 4.3. Etapes de LIGMotif. (1) Le module d'extraction d'information et d'initialisation permet de récupérer la séquence à analyser et de cibler l'ensemble des motifs à utiliser pour annoter la séquence. (2) Le module de prédiction de gènes permet d'identifier les gènes par la recherche de leurs motifs. (3) Les gènes variables sont soumis à un module spécifique dans lequel un score est calculé dans la V-REGION et en déduit la vraisemblance du gène variable prédit. (4) Le module de description permet de décrire les gènes prédits et (5) le module d'affichage permet d'afficher les résultats de la description des gènes ainsi que les scores.

La première étape effectuée par LIGMotif est l'extraction d'informations contenues dans le fichier EMBL. Ces informations vont permettre de cibler les gènes à rechercher dans la séquence. La deuxième étape, la prédiction (localisation) de gènes, se base sur une recherche de motifs caractérisant les gènes recherchés. Dans le cas où des gènes variables sont prédits, un score est calculé. Cette étape donne une idée de la pertinence de la prédiction. Puis, les gènes sont décrits, c'est-à-dire structurés en fonction des concepts de description d'IMGT-ONTOLOGY. Finalement, les fichiers d'aide à l'annotation et de validation sont produits à l'étape d'affichage de sortie. Toutes ces fonctionnalités ont été analysées précisément dans le but d'obtenir les informations nécessaires au développement d'un nouveau programme.

CHAPITRE 5
IMGT/LIGMotif

La modélisation du système d'annotation des IG et TR d'IMGT® et l'analyse du logiciel LIGMotif ont permis de poser les bases pour le développement de IMGT/LIGMotif. A notre connaissance, aucun logiciel identifiant les gènes codants ne sont adaptés à l'annotation des gènes V, D et J. Les méthodes existantes sont basées sur un modèle de gène conventionnel, ce qui exclut la détection des signaux de recombinaison (RS). IMGT/LIGMotif sera donc conçu à partir du système d'annotation des IG et TR en intégrant IMGT/V-QUEST, BLASTN et la recherche du pattern des types de gènes qui est inspirée de LIGMotif. Nous verrons dans ce chapitre le modèle d'IMGT/LIGMotif, son algorithme et son évaluation.

5.1 Modèle

5.1.1 Prototypes, labels et patterns

IMGT/LIGMotif a pour objectif d'identifier les instances V-gene, D-gene et J-gene (instances du concept 'Molecule_EntityType') et de décrire les gènes V, D et J correspondant dans de larges séquences génomiques. Ainsi, les V-gene, D-gene et J-gene identifiés sont respectivement décrits par trois instances du concept 'Molecule_EntityPrototype': V-GENE, D-GENE, J-GENE et C-GENE. Leur représentation graphique, ou prototype, et les labels qui les décrivent sont illustrés dans les Figure 5.1, Figure 5.2 et Figure 5.3. Parmi les 242 labels IMGT® définis pour les séquences nucléotidiques, 47 sont utilisés pour représenter ces prototypes (Figure 5.1B, Figure 5.2B et Figure 5.3B). De ces 47 labels 43 sont spécifiques d'un prototype (23 pour un V-GENE, 11 pour un D-GENE et 9 pour 1 J-GENE). Deux labels (5'UTR et 3'UTR) sont communs à tous les prototypes, alors que 2 labels (ACCEPTOR-SPLICE et DONOR-SPLICE) sont partagés par plusieurs prototypes (associés à des ronds noirs dans la Figure 5.1B, Figure 5.2B et Figure 5.3B).

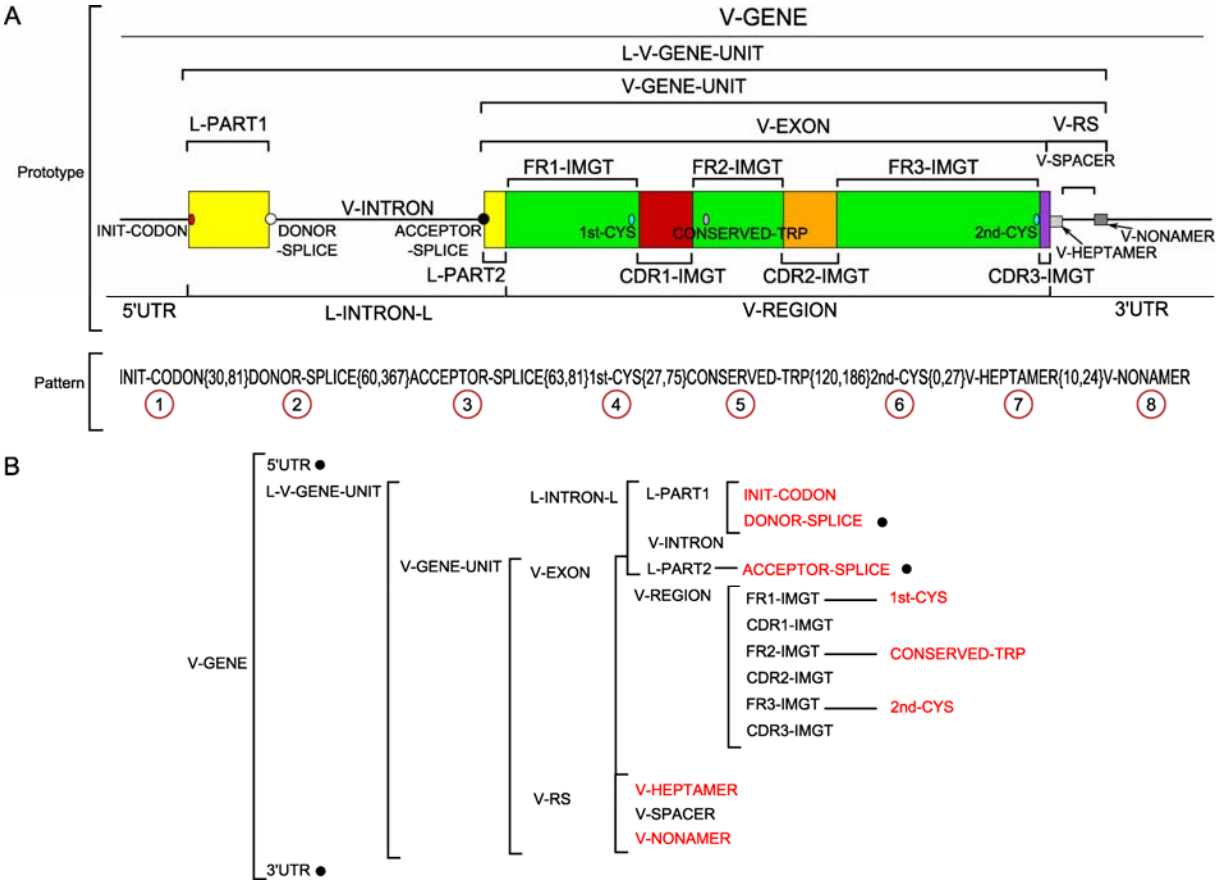


Figure 5.1. Prototypes, labels et patterns d'un V-GENE. A. Le prototype V-GENE est représenté avec ses labels et son pattern. B. Présentation des labels d'un V-GENE. Le L-V-GENE-UNIT est décrit par 24 labels (22 spécifiques et 2 partagés). Trois labels additionnels dont un spécifique et 2 autres communs (5'UTR et 3'UTR), permettent de décrire le V-GENE (27 labels). Les labels partagés et communs sont représentés en cercles noirs et les labels représentant les motifs conservés en rouge.

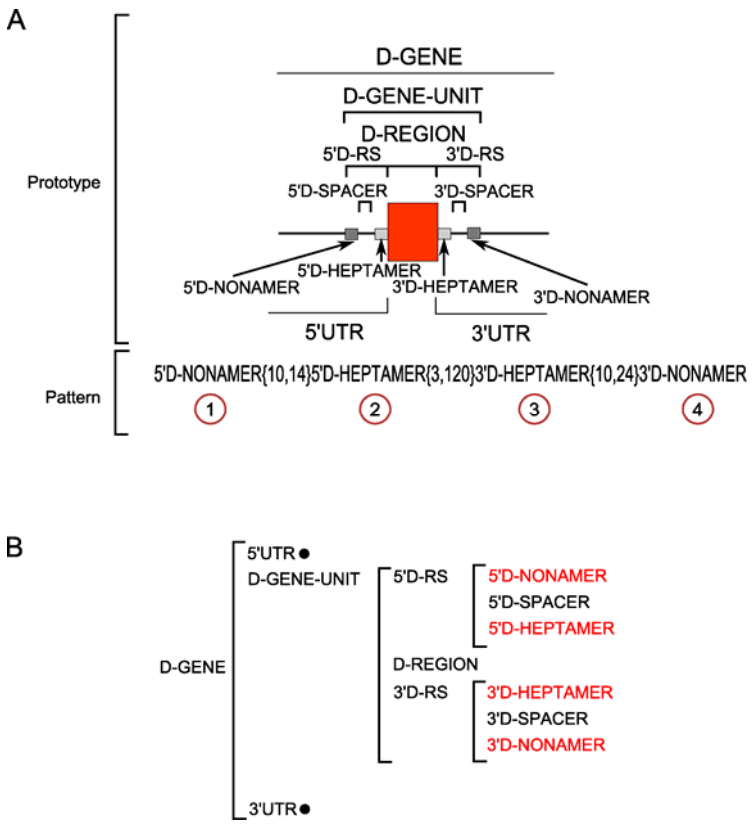


Figure 5.2. Prototypes, labels et patterns d'un D-GENE. A. Le prototype D-GENE est représenté avec ses labels et son pattern. B. Présentation des labels d'un D-GENE. Le D-GENE-UNIT est décrit avec 10 labels (tous spécifiques). Trois labels additionnels dont un spécifique et 2 autres communs (5'UTR et 3'UTR), permettent de décrire le D-GENE (13 labels). Les labels partagés et communs sont représentés en cercle noirs et les labels représentant les motifs conservés en rouge.

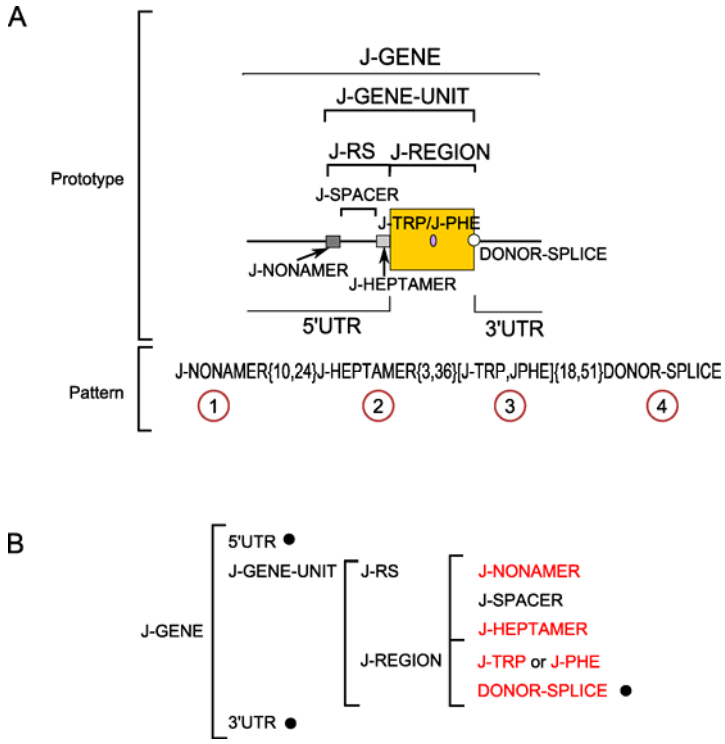


Figure 5.3. Prototypes, labels et patterns d'un J-GENE. A. Le prototype J-GENE est représenté avec ses labels et son pattern. B. Présentation des labels d'un J-GENE. Un J-GENE-UNIT est décrit par 8 labels (7 labels spécifiques, J-PHE et J-TRP sont mutuellement exclusifs et un unique label partagé). Trois labels additionnels dont un spécifique et 2 autres communs (5'UTR et 3'UTR), permettent de décrire le J-GENE (11 labels). Les labels partagés et communs sont représentés en cercles noirs et les labels représentant les motifs conservés en rouge.

Le modèle intègre les patterns des gènes V, D et J, tout comme LIGMotif. Le pattern peut être considéré comme un assemblage de séquences consensus ou motifs. Il considère deux séquences S1 et S2 avec S1 = {ACATTTACGGGGT} et S2 = {AAGTTAAACGGGGT} appartenant à la même famille. Les tirets qui séparent certains nucléotides dans les séquences alignées sont appelés gaps. Ils sont placés de manière à garder le maximum de similarité entre les séquences. L'alignement fait ressortir deux motifs entre les séquences. Le motif 1 « **TT** » et le motif 2 « **ACGGGGT** ».

```
S1     : (5′) A C A - T T - T A C G G G G T (3′)
S2     :      A - A G T T A A A C G G G G T
motifs :             T T     A C G G G G T
S3     :      A - A G T T - - - C G G G G T
motifs :             T T       C G G G G T
S4     :      A - A G T T - - - C C G G C T
motifs :             T T       C . G G . T
```

Figure 5.4. Alignement de deux séquences nucléotidiques.

Les motifs sont ordonnés selon une ***syntaxe*** qui correspond à l'ordre dans lequel apparaissent les motifs pour identifier un gène. Comme la lecture d'un gène se fait **toujours** dans le sens 5' (c'est le nom que l'on donne conventionnellement à l'extrémité gauche d'une séquence) vers 3', le motif 1 est toujours avant le motif 2. Les gaps font varier les ***intervalles*** de distances de nucléotides entre les motifs. En effet, le motif 1 et le motif 2 de S1 sont séparés par un nucléotide alors que les deux motifs de S2 sont séparés par un intervalle de deux nucléotides. Ce qui nous donne un intervalle de distances de nucléotides minimal de 1 et maximal de 2.

Comme nous venons de le voir, les motifs sont associés à une syntaxe et à des intervalles. Cet ensemble peut être représenté informatiquement par des expressions dites régulières. On parle aussi de « ***pattern*** »: c'est d'ailleurs le terme qui sera utilisé par la suite. Le pattern désignant les séquences S1 et S2 est « **TT**{1,2}**ACGGGGT** ». Nous retrouvons les deux motifs en rouge ainsi que l'intervalle de distances minimum et maximum, respectivement de 1 et de 2, et la syntaxe où le motif 1 précède le motif 2. Le ***lexique*** de ce pattern correspond à l'ensemble des motifs qu'il contient. En l'occurrence il s'agit des motifs 1 et 2 des séquences. Maintenant, si l'on ajoute une nouvelle séquence S3 appartenant à la même famille telle que S3 = {AAGTTCGGGGT}, alors le pattern défini ne correspond plus aux trois séquences car il est trop restrictif. Le nouveau pattern représentatif des trois séquences est « **TT**{1,3}**CGGGGT** ». Ce pattern peut aussi être représenté par un autre tel que « **TT**{1,3}**C.GG.T** ». Le point du pattern représente n'importe quel acide nucléotidique. Les séquences (ici S4 = {AAGTTCCGGCT}) appartenant à la même famille que les autres séquences peuvent être détectées par ce pattern. Mais il faut éviter de rendre le motif trop permissif car il peut détecter certaines séquences qui n'appartiennent pas à la même famille de séquences.

Les unités de gènes L-V-GENE-UNIT (Figure 5.1), D-GENE-UNIT (Figure 5.2) et J-GENE-UNIT (Figure 5.3) ont été créées afin de délimiter précisément les extrémités en 5' et 3' des entités par les extrémités 5' et 3' de labels codants (L-PART1 et V-EXON pour les gènes V, D-REGION pour les gènes D, J-REGION pour les gènes J). De plus, la partie des prototypes que ces labels recouvrent peut être définie par un pattern (Figure 5.1A, Figure 5.2A, Figure 5.3A). Les motifs conservés sont séparés les uns des autres par une distance en paire de bases comprise entre une longueur minimale et maximale (entre accolades dans les Figure 5.1A, Figure 5.2A et Figure 5.3A). Les motifs sont ordonnés de l'extrémité 5' à 3'

selon leur rang (dans un cercle Figure 5.1A, Figure 5.2A et Figure 5.3A) qui correspond à leur localisation relative dans un pattern. Le motif le plus en 3' possède un rang égal au nombre de motifs dans le pattern, c'est-à-dire 8 pour les V, 4 pour les gènes D et 4 pour les gènes J. Les deux motifs J-TRP et J-PHE peuvent être retrouvés pour le même rang (ils sont ensuite représentés entre accolades dans le pattern et séparés par une virgule). Les acides aminés conservés J-TRP et J-PHE font partie d'un motif conservé '{W,F}{G,A}XG' où W=tryptophane (J-TRP), F=phénylalanine (J-PHE), G=glycine, A=alanine et X=n'importe quel acide aminé excepté la proline.

5.1.2 Matrice de scores position spécifique

Les matrices de scores position spécifique [188] (position-specific scoring matrix, PSSM) sont utilisées pour reconnaître les heptamères et les nonamères des signaux de recombinaison, si nécessaire. Une PSSM est une matrice à deux dimensions dans laquelle chaque ligne correspond à un des 4 nucléotides et chaque colonne à une position du site décrit. Les éléments de la matrice sont la fréquence de chaque nucléotide à chaque position calculée à partir d'un alignement de séquences d'ADN du site (Figure 5.5).

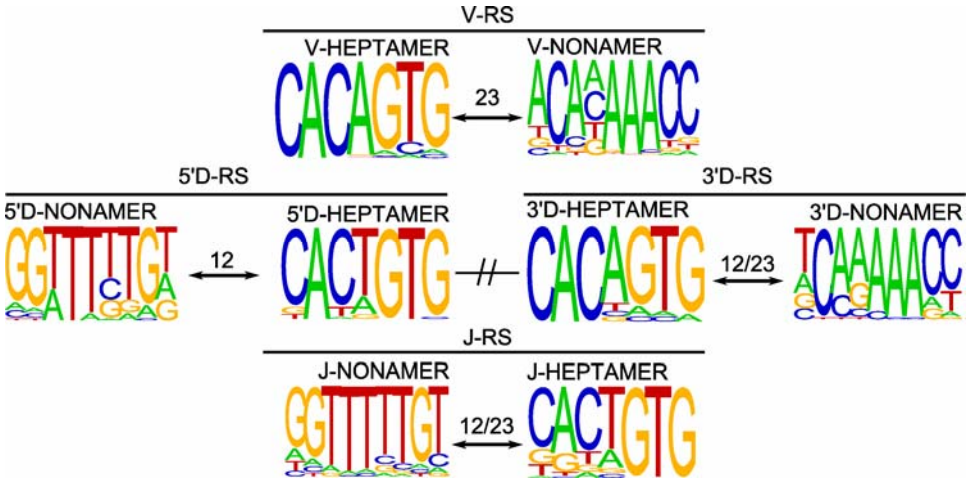


Figure 5.5. Représentation en logo des signaux de recombinaison (RS) des gènes d'IG et TR chez l'homme. Les logos ont été générés à l'aide du programme WebLogo [189]. Un logo représente chaque colonne d'un alignement des heptamères et nonamères par pile de lettres, à la hauteur de chaque lettre proportionnelle à la fréquence observée du nucléotidique correspondant à cette position. Les flèches numérotées représentent les espaceurs et leur longueur respective en nucléotides (12 ou 23 nucléotides (12/23), selon les locus).

Une fonction prenant en paramètre les fréquences de la matrice est appliquée sur une fenêtre glissante, dont la taille est celle du site (soit 7 pour les heptamères, soit 9 pour les nonamères), le long d'une séquence pour déterminer la probabilité qu'elle soit un motif (un heptamère ou un nonamère en fonction de la matrice utilisée). Cette fonction calcule un score

en additionnant la fréquence de chaque nucléotide de la matrice retrouvé dans la séquence. De manière plus précise le score peut se calculer sous la forme du logarithme de probabilités (log-likelihoods):

$$S = \sum_{i=1}^{L} \log(f_{i,A_i})$$

où *A* est une séquence à pondérer, *L* est sa longueur en nucléotides et f_{i,A_i} est la fréquence du $i^{ème}$ nucléotide dans *A* à la position *i* de la PSSM.

Une PSSM peut être utilisée pour scanner l'ADN génomique et déterminer statistiquement les appariements significatifs. C'est le score qui permet d'évaluer la significativité d'une séquence. C'est pourquoi, un seuil score définissant un site fonctionnel est déterminé en jouant avec le nombre de faux positifs et de faux négatifs.

5.1.3 Organisation du modèle

Le modèle d'IMGT/LIGMotif se compose de 4 modules (Figure 5.6). Trois d'entre eux ('Identification de gène(s)', 'Description de gène(s)' et 'Identification de la fonctionnalité') tiennent compte des caractéristiques des gènes d'IG et TR, pour traiter les unités de gène, tandis que le quatrième module ('Délimitation des gènes et leur assemblage en cluster') traite de la délimitation des gènes et de l'assemblage de gènes dans un cluster, et fournit une séquence génomique annotée.

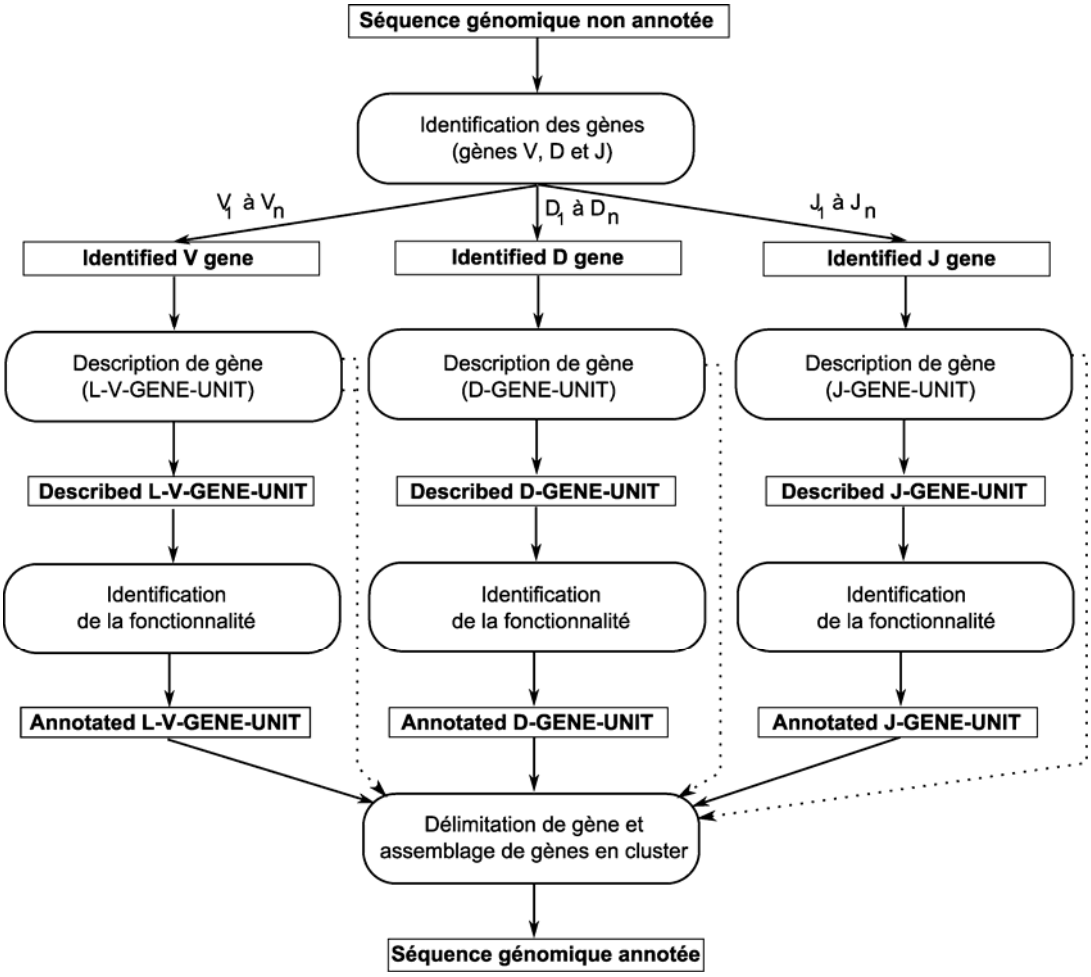


Figure 5.6. Aperçu du modèle d'IMGT/LIGMotif. Les 4 modules d'IMGT/LIGMotif comprennent 'Identification des gènes', 'Description de gène', 'Identification de la fonctionnalité' et 'Délimitation de gène et assemblage des gènes en cluster'. Les pointillés indiquent les sorties partiellement décrites et non décrites (pour les gènes V, D et J) du module 'Description de gène' lesquelles sont réentrées, dans le dernier module pour leur inclusion dans la description finale de la séquence.

5.1.3.1 Identification des gènes

Le module 'Identification de gène(s)' identifie les gènes V, D et J potentiels le long de la séquence génomique analysée. Tout d'abord, une recherche heuristique des alignements locaux est effectuée contre des bases de données de motifs de référence incluses dans IMGT/LIGMotif (Tableau 5.1). Ces bases de données comprennent des séquences de nucléotides qui correspondent aux labels d'unité de gène d'IG et TR (L-V-GENE-UNIT, D-GENE-UNIT, J-GENE-UNIT) et à des motifs qui les composent. Ces bases de données sont créées dynamiquement à partir des séquences de IMGT/LIGM-DB [8], en utilisant l'interface IMGT/GENE-DB [31] qui permet d'effectuer des requêtes par un à plusieurs labels. Des séquences d'IG et TR correspondant à 34 labels ont été récupérées et proviennent de l'homme (*Homo sapiens*) et de la souris (*Mus musculus*; quelques séquences de *Mus Pahari*, *Mus*

saxicola, *Mus spretus* ont également été incluses). Les pseudogènes qui sont trop peu conservés pour pouvoir être affectés à un sous-groupe ont été exclus.

Les alignements obtenus dans cette première étape fournissent des paires de segments à score élevé (en anglais, High-scoring Segment Pairs ou HSP) pour un label particulier. Ensuite, les HSP labellisés sont sélectionnés afin de ne garder que les HSP les plus représentatifs d'un gène. Pour ce faire, ces HSP sont groupés en prenant en compte leur type de gène (V, D ou J) et leur topologie (localisation dans la séquence et dans le prototype). Ainsi, le module 'Identification des gènes' fournit les gènes V, D et J potentiels des gènes identifiés comme des groupes de HSP labellisés le long de la séquence HSP analysée.

Tableau 5.1. Bases de données de référence de motifs d'IMGT/LIGMotif.

Prototype	Nombre de bases de données	Base de données de motifs de références	Nombre de séquences		Identification des gènes (HSP provenant du Blast)	Description des gènes et identification de la fonctionnalité
			Homme	Souris		
V-GENE	16	L-V-GENE-UNIT	368	204	+	
		V-GENE-UNIT	385	221	+	
		L-PART1	534	550	+	
		V-INTRON	575	518	+	
		L-INTRON-L	470	288	+	
		V-EXON	660	550	+	+
		L-PART2	591	580	+	
		V-REGION	887	1036	+	
		FR1-IMGT	785	730	+	
		CDR1-IMGT	790	731	+	
		FR2-IMGT	790	739	+	
		CDR2-IMGT	788	737	+	
		FR3-IMGT	788	737	+	
		CDR3-IMGT	674	607	+	
		V-RS	378	222	+	
		V-SPACER	485	449	+	
	2	V-HEPTAMER	326	317		+
		V-NONAMER	262	298		+
D-GENE	6	D-GENE-UNIT	36	28	+	
		5'D-RS	40	29	+	
		5'D-SPACER	50	29	+	
		D-REGION	50	38	+	
		3'D-RS	36	32	+	
		3'D-SPACER	47	32	+	
	4	5'D-NONAMER	36	22		+
		5'D-HEPTAMER	36	22		+
		3'D-HEPTAMER	36	21		+
		3'D-NONAMER	33	21		+
J-GENE	4	J-GENE-UNIT	120	107	+	
		J-RS	120	107	+	
		J-SPACER	120	108	+	
		J-REGION	130	121	+	
	2	J-NONAMER	101	76		+
		J-HEPTAMER	101	77		+

5.1.3.2 Description des gènes

Le module ‘Description des gènes’ fournit la description de chacun des gènes identifiés dans le premier module. Il comprend une recherche de motifs conservés fondée sur des prototypes et des modèles. Les codons d’acides aminés conservés dans les patterns (‘tgg’ pour CONSERVED-TRP et J-TRP, ‘tgc’ et ‘tgt’ pour la 1st-CYS et 2nd-CYS, et ‘ttt’ et ‘ttc’ pour J-PHE) sont difficiles à identifier par des algorithmes classiques puisque les motifs (triplets ‘tgg’, ‘tgc’, ‘tgt’, ‘ttt’ et ‘ttc’) sont très fréquents dans les séquences. Pour cette raison, les codons des acides aminés conservés de la V-REGION et du V-EXON (qui comprend 1st-CYS, CONSERVED-TRP et 2nd-CYS) sont identifiés par le logiciel IMGT/V-QUEST [30]. Le résultat attendu du module ‘Description des gènes’ décrit les unités de gène, bien que, comme discuté dans la section algorithme, des unités de gène partiellement décrites et non décrites peuvent également être obtenues.

5.1.3.3 Identification de la fonctionnalité

Le module ‘Identification de la fonctionnalité’ inclut le contrôle des éléments nécessaires à l’attribution de la fonctionnalité et permet d’obtenir des unités de gènes annotées. La fonctionnalité des gènes peut être identifiée pour les gènes V, D et J non réarrangés à partir des règles et des concepts de l’IMGT-ONTOLOGY. Un gène est qualifié de fonctionnel si la région codante possède un cadre de lecture ouvert sans codon stop et s’il n’y a pas de défaut(s) décrit(s) dans les sites d’épissage, les RS et/ou les éléments de régulation. Un gène est qualifié d’ORF si la région codante possède un cadre de lecture ouvert, mais des altérations ont été décrites dans les sites d’épissage, les RS et/ou les éléments de régulation et/ou des substitutions entre acides aminés ont été suggérées par les auteurs pouvant entraîner une structure incorrecte et/ou une entité « est » orphon. Un gène est qualifié de pseudogène si la région codante contient des codons stop et/ou des insertions ou délétions entraînant un saut du cadre de lecture. En particulier, un V-GENE (ou L-V-GENE-UNIT) est considéré comme pseudogène si ces défauts sont dans le L-PART1 et/ou V-EXON, ou s’il y a une mutation dans l’INIT-CODON (‘atg’) du L-PART1. Un J-GENE (ou J-GENE-UNIT) est considéré comme un ‘pseudogène’ s’il a été identifié par la présence d’un SR en amont du cadre de lecture ouvert, mais n’a pas de site donneur d’épissage en 5’ ou le site donneur d’épissage n’a pas le cadre d’épissage sf1 ou s’il n’a pas le motif conservé ‘{W,F}{G,A}XG’.

5.1.3.4 Délimitation des gènes et assemblage en cluster

Dans ce dernier module, les gènes (V-GENE, D-GENE et J -GENE) sont délimités et assemblés en un cluster si la séquence génomique analysée contient plusieurs gènes. Le résultat final d'IMGT/LIGMotif est la séquence génomique annotée.

5.2 Algorithme

5.2.1 Extraction d'informations

La séquence contenue dans le fichier au niveau du repère « SQ » (Figure 5.7) est extraite par IMGT/LIGMotif pour les fichiers EMBL mais cette fonctionnalité est aussi présente pour les séquences au format FASTA.

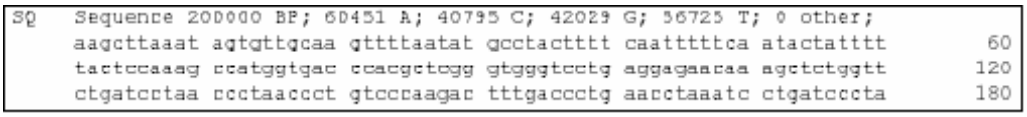


Figure 5.7. Portion du fichier EMBL correspondant à une partie de la séquence analysée.

5.2.2 Identification des gènes V, D et J

5.2.2.1 BLAST

L'algorithme commence par l'alignement de la séquence génomique en utilisant BLASTN [187]. C'est une heuristique de l'alignement de Smith-Waterman qui permet d'identifier efficacement dans une séquence d'ADN des exons ou d'autres séquences conservées à partir d'une base de séquences de référence de protéines ou de nucléotides. Le point clef de cette méthode repose sur le découpage d'une séquence analysée en mots de petite taille (quelques acides aminés ou nucléotides) identique et l'indexation de leurs position. Une séquence de taille L contient un nombre de mots w maximum égal à L-w+1. Les mots dont le potentiel de score (on se place dans le cas ou le mot s'est apparié avec une séquence identique) est supérieur à un seuil déterminé empiriquement sont indexés dans une liste. La base de données est fouillée pour retrouver les appariements des mots de la liste. Les appariements des mots dans les séquences de référence vont permettre de cibler les régions les plus conservées à partir desquelles les alignements pourront être étendus. Seuls les ensembles d'au moins deux hits (un hit est un séparés par une distance maximale donnée et qui doivent être sur la même diagonale d'une matrice de points (technique de « dot plot ») sont utilisés par BLAST. Les mots vont servir de base pour étendre l'alignement d'un côté et de l'autre. L'extension s'arrête lorsque le score de l'alignement devient inférieur à un seuil de confiance

qui correspond au score maximum rencontré lors de l'alignement diminué d'une valeur de tolérance définie arbitrairement. Les alignements sont filtrés de manière à ne conserver que les plus significatifs qui doivent posséder un score supérieur à un seuil. Les alignements sélectionnés sont appelés HSP. La position des HSP provenant des deux mêmes séquences doit être consistante. Le cas idéal correspond à l'alignement des HSP sur une même diagonale dans une matrice de points. Les HSP ne doivent pas se chevaucher. Deux HSP non chevauchants, mais dont l'un est placé, avant l'autre dans une séquence et inversement dans l'autre, sont considérés comme inconsistants. En effet, la région 5' d'une séquence codante devrait s'aligner avec la partie N-terminale d'une protéine et la région 3' avec la partie C-terminale et pas l'inverse.

La vraisemblance de chaque alignement est évaluée statistiquement en calculant leur E-value et leur P-value. La probabilité que l'alignement produit par le programme soit dû au hasard pour un score donné, est représentée par le paramètre E-value, alors que la P-value est la probabilité de trouver au moins un alignement avec un score supérieur ou égal à celui dont on teste la vraisemblance. Le calcul vérifie que la séquence analysée n'a pas de similarité fortuite avec les séquences de la base de données. La fréquence des acides aminés ou nucléotides présents dans la base de séquences de référence a donc une influence sur l'E-value. Plus la E-value est faible plus l'alignement est vraisemblable. Des valeurs inférieures à 10^{-20} sont hautement significatives et des valeurs, des valeurs supérieures à 10^{-5} nécessitent un contrôle. Les alignements vraisemblables sont appelés 'high scoring segment pair' (HSP) car ils possèdent une E-value ou P-value supérieure à une valeur qui peut être fixée par l'utilisateur.

5.2.2.2 Recherche des alignements labellisés

Les alignements du BLAST se font avec les bases de données de motifs de référence d'IMGT/LIGMotif (Tableau 5.1). La possibilité est donnée au biocurateur de sélectionner les bases de données en fonction de l'espèce (humain et/ou souris), du locus (IGH, IGK, IGL, TRA, TRB, TRG et/ou TRD), du type de gène (V, D et/ou J), de la fonctionnalité (F, P et/ou ORF), et de choisir n'importe quelle combinaison de cette sélection. IMGT/BLAST fournit des HSP qui informent sur la similarité de la séquence analysée (query) avec les motifs labellisés de la base de données de référence (subject). Ces HSP labellisés sont obtenus sur les deux brins de l'ADN de la séquence analysée. Pratiquement, NCBI-BLASTN (version 2.2.18)

est utilisée avec la longueur minimum possible du mot hit (ici 4), une E-value seuil de 0,01 (sauf pour les D-GENE-UNIT et leurs motifs, où une E-value seuil de 5 est utilisée en raison de leur très courte longueur). Le BLAST a été préféré aux méthodes basées sur le modèle de Markov caché (HMM) et aux logiciels tels que HMMER [190-191] pour des raisons pratiques. En effet, les HMM nécessitent des alignements multiples de séquences de référence avant de construire le modèle ce qui aurait pris un temps considérable en raison du nombre élevé de séquences de référence. Cependant, les HMM ne sont pas à exclure car ils sont plus sensibles que le BLAST.

5.2.2.3 Sélection des HSP labellisés

IMGT/BLAST produit une énorme quantité de HSP mais tous n'ont pas la même importance. Les HSP obtenus avec les différentes bases de données de motifs de référence peuvent se chevaucher, car ils sont les composants d'un même prototype (par exemple, un V-EXON chevauchant une V-REGION). Ces chevauchements de HSP n'ont pas besoin d'être filtrés puisqu'ils délimitent des labels différents et attendus. En raison de la duplication des gènes dans les locus IG et TR, les HSP obtenus avec la même base de données de motifs de références peuvent se chevaucher à un même endroit, alors que les HSP appartiennent à différents gènes. En conséquence, la qualité des HSP est évaluée pour une région donnée sur le score, E-value, la longueur et le pourcentage d'identité du BLAST. La méthode qui filtre les HSP se chevauchant et obtenus avec la même base de données de motifs de référence est décrite de la façon suivante: *score(), length(), identity(), evalue()* sont les fonctions qui retournent 1 si un HSP donné (hsp1) est plus performant que l'autre (hsp2) pour les paramètres testés (le score, la longueur et le pourcentage d'identité les plus élevés et la valeur de E-value la plus faible), 0 si les deux HSP sont égaux et -1 si hsp1 est moins performant que hsp2. Le paramètre g1 (somme de *score()* et *evalue()*, avec $-2 \leq g1 \leq 2$) et le paramètre g2 (somme de *length()* et *identity()*, avec $-2 \leq g2 \leq 2$) sont calculés entre deux HSP se chevauchant et obtenus avec la même base de données de motifs. Si $g1>0$, ou si $g1 = 0$ et $g2>0$, hsp2 est supprimé. Dans les autres cas ($g1<0$, ou $g1= 0$ et $g2<0$), hsp1 est supprimé.

5.2.2.4 Groupement des HSP sélectionnés en gènes V, D ou J

L'objectif de cette étape est de grouper les HSP sélectionnés qui peuvent appartenir à un même gène. A cette fin, les positions de certains HSP provenant du même brin d'ADN et avec le même label d'un même type de gène (V, D ou J) sont, à cette étape, comparés les uns

aux autres (Figure 5.8). Si la relation topologique entre deux HSP comparés n'est pas cohérente, les HSP sont considérés comme appartenant à différents gènes. Si la relation topologique est cohérente, une longueur arbitraire, spécifique à chaque label, est ajoutée aux deux extrémités de chaque HSP comparé (Figure 5.8) et ces nouvelles régions sont examinées en vue d'un éventuel chevauchement. Si un chevauchement est découvert, les deux HSP sont considérés comme appartenant à un même gène. Sinon, les deux HSP sont considérés comme appartenant à différents gènes. Pour un groupe de HSP, la position la plus en 5' et la position la plus en 3' définissent la zone qui contient un gène potentiel (indiqué par des flèches dans la Figure 5.8). Les HSP labellisés qui composent ce groupe définissent le type du gène (V, D ou J) et, à l'aide de leur localisation sur les brins d'ADN, l'orientation du gène. Chaque groupe de HSP correspond à un gène potentiel identifié le long de la séquence (indiqué par V_1 à V_n, D_1 à D_n et J_n à J_1 dans la Figure 5.6).

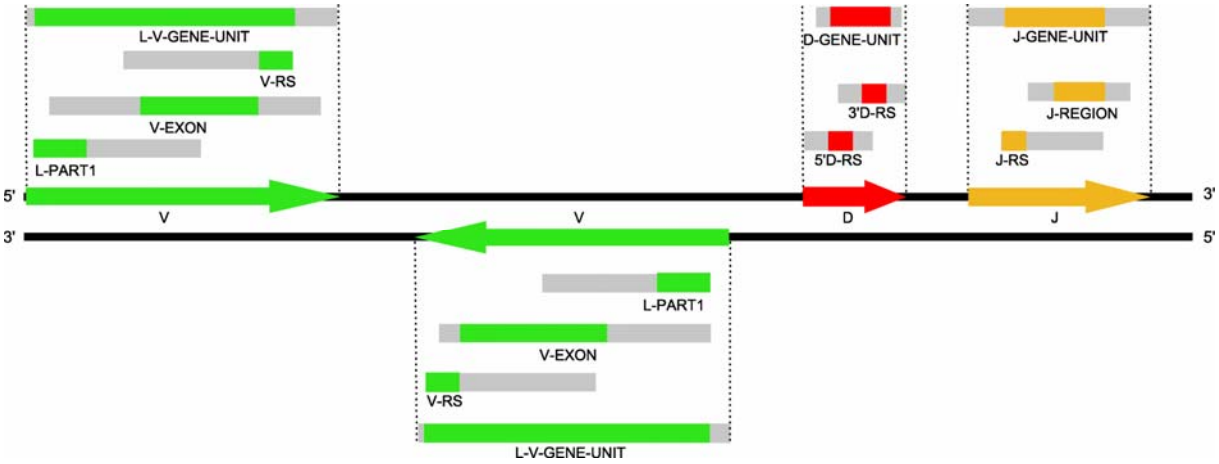


Figure 5.8. Groupement des HSP labellisés et sélectionnés en gènes V, D et J. Les HSP labellisés et sélectionnés sont groupés en gènes V, D et J. Les rectangles en gris représentent la longueur ajoutée en 5' et 3' des HSP. Une recherche des positions chevauchantes avec les autres HSP sélectionnés et étendus a permis leur groupement. Les flèches indiquent l'orientation des gènes sur les brins d'ADN, telle qu'elle est déduite par l'algorithme des positions respectives des HSP labellisés. Il faut noter que les labels, à ce stade, indiquent les HSP labellisés et se réfèrent à l'identification des gènes, et non pas à une description détaillée de ces gènes.

5.2.3 Description des L-V-GENE-UNIT, D-GENE-UNIT et J-GENE-UNIT

Le deuxième module de IMGT/LIGMotif, 'Description de gène', décrit en détail chaque gène identifié, individuellement (Figure 5.9). Cette analyse est effectuée à partir de l'extrémité 5' vers l'extrémité 3', en explorant les deux brins de l'ADN. IMGT/LIGMotif commence par la recherche de motifs conservés des patterns décrits à la figure 1A. La

description des gènes est réalisée par la recherche et la délimitation de motifs conservés qui sont caractéristiques de chaque type de gène (V, D et J).

5.2.3.1 Délimitation des zones de recherche des motifs conservés (CMSA)

En vue de réduire les temps d'exécution de l'algorithme, la recherche de motifs conservés est limitée à des zones de recherche des motifs conservés (en anglais, conserved motif searching areas ou CMSA) qui sont délimitées à partir des positions de la combinaison la plus informative de chacun des HSP du groupe. La meilleure combinaison de HSP dépend du type de gène (Tableau 5.2): par exemple, pour un gène V, la meilleure combinaison est L-PART1 + V-EXON + V-RS, pour un gène D, c'est 5'D-RS + 3'D-RS, alors que, pour un gène J, c'est J-RS + J-REGION. Si la meilleure combinaison est absente, d'autres combinaisons sont utilisées dans l'ordre indiqué dans le Tableau 5.2. Une longueur arbitraire est ajoutée aux deux extrémités (5' et 3') de la combinaison de HSP afin de délimiter les CMSA. Par exemple, une longueur de 40 nt (la longueur maximale d'un RS) est ajoutée aux extrémités 5' et 3' d'un HSP labellisé RS.

Tableau 5.2. Les combinaisons de HSP labellisés utilisées pour la délimitation des zones de recherche des motifs conservés (CMSA).

Prototype	L-V-GENE-UNIT	D-GENE-UNIT	J-GENE-UNIT
Combinaison des HSP labellisés	L-PART1 + V-EXON + V-RS	5'D-RS + 3'D-RS(2)	J-RS + J-REGION
	L-PART1 + V-EXON	5'D-RS + D-REGION	J-REGION
	V-EXON(1) + V-RS	D-REGION + 3'D-RS	J-RS
	V-EXON(1)	5'D-RS	J-GENE-UNIT
	L-V-GENE-UNIT	3'D-RS	
		D-REGION	
		D-GENE-UNIT	

Les combinaisons utilisées en priorité sont soulignées. Si ces combinaisons ne sont pas présentes, les combinaisons situées en-dessous dans les colonnes sont choisies dans l'ordre, de haut en bas.
(1) Si le L-PART1 est absent, ses motifs (INIT-CODON et DONOR-SPLICE) ne sont pas délimités. (2) D-REGION n'est pas utilisé, même si un HSP a été identifié parce que les HSP 5'D-RS et 3'D-RS ainsi que ceux de leur heptamères et nonamères sont suffisants pour sa délimitation précise.

5.2.3.2 Recherche des motifs conservés dans les CMSA

Les motifs conservés incluent les acides aminés conservés (INIT-CODON pour le gène V, J-PHE et J-TRP du gène J), les sites d'épissage (DONOR-SPLICE, ACCEPTOR-SPLICE), les heptamères et les nonamères. Ces motifs sont recherchés dans les CMSA qui sont connus pour les inclure. Heptamères et nonamères sont recherchés par alignement avec les motifs des bases de données de référence (Annexe 5). Si aucune correspondance exacte

n’est trouvée, une forme approximative de la recherche est effectuée en utilisant une PSSM non intercalé par des insertions délétions [188]. A la suite de la recherche de motifs conservés dans les CMSA, les appariements sont groupés et retenus sous la forme d’un ensemble nommé ‘solution’ si la distance entre les motifs est autorisée par le modèle. Un minimum de deux motifs conservés est nécessaire pour maintenir la solution. Si cette condition n’est pas remplie, les gènes ne peuvent pas être décrits à l’aide d’IMGT/LIGMotif et, ils sont alors définis comme ‘V undescribed’, ‘D undescribed’ et ‘J undescribed’ (Figure 5.9).

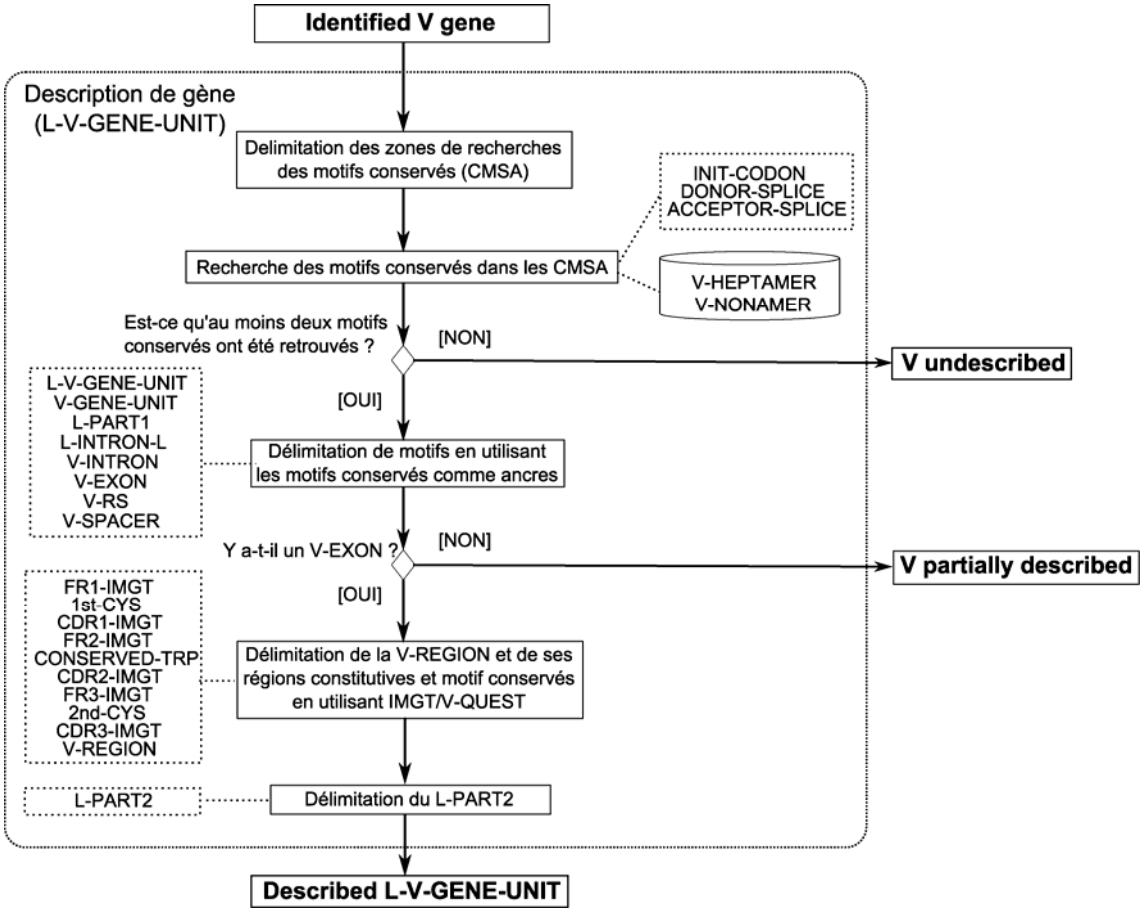


Figure 5.9. Module ‘Description de gène’ V. La sortie du module est ‘Described L-V-GENE-UNIT’. La délimitation des zones de recherche des motifs conservés (CMSA) est expliquée dans le texte. Au moins deux motifs conservés doivent être trouvés. Sinon, la sortie est ‘V undescribed’. L’absence de V-EXON d’un gène V conduit à un ‘V partially described’. La délimitation des motifs en utilisant les motifs conservés comme ancres est illustrée à la Figure 5.12.

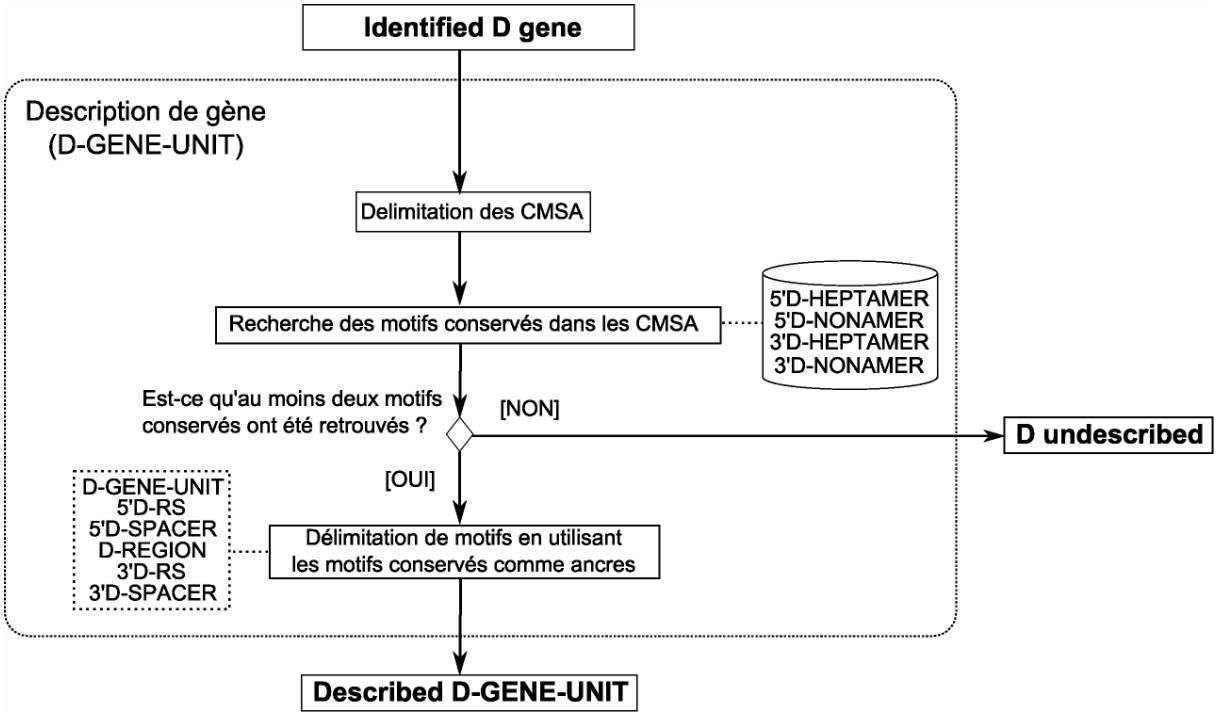


Figure 5.10. Module 'Description de gène' D. La sortie du module est 'Described D-GENE-UNIT'. La délimitation des zones de recherche des motifs conservés (CMSA) est expliquée dans le texte. Au moins deux motifs conservés doivent être trouvés. Sinon, la sortie est 'D undescribed'. La délimitation des motifs en utilisant les motifs conservés comme ancres est illustrée à la Figure 5.12.

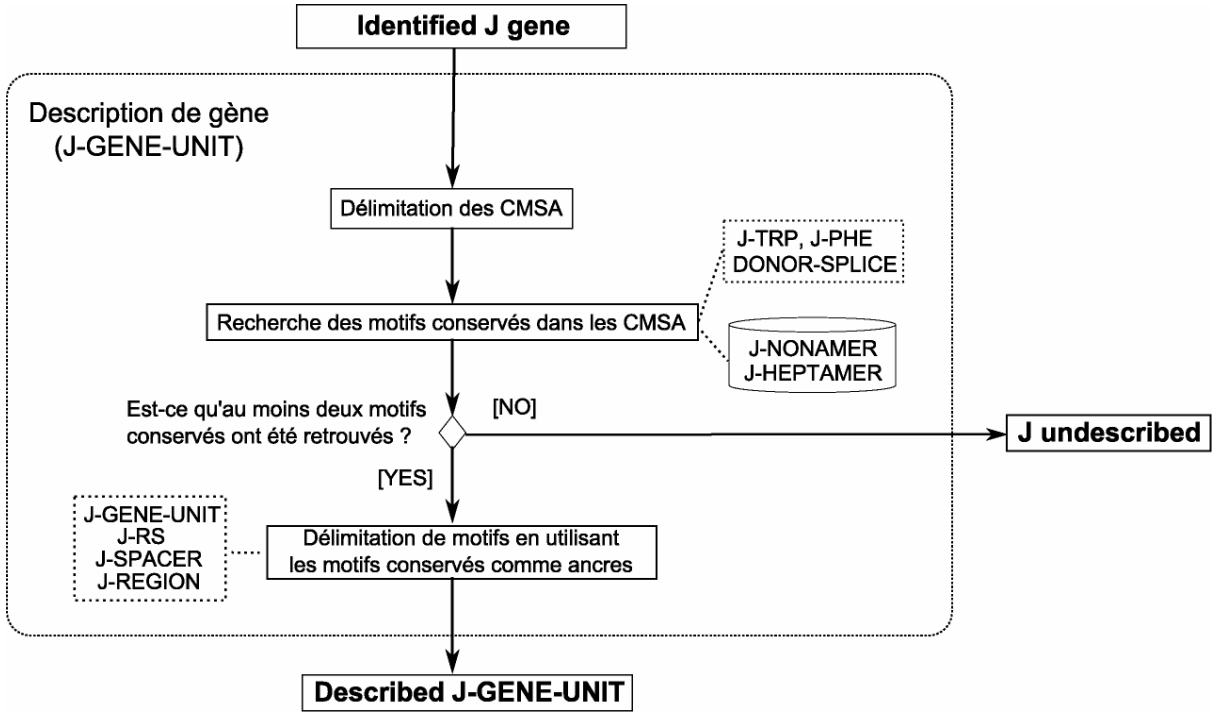


Figure 5.11. Module 'Description de gène' J. La sortie du module est 'Described J-GENE-UNIT'. La délimitation des zones de recherche des motifs conservés (CMSA) est expliquée dans le texte. Au moins deux motifs conservés doivent être trouvés. Sinon, la sortie est 'J undescribed'. La délimitation des motifs en utilisant les motifs conservés comme ancres est illustrée à la Figure 5.12.

5.2.3.3 Délimitation des motifs à partir des motifs conservés

Au cours de cette étape, d'autres motifs du prototype sont délimités précisément en utilisant les motifs conservés comme ancres (Figure 5.12).

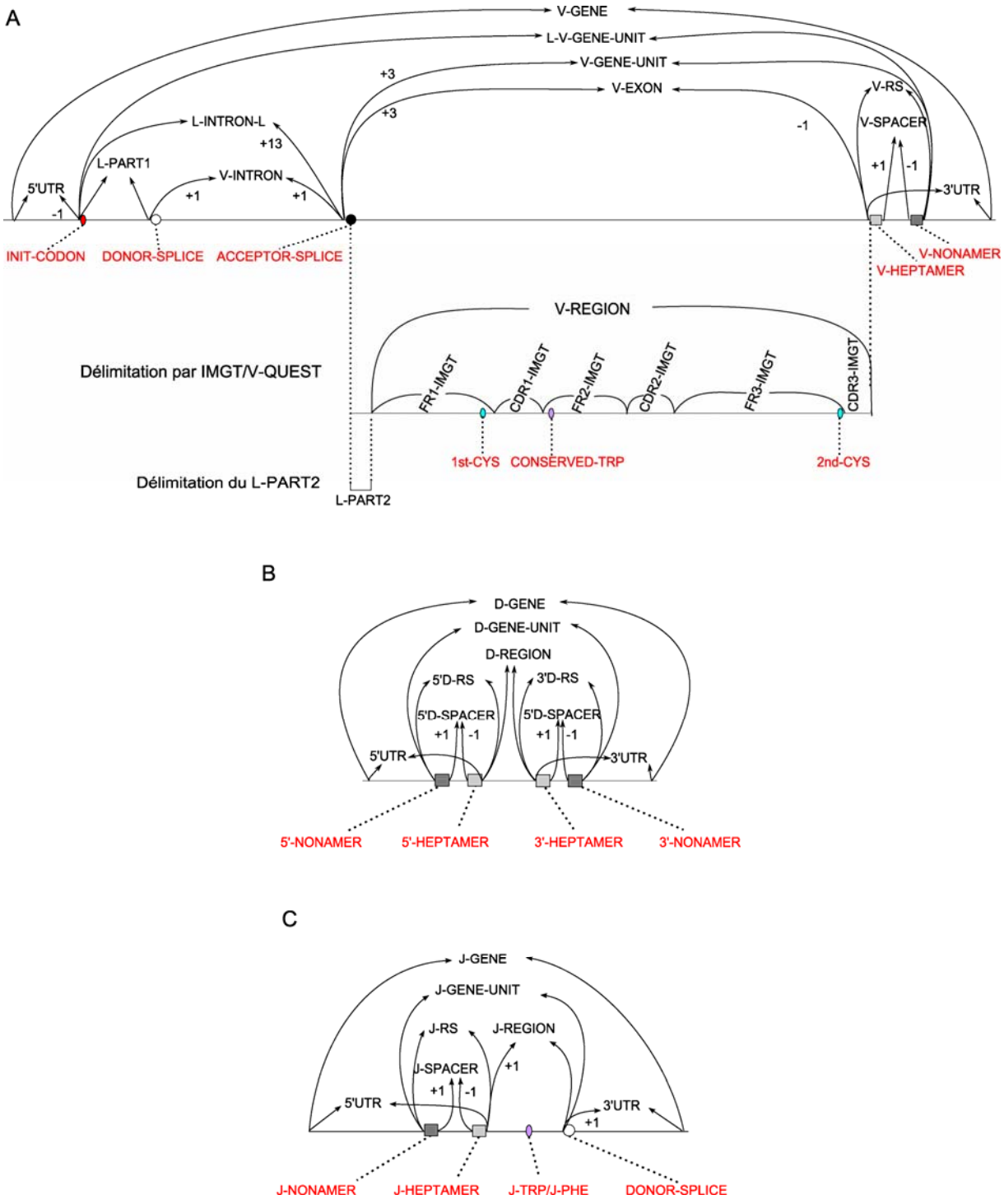


Figure 5.12. Délimitation des motifs en utilisant les motifs conservés comme ancres. La délimitation des motifs en utilisant les motifs conservés comme ancres est utilisée pour la description d'un V-GENE-UNIT (A), D-GENE-UNIT (B) et J-GENE-UNIT (C). Cette approche est nécessaire et suffisante pour la description d'un D-GENE-UNIT ou d'un J-GENE-UNIT. La description d'un V-GENE-UNIT nécessite deux étapes: la délimitation de la V-REGION et de ses régions constitutives et motifs conservés par IMGT/V-QUEST [178-179], et la délimitation du L-PART2.

Les flèches qui prennent racine à partir des ancres délimitent précisément les nouveaux motifs labellisés. Par exemple, ACCEPTOR-SPLICE et V-HEPTAMER (motifs conservés) permettent de délimiter le V-EXON. Une flèche arrivant de la gauche délimite l'extrémité 5' d'un motif alors qu'une flèche arrivant de la droite délimite l'extrémité 3'. Le numéro associé à une flèche indique le nombre de nucléotides qui doit être ajouté (+) ou soustrait (-) à la position d'un motif conservé afin de délimiter précisément le nouveau motif labellisé. Les J-TRP et J-PHE sont les seuls motifs conservés qui ne délimitent pas de nouveaux motifs, et, par conséquent, aucune flèche n'y prend racine.

5.2.3.4 Etapes supplémentaires de la description d'un L-V-GENE-UNIT

La description d'un L-V-GENE-UNIT nécessite deux étapes supplémentaires. La première est la délimitation de la V-REGION avec ses régions constitutives (FR-IMGT et CDR-IMGT) et ses motifs conservés (1st-CYS, CONSERVED-TRP et 2nd-CYS) à l'aide d'IMGT/V-QUEST [178-179]. Cette étape est uniquement effectuée si un V-EXON a été identifié. Si un V-EXON manque, le gène est défini comme 'V partially described'. L'étape finale pour la description d'un L-V-GENE-UNIT est la délimitation du L-PART2, une région délimitée par le site accepteur d'épissage du V-EXON et l'extrémité 5' de la V-REGION déterminée par IMGT/V-QUEST.

5.2.4 Identification de la fonctionnalité

Le troisième module d'IMGT/LIGMotif 'Identification de la fonctionnalité' identifie les fonctionnalités de chaque GENE-UNIT décrit (Figure 5.15). Un L-V-GENE-UNIT est identifié comme fonctionnel s'il possède l'ensemble des 22 labels spécifiques et sites d'épissage attendus (2 labels), aucun codon stop dans le L-PART1 et V-EXON, un cadre d'épissage de type sf1 entre le L-PART1 et V-EXON, aucun saut de cadre de lecture (même cadre de lecture pour le 1st-CYS, CONSERVED-TRP et 2nd-CYS), un V-SPACER de taille attendue, et un V-HEPTAMER et V-NONAMER identiques à ceux trouvés dans des gènes fonctionnels (Figure 5.13).

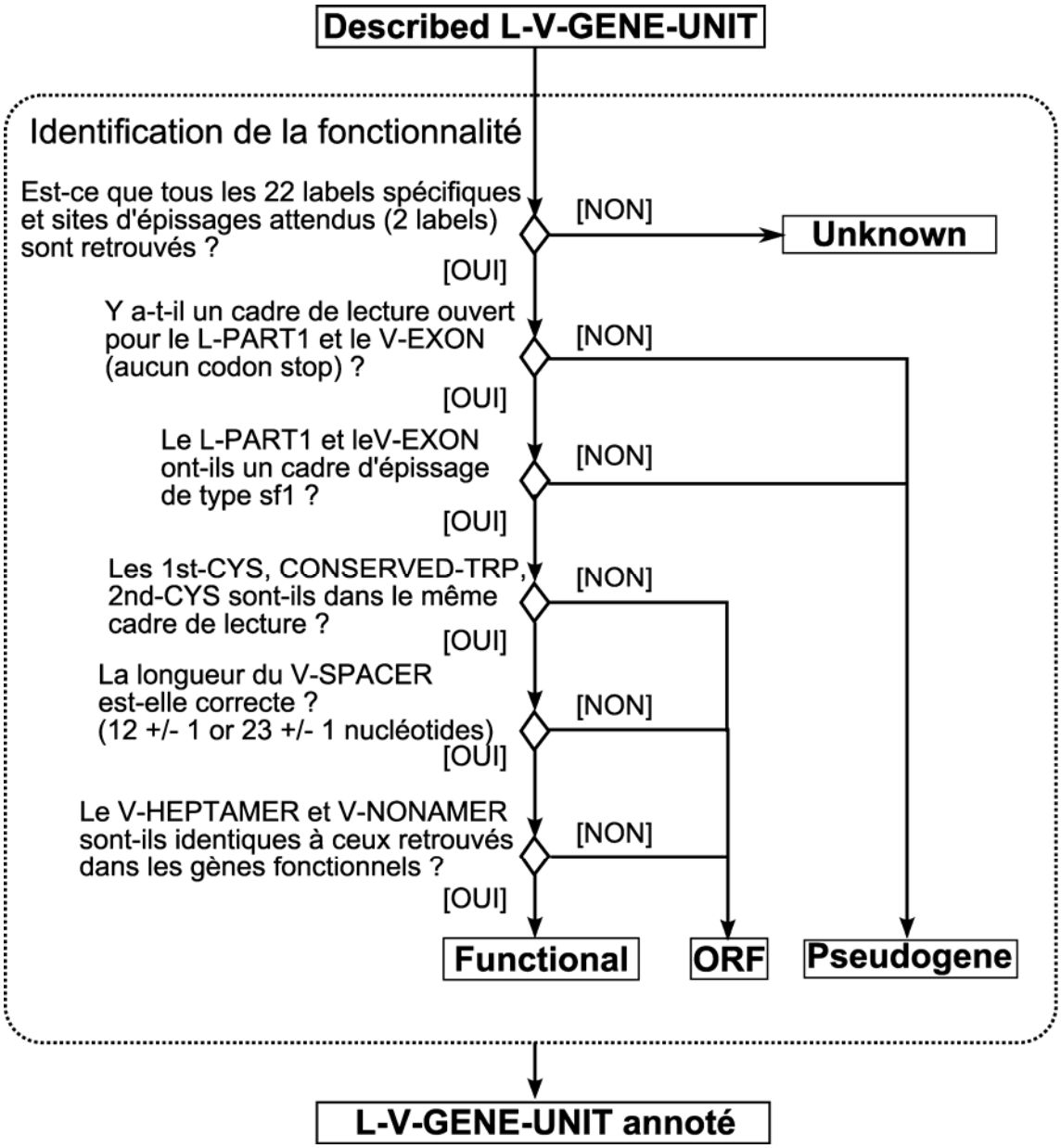


Figure 5.13. Identification de la fonctionnalité d'un L-V-GENE-UNIT. Le résultat du module d'identification de la fonctionnalité' est un L-V-GENE-UNIT annoté ('Annotated L-V-GENE-UNIT'). L'identification de la fonctionnalité permet d'attribuer aux unités de gènes les classes 'Functional', 'ORF' et 'Pseudogene'. Si le premier critère n'est pas rempli, la fonctionnalité est 'Unknown', car elle ne peut pas être identifiée automatiquement. Dans la figure, les heptamères et nonamères sont repérés de 5' à 3' par rapport aux unités gènes.

Un D-GENE-UNIT est identifié comme fonctionnel s'il possède tous les 10 labels spécifiques, au moins un cadre de lecture ouvert sans codon stop, un 5'D-SPACER et 3'D-SPACER de longueurs attendues, et des heptamères et nonamères identiques à ceux trouvés dans des gènes fonctionnels (Figure 5.14). Un gène est identifié comme un cadre de lecture ouvert (ORF), si les 3 derniers critères des gènes V et J (Figure 5.13, Figure 5.15) et si les 2 derniers d'un gène D (Figure 5.14) ne sont pas remplies. Dans les autres cas, le gène est identifié comme un pseudogene (P). Il faut noter que si le premier critère (nombre de labels

spécifiques et de sites d'épissage) n'est pas rempli, la fonction est inconnue 'Unknown', car elle ne peut pas être déterminée automatiquement et donc son identification requiert une expertise manuelle.

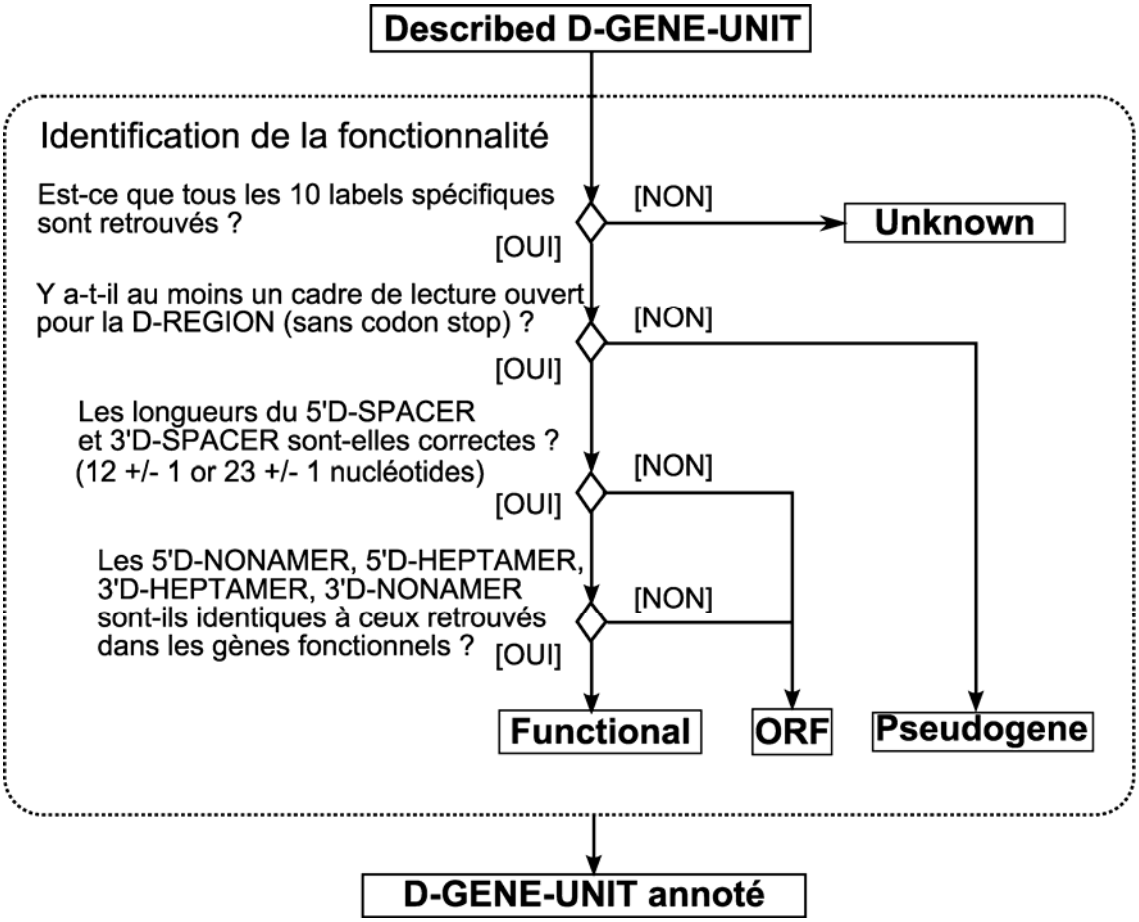


Figure 5.14. Identification de la fonctionnalité d'un D-GENE-UNIT. Le résultat du module d'identification de la fonctionnalité est le D-GENE-UNIT annoté ('Annotated D-GENE-UNIT'). L'identification de la fonctionnalité permet d'attribuer aux unités de gènes les classes 'Functional', 'ORF' et 'Pseudogene'. Si le premier critère n'est pas rempli, la fonctionnalité est 'Unknown', car elle ne peut pas être identifiée automatiquement. Dans la figure, les heptamères et nonamères sont repérés de 5' à 3' par rapport aux unités gènes.

Un J-GENE-UNIT est identifié comme fonctionnel s'il possède tous les 7 labels spécifiques et le site d'épissage attendu (1 label), aucun codon stop dans la J-REGION, un site donneur d'épissage de type sf1, le motif conservé '[W,F]-[G,A]-X-G' servant d'indicateur de l'absence de saut de cadre de lecture, de taille attendue, un J-SPACER, et un J-NONAMER et J-HEPTAMER identiques à ceux trouvés dans des gènes fonctionnels (Figure 5.15).

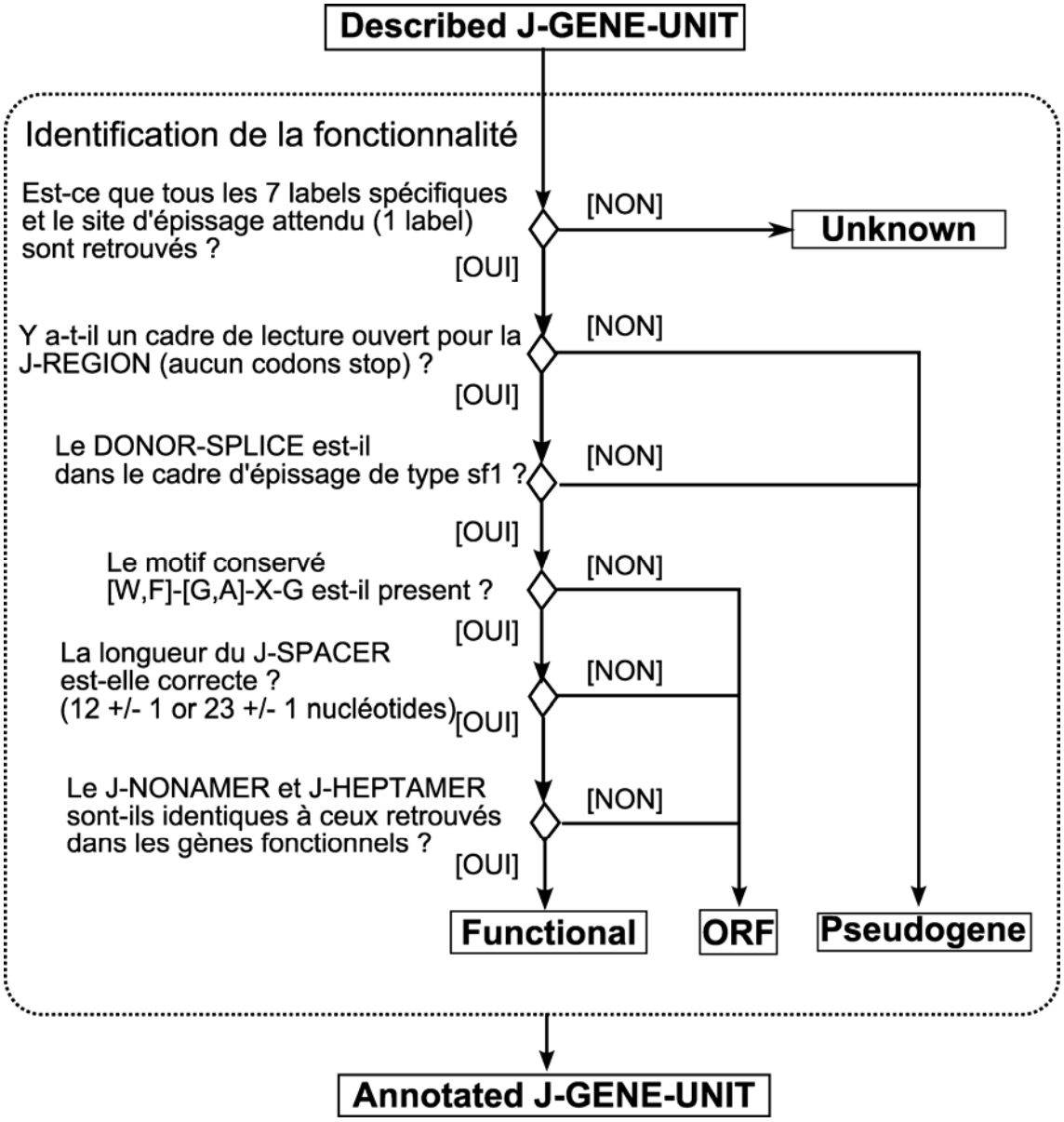


Figure 5.15. Identification de la fonctionnalité d'un J-GENE-UNIT. Le résultat du module d''identification de la fonctionnalité' est un J-GENE-UNIT annoté ('Annotated J-GENE-UNIT'). L'identification de la fonctionnalité permet d'attribuer aux unités de gènes les classes 'Functional', 'ORF' et 'Pseudogene'. Si le premier critère n'est pas rempli, la fonctionnalité est 'Unknown', car elle ne peut pas être identifiée automatiquement. Dans la figure, les heptamères et nonamères sont repérés de 5' à 3' par rapport aux unités gènes.

5.2.5 Délimitation de gène et assemblage en cluster

Pour chaque unité de gène, et quel que soit son état de description, annotée (L-V-GENE-UNIT, D-GENE-UNIT, J-GENE-UNIT), partiellement décrits ('V partially described') ou partiellement décrites ('V undescribed', 'D undescribed', 'J undescribed'), 5'UTR et 3'UTR délimitations correspondant à un V-GENE, D-GENE ou J-GENE sont déterminées en répartissant équitablement la même distance entre les deux unités de gènes. Ainsi, la séquence peut être considérée comme une succession de gènes qui constituent un

groupe. Le cluster est défini sur la base des instances du concept ‘Molecule_EntityPrototype’ trouvées dans la séquence et un label de cluster d’IMGT® est attribué (Tableau 2.2), par exemple V-D-J-CLUSTER, si la séquence contient au moins un V-GENE, D-GENE et un J-GENE. Enfin, le nombre de gènes dans la séquence analysée est calculé par brin d’ADN, par type de gène et par fonctionnalité.

5.3 Evaluation des solutions

Les solutions d’un gène sont évaluées par le calcul d’un score. Les motifs conservés ont une probabilité élevée pour apparier par hasard en raison de leur courte longueur, ce qui entraîne donc le chevauchement de plusieurs appariements de pattern qui doivent être sélectionnés. Ainsi des solutions sont évaluées par un score pour leur correspondance au modèle des gènes V, D ou J. Le score est calculé pour les motifs conservés labellisés et est la somme de chaque partition conservée marquée motif présent dans une solution (valable pour une solution V-GENE, D-GENE ou J-GENE). Le score des motifs conservés labellisés dépend de la probabilité $p(m)$ pour trouver le motif m par hasard dans une séquence de nucléotides (Tableau 5.3). La probabilité $p(m)$ est calculée comme suit:

$$p(m) = \left(\frac{1}{4}\right)^{Lm} \times Slex$$

où Lm est la longueur de motif en nucléotides et $Slex$ le nombre de séquences différentes possibles pour m. Ainsi, $Slex$ est le nombre de codons pour les acides aminés conservés, $Slex$ est égal à 4 pour le site donneur d’épissage et de 64 pour le site accepteur d’épissage. Les probabilités sont calculées pour les heptamères et nonamères en supposant que le nombre de séquences différentes possible est contenu dans les bases de données de motifs de référence: plus forte est la probabilité, plus faible est le score (Tableau 5.3). Les probabilités peuvent être classées en tant que p(splicingSites)> p(cys/phe)> p(met/trp)> p(heptamère)> p(nonamère). Les valeurs des scores sont comprises entre 1 et 5, de la plus haute probabilité à la plus basse.

Tableau 5.3. Probabilités de retrouver des motifs conservés et leur score.

Motif conservé	Séquence nucléotidique	Probabilité	Score
donor-splice	NGT	6,25E-02	1
acceptor-splice	NAGNN	6,25E-02	1
cystéine	TGC et TGT	3,13E-02	2
phénylalanine	TTT et TTC	3,13E-02	2
méthionine	ATG	1,56E-02	3
tryptophane	TGG	1,56E-02	3
heptamères	CACAGTG (1)	1,09E-02	4
nonamères	ACACAAACC (1)	1,54E-03	5

(1) La base de séquences de références de IMGT/LIGMotif contient 179 heptamères différents (*Slex* = 179) et 404 nonamères différents (*Slex* = 404) pour l'homme et la souris. Comme la longueur d'un heptamère est de 7 nucléotides (*Lm*= 7) et celle d'un nonamère de 9 nucléotides (*Lm* = 9), la probabilité de retrouver au hasard un heptamère *p(*heptamère*)*= $(1/4)^7$ x 179 et celle d'un nonamère *p(*nonamère*)*= $(1/4)^9$ x 404 (Annexe 5).

D'autres fonctionnalités sont évaluées par un score. Pour un gène V et un gène J, une solution avec un sf1 a un score de 1 pour cette fonctionnalité. Pour les gènes V, une solution avec des acides aminés conservés dans un même cadre a un score de 1 pour cette fonctionnalité. Une solution avec un nonamère ou un heptamère exact est pondérée par 1 pour chaque motif. Une solution sans aucun codon stop a un score de 1 pour cette fonctionnalité. Uniquement pour les J-GENE, la présence d'un environnement en acides aminés pour le J-PHE et J-TRP dans une solution a un score de 4. Une longueur d'espaceur d'un RS égale à 12 ou 23 nucléotides est pondérée par un score de 2, pour une longueur de 11, 13, 22 ou 24 nucléotides, par un score de 1. Tous ces scores de caractéristiques de gène sont additionnés dans un score global de gène. Ainsi, un V-GENE a un score maximum de 24, un D-GENE un maximum de 22 a J-GENE un maximum de 19 si J-PHE ou 20 si J-TRP est présent (voir l'annexe 5 pour le détail du calcul du score). Le score de la correspondance du HSP au modèle est également un paramètre utilisé dans la sélection d'une solution. Pour chaque solution, le score de la correspondance du HSP est calculé. Chaque région de motif de la solution (*motif_region*) qui exclut des motifs conservés est comparée aux motifs qui se chevauchent provenant de HSP (*hsp_region*) du même label et brin (voir formule ci-dessous (*s1*) et (*s2*)). La correspondance du HSP est définie par le pourcentage de couverture entre les longueurs des *hsp_region* et *motif_region*, et inversement, suivant les formules:

$$s1 = \frac{i \times 100}{hsp\_region}, \; s2 = \frac{i \times 100}{motif\_region}$$

où *i* est la longueur des intersections entre une *hsp_region* et *motif_region* donnée. Pour chaque *motif_region* et *hsp_region* comparées qui se chevauchent, *s1* et *s2* sont additionnés

au score de la correspondance du HSP. Plus une *hsp_region* correspond à une *motif_region*, plus grand est le score. Ainsi, les prototypes V-GENE, D-GENE et J-GENE ont un score maximum de 3200, 1200 et 800, respectivement. Le score maximum est lié au nombre de bases de données de motifs de références non conservées (16, 6, 4, pour les prototypes V-GENE, D-GENE et J-GENE, respectivement) utilisés par IMGT/BLAST, plus le nombre de bases est élevé, plus le score maximum est élevé.

5.4 Implémentation: application Web

L'algorithme d'IMGT/LIGMotif est implémenté en Java (http://www.java.com/fr/). Actuellement, il est possible d'exécuter IMGT/LIGMotif en ligne de commande et de visualiser les résultats à travers une interface d'annotation. Cependant il est bien plus pratique d'utiliser l'outil à travers une interface web. L'application web d'IMGT/LIGMotif est disponible sur un serveur Tomcat (http://tomcat.apache.org/). La séquence génomique à analyser peut être copiée/collée par le biocurateur ou téléchargée au format FASTA ou EMBL (Figure 5.16). Les paramètres de la requête peuvent être modifiés afin d'optimiser l'efficacité de l'analyse. Par exemple, les bases de données de motifs de référence peuvent être sélectionnées en fonction du type de gène, du locus, de la fonctionnalité et de l'organisme. L'interface de requête permet de paramétrer la base de séquence de référence de motifs avec 26670 combinaisons possibles entre les 3 types de gènes, les 3 fonctionnalités, les 7 locus et les 2 espèces. Les bases de motifs conservés d'heptamères et de nonamères peuvent être combinées de 2667 façons différentes en excluant le rat. Le jeu de séquences de référence d'IMGT/V-QUEST donne une possibilité de 2047 combinaisons avec ses 11 espèces. Le temps d'exécution dépend du type de gène et du nombre de gènes existant dans la séquence. L'analyse d'une séquence ne contenant qu'un seul gène prend seulement quelques secondes tandis que celle d'un locus qui contient plus de 100 gènes, prend de 30 minutes à 1 heure, en utilisant les paramètres standards.

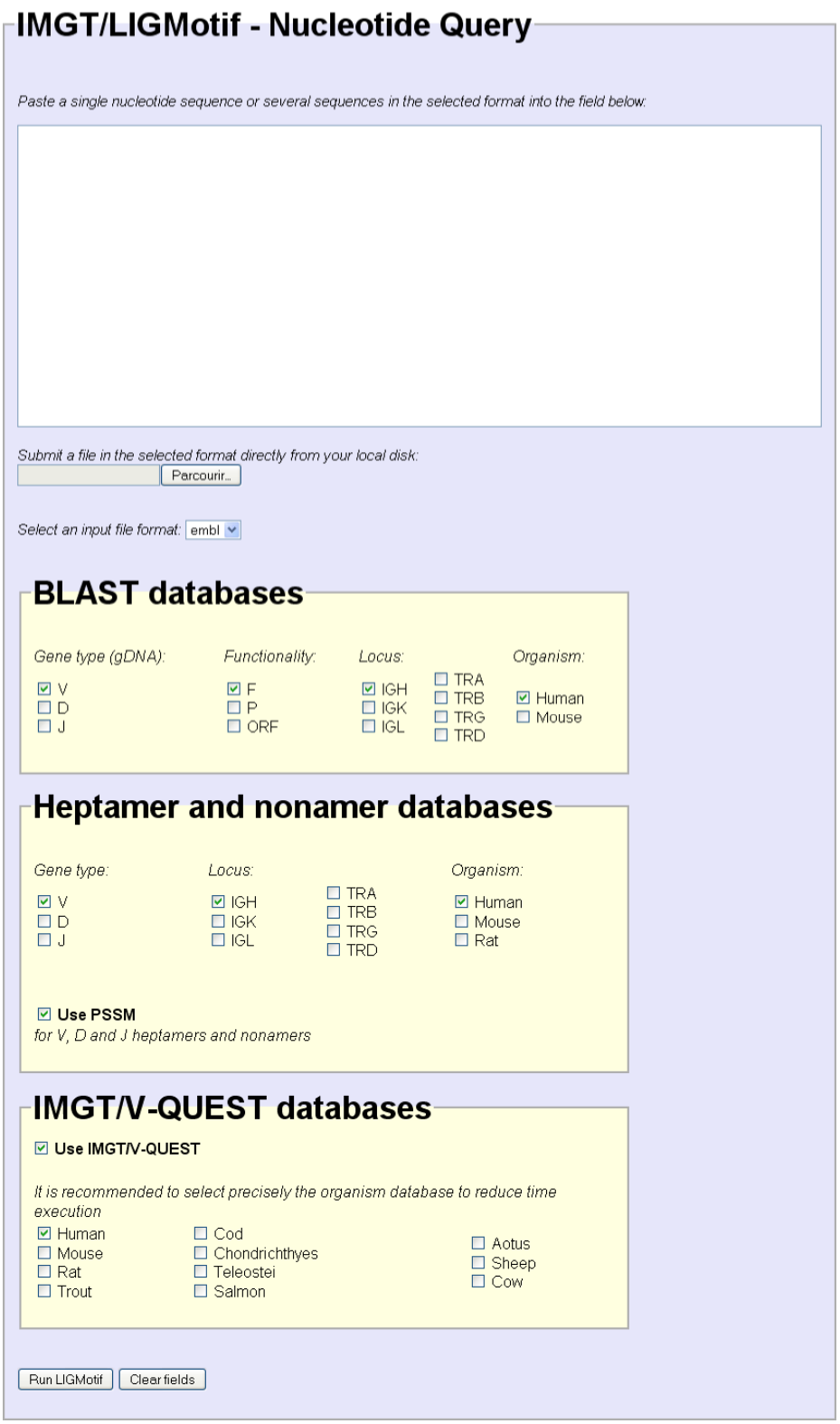


Figure 5.16. Affichage initial de l'interface de requête d'IMGT/LIGMotif. Le premier cadre permet d'y copier/coller une séquence à analyser. L'analyse peut aussi se faire à partir d'un fichier contenant la séquence placée sur le disque dur de l'utilisateur. Le format de la séquence doit être de type EMBL ou FASTA. Un menu déroulant permet de spécifier le type de format choisi. Le 2^{ième} cadre permet de configurer les bases de séquences de références destinées au BLAST. Le 3^{ième} cadre situé en dessous permet de paramétrer la base d'heptamères et de nonamères provenant de gènes fonctionnels. Il est aussi possible d'autoriser ou non l'utilisation de PSSM à partir de la base des heptamères et nonamères configurés. Le cadre du bas permet de paramétrer IMGT/V-QUEST en sélectionnant le jeu de séquences de références d'un ou plusieurs organismes.

En haut de la page de résultats d'IMGT/LIGMotif (Figure 5.17) sont affichés le temps d'exécution, la longueur de la séquence analysée, le nombre total de gènes par brin d'ADN (direct et inversé complémentaire) et deux tableaux: le premier indique le nombre de gènes selon le statut de description ('GENE-UNIT', 'Partially described', 'Undescribed') et par type de gène, le second indique le nombre de GENE-UNIT annotés par fonctionnalité ('Functional', 'ORF', 'Pseudogene', 'Unknown').

Statistics

Execution time (h:m:s):
0:4:14

Sequence length:
232650 bp

Total number of genes per DNA strand

	Plus	Minus
Total	25	1

Description status and gene type

	V	D	J	Total
GENE-UNIT	11	2	13	26
Partially described	0			0
Undescribed	0	0	0	0
Total	11	2	13	26

GENE-UNIT and functionality status

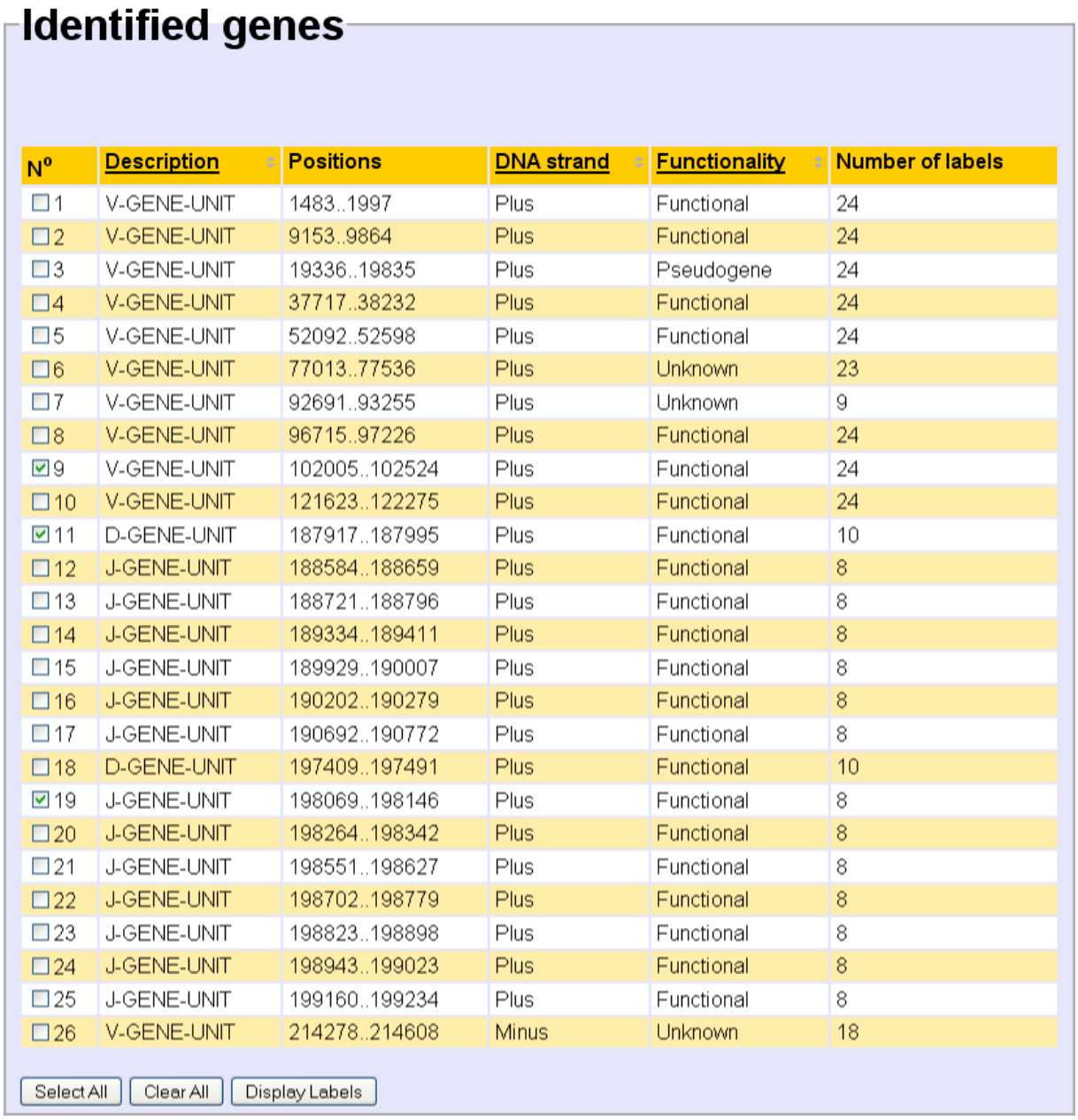
	Functional	ORF	Pseudogene	Unknown	Total
V-GENE-UNIT	7	0	1	3	11
D-GENE-UNIT	2	0	0	0	2
J-GENE-UNIT	13	0	0	0	13
Total	22	0	1	3	26

Figure 5.17. Affichage des statistiques de l'analyse faite par IMGT/LIGMotif de la séquence U66061 provenant du locus TRB humain. Le programme affiche le temps d'exécution (4 minutes et 14 secondes) et la longueur de la séquence analysée 232650 bp. Vingt-six gènes sont identifiés par IMGT/LIGMotif, 25 dans le brin direct (Plus) et 1 seul dans le brin inversé complémentaire (Minus). Tous les gènes identifiés sont de type GENE-UNIT. Parmi les 11 gènes V, 7 sont identifiés comme fonctionnels, un seul pseudogène et les 3 derniers ont une fonctionnalité qui n'a pu être déterminée ('Unknown'). Deux gènes D et 13 gènes J sont fonctionnels.

Le tableau principal affiche le contenu de la séquence analysée, à partir du numéro 1, le gène le plus en 5'. Ce tableau donne pour chacun des gènes identifiés: Description (le label

correspondant au statut GENE-UNIT ou état de description), Positions (dans la séquence analysée), le brin d'ADN, la fonctionnalité et le nombre de labels.

La sélection des gènes affiche leurs labels et leurs positions, leurs séquences de nucléotides et, si la région est codante, la séquence d'acides aminés (Figure 5.18).



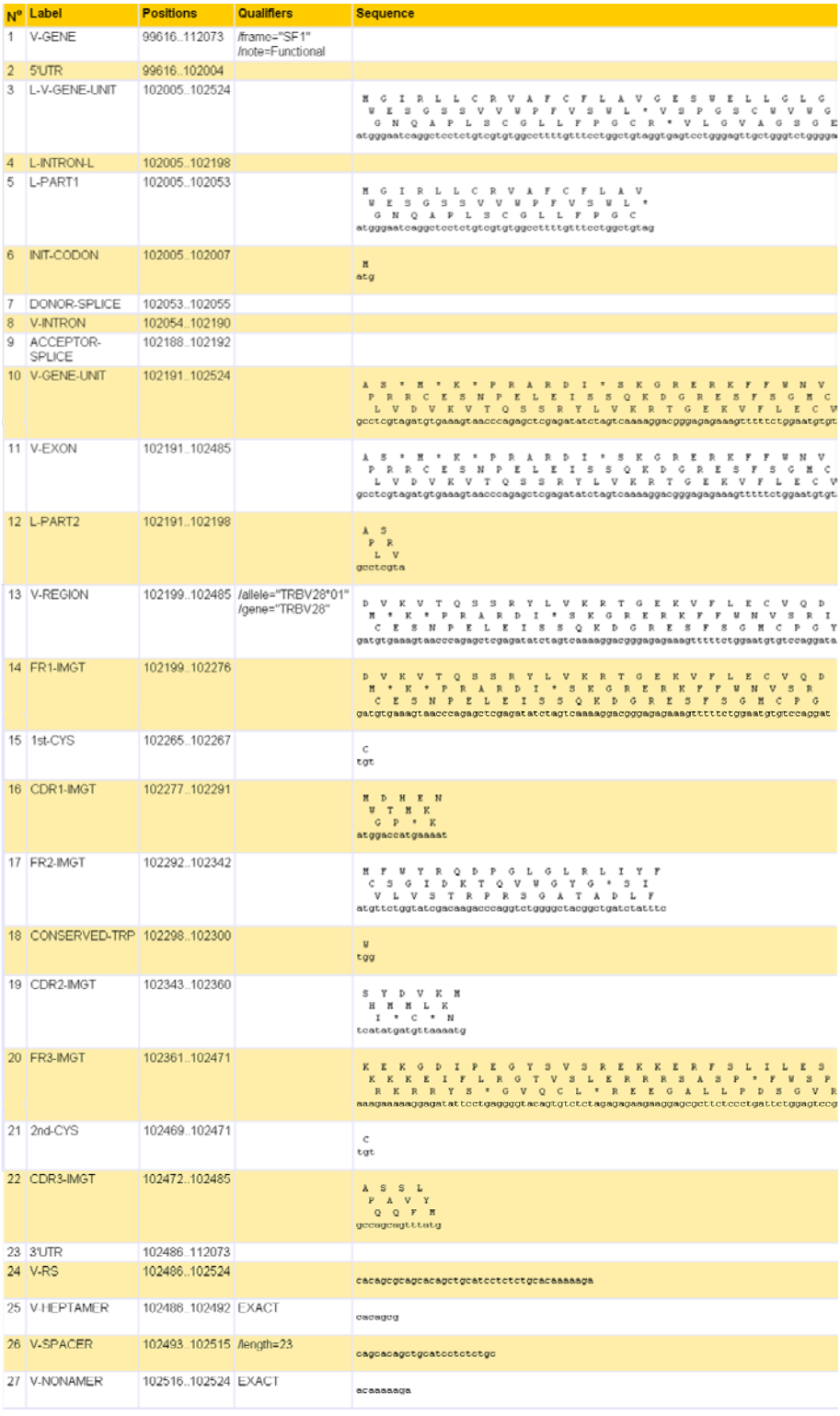
Identified genes

N°	Description	Positions	DNA strand	Functionality	Number of labels
☐ 1	V-GENE-UNIT	1483..1997	Plus	Functional	24
☐ 2	V-GENE-UNIT	9153..9864	Plus	Functional	24
☐ 3	V-GENE-UNIT	19336..19835	Plus	Pseudogene	24
☐ 4	V-GENE-UNIT	37717..38232	Plus	Functional	24
☐ 5	V-GENE-UNIT	52092..52598	Plus	Functional	24
☐ 6	V-GENE-UNIT	77013..77536	Plus	Unknown	23
☐ 7	V-GENE-UNIT	92691..93255	Plus	Unknown	9
☐ 8	V-GENE-UNIT	96715..97226	Plus	Functional	24
☑ 9	V-GENE-UNIT	102005..102524	Plus	Functional	24
☐ 10	V-GENE-UNIT	121623..122275	Plus	Functional	24
☑ 11	D-GENE-UNIT	187917..187995	Plus	Functional	10
☐ 12	J-GENE-UNIT	188584..188659	Plus	Functional	8
☐ 13	J-GENE-UNIT	188721..188796	Plus	Functional	8
☐ 14	J-GENE-UNIT	189334..189411	Plus	Functional	8
☐ 15	J-GENE-UNIT	189929..190007	Plus	Functional	8
☐ 16	J-GENE-UNIT	190202..190279	Plus	Functional	8
☐ 17	J-GENE-UNIT	190692..190772	Plus	Functional	8
☐ 18	D-GENE-UNIT	197409..197491	Plus	Functional	10
☑ 19	J-GENE-UNIT	198069..198146	Plus	Functional	8
☐ 20	J-GENE-UNIT	198264..198342	Plus	Functional	8
☐ 21	J-GENE-UNIT	198551..198627	Plus	Functional	8
☐ 22	J-GENE-UNIT	198702..198779	Plus	Functional	8
☐ 23	J-GENE-UNIT	198823..198898	Plus	Functional	8
☐ 24	J-GENE-UNIT	198943..199023	Plus	Functional	8
☐ 25	J-GENE-UNIT	199160..199234	Plus	Functional	8
☐ 26	V-GENE-UNIT	214278..214608	Minus	Unknown	18

Select All | Clear All | Display Labels

Figure 5.18. Affichage des résultats synthétiques de l'analyse faite par IMGT/LIGMotif de la séquence U66061 provenant du locus TRB humain. Vingt-six gènes sont identifiés, le plus en 3' est le seul localisé dans le brin inversé complémentaire (Minus). Tous les gènes identifiés sont de type GENE-UNIT. La fonctionnalité des gènes est affichée. Le nombre de labels est celui de ceux qui décrivent l'unité de gène identifiée, soit pour un gène fonctionnel le nombre de tous les labels d'un type de gène diminué des deux UTR (5'UTR et 3'UTR) et du label du prototype (V-GENE, D-GENE ou J-GENE). Trois gènes ont été sélectionnés : un V-GENE-UNIT (N°9), un D-GENE-UNIT (N°11) et un J-GENE-UNIT (N°19).

Dans cet affichage détaillé, les labels sont ordonnés de l'extrémité 5' à l'extrémité 3' de la séquence analysée. Si deux labels ont la même position à leur extrémité 5', le label le plus long est affiché en premier. Les résultats produits par IMGT/LIGMotif peuvent être exportés vers un fichier de feuille de calcul qui peut être modifié par le biocurateur si nécessaire. A l'avenir, il serait utile de faire apparaître le numéro d'accès de la séquence soumise dans les résultats d'IMGT/LIGMotif car il permet de se remémorer d'où proviennent les résultats d'une séquence parmi plusieurs autres résultats.



N°	Label	Positions	Qualifiers	Sequence
1	V-GENE	99616..112073	/frame="SF1" /note=Functional	
2	5'UTR	99616..102004		
3	L-V-GENE-UNIT	102005..102524		M G I R L L C R V A F C F L A V G E S W E L L G L G W E S G S S V V W P F V S W L * V S P G S C W V W G G N Q A P L S C G L L F P G C R * V L G V A G S G E atgggaatcaggctcctctgtcgtgtggcttttgtttcctggctgtaggtgagtcctgggagttgctgggtctgggga
4	L-INTRON-L	102005..102198		
5	L-PART1	102005..102053		M G I R L L C R V A F C F L A V W E S G S S V V W P F V S W L * G N Q A P L S C G L L F P G C atgggaatcaggctcctctgtcgtgtggcttttgtttcctggctgtag
6	INIT-CODON	102005..102007		M atg
7	DONOR-SPLICE	102053..102055		
8	V-INTRON	102054..102190		
9	ACCEPTOR-SPLICE	102188..102192		
10	V-GENE-UNIT	102191..102524		A S * M * K * P R A R D I * S K G R E R K F F W N V P R R C E S N P E L E I S S Q K D G R E S F S G M C L V D V K V T Q S S R Y L V K R T G E K V F L E C V gcctcgtagatgtgaaagtaacccagagctcgagatatctagtcaaaaggacgggagagaaagtttttctggaatgtgt
11	V-EXON	102191..102485		A S * M * K * P R A R D I * S K G R E R K F F W N V P R R C E S N P E L E I S S Q K D G R E S F S G M C L V D V K V T Q S S R Y L V K R T G E K V F L E C V gcctcgtagatgtgaaagtaacccagagctcgagatatctagtcaaaaggacgggagagaaagtttttctggaatgtgt
12	L-PART2	102191..102198		A S P R L V gcctcgta
13	V-REGION	102199..102485	/allele="TRBV28*01" /gene="TRBV28"	D V K V T Q S S R Y L V K R T G E K V F L E C V Q D M * K * P R A R D I * S K G R E R K F F W N V S R I C E S N P E L E I S S Q K D G R E S F S G M C P G Y gatgtgaaagtaacccagagctcgagatatctagtcaaaaggacgggagagaaagtttttctggaatgtgtccaggata
14	FR1-IMGT	102199..102276		D V K V T Q S S R Y L V K R T G E K V F L E C V Q D M * K * P R A R D I * S K G R E R K F F W N V S R C E S N P E L E I S S Q K D G R E S F S G M C P G gatgtgaaagtaacccagagctcgagatatctagtcaaaaggacgggagagaaagtttttctggaatgtgtccaggat
15	1st-CYS	102265..102267		C tgt
16	CDR1-IMGT	102277..102291		M D N E N W T M K G P * K atggaccatgaaaat
17	FR2-IMGT	102292..102342		M F W Y R Q D P G L G L R L I Y F C S G I D K T Q V W G Y G * S I V L V S T R P R S G A T A D L F atgttctggtatcgacaagacccaggtctggggctacggctgatctatttc
18	CONSERVED-TRP	102298..102300		W tgg
19	CDR2-IMGT	102343..102360		S Y D V K N H M M L K I * C * N tcatatgatgttaaaatg
20	FR3-IMGT	102361..102471		K E K G D I P E G Y S V S R E K K E R F S L I L E S K K K E I F L R G T V S L E R R R S A S P * F W S P R K R R Y S * G V Q C L * R K E G A L L P D S G V R aaagaaaaaaggagatattcctgaggggtacagtgtctctagagagaaggaggagcgcttctccctgattctggagtccg
21	2nd-CYS	102469..102471		C tgt
22	CDR3-IMGT	102472..102485		A S S L P A V Y Q Q F M gccagcagtttatg
23	3'UTR	102486..112073		
24	V-RS	102486..102524		cacagcgcagcacagctgcatcctctctgcacaaaaaga
25	V-HEPTAMER	102488..102492	EXACT	cacagcg
26	V-SPACER	102493..102515	/length=23	cagcacagctgcatcctctctgc
27	V-NONAMER	102516..102524	EXACT	acaaaaaga

Figure 5.19. Affichage de la description du V-GENE fonctionnel N°9 identifié par IMGT/LIGMotif provenant de la séquence U66061 du locus TRB humain. La quatrième colonne en partant de la gauche indique les qualifieurs. Le qualifieur /note=Functional du label V-GENE indique que le gène possède toutes les caractéristiques d'un gène fonctionnel. Les autres qualifieurs signalent le type du cadre d'épissage qui doit être de type 1 (/frame= "SF1"), le type de l'heptamère et du nonamère qui doit être exact et la longueur de l'espaceur (V-SPACER) qui doit être de 12±1 et/ou 23±1 nucléotides pour les gènes V fonctionnels. Le qualifieur /allele et /gene indique l'allèle et le gène le plus proche retrouvés dans la base de séquences de références utilisée par IMGT/V-QUEST, respectivement.

28	D-GENE	155096..188289	/note=Functional	
29	5'UTR	155096..187944		
30	D-GENE-UNIT	187917..187995		C F C T K L * H C G D R G P Q * F N S T G N L Y K N V F V Q S C N I V G T G G H N D S T L R E T F T K T F L Y K A V T L W G Q G A T M I Q L Y G K P L Q K tgtttttgtacaaagctgtaacattgtggggacaggggccacaatgattcaactctacgggaaacctttacaaaaacc
31	5'D-RS	187917..187944		tgtttttgtacaaagctgtaacattgtg
32	5'D-NONAMER	187917..187925	EXACT	tgtttttgt
33	5'D-SPACER	187926..187937	/length=12	acaaagctgtaa
34	5'D-HEPTAMER	187938..187944	EXACT	cattgtg
35	D-REGION	187945..187956		G T G G G Q G D R G gggacaggggc
36	3'UTR	187957..188289		
37	3'D-RS	187957..187995		cacaatgattcaactctacgggaaacctttacaaaaacc
38	3'D-HEPTAMER	187957..187963	EXACT	cacaatg
39	3'D-SPACER	187964..187986	/length=23	attcaactctacgggaaacctttt
40	3'D-NONAMER	187987..187995	EXACT	acaaaaacc

Figure 5.20. Affichage de la description d'un D-GENE fonctionnel N°11 identifié par IMGT/LIGMotif provenant de la séquence U66061 du locus TRB humain. Le qualifieur /note=Functional du label D-GENE indique que le gène possède toutes les caractéristiques d'un gène fonctionnel. Les autres qualifieurs signalent le type de l'heptamère et du nonamère qui doit être exact et la longueur des espaceurs (5'D-SPACER et 3'D-SPACER) qui doit être de 12±1 et/ou 23±1 nucléotides pour les gènes D fonctionnels.

41	J-GENE	197780..198205	/frame="SF1" /note=Functional	
42	5'UTR	197780..198096		
43	J-GENE-UNIT	198069..198146		E F W A A P S H C A P T M S S S S G Q G H G S P C * N S G Q P L P T V L L Q * A V L R A R D T A H R A I L G S P F P L C S Y N E Q F F G P G T R L T V L gaattctgggcagccccttcccactgtgctcctacaatgagcagttcttcgggccagggacacggctcaccgtgctag
44	J-RS	198069..198096		gaattctgggcagccccttcccactgtg
45	J-NONAMER	198069..198077	EXACT	gaattctgg
46	J-SPACER	198078..198089		gcagccccttcc
47	J-HEPTAMER	198090..198096	EXACT	cactgtg
48	J-REGION	198097..198146		L L Q * A V L R A R D T A H R A S Y N E Q F F G P G T R L T V L P T M S S S S G Q G H G S P C * ctcctacaatgagcagttcttcgggccagggacacggctcaccgtgctag
49	J-PHE	198116..198118	/motif=[GPG]	F ttc
50	DONOR-SPLICE	198146..198148		
51	3'UTR	198147..198205		

Figure 5.21. Affichage de la description d'un J-GENE fonctionnel N°19 identifié par IMGT/LIGMotif provenant de la séquence U66061 du locus TRB humain. Le qualifieur /note=Functional du label J-GENE indique que le gène possède toutes les caractéristiques d'un gène fonctionnel. Les autres qualifieurs signalent le type du cadre d'épissage qui doit être de type 1 (/frame= "SF1"), le type de l'heptamère et du nonamère qui doit être exact et le motif « G-X-G » qui doit suivre le J-PHE (/motif=[GPG]) pour les gènes J fonctionnels.

5.5 Analyse du locus IGL et TRB de l'homme, TRG de la souris et IGK du rat

L'évaluation d'IMGT/LIGMotif peut se faire automatiquement par comparaison des fichiers annotés par les experts et les fichiers contenant les résultats produits automatiquement par le programme. Cependant, l'évaluation d'un annotateur est toujours nécessaire. C'est pourquoi nous avons choisi d'évaluer dans un premier temps des séquences ne dépassant pas les 300000 pb.

L'analyse de 40778 pb du locus IGL humain (dont le numéro d'accession est D86999) a pris 8 secondes. Cette région du locus contient 5 gènes V identifiés dans l'annotation du fichier à plat LIGM-DB. IMGT/LIGMotif a détecté un gène V supplémentaire en plus de ces 5 gènes. L'analyse ultérieure et plus poussée de ce gène permettra de déterminer s'il s'agit d'un artefact ou s'il doit être intégré dans les annotations du locus (Tableau 5.4).

Tableau 5.4. Récapitulatif de l'identification des gènes et de leur fonctionnalité.

Espèces	Locus	Identification												
		Gènes								Fonctionnalités				
		IMGT/LIGM-DB				IMGT/LIGMotif				IMGT/LIGMotif				
		V	D	J	Total	V	D	J	Total	V	D	J	Total	Correctement identifiées
Homme	IGL	5	0	0	5	6	0	0	6	3	0	0	3	3
Souris	TRG	2	0	2	4	2	0	2	4	1	0	0	1	0
Homme	TRB	14	2	14	30	11	2	13	26	8	2	13	23	23
Rat	IGK	1	0	0	1	1	0	0	1	0	0	0	0	0
Total		22	2	16	40	20	2	15	37	12	2	13	27	26

Toutes les fonctionnalités identifiées par IMGT/LIGMotif sont correctes. Le reste des fonctionnalités des gènes est classé comme 'Unknown'. Trois des 4 gènes fonctionnels ont leur fonctionnalité correctement identifiée par IMGT/LIGMotif. Les L-PART1 du gène fonctionnel IGLV5-45 et du pseudogène IGLV(I)-42 n'ont pas été délimités par IMGT/LIGMotif. Il est attendu, que pour le gène IGLV(I)-42, l'outil ne trouve pas le L-PART1 car ce gène appartient à un clan et les clans sont exclus de la base des séquences de références. Le L-INTRON-L du gène IGLV5-45 est mal délimité en 3' car l'INIT-CODON n'a pas été retrouvé. Donc le L-INTRON-L n'a pu être délimité de façon standardisée. Le HSP obtenu par le BLAST a donc servi à sa délimitation. L'ordre des labels est correct.

L'analyse des 232650 pb du locus TRB humain (dont le numéro d'accession est U66061) a pris 5 minutes. Cette région du locus contient 14 gènes V, 2 gènes D et 14 gènes J. Trois gènes V (TRBV22-1 (P), TRBV23-1 (P), TRBVA (P)) et 1 gène J (TRBJ2-2P) n'ont pas été détectés par IMGT/LIGMotif (Tableau 5.4). Toutes les fonctionnalités ainsi que tous les labels IMGT® des gènes identifiés par IMGT/LIGMotif sont correctes à l'exception d'un V-EXON dont l'ACCEPTOR-SPLICE a été mal positionné (Tableau 5.5). Parmi les 11 gènes V identifiés, seule la fonctionnalité de 3 gènes V n'a pas été identifiée.

Tableau 5.5. Mauvais positionnement des labels obtenus avec IMGT/LIGMotif.

IMGT/LIGM-DB	IMGT/LIGMotif	Explications éventuelles
TRBV20-1 (F) :V-EXON 9522-9825	V-EXON 9519-9825	Il y a 2 ACCEPTOR-SPLICE possibles dans la séquence, IMGT/LIGMotif a pris par défaut le 1er

L'ordre des labels est correct à l'exception du gène TRBV30 qui est localisé dans le brin inversé complémentaire. IMGT/LIGMotif ordonne les labels des gènes retrouvés sur le brin direct de l'ADN de la même manière que pour les gènes retrouvés sur le brin inversé complémentaire.

L'analyse des 62031 pb du locus TRG de la souris (dont le numéro d'accession est AF021335) a pris 31 secondes. Cette région du locus contient 2 gènes V et 2 gènes J tous retrouvés par IMGT/LIGMotif (Tableau 5.4). Sur les 4 gènes fonctionnels seule la fonctionnalité du gène TRGV1*02 a été identifiée par IMGT/LIGMotif comme ORF (Tableau 5.6). Cette erreur est liée à un mauvais positionnement de l'heptamère. La fonctionnalité des autres gènes n'a pu être identifiée.

Tableau 5.6. Mauvaise identification de la fonctionnalité des gènes par IMGT/LIGMotif.

IMGT/LIGM-DB	Fonctionnalité	IMGT/LIGMotif	Origine de l'erreur
TRGV1*02	F	ORF	Le V-HEPTAMER, correctement délimité par IMGT/LIGMotif, a une forme non canonique 'cacaaca'.

Les labels IMGT® n'ont pas tous été retrouvés pour le gène TRGJ2 dont il manque le J-HEPTAMER, J-NONAMER et J-SPACER, pour le gène TRGV2 dont il manque le V-RS et le V-HEPTAMER et pour le gène TRGJ4 dont il manque le DONOR-SPLICE (
Tableau 5.7).

Tableau 5.7. Labels manquants dans description des gènes obtenue avec IMGT/LIGMotif.

IMGT/LIGM-DB	Fonctionnalité	Label(s) manquant(s)
TRGJ2	F	J-HEPTAMER, J-SPACER, J-NONAMER
TRGV2	F	V-RS, V-HEPTAMER
TRGJ4	F	DONOR-SPLICE

Tous les labels décrits par IMGT/LIGMotif sont bien positionnés. De plus, IMGT/LIGMotif indique que des labels sont absents ou omis de l'annotation, tels que le 5'-UTR entre TRGJ4 et TRGV1. Ainsi, IMGT/LIGMotif a permis de révéler des erreurs commises lors de l'annotation manuelle de cette séquence. Cependant, l'ordre des labels est mauvais pour les gènes TRGJ2 et TRGV2 qui sont situés dans le brin inversé complémentaire.

L'analyse des 32640 pb du locus IGK du rat (dont le numéro d'accession est AABR03033517) a pris 13 secondes. Cette région du locus contient 1 gène V en sens inversé complémentaire identifié par IMGT/LIGMotif (Tableau 5.4). La fonctionnalité du V n'a pas été identifiée. Le L-PART1 de ce pseudogène (IGKV16-fb) n'a pas été retrouvé par IMGT/LIGMotif. L'ordre des labels est inversé pour le gène V car il est sur le brin inversé complémentaire.

Pour l'ensemble de ces 4 évaluations, IMGT/LIGMotif a identifié 20 gènes V dont un supplémentaire par rapport aux 22 gènes connus (IMGT/LIGM-DB). L'existence réelle de ce gène V devra être vérifiée pour pouvoir l'intégrer aux annotations. Les 2 gènes D ont tous été identifiés et 15 gènes J sur les 16 connus ont été identifiés (Tableau 5.4). Tous les gènes ne sont pas identifiés car ils n'étaient pas assez proches des gènes contenus dans la base de séquences de référence. Le programme est capable d'identifier les gènes dans les deux brins de l'ADN. Certains labels n'ont pas été décrits, soit parce qu'aucun alignement du label n'a été obtenu pour la séquence analysée avec la base de données de séquences de références, soit parce que les motifs conservés délimitant le label n'ont pas été retrouvés. Les erreurs de positionnement des labels sont liées à la description d'artefacts de motifs conservés (par exemple l'ACCEPTOR-SPLICE, l'INIT-CODON). L'identification de la fonctionnalité de 26 gènes sur 27 retrouvés par IMGT/LIGMotif, et dans les annotations d'IMGT/LIGM-DB, est correcte. L'identification de la fonctionnalité est limitée par la conservation de la séquence. Si le gène est trop muté le programme n'identifie pas la fonctionnalité correctement et considère

que la fonctionnalité ne peut être identifiée en lui attribuant la classe ‘Unknown’. La rapidité d’exécution du programme est satisfaisante puisque le temps d’attente ne dépasse pas les 5 minutes pour une séquence longue de 232650 pb. Les gènes inversés complémentaires sont présentés dans l’ordre inversé. Il faudra donc, à l’avenir, faire en sorte que les labels d’un gène inversé complémentaire s’affichent dans l’ordre d’annotation et non en fonction de la position des labels sur la séquence.

5.6 Conclusion

L’analyse du système d’annotation manuelle des IG et TR existant à IMGT®, la conception et l’implémentation d’IMGT/LIGMotif, ainsi que son application Web, ont constitué la majeure partie de mon travail de thèse. Le logiciel intègre les programmes spécialisés BLAST dans l’identification des gènes V, D et J, et IMGT/V-QUEST dans la description des V-REGION. L’analyse du processus d’annotation des séquences génomiques des IG et TR a permis de mieux comprendre quelle était la place de ces programmes pour mieux les intégrer dans IMGT/LIGMotif qui forme un système cohérent dédié à l’annotation des gènes V, D et J.

L’algorithme est compartimenté en 4 modules, chacun ayant pour tâche d’identifier et/ou de décrire la séquence. Le premier module identifie les gènes V, D et J, le second les décrit, le quatrième leur attribue une fonctionnalité et le dernier décrit les gènes entre eux en les assemblant et identifie leur cluster.

L’interface Web constitue un moyen efficace, simple et convivial de paramétrer IMGT/LIGMotif et de visualiser les résultats. La séquence peut être soumise en format FASTA ou EMBL et les bases de séquences de références sont paramétrables via l’interface. L’interface des résultats permet de visualiser les statistiques générales en fonction du nombre de gènes sur les deux brins de l’ADN, du statut de description et de la fonctionnalité des gènes les mieux décrits (statut ‘GENE-UNIT’). Tous les gènes identifiés sont affichés et peuvent être sélectionnés pour obtenir le détail de leurs labels constitutifs. Il est possible d’exporter les résultats de la description dans un format ‘CSV’ ou ‘XML’.

L’analyse des résultats d’IMGT/LIGMotif a permis de vérifier ses performances et de voir comment les annotateurs pouvaient utiliser le programme et ses résultats. La plupart des

gènes sont identifiés (37 sur 40) dans chacun des brins de l'ADN. La plupart des labels des gènes identifiés sont décrits correctement. De plus, l'identification de la fonctionnalité de 26 gènes sur 27 retrouvés par IMGT/LIGMotif et dans les annotations d'IMGT/LIGM-DB est correcte. IMGT/LIGMotif a fourni dans un temps tout-à-fait respectable une annotation exploitable rapidement et qui a permis de mettre en avant des erreurs d'annotation résultant d'une mauvaise saisie des positions. Ces résultats serviront de base pour corriger les séquences étudiées.

CONCLUSIONS ET PERSPECTIVES

IMGT® est le système d'information international en ImMunoGénétique, spécialisé dans la gestion des séquences et des structures 3D des récepteurs d'antigènes IG et TR, des vertébrés à mâchoire. Ces récepteurs d'antigènes assurent la reconnaissance antigénique et la spécificité du système immunitaire adaptatif. Compte tenu de la nécessité d'analyser la masse de données provenant du séquençage des génomes et de produire des connaissances biologiques et des standards de haute qualité utilisés tant en recherche fondamentale que médicale, nous avons développé IMGT/LIGMotif, un logiciel dédié à l'annotation des gènes des IG et TR dans des séquences génomiques de grande taille. L'originalité de cet outil est d'être basé sur les axiomes et concepts d'IMGT-ONTOLOGY et de permettre ainsi l'annotation automatique standardisée (identification et description) des gènes variables (V), de diversité (D) et de jonction (J) dans des séquences génomiques. Cette approche s'avère originale car c'est, à notre connaissance, la première dédiée à l'annotation automatique des gènes V, D et J des récepteurs d'antigènes dans les séquences génomiques.

La première étape de ma thèse a consisté à analyser l'organisation des locus tels qu'ils sont gérés par IMGT®, à approfondir tous les aspects d'IMGT-ONTOLOGY et à analyser l'existant, préalables nécessaires pour concevoir et développer un outil Java capable d'annoter les gènes V, D et J des IG et TR dans l'ADN génomique et, ainsi, d'accélérer le processus d'annotation d'IMGT®.

L'annotation des génomes est une tâche ardue et soumise à des erreurs possibles. Une analyse rigoureuse est nécessaire pour l'identification des gènes et des allèles d'IG et TR dans le génome d'autant plus que ces gènes appartiennent à des familles multigéniques et sont caractérisées par de nombreux polymorphismes. L'analyse des locus de l'homme et de la souris ont permis de répertorier les gènes existants et la taille des locus. Les locus des IG et TR contiennent plusieurs centaines de gènes et ce nombre varie en fonction de l'haplotype. L'homme possède de 608 à 665 gènes d'IG et TR alors que la souris possède de 619 à 628 gènes. Chez l'homme, la taille des locus varie de 60 kb (pour le locus TRD) à 1820 kb (pour le locus IGK), et chez la souris de 205 kb (pour le locus TRG) à 3200 kb (pour le locus IGK). Cette analyse a démontré les difficultés de l'annotation des gènes V, D et J des IG et TR

auxquelles IMGT/LIGMotif doit répondre notamment la taille des locus et le nombre de gènes à traiter et les requis informatiques (espace mémoire et temps d'exécution).

Il a ensuite été nécessaire d'analyser le système d'annotation des gènes V, D et J des IG et TR afin de le formaliser et d'en identifier les points automatisables. L'étude du système d'annotation des IG et TR d'IMGT® m'a également permis de répertorier les logiciels utilisés et d'identifier leurs fonctions. BLAST, LIGMotif (« l'ancêtre » d'IMGT/LIGMotif) et IMGT/V-QUEST ressortent du processus. BLAST permet aux annotateurs d'identifier et de classer les gènes. LIGMotif identifie et décrit les gènes d'IG et TR. IMGT/V-QUEST décrit précisément les V-REGION et identifie l'allèle le plus proche.

Le modèle d'IMGT/LIGMotif a été conçu pour prendre en compte les particularités structurales des gènes V, D et J mais aussi la variabilité des motifs des gènes d'IG et TR en fonction de l'espèce, du locus et du type de gènes, notamment pour les heptamères et les nonamères. La conception de bases de données de séquences de références classées en fonction de l'espèce, du locus et du type de gène pour chaque motif permet de prendre en compte cette variabilité. Les concepts d'identification, de classification, de description, de localisation et d'orientation d'IMGT-ONTOLOGY ont été utilisés pour la standardisation du programme. Ainsi l'outil gère l'identification du type des gènes ('variable', 'diversity' et 'junction') et la fonctionnalité des séquences d'ADNg ('functional', 'ORF' et 'pseudogene'). IMGT/LIGMotif gère la description de chaque type de gène identifié (V, D et J) en le décrivant à l'aide de prototype V-GENE, D-GENE et J-GENE, et des patterns et labels correspondants. Le modèle d'IMGT/LIGMotif intègre les relations qui existent entre labels pour un prototype donné. Alors que dix relations sont nécessaires et suffisantes pour la description d'un prototype. Deux nouvelles relations ont été ajoutées afin de décrire la localisation respective de deux motifs distants l'un de l'autre (« is_in_5_prime_of » et « is_in_3_prime_of »). Chaque séquence analysée contenant plusieurs gènes est décrite par un label désignant un cluster. Les gènes sont orientés dans la séquence en fonction du brin dans lequel ils ont été trouvés ('Plus' ou 'Minus'). L'utilisation d'IMGT/V-QUEST enrichit IMGT/LIGMotif par d'autres concepts provenant de la classification et de la numérotation. En particulier, la numérotation standardisée et la description de la V-REGION et de ses différents motifs qui sont déduits des alignements réalisés par IMGT/V-QUEST.

L'algorithme d'IMGT/LIGMotif débute par la recherche des similarités en réalisant des alignements locaux (BLAST) entre la séquence à analyser et les séquences de références. Les gènes sont identifiés en regroupant des alignements locaux de motifs appartenant à un même type de gène. Les motifs des patterns sont ensuite recherchés, par alignements, dans une zone restreinte à proximité et/ou dans les parties du gène identifié (car plus difficiles à retrouver). Les gènes V, D et J identifiés sont ainsi décrits à partir des motifs et des patterns, à l'exception de la description de la V-REGION et de ses motifs constitutifs qui est réalisée par IMGT/V-QUEST. La description totale des motifs constitutifs du V-EXON est réalisée à partir de la V-REGION qui permet de délimiter le L-PART2, le dernier motif nécessaire pour une description complète du V-EXON. La fonctionnalité est identifiée pour chaque gène en utilisant les critères de la charte scientifique d'IMGT®. Finalement, les labels V-GENE, D-GENE et J-GENE ainsi que les 5'UTR et 3'UTR, sont attribués et assemblés en clusters. Une priorité sera l'enrichissement des bases de données d'IMGT/LIGMotif. Nous souhaitons ainsi développer la mise à jour automatique des séquences de références contenues dans IMGT/GENE-DB vers les bases d'IMGT/LIGMotif. Toute modification apportée à IMGT/GENE-DB serait ainsi reportée systématiquement dans la base d'IMGT/LIGMotif. Cette mise à jour est particulièrement importante car l'interface Web d'IMGT/LIGMotif permet à l'annotateur de sélectionner les bases de motifs de références. A l'heure actuelle, l'interface comprend une vue synthétique qui donne accès à une description détaillée des gènes sélectionnés avec leurs labels et à leur positionnement précis dans la séquence. L'exportation des résultats peut se faire sous le format CSV ou XML (eXtensible Markup Language). Actuellement, il est possible d'exécuter IMGT/LIGMotif en ligne de commande et de visualiser les résultats à travers une interface d'annotation. Nous envisageons d'associer l'application web d'IMGT/LIGMotif avec une interface d'annotation en ligne qui est dans sa phase finale de développement à IMGT®. Cette interface devrait permettre de modifier et compléter les annotations produites par IMGT/LIGMotif en ligne.

IMGT/LIGMotif a été évalué pour les locus IGL et TRB de l'homme, TRG de la souris, et IGK du rat. Des évaluations plus ponctuelles des locus de l'homme et de la souris et de gènes de Téléostei ont également été réalisées. Les résultats ont montré qu'IMGT/LIGMotif facilitait le travail des annotateurs en identifiant les gènes, leur fonctionnalité et leur cluster. Les annotations d'IMGT/LIGMotif sont produites automatiquement, ce qui évite certaines omissions possibles. Le temps d'exécution du programme de l'ordre de la minute est suffisant pour pouvoir annoter des séquences de

plusieurs centaines de milliers de nucléotides. IMGT/LIGMotif représente ainsi une étape importante dans l'annotation automatique de séquences génomiques de grande taille des IG et TR.

IMGT/LIGMotif a souligné l'importance d'une approche standardisée des annotations. L'annotation d'une séquence génomique fournit des informations essentielles à l'analyse bioinformatique et moléculaire. Cependant, la terminologie biologique est notoirement ambiguë puisque le même mot est souvent utilisé pour décrire plusieurs choses. Si les annotations étaient toujours décrites avec le même langage, alors l'analyse comparative de l'information serait énormément simplifiée. Les ontologies œuvrent dans ce sens. Elles facilitent l'accès, l'échange, l'analyse et la gestion des données biologiques. En effet, la description des données d'un domaine biologique est plus précise, pertinente, standardisée et rigoureuse dans une ontologie que dans une description libre. De plus, la description d'une ontologie de domaine spécifie formellement les relations qui existent entre les concepts.

La consistance des annotations est cruciale quand les annotations de différentes sources doivent être comparées. En conséquence, et avant que les annotations soient validées, leur consistance peut être évaluée par comparaison aux connaissances contenues dans l'ontologie de référence. Dans le processus de validation de Sequence Ontology (SO) [12, 173], chaque affirmation faite dans une annotation doit être trouvée dans l'ontologie pour que l'annotation soit validée.

L'annotation basée sur une ontologie facilite l'association des données biologiques. La standardisation des annotations par l'ontologie permet le traitement automatique des données biologiques brutes en répercutant les annotations des données similaires. C'est ainsi, par exemple le cas de l'annotation électronique des données du transcriptome avec Gene Ontology, http://www.geneontology.org/GO.evidence.shtml. L'association entre des ontologies spécialisées et des ontologies générales est délicate. Ainsi, l'association d'IMGT-ONTOLOGY, spécialisée et très détaillée pour les IG, TR et MHC, avec Gene Ontology se fait uniquement au niveau des identifiants de ces types de molécules, tandis qu'avec SO, 64 labels sont du commun.

L'existence de relations entre concepts permet de naviguer au sein d'une ontologie. Par exemple, la délimitation des domaines (données structurales) définie dans IMGT-

ONTOLOGY, tient compte de la structure des gènes en exons et par suite, la définition de domaine peut ainsi être transposée au niveau des protéines, des transcrits et des séquences génomiques. L'annotation issue des concepts d'IMGT-ONTOLOGY peut donc être utilisée afin de passer des structures 3D des protéines à l'ADN génomique, et inversement, facilitant par exemple, les travaux d'ingénierie moléculaire tels que l'humanisation des anticorps [192-193].

Les concepts d'une ontologie peuvent aussi servir de base à la construction d'autres ontologies et viennent compléter les connaissances qui existent. Par exemple, les concepts d'IMGT-ONTOLOGY sont diffusés dans d'autres ontologies, telles que 'Immunome Epitope Database and Analysis Resource' (IEDB) [194]. Les concepts d'interaction d'IMGT-ONTOLOGY élaborés au cours du projet ImmunoGrid [195-196] ont permis en les généralisant aux composants cellulaires, de les faire évoluer avec le modèle Catania Mouse Model (CMM) (un protocole de vaccination chez la souris [197]). Les ontologies sont donc utilisées pour l'échange et le partage des connaissances dans des domaines de recherche très divers.

L'annotation est un processus de recherche d'information qui peut nécessiter l'utilisation de plusieurs outils à la suite. La conception d'un système d'interaction entre outils (workflow) basé sur une ontologie simplifie la recherche d'information. IMGT-ONTOLOGY a permis de concevoir IMGT-Choreography [198], un système permettant de gérer au mieux les requêtes à travers les logiciels et bases de données d'IMGT® en utilisant un format d'échange standard XML (IMGT-ML [199-200]).

Les annotations sont conservées et organisées dans des bases de données qui peuvent aussi bénéficier de l'apport des ontologies. Bien qu'initialement conçue pour gérer les connaissances du domaine complexe de l'immunologie, IMGT-ONTOLOGY a été le modèle d'une base de données PenBase [201] relative aux peptides antimicrobiens ou *Penaeidins*, de six espèces de crevette, en reprenant les concepts de description, de classification et de numérotation. Les concepts de numérotation (Colliers de Perles) ont été adoptés dans deux bases de données, de peptides cycliques, Knottin [202] et CyBase [203-204].

Les ontologies offrent à la communauté de programmation en bioinformatique une opportunité significative d'améliorer la conception de logiciels d'annotation (par exemple de

visualisation des annotations ou d'annotation automatique comme IMGT/LIGMotif) et la rapidité du cycle de développement. Si un logiciel est implémenté avec un vocabulaire contrôlé (et donc avec des termes qui ne spécifient pas eux-mêmes leurs relations contrairement aux concepts dans une ontologie), les relations (de type hiérarchique, topologique, méréologique…) doivent être codées dans le programme. Dans une ontologie l'ajout d'un nouveau concept et/ou le changement des relations entre les concepts nécessite de revoir les applications qui font l'usage de ces concepts. Ainsi, un logiciel conçu pour intégrer les modifications d'une ontologie aurait seulement besoin qu'elle soit mise à jour et le reste suivrait automatiquement. Néanmoins, l'utilisation d'une ontologie requiert que le logiciel soit capable de lire un fichier contenant l'ontologie aux formats tels que le langage d'ontologie web (OWL, http://www.w3.org/TR/owl-guide/) ou Open Biomedical Ontologies (OBO) [205] et puisse naviguer dans le réseau de relations. Actuellement IMGT-ONTOLOGY est développé avec le logiciel Protégé à partir duquel on obtient un fichier de l'ontologie au format OWL. Il est donc envisageable de faire évoluer IMGT/LIGMotif de manière à ce qu'il puisse lire les fichiers OWL, et ainsi d'intégrer par le biais de l'ontologie entre concepts.

BIBLIOGRAPHIE

1. Sanger, F., Nicklen, S. and Coulson, A.R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74, 5463-5467 (1977).
2. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., *et al.* Initial sequencing and analysis of the human genome. Nature. 409, 860-921 (2001).
3. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. and et al. The sequence of the human genome. Science. 291, 1304-1351 (2001).
4. Mardis, E.R. The impact of next-generation sequencing technology on genetics. Trends Genet. 24, 133-141 (2008).
5. Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P. and et al. EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res. 35, D16-20 (2007).
6. Sugawara, H., Ikeo, K., Fukuchi, S., Gojobori, T. and Tateno, Y. DDBJ dealing with mass data produced by the second generation sequencer. Nucleic Acids Res. 37, D16-18 (2009).
7. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. GenBank. Nucleic Acids Res. 37, D26-31 (2009).
8. Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D. and Lefranc, M.-P. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 34, D781-784 (2006).
9. Chang, A., Scheer, M., Grote, A., Schomburg, I. and Schomburg, D. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res. 37, D588-592 (2009).
10. Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R. and Zhang, H. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 37, D555-559 (2009).
11. The Gene Ontology, c. The Gene Ontology project in 2008. Nucleic Acids Res. 36, D440-444 (2008).
12. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 6, R44 (2005).
13. Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J., Chang, T.-C., Hu, Z., Liu, H., Smith, B. and Wu, C.H. Framework for a protein ontology. BMC Bioinformatics. 8 Suppl 9, S1 (2007).
14. Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bosc, N., Folch, G., Guiraudou, D., Jabado-Michaloud, J., Magris, S., Scaviner, D., Thouvenin, V., Combres, K., Girod, D., Jeanjean, S., Protat, C., Yousfi-Monod, M., Duprat, E., Kaas, Q., Pommié, C., Chaume, D. and Lefranc, G. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. In Silico Biol. 4, 17-29 (2004).

15. Duroux, P., Kaas, Q., Brochet, X., Lane, J., Ginestoux, C., Lefranc, M.-P. and Giudicelli, V. IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie. 90, 570-583 (2008).
16. Lloyd, C.M., Halstead, M.D.B. and Nielsen, P.F. CellML: its future, present and past. Prog Biophys Mol Biol. 85, 433-450 (2004).
17. Lloyd, C.M., Lawson, J.R., Hunter, P.J. and Nielsen, P.F. The CellML Model Repository. Bioinformatics. 24, 2122-2123 (2008).
18. Halling-Brown, M.D., Moss, D.S., Sansom, C.E. and Shepherd, A.J. A computational Grid framework for immunological applications. Philos Transact A Math Phys Eng Sci. 367, 2705-2716 (2009).
19. Lefranc M-P, L.G. The Immunoglobulin FactsBook, 2001.
20. Lefranc M-P, L.G. The T cell receptor FactsBook, 2001.
21. Sakano, H., Hüppi, K., Heinrich, G. and Tonegawa, S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature. 280, 288-94 (1979).
22. Bleakley, K., Lefranc, M.-P. and Biau, G. Recovering probabilities for nucleotide trimming processes for T cell receptor TRA and TRG V-J junctions analyzed with IMGT tools. BMC Bioinformatics. 9, 408 (2008).
23. Alt, F.W. and Baltimore, D. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. Proc Natl Acad Sci U S A. 79, 4118-4122 (1982).
24. Neuberger, M.S. and Rada, C. Somatic hypermutation: activation-induced deaminase for C/G followed by polymerase eta for A/T. J Exp Med. 204, 7-10 (2007).
25. Gearhart, P.J., Johnson, N.D., Douglas, R. and Hood, L. IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. Nature. 291, 29-34 (1981).
26. Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. and Duroux, P. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 37, D1006-1012 (2009).
27. Giudicelli, V. and Lefranc, M.-P. Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics. 15, 1047-1054 (1999).
28. Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W. and Povey, S. Guidelines for human gene nomenclature. Genomics. 79, 464-470 (2002).
29. 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. 32, 461-463 (2008).
30. Lefranc, M.-P. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. Immunogenetics. 59, 899-902 (2007).
31. Giudicelli, V., Chaume, D. and Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 33, D256-261 (2005).
32. Letovsky, S.I., Cottingham, R.W., Porter, C.J. and Li, P.W. GDB: the Human Genome Database. Nucleic Acids Res. 26, 94-99 (1998).
33. Pruitt, K.D. and Maglott, D.R. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29, 137-140 (2001).
34. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 33, D54-58 (2005).
35. Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K.,

Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., *et al.* Ensembl 2009. Nucleic Acids Res. 37, D690-697 (2009).
36. Loveland, J. VEGA, the genome browser with a difference. Brief Bioinform. 6, 189-193 (2005).
37. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. and Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33, 6494-6506 (2005).
38. Burge, C. and Karlin, S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 268, 78-94 (1997).
39. Gross, S.S. and Brent, M.R. Using multiple alignments to improve gene prediction. J Comput Biol. 13, 379-393 (2006).
40. Tonegawa, S. Somatic generation of antibody diversity. Nature. 302, 575-581 (1983).
41. Brack, C., Hirama, M., Lenhard-Schuller, R. and Tonegawa, S. A complete immunoglobulin gene is created by somatic recombination. Cell. 15, 1-14 (1978).
42. Weigert, M., Perry, R., Kelley, D., Hunkapiller, T., Schilling, J. and Hood, L. The joining of V and J gene segments creates antibody diversity. Nature. 283, 497-499 (1980).
43. Croce, C.M., Shander, M., Martinis, J., Cicurel, L., D'Ancona, G.G., Dolby, T.W. and Koprowski, H. Chromosomal location of the genes for human immunoglobulin heavy chains. Proc Natl Acad Sci U S A. 76, 3416-3419 (1979).
44. Kirsch, I.R., Morton, C.C., Nakahara, K. and Leder, P. Human immunoglobulin heavy chain genes map to a region of translocations in malignant B lymphocytes. Science. 216, 301-303 (1982).
45. McBride, O.W., Battey, J., Hollis, G.F., Swan, D.C., Siebenlist, U. and Leder, P. Localization of human variable and constant region immunoglobulin heavy chain genes on subtelomeric band q32 of chromosome 14. Nucleic Acids Res. 10, 8155-8170 (1982).
46. Shin, E.K., Matsuda, F., Nagaoka, H., Fukita, Y., Imai, T., Yokoyama, K., Soeda, E. and Honjo, T. Physical map of the 3' region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. EMBO J. 10, 3641-3645 (1991).
47. Matsuda, F., Shin, E.K., Nagaoka, H., Matsumura, R., Haino, M., Fukita, Y., Taka-ishi, S., Imai, T., Riley, J.H., Anand, R. and et al. Structure and physical map of 64 variable segments in the 3'0.8-megabase region of the human immunoglobulin heavy-chain locus. Nat Genet. 3, 88-94 (1993).
48. Cook, G.P., Tomlinson, I.M., Walter, G., Riethman, H., Carter, N.P., Buluwela, L., Winter, G. and Rabbitts, T.H. A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome 14q. Nat Genet. 7, 162-168 (1994).
49. Cook, G.P. and Tomlinson, I.M. The human immunoglobulin VH repertoire. Immunol Today. 16, 237-242 (1995).
50. Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K.i., Hayashida, H., Miyata, T. and Honjo, T. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med. 188, 2151-2162 (1998).
51. Pallarès, N., Lefebvre, S., Contet, V., Matsuda, F. and Lefranc, M.P. The human immunoglobulin heavy variable genes. Exp Clin Immunogenet. 16, 36-60 (1999).
52. Siebenlist, U., Ravetch, J.V., Korsmeyer, S., Waldmann, T. and Leder, P. Human immunoglobulin D segments encoded in tandem multigenic families. Nature. 294, 631-635 (1981).

53. Buluwela, L., Albertson, D.G., Sherrington, P., Rabbitts, P.H., Spurr, N. and Rabbitts, T.H. The use of chromosomal translocations to study human immunoglobulin gene organization: mapping DH segments within 35 kb of the C mu gene and identification of a new DH locus. EMBO J. 7, 2003-2010 (1988).
54. Corbett, S.J., Tomlinson, I.M., Sonnhammer, E.L., Buck, D. and Winter, G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. J Mol Biol. 270, 587-597 (1997).
55. Ruiz, M., Pallarès, N., Contet, V., Barbi, V. and Lefranc, M.-P. The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. Exp Clin Immunogenet. 16, 173-184 (1999).
56. Ravetch, J.V., Siebenlist, U., Korsmeyer, S., Waldmann, T. and Leder, P. Structure of the human immunoglobulin mu locus: characterization of embryonic and rearranged J and D genes. Cell. 27, 583-591 (1981).
57. Ellison, J., Buxbaum, J. and Hood, L. Nucleotide sequence of a human immunoglobulin C gamma 4 gene. DNA. 1, 11-18 (1981).
58. Rabbitts, T.H., Forster, A. and Milstein, C.P. Human immunoglobulin heavy chain genes: evolutionary comparisons of C mu, C delta and C gamma genes and associated switch sequences. Nucleic Acids Res. 9, 4509-4524 (1981).
59. Ellison, J. and Hood, L. Linkage and sequence homology of two human immunoglobulin gamma heavy chain constant region genes. Proc Natl Acad Sci U S A. 79, 1984-1988 (1982).
60. Ellison, J.W., Berson, B.J. and Hood, L.E. The nucleotide sequence of a human immunoglobulin C gamma1 gene. Nucleic Acids Res. 10, 4071-4079 (1982).
61. Flanagan, J.G. and Rabbitts, T.H. Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing gamma, epsilon and alpha genes. Nature. 300, 709-713 (1982).
62. Lefranc, M.-P., Lefranc, G. and Rabbitts, T.H. Inherited deletion of immunoglobulin heavy chain constant region genes in normal human individuals. Nature. 300, 760-762 (1982).
63. Lefranc, M.-P., Lefranc, G., de Lange, G., Out, T.A., van den Broek, P.J., van Nieuwkoop, J., Radl, J., Helal, A.N., Chaabani, H. and van Loghem, E. Instability of the human immunoglobulin heavy chain constant region locus indicated by different inherited chromosomal deletions. Mol Biol Med. 1, 207-217 (1983).
64. Flanagan, J.G., Lefranc, M.-P. and Rabbitts, T.H. Mechanisms of divergence and convergence of the human immunoglobulin alpha 1 and alpha 2 constant region gene sequences. Cell. 36, 681-688 (1984).
65. White, M.B., Shen, A.L., Word, C.J., Tucker, P.W. and Blattner, F.R. Human immunoglobulin D: genomic sequence of the delta heavy chain. Science. 228, 733-737 (1985).
66. Huck, S., Fort, P., Crawford, D.H., Lefranc, M.-P. and Lefranc, G. Sequence of a human immunoglobulin gamma 3 heavy chain constant region gene: comparison with the other human C gamma genes. Nucleic Acids Res. 14, 1779-1789 (1986).
67. Bensmana, M., Huck, S., Lefranc, G. and Lefranc, M.-P. The human immunoglobulin pseudo-gamma IGHGP gene shows no major structural defect. Nucleic Acids Res. 16, 3108 (1988).
68. Huck, S., Lefranc, G. and Lefranc, M.-P. A human immunoglobulin IGHG3 allele (Gmb0,b1,c3,c5,u) with an IGHG4 converted region and three hinge exons. Immunogenetics. 30, 250-257 (1989).

69. Malcolm, S., Barton, P., Murphy, C., Ferguson-Smith, M.A., Bentley, D.L. and Rabbitts, T.H. Localization of human immunoglobulin kappa light chain variable region genes to the short arm of chromosome 2 by in situ hybridization. Proc Natl Acad Sci U S A. 79, 4957-61 (1982).
70. McBride, O.W., Hieter, P.A., Hollis, G.F., Swan, D., Otey, M.C. and Leder, P. Chromosomal location of human kappa and lambda immunoglobulin light chain constant region genes. J Exp Med. 155, 1480-1490 (1982).
71. Barbié, V. and Lefranc, M.-P. The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp Clin Immunogenet. 15, 171-183 (1998).
72. Cox, J.P., Tomlinson, I.M. and Winter, G. A directory of human germ-line V kappa segments reveals a strong bias in their usage. Eur J Immunol. 24, 827-836 (1994).
73. Huber, C., Schäble, K.F., Huber, E., Klein, R., Meindl, A., Thiebe, R., Lamm, R. and Zachau, H.G. The V kappa genes of the L regions and the repertoire of V kappa gene sequences in the human germ line. Eur J Immunol. 23, 2868-2875 (1993).
74. Scaviner, D., Barbié, V., Ruiz, M. and Lefranc, M.-P. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp Clin Immunogenet. 16, 234-240 (1999).
75. Schäble, K., Thiebe, R., Flügel, A., Meindl, A. and Zachau, H.G. The human immunoglobulin kappa locus: pseudogenes, unique and repetitive sequences. Biol Chem Hoppe Seyler. 375, 189-199 (1994).
76. Schäble, K.F. and Zachau, H.G. The variable genes of the human immunoglobulin kappa locus. Biol Chem Hoppe Seyler. 374, 1001-1022 (1993).
77. Zachau, H.G. The immunoglobulin kappa locus-or-what has been learned from looking closely at one-tenth of a percent of the human genome. Gene. 135, 167-173 (1993).
78. Hieter, P.A., Maizel, J.V., Jr. and Leder, P. Evolution of human immunoglobulin kappa J region genes. J Biol Chem. 257, 1516-1522 (1982).
79. Hieter, P.A., Max, E.E., Seidman, J.G., Maizel, J.V., Jr. and Leder, P. Cloned human and mouse kappa immunoglobulin constant and J region genes conserve homology in functional segments. Cell. 22, 197-207 (1980).
80. Erikson, J., Martinis, J. and Croce, C.M. Assignment of the genes for human lambda immunoglobulin chains to chromosome 22. Nature. 294, 173-175 (1981).
81. Emanuel, B.S., Cannizzaro, L.A., Magrath, I., Tsujimoto, Y., Nowell, P.C. and Croce, C.M. Chromosomal orientation of the lambda light chain locus: V lambda is proximal to C lambda in 22q11. Nucleic Acids Res. 13, 381-387 (1985).
82. Frippiat, J.P., Williams, S.C., Tomlinson, I.M., Cook, G.P., Cherif, D., Le Paslier, D., Collins, J.E., Dunham, I., Winter, G. and Lefranc, M.-P. Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2. Hum Mol Genet. 4, 983-991 (1995).
83. Kawasaki, K., Minoshima, S., Schooler, K., Kudoh, J., Asakawa, S., de Jong, P.J. and Shimizu, N. The organization of the human immunoglobulin lambda gene locus. Genome Res. 5, 125-135 (1995).
84. Williams, S.C., Frippiat, J.P., Tomlinson, I.M., Ignatovich, O., Lefranc, M.-P. and Winter, G. Sequence and evolution of the human germline V lambda repertoire. J Mol Biol. 264, 220-232 (1996).
85. Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J.L., Wang, J. and Shimizu, N. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. Genome Res. 7, 250-261 (1997).

86. Pallarès, N., Frippiat, J.P., Giudicelli, V. and Lefranc, M.-P. The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp Clin Immunogenet. 15, 8-18 (1998).
87. Hieter, P.A., Hollis, G.F., Korsmeyer, S.J., Waldmann, T.A. and Leder, P. Clustered arrangement of immunoglobulin lambda constant region genes in man. Nature. 294, 536-540 (1981).
88. Taub, R.A., Hollis, G.F., Hieter, P.A., Korsmeyer, S., Waldmann, T.A. and Leder, P. Variable amplification of immunoglobulin lambda light-chain genes in human populations. Nature. 304, 172-174 (1983).
89. Dariavach, P., Lefranc, G. and Lefranc, M.-P. Human immunoglobulin C lambda 6 gene encodes the Kern+Oz-lambda chain and C lambda 4 and C lambda 5 are pseudogenes. Proc Natl Acad Sci U S A. 84, 9074-9078 (1987).
90. Vasicek, T.J. and Leder, P. Structure and expression of the human immunoglobulin lambda genes. J Exp Med. 172, 609-20 (1990).
91. Rabbitts, T.H., Lefranc, M.-P., Stinson, M.A., Sims, J.E., Schroder, J., Steinmetz, M., Spurr, N.L., Solomon, E. and Goodfellow, P.N. The chromosomal location of T-cell receptor genes and a T cell rearranging gene: possible correlation with specific translocations in human T cell leukaemia. EMBO J. 4, 1461-1465 (1985).
92. Arden, B., Clark, S.P., Kabelitz, D. and Mak, T.W. Human T-cell receptor variable gene segment families. Immunogenetics. 42, 455-500 (1995).
93. Scaviner, D. and Lefranc, M.-P. The human T cell receptor alpha variable (TRAV) genes. Exp Clin Immunogenet. 17, 83-96 (2000).
94. Koop, B.F., Rowen, L., Wang, K., Kuo, C.L., Seto, D., Lenstra, J.A., Howard, S., Shan, W., Deshpande, P. and Hood, L. The human T-cell receptor TCRAC/TCRDC (C alpha/C delta) region: organization, sequence, and evolution of 97.6 kb of DNA. Genomics. 19, 478-493 (1994).
95. Baer, R., Lefranc, M.-P., Minowada, J., Forster, A., Stinson, M.A. and Rabbitts, T.H. Organization of the T-cell receptor alpha-chain gene and rearrangement in human T-cell leukaemias. Mol Biol Med. 3, 265-277 (1986).
96. Yoshikai, Y., Clark, S.P., Taylor, S., Sohn, U., Wilson, B.I., Minden, M.D. and Mak, T.W. Organization and sequences of the variable, joining and constant region genes of the human T-cell receptor alpha-chain. Nature. 316, 837-40 (1985).
97. Loh, E.Y., Cwirla, S., Serafini, A.T., Phillips, J.H. and Lanier, L.L. Human T-cell-receptor delta chain: genomic organization, diversity, and expression in populations of cells. Proc Natl Acad Sci U S A. 85, 9714-9718 (1988).
98. Takihara, Y., Tkachuk, D., Michalopoulos, E., Champagne, E., Reimann, J., Minden, M. and Mak, T.W. Sequence and organization of the diversity, joining, and constant region genes of the human T-cell delta-chain locus. Proc Natl Acad Sci U S A. 85, 6097-6101 (1988).
99. Satyanarayana, K., Hata, S., Devlin, P., Roncarolo, M.G., De Vries, J.E., Spits, H., Strominger, J.L. and Krangel, M.S. Genomic organization of the human T-cell antigen-receptor alpha/delta locus. Proc Natl Acad Sci U S A. 85, 8166-8170 (1988).
100. Barker, P.E., Ruddle, F.H., Royer, H.D., Acuto, O. and Reinherz, E.L. Chromosomal location of human T-cell receptor gene Ti beta. Science. 226, 348-349 (1984).
101. Caccia, N., Kronenberg, M., Saxe, D., Haars, R., Bruns, G.A., Goverman, J., Malissen, M., Willard, H., Yoshikai, Y., Simon, M. and et al. The T cell receptor beta chain genes are located on chromosome 6 in mice and chromosome 7 in humans. Cell. 37, 1091-1099 (1984).

102. Isobe, M., Erikson, J., Emanuel, B.S., Nowell, P.C. and Croce, C.M. Location of gene for beta subunit of human T-cell receptor at band 7q35, a region prone to rearrangements in T cells. Science. 228, 580-582 (1985).
103. Folch, G. and Lefranc, M.-P. The human T cell receptor beta variable (TRBV) genes. Exp Clin Immunogenet. 17, 42-54 (2000).
104. Lai, E., Concannon, P. and Hood, L. Conserved organization of the human and murine T-cell receptor beta-gene families. Nature. 331, 543-546 (1988).
105. Rowen, L., Koop, B.F. and Hood, L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. Science. 272, 1755-1762 (1996).
106. Slightom, J.L., Siemieniak, D.R., Sieu, L.C., Koop, B.F. and Hood, L. Nucleotide sequence analysis of 77.7 kb of the human V beta T-cell receptor gene locus: direct primer-walking using cosmid template DNAs. Genomics. 20, 149-168 (1994).
107. Wei, S., Charmley, P., Robinson, M.A. and Concannon, P. The extent of the human germline T-cell receptor V beta gene segment repertoire. Immunogenetics. 40, 27-36 (1994).
108. Wilson, R.K., Lai, E., Concannon, P., Barth, R.K. and Hood, L.E. Structure, organization and polymorphism of murine and human T-cell receptor alpha and beta chain gene families. Immunol Rev. 101, 149-172 (1988).
109. Toyonaga, B., Yoshikai, Y., Vadasz, V., Chin, B. and Mak, T.W. Organization and sequences of the diversity, joining, and constant region genes of the human T-cell receptor beta chain. Proc Natl Acad Sci U S A. 82, 8624-8628 (1985).
110. Folch, G. and Lefranc, M.-P. The human T cell receptor beta diversity (TRBD) and beta joining (TRBJ) genes. Exp Clin Immunogenet. 17, 107-114 (2000).
111. Lefranc, M.-P., Chuchana, P., Dariavach, P., Nguyen, C., Huck, S., Brockly, F., Jordan, B. and Lefranc, G. Molecular mapping of the human T cell receptor gamma (TRG) genes and linkage of the variable and constant regions. Eur J Immunol. 19, 989-994 (1989).
112. Bensmana, M., Mattei, M.G. and Lefranc, M.-P. Localization of the human T-cell receptor gamma locus (TCRG) to 7p14----p15 by in situ hybridization. Cytogenet Cell Genet. 56, 31-32 (1991).
113. Lefranc, M.-P., Forster, A., Baer, R., Stinson, M.A. and Rabbitts, T.H. Diversity and rearrangement of the human T cell rearranging gamma genes: nine germ-line variable genes belonging to two subgroups. Cell. 45, 237-246 (1986).
114. Forster, A., Huck, S., Ghanem, N., Lefranc, M.-P. and Rabbitts, T.H. New subgroups in the human T cell rearranging V gamma gene locus. EMBO J. 6, 1945-1950 (1987).
115. Zhang, X.M., Cathala, G., Soua, Z., Lefranc, M.-P. and Huck, S. The human T-cell receptor gamma variable pseudogene V10 is a distinctive marker of human speciation. Immunogenetics. 43, 196-203 (1996).
116. Huck, S. and Lefranc, M.-P. Rearrangements to the JP1, JP and JP2 segments in the human T-cell rearranging gamma gene (TRG gamma) locus. FEBS Lett. 224, 291-296 (1987).
117. Lefranc, M.-P., Forster, A. and Rabbitts, T.H. Genetic polymorphism and exon changes of the constant regions of the human T-cell rearranging gene gamma. Proc Natl Acad Sci U S A. 83, 9596-9600 (1986).
118. Buresi, C., Ghanem, N., Huck, S., Lefranc, G. and Lefranc, M.-P. Exon duplication and triplication in the human T-cell receptor gamma constant region genes and RFLP in French, Lebanese, Tunisian, and black African populations. Immunogenetics. 29, 161-172 (1989).
119. Almagro, J.C., Hernandez, I., del Carmen Ramirez, M. and Vargas-Madrazo, E. The differences between the structural repertoires of VH germ-line gene segments of mice

and humans: implication for the molecular mechanism of the immune response. Mol Immunol. 34, 1199-1214 (1997).
120. Brodeur, P.H. and Riblet, R. The immunoglobulin heavy chain variable region (Igh-V) locus in the mouse. I. One hundred Igh-V genes comprise seven families of homologous genes. Eur J Immunol. 14, 922-930 (1984).
121. Crews, S., Griffin, J., Huang, H., Calame, K. and Hood, L. A single VH gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of the antibody. Cell. 25, 59-66 (1981).
122. Kofler, R. A new murine Ig VH gene family. J Immunol. 140, 4031-4034 (1988).
123. Kofler, R., Geley, S., Kofler, H. and Helmberg, A. Mouse variable-region gene families: complexity, polymorphism and use in non-autoimmune responses. Immunol Rev. 128, 5-21 (1992).
124. Mainville, C.A., Sheehan, K.M., Klaman, L.D., Giorgetti, C.A., Press, J.L. and Brodeur, P.H. Deletional mapping of fifteen mouse VH gene families reveals a common organization for three Igh haplotypes. J Immunol. 156, 1038-1046 (1996).
125. Pennell, C.A., Sheehan, K.M., Brodeur, P.H. and Clarke, S.H. Organization and expression of VH gene families preferentially expressed by Ly-1+ (CD5) B cells. Eur J Immunol. 19, 2115-2121 (1989).
126. Sheehan, K.M., Mainville, C.A., Willert, S. and Brodeur, P.H. The utilization of individual VH exons in the primary repertoire of adult BALB/c mice. J Immunol. 151, 5364-5375 (1993).
127. Sims, M.J., Krawinkel, U. and Taussig, M.J. Characterization of germ-line genes of the VGAM3.8 VH gene family from BALB/c mice. J Immunol. 149, 1642-1648 (1992).
128. Tutter, A., Brodeur, P., Shlomchik, M. and Riblet, R. Structure, map position, and evolution of two newly diverged mouse Ig VH gene families. J Immunol. 147, 3215-3223 (1991).
129. Winter, E., Radbruch, A. and Krawinkel, U. Members of novel VH gene families are found in VDJ regions of polyclonally activated B-lymphocytes. EMBO J. 4, 2861-2867 (1985).
130. Feeney, A.J. and Riblet, R. DST4: a new, and probably the last, functional DH gene in the BALB/c mouse. Immunogenetics. 37, 217-221 (1993).
131. Kurosawa, Y. and Tonegawa, S. Organization, structure, and assembly of immunoglobulin heavy chain diversity DNA segments. J Exp Med. 155, 201-218 (1982).
132. Wood, C. and Tonegawa, S. Diversity and joining segments of mouse immunoglobulin heavy chain genes are closely linked and in the same orientation: implications for the joining mechanism. Proc Natl Acad Sci U S A. 80, 3030-3034 (1983).
133. Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. Cell. 19, 981-992 (1980).
134. Newell, N., Richards, J.E., Tucker, P.W. and Blattner, F.R. J genes for heavy chain immunoglobulins of mouse. Science. 209, 1128-1132 (1980).
135. Sakano, H., Maki, R., Kurosawa, Y., Roeder, W. and Tonegawa, S. Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. Nature. 286, 676-683 (1980).
136. Honjo, T. and Kataoka, T. Organization of immunoglobulin heavy chain genes and allelic deletion model. Proc Natl Acad Sci U S A. 75, 2140-2144 (1978).

137. Yamawaki-Kataoka, Y., Nakai, S., Miyata, T. and Honjo, T. Nucleotide sequences of gene segments encoding membrane domains of immunoglobulin gamma chains. Proc Natl Acad Sci U S A. 79, 2623-2627 (1982).
138. Shimizu, A., Takahashi, N., Yaoita, Y. and Honjo, T. Organization of the constant-region gene family of the mouse immunoglobulin heavy chain. Cell. 28, 499-506 (1982).
139. Kirschbaum, T., Pourrajabi, S., Zocher, I., Schwendinger, J., Heim, V., Roschenthaler, F., Kirschbaum, V. and Zachau, H.G. The 3' part of the immunoglobulin kappa locus of the mouse. Eur J Immunol. 28, 1458-1466 (1998).
140. Kirschbaum, T., Roschenthaler, F., Bensch, A., Holscher, B., Lautner-Rieske, A., Ohnrich, M., Pourrajabi, S., Schwendinger, J., Zocher, I. and Zachau, H.G. The central part of the mouse immunoglobulin kappa locus. Eur J Immunol. 29, 2057-2064 (1999).
141. Roschenthaler, F., Kirschbaum, T., Heim, V., Kirschbaum, V., Schable, K.F., Schwendinger, J., Zocher, I. and Zachau, H.G. The 5' part of the mouse immunoglobulin kappa locus. Eur J Immunol. 29, 2065-2071 (1999).
142. Schable, K.F., Thiebe, R., Bensch, A., Brensing-Kuppers, J., Heim, V., Kirschbaum, T., Lamm, R., Ohnrich, M., Pourrajabi, S., Roschenthaler, F., Schwendinger, J., Wichelhaus, D., Zocher, I. and Zachau, H.G. Characteristics of the immunoglobulin Vkappa genes, pseudogenes, relics and orphons in the mouse genome. Eur J Immunol. 29, 2082-2086 (1999).
143. Thiebe, R., Schable, K.F., Bensch, A., Brensing-Kuppers, J., Heim, V., Kirschbaum, T., Mitlohner, H., Ohnrich, M., Pourrajabi, S., Roschenthaler, F., Schwendinger, J., Wichelhaus, D., Zocher, I. and Zachau, H.G. The variable genes and gene families of the mouse immunoglobulin kappa locus. Eur J Immunol. 29, 2072-2081 (1999).
144. Sakano, H., Huppi, K., Heinrich, G. and Tonegawa, S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature. 280, 288-294 (1979).
145. Max, E.E., Maizel, J.V., Jr. and Leder, P. The nucleotide sequence of a 5.5-kilobase DNA segment containing the mouse kappa immunoglobulin J and C region genes. J Biol Chem. 256, 5116-5120 (1981).
146. Max, E.E., Seidman, J.G. and Leder, P. Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene. Proc Natl Acad Sci U S A. 76, 3450-3454 (1979).
147. Tonegawa, S., Maxam, A.M., Tizard, R., Bernard, O. and Gilbert, W. Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. Proc Natl Acad Sci U S A. 75, 1485-1489 (1978).
148. Arp, B., McMullen, M.D. and Storb, U. Sequences of immunoglobulin lambda 1 genes in a lambda 1 defective mouse strain. Nature. 298, 184-187 (1982).
149. Weiss, S. and Wu, G.E. Somatic point mutations in unrearranged immunoglobulin gene segments encoding the variable region of lambda light chains. EMBO J. 6, 927-932 (1987).
150. Storb, U., Haasch, D., Arp, B., Sanchez, P., Cazenave, P.A. and Miller, J. Physical linkage of mouse lambda genes by pulsed-field gel electrophoresis suggests that the rearrangement process favors proximate target sequences. Mol Cell Biol. 9, 711-718 (1989).
151. Miller, J., Selsing, E. and Storb, U. Structural alterations in J regions of mouse immunoglobulin lambda genes are associated with differential gene expression. Nature. 295, 428-430 (1982).

152. Bernard, O., Hozumi, N. and Tonegawa, S. Sequences of mouse immunoglobulin light chain genes before and after somatic changes. Cell. 15, 1133-1144 (1978).
153. Selsing, E., Miller, J., Wilson, R. and Storb, U. Evolution of mouse immunoglobulin lambda genes. Proc Natl Acad Sci U S A. 79, 4681-4685 (1982).
154. Mami, F. and Kindt, T.J. C lambda 2 and C lambda 4 immunoglobulin light chain genes in a wild-derived inbred mouse strain. J Immunol. 138, 3980-3985 (1987).
155. Arden, B., Clark, S.P., Kabelitz, D. and Mak, T.W. Mouse T-cell receptor variable gene segment families. Immunogenetics. 42, 501-530 (1995).
156. Wilson, R.K., Koop, B.F., Chen, C., Halloran, N., Sciammis, R. and Hood, L. Nucleotide sequence analysis of 95 kb near the 3' end of the murine T-cell receptor alpha/delta chain locus: strategy and methodology. Genomics. 13, 1198-1208 (1992).
157. Koop, B.F., Wilson, R.K., Wang, K., Vernooij, B., Zallwer, D., Kuo, C.L., Seto, D., Toda, M. and Hood, L. Organization, structure, and function of 95 kb of DNA spanning the murine T-cell receptor C alpha/C delta region. Genomics. 13, 1209-1230 (1992).
158. Seto, D., Koop, B.F., Deshpande, P., Howard, S., Seto, J., Wilk, E., Wang, K. and Hood, L. Organization, sequence, and function of 34.5 kb of genomic DNA encompassing several murine T-cell receptor alpha/delta variable gene segments. Genomics. 20, 258-266 (1994).
159. Azuara, V., Lembezat, M.P. and Pereira, P. The homogeneity of the TCRdelta repertoire expressed by the Thy-1dull gammadelta T cell population is due to cellular selection. Eur J Immunol. 28, 3456-3467 (1998).
160. Chien, Y.H., Iwashima, M., Wettstein, D.A., Kaplan, K.B., Elliott, J.F., Born, W. and Davis, M.M. T-cell receptor delta gene rearrangements in early thymocytes. Nature. 330, 722-727 (1987).
161. Toda, M., Fujimoto, S., Iwasato, T., Takeshita, S., Tezuka, K., Ohbayashi, T. and Yamagishi, H. Structure of extrachromosomal circular DNAs excised from T-cell antigen receptor alpha and delta-chain loci. J Mol Biol. 202, 219-231 (1988).
162. Iwashima, M., Green, A., Davis, M.M. and Chien, Y.H. Variable region (V delta) gene segment most frequently utilized in adult thymocytes is 3' of the constant (C delta) region. Proc Natl Acad Sci U S A. 85, 8161-8165 (1988).
163. Bosc, N. and Lefranc, M.-P. The mouse (Mus musculus) T cell receptor beta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. Exp Clin Immunogenet. 17, 216-228 (2000).
164. Louie, M.C., Nelson, C.A. and Loh, D.Y. Identification and characterization of new murine T cell receptor beta chain variable region (V beta) genes. J Exp Med. 170, 1987-1998 (1989).
165. Six, A., Jouvin-Marche, E., Loh, D.Y., Cazenave, P.A. and Marche, P.N. Identification of a T cell receptor beta chain variable region, V beta 20, that is differentially expressed in various strains of mice. J Exp Med. 174, 1263-1266 (1991).
166. Hayday, A.C., Saito, H., Gillies, S.D., Kranz, D.M., Tanigawa, G., Eisen, H.N. and Tonegawa, S. Structure, organization, and somatic rearrangement of T cell gamma genes. Cell. 40, 259-269 (1985).
167. Garman, R.D., Doherty, P.J. and Raulet, D.H. Diversity, rearrangement, and expression of murine T cell gamma genes. Cell. 45, 733-742 (1986).
168. Pelkonen, J., Traunecker, A. and Karjalainen, K. A new mouse TCR V gamma gene that shows remarkable evolutionary conservation. EMBO J. 6, 1941-1944 (1987).
169. Traunecker, A., Oliveri, F., Allen, N. and Karjalainen, K. Normal T cell development is possible without 'functional' gamma chain genes. EMBO J. 5, 1589-1593 (1986).

170. Vernooij, B.T., Lenstra, J.A., Wang, K. and Hood, L. Organization of the murine T-cell receptor gamma locus. Genomics. 17, 566-574 (1993).
171. Eppig, J.T., Bult, C.J., Kadin, J.A., Richardson, J.E., Blake, J.A., Anagnostopoulos, A., Baldarelli, R.M., Baya, M., Beal, J.S., Bello, S.M., Boddy, W.J., Bradt, D.W., Burkart, D.L., Butler, N.E., Campbell, J., Cassell, M.A., Corbani, L.E., Cousins, S.L., Dahmen, D.J., Dene, H., Diehl, A.D., Drabkin, H.J., Frazer, K.S., Frost, P., Glass, L.H., Goldsmith, C.W., Grant, P.L., Lennon-Pierce, M., Lewis, J., Lu, I., *et al.* The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Res. 33, D471-5 (2005).
172. Kaas, Q., Ruiz, M. and Lefranc, M.-P. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res. 32, D208-210 (2004).
173. Eilbeck, K. and Lewis, S.E. Sequence ontology annotation guide. Comp Funct Genomics. 5, 642-647 (2004).
174. Duprat, E., Kaas, Q., Garelle, V., Lefranc, G. and Lefranc, M.-P. IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily, 2004.
175. Lefranc, M.-P., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A. and Lefranc, G. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev Comp Immunol. 29, 917-938 (2005).
176. Lefranc, M.-P., Pommie, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piedade, I., Rouard, M., Foulquier, E., Thouvenin, V. and Lefranc, G. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol. 29, 185-203 (2005).
177. Lefranc, M.-P., Pommie, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V. and Lefranc, G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol. 27, 55-77 (2003).
178. Brochet, X., Lefranc, M.-P. and Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res. 36, W503-508 (2008).
179. Giudicelli, V., Chaume, D. and Lefranc, M.-P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res. 32, W435-440 (2004).
180. Yousfi Monod, M., Giudicelli, V., Chaume, D. and Lefranc, M.-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics. 20 Suppl 1, i379-385 (2004).
181. Elemento, O. and Lefranc, M.-P. IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. Dev Comp Immunol. 27, 763-779 (2003).
182. Baum, T.P., Hierle, V., Pasqual, N., Bellahcene, F., Chaume, D., Lefranc, M.-P., Jouvin-Marche, E., Marche, P.N. and Demongeot, J. IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. BMC Bioinformatics. 7, 224 (2006).
183. Baum, T.P., Pasqual, N., Thuderoz, F., Hierle, V., Chaume, D., Lefranc, M.-P., Jouvin-Marche, E., Marche, P.N. and Demongeot, J. IMGT/GeneInfo: enhancing V(D)J recombination database accessibility. Nucleic Acids Res. 32, D51-54 (2004).

184. Pommie, C., Levadoux, S., Sabatier, R., Lefranc, G. and Lefranc, M.-P. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit. 17, 17-32 (2004).
185. Brochet, X. Conception et intégration d'un système d'information dédié à l'analyse et à la gestion des séquences réarrangées des récepteurs d'antigènes au sein d'IMGT: application à la leucémie lymphoïde chronique 188 (2008).
186. Giudicelli V, P.C.a.L.M.-P. The IMGT strategy for the automatic annotation of IG and TR cDNA: IMGT/Automat 103-104 (2003).
187. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. J Mol Biol. 215, 403-410 (1990).
188. Mitrophanov, A.Y. and Borodovsky, M. Statistical significance in biological sequence analysis. Brief Bioinform. 7, 2-24 (2006).
189. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. WebLogo: a sequence logo generator. Genome Res. 14, 1188-1190 (2004).
190. Eddy, S.R. HMMER - biosequence analysis using profile hidden Markov models. Available: http://hmmer.janelia.org/2007.
191. Durbin R, E.S.K.A.M.G. biological sequence analysis: probabilistic models of proteins and nucleic acids, 1998.
192. Pelat, T., Bedouelle, H., Rees, A.R., Crennell, S.J., Lefranc, M.P. and Thullier, P. Germline humanization of a non-human primate antibody that neutralizes the anthrax toxin, by in vitro and in silico engineering. J Mol Biol. 384, 1400-7 (2008).
193. Pelat, T., Hust, M., Laffly, E., Condemine, F., Bottex, C., Vidal, D., Lefranc, M.P., Dubel, S. and Thullier, P. High-affinity, human antibody-like antibody fragment (single-chain variable fragment) neutralizing the lethal factor (LF) of Bacillus anthracis by inhibiting protective antigen-LF complex formation. Antimicrob Agents Chemother. 51, 2758-64 (2007).
194. Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P.E., Bui, H.H., Buus, S., Frankild, S., Greenbaum, J., Lund, O., Lundegaard, C., Nielsen, M., Ponomarenko, J., Sette, A., Zhu, Z. and Peters, B. Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res. 36, W513-518 (2008).
195. Emerson, A. and Rossi, E. ImmunoGrid - the virtual human immune system project. Stud Health Technol Inform. 126, 87-92 (2007).
196. Pappalardo, F., Halling-Brown, M.D., Rapin, N., Zhang, P., Alemani, D., Emerson, A., Paci, P., Duroux, P., Pennisi, M., Palladini, A., Miotto, O., Churchill, D., Rossi, E., Shepherd, A.J., Moss, D.S., Castiglione, F., Bernaschi, M., Lefranc, M.P., Brunak, S., Motta, S., Lollini, P.L., Basford, K.E. and Brusic, V. ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization. Brief Bioinform. 10, 330-340 (2009).
197. Pappalardo, F., Pennisi, M., Castiglione, F. and Motta, S. Vaccine protocols optimization: *In silico* experiences. Biotechnology Advances. (2009).
198. Lefranc, M.-P., Clement, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D. and Lefranc, G. IMGT-Choreography for immunogenetics and immunoinformatics. In Silico Biol. 5, 45-60 (2005).
199. Chaume, D., Combres, K., Giudicelli, V. and Lefranc, M.-P. Retrieving factual data and documents using IMGT-ML in the IMGT information system® 47-51 (2005).
200. Chaume, D., Giudicelli, V. and Lefranc, M.-P. IMGT-ML a XML language for IMGT-ONTOLOGY and IMGT/LIGM-DB data 71-75 (2001).
201. Gueguen, Y., Garnier, J., Robert, L., Lefranc, M.-P., Mougenot, I., de Lorgeril, J., Janech, M., Gross, P.S., Warr, G.W., Cuthbertson, B., Barracco, M.A., Bulet, P., Aumelas, A., Yang, Y., Bo, D., Xiang, J., Tassanakajon, A., Piquemal, D. and

Bachere, E. PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. Dev Comp Immunol. 30, 283-288 (2006).
202. Gelly, J.C., Gracy, J., Kaas, Q., Le-Nguyen, D., Heitz, A. and Chiche, L. The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. Nucleic Acids Res. 32, D156-159 (2004).
203. Mulvenna, J.P., Wang, C. and Craik, D.J. CyBase: a database of cyclic protein sequence and structure. Nucleic Acids Res. 34, D192-194 (2006).
204. Wang, C.K., Kaas, Q., Chiche, L. and Craik, D.J. CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. Nucleic Acids Res. 36, D206-210 (2008).
205. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L. and Lewis, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 25, 1251-1255 (2007).

ANNEXES

Annexe 1. Labels du V-GENE, D-GENE et J-GENE.

Label	Description
V-GENE	ADN génomique germinal regroupant un L-PART1, un V-INTRON et un V-EXON, ainsi qu'un 5'UTR et un 3'UTR.
L-INTRON-L	Séquence d'ADN génomique regroupant le L-PART1, V-INTRON et L-PART2.
INIT-CODON	Codon d'initiation de la traduction 'ATG'.
L-PART1	Exon traduit dans sa première partie en peptide signal d'un V-, V-D-, V-D-J- ou V-J-GENE.
V-INTRON	Séquence non codante comprise entre le L-PART1 et V-EXON, dans l'ADN génomique.
ACCEPTOR-SPLICE	Site d'épissage en 5' d'une région codante dont la séquence consensus est NAGNN avec N correspondent à n'importe quel nucléotide, l'épissage s'effectue après le nucléotide G.
L-PART2	Région 5' d'un V-EXON traduit dans sa seconde partie en peptide signal d'un V-, V-D-, V-D-J- or V-J-GENE.
L-V-GENE-UNIT	ADN génomique germinal regroupant un L-PART1, V-INTRON, V-EXON et V-RS.
V-GENE-UNIT	ADN génomique germinal regroupant un V-EXON et V-RS.
V-EXON	ADN génomique germinal regroupant un L-PART2 et une V-REGION.
V-REGION	Région codante d'un V-GENE sans le peptide signal dont le cadre de lecture codant est situé à 1 ou 2 nucléotides avant le V-HEPTAMER.
FR1-IMGT	Premier framework selon les standards IMGT®.
1st-CYS	Codon de la cystéine conservée localisée dans le FR1-IMGT intervenant dans un pont intra-disulfure avec la 2nd-CYS.
CDR1-IMGT	Première région déterminant la complémentarité selon la numérotation unique IMGT®.
FR2-IMGT	Second framework selon les standards IMGT®.
CONSERVED-TRP	Codon du tryptophane conservé localisé dans le FR2-IMGT.
CDR2-IMGT	Deuxième région déterminant la complémentarité selon la numérotation unique IMGT®.
FR3-IMGT	Troisième framework selon les standard IMGT®.
2nd-CYS	Codon de la cystéine conserves localise dans le FR3-IMGT intervenant dans un pont intra-disulfure avec la 1st-CYS.
CDR3-IMGT	Troisième région déterminant la complémentarité selon la numérotation unique IMGT®.
V-RS	Signal de recombinaison regroupant en 5' un V-HEPTAMER, en 3 un V-NONAMER qui délimité un V-SPACER. Le V-RS est en 3' de la V-REGION.
V-HEPTAMER	Site de recombinaison constitué d'un heptamère situé en 5' du V-RS et dont la séquence consensus est 'CACAGTG'.
V-SPACER	Espaceur de 12 ± 1ou 23 ± 1nucléotides compris entre le V-HEPTAMER et le V-NONAMER d'un V-RS.
V-NONAMER	Site de recombinaison constitué d'un nonamère situé en 3' du V-RS et dont la séquence consensus est 'ACAAAAACC'.
D-GENE	ADN génomique germinal regroupant un 5'D-RS, une D-REGION, un 3'D-RS ainsi qu'un 5'UTR et un 3'UTR.
5'D-RS	Signal de recombinaison qui inclue le 5'D-HEPTAMER, 5'D-SPACER, et 5'D-NONAMER du côté 5' de la D-REGION d'un D-GENE.
5'D-SPACER	Espaceur de 12 ± 1 or 23 ± 1 nucléotides compris entre le 5'D-HEPTAMER et le 5'D-NONAMER d'un 5'D-RS.
D-GENE-UNIT	ADN génomique en configuration contenant un 5'D-RS, D-REGION et 3'D-RS.
D-REGION	Région codante d'un D-GENE. La phase codante n'est pas forcément position de début de la région. Après réarrangement et épissage la phase codante région est rétablit.
3'D-RS	Signal de recombinaison qui inclue le 3'D-HEPTAMER, 3'D-SPACER, et 3'D-NONAMER du côté 3' de la D-REGION d'un D-GENE.
3'D-SPACER	Espaceur de 12 ± 1 or 23 ± 1 nucléotides compris entre le 3'D-HEPTAMER et le 3'D-NONAMER d'un 3'D-RS.
3'D-HEPTAMER	Motif conservé de 7 nucléotides dont la séquence consensus est 'CACAGTG'. Il fait parti du même signal de recombinaison que le 5'D-HEPTAMER (3'D-RS).
3'D-NONAMER	Motif conservé de 9 nucléotides dont la séquence consensus est 'ACAAAAACC'. Il fait parti du même signal de recombinaison que le 3'D-HEPTAMER (3'D-RS).
J-GENE	ADN génomique germinal regroupant une J-REGION ainsi qu'un 5' UTR et un 3' UTR.
J-GENE-UNIT	ADN génomique germinal composé en 5' d'un J-RS suivit d'une J-REGION.
J-RS	Signal de recombinaison regroupant un J-HEPTAMER, J-SPACER et J-NONAMER en 5' de la J-REGION.
J-NONAMER	Site de recombinaison de 9 nucléotides dont la séquence consensus est 'GGTTTTTGT'. Ce nonamère fait parti du J-RS.
J-SPACER	Espaceur de 12 ± 1 ou 23 ± 1 nucléotides compris entre le J-NONAMER et le J-HEPTAMER d'un J-RS.
J-HEPTAMER	Site de recombinaison de 7 nucléotides, dont la séquence consensus est 'CACAGTG'. Cet heptamère fait parti du J-RS.
J-REGION	Région codante du J-GENE dont le cadre de lecture codant peut être à 1 ou 2 nucléotide(s) après le J-HEPTAMER.
J-PHE	Codon de la phénylalanine conservée de la J-REGION.
J-TRP	Codon du tryptophane conservé de la J-REGION.
5'UTR	Séquence non traduite en 5' d'un V-GENE, D-GENE et J-GENE.
3'UTR	Séquence non traduite en 3' d'un V-GENE, D-GENE et J-GENE.
DONOR-SPLICE	Site d'épissage présent à l'extrémité 3' d'une région codante (L-PART1 et J-REGION).
STOP-CODON	Codon de terminaison de la traduction 'TAG', 'TGA', 'TAA'.

Annexe 2. Alphabet dégénéré de l'ADN selon le code IUPAC-IUB

Lorsque nous parlons de comparaison de séquences, nous devons envisager le fait qu'à une position donnée, il puisse exister des incertitudes sur la détermination de la base. On peut par exemple vouloir autoriser à une position soit une adénine, soit une guanine: on dit alors qu'à cette position, on est en présence d'une base dégénérée.

Nous avons utilisé comme notation pour les bases dégénérées, le code IUPAC-IUB, résumé dans la figure suivante :

Code	**Base**	**Origine du choix de la lettre**
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A ou G	puRine
Y	C ou T	pYrimidine
S	G ou C	Interaction forte (3 ponts H) Strong
W	A ou T	Interaction faible (2 ponts H) Weak
K	G ou T	Keto
M	A ou C	aMino
B	C ou G ou T	Pas de A
D	A ou G ou T	Pas de C
H	A ou C ou T	Pas de G
V	A ou C ou G	Pas de T
N	N'importe qu'elle base	aNy
.	gap	

Ce tableau se lit ainsi: un R correspond à une adénine ou à une guanine.

Annexe 3. Matrice de substitution utilisée pour les alignements sans insertions et délétions

Matrice de Substitution, utilisée pour le calcul du score d'alignement par IMGT/VQUEST. Elle prend en compte le code dégénéré de l'ADN. Pour une substitution d'un nucléotide par un même nucléotide la valeur est de 0 (ex : A en A), pour une substitution d'un nucléotide par un autre nucléotide la valeur est de 2 (ex : T en C ou W en S), et pour une substitution d'un nucléotide dégénéré par un nucléotide dégénéré identique ou différents la valeur est de 1 (ex : R en S).

	.	x	T	c	A	G	R	Y	K	M	S	W	B	D	H	V
V	0	0	2	0	0	0	0	1	1	0	0	1	1	1	1	0
H	0	0	0	0	0	2	1	0	1	0	1	0	1	1	0	1
D	0	0	0	2	0	0	0	1	0	1	1	0	1	0	1	1
B	0	0	0	0	2	0	1	0	0	1	0	1	0	1	1	1
W	0	0	0	2	0	2	1	1	1	1	2	0	1	0	0	1
S	0	0	2	0	2	0	1	1	1	1	0	2	0	1	1	0
M	0	0	2	0	0	2	1	1	2	0	1	1	1	1	0	0
K	0	0	0	2	2	0	1	1	0	2	1	1	0	0	1	1
Y	0	0	0	0	2	2	2	0	1	1	1	1	0	1	0	1
R	0	0	2	2	0	0	0	2	1	1	1	1	1	0	1	0
G	0	0	2	2	2	0	0	2	0	2	0	2	0	0	2	0
A	0	0	2	2	0	2	0	2	2	0	2	0	2	0	0	0
C	0	0	2	0	2	2	2	0	2	0	0	2	0	2	0	0
T	0	0	0	2	2	2	2	0	0	2	2	0	0	0	0	2
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Annexe 4. Valeurs du seuil et de l'overlap utilisées pour l'alignement global par IMGT/V-QUEST

Valeurs des seuils utilisés par IMGT/V-QUEST lors de différentes étapes d'alignement par IMGT/V-QUEST. Une valeur seuil correspond au score d'alignement minimal pour que celui-ci soit considéré comme significatif.

Valeurs des overlaps utilisées lors des différentes étapes d'alignement par IMGT/VQUEST. La valeur d'un overlap définit la longueur minimale de l'alignement pour qu'il soit considéré, en fonction de la longueur de la plus petite des deux séquences à aligner. Par exemple un overlap de 1/3 correspond à une longueur minimale égale à au moins 1/3 de la longueur de la plus petite des deux séquences à aligner.

Etape d'alignement	**seuil**	**overlap**
Identification du type de chaîne	**100**	**1/2**
Identification et description de la V-REGION	**600**	**1/3**
Identification et description de la J-REGION	**600**	**1/4**
Identification et description de la D-REGION	**500**	**1/2**

Annexe 5. Séquences et nombres d'héptamères et de nonamères différents et fonctionnels chez l'homme et la souris dans la base de IMGT/LIGMotif

Les heptamères et nonamères intégrés dans IMGT/LIGMotif proviennent d'IMGT/GENE-DB. L'occurrence de chaque séquence est calculée.

V-NONAMER											
séquence	nombre										
tcacaaact	1	atataagca	1	atacaaact	1	gacaaaatc	1	acacagaat	2	cctcaaact	4
accacaacc	1	atataaacc	1	caagaacac	1	acacaacat	1	tcaaaaact	2	tcagaaaac	4
agaaaaact	1	tcagaaatt	1	acccaacct	1	atcaaaacc	1	ctgaaaacc	2	actcaaacc	4
acgtaaaca	1	acacaaaaa	1	ccagaaatc	1	atggctcta	1	cacagactc	2	ataagaacc	4
acaaaaaaa	1	acagtattt	1	acagaaagt	1	atccaaaca	1	acgcaaacc	2	acaaaaatc	4
acaaaatgc	1	ttacagcag	1	ccgttttcc	1	gtgggctta	1	acataaatg	2	atataaact	4
aaacaaata	1	attcgtaag	1	gcaaaaaca	1	acatcaacc	1	aaaaaaacc	2	acaagaaca	4
gaataagta	1	accaaaatt	1	atggatagg	1	agaaagtga	1	cacataaaa	2	atgcaaacc	4
gtacaaaca	1	acacaaaat	1	cacccaaaa	1	gcaatattt	1	aaataaacc	2	tctaaaacc	4
acaaaaaag	1	gcacaaaac	1	acaaaacct	1	gcacaaagc	1	aaaaaaaat	2	acacaaaca	4
gcataaaac	1	acagaaaca	1	tctctcagt	1	caaaaggac	1	tcaataaat	2	acataaaca	4
catagatgg	1	cacacacac	1	aaccaaacc	1	attccatca	1	ccatgcata	2	gcaagaacc	5
agaagaacc	1	ggtttgggt	1	gacacaaaa	1	acaaaaaga	1	gcataaacc	2	acctaaacc	5
atctaaaac	1	acagaaatt	1	tcaagaaag	1	acacatacc	1	cctcaaacc	2	cagaaaccc	5
aaacaaacc	1	ggcaagacc	1	caaaatccc	1	acaaaaagg	1	ctgaacatc	2	tcaaaaacc	5
aaaccaagg	1	tgcaaacat	1	ctcaaacaa	1	acataaaga	1	caatatctc	2	gcaaaaacc	5
acgaaaacc	1	tcagtaacc	1	ggcaaaagc	1	ctcaaacct	1	acatcaact	2	tcagaaatc	6
cctcaaacg	1	attaacctg	1	caaacctcc	1	caaaaagat	1	ttagaaacc	2	acagaaagg	6
cgagaactc	1	cttcaaacc	1	gcaaaacca	1	gcactaatc	2	acaaaccca	2	acaaaaata	6
acaaatact	1	tgcaaaccc	1	acttgaact	1	tcagaaacg	2	acctaaact	2	actaaaacc	7
tcataaact	1	acaaacacc	1	acaaaactg	1	acaagaccc	2	acaaacccc	2	ctgaaaatc	7
cagaacctc	1	ggaaaatat	1	caagaaccc	1	acccaaaca	2	aacttaatc	2	tcacaaacc	8
acggaaatg	1	gaagaaacc	1	acaataact	1	cacaaaccc	2	aacaaaccc	2	gcagaaacc	9
ggtccacag	1	cacaagcca	1	cagaaatcc	1	cagaaaact	2	tgcggaaat	2	acataaagg	9
acaaatacc	1	acctaaagc	1	ccagaaacc	1	aagaaaacc	2	atacaaacc	2	acatgaacc	10
ctcccagag	1	acactgacc	1	acacaaacg	1	acacaaaac	2	acaaaatcc	3	gcacaaact	10
actcaaact	1	taaaaaaac	1	ctgtcctca	1	cataaactt	2	acacagacc	3	acaagaacc	10
acataaacg	1	taaaaaaaa	1	tgtcctcaa	1	ataaaaact	2	gcataaatg	3	acacaaact	12
tgttctgac	1	gcaaaaatt	1	gcaaaatcc	1	aacacaaac	2	acacaaatc	3	gcacaaacc	12
actcaaatt	1	acaaaatct	1	cccaaaccc	1	acataagcc	2	agagaaacc	3	accaaaacc	14
acatgaaac	1	ctccaaatc	1	gcagaaact	1	ggaaactat	2	gcagaaaac	3	acaaaaact	14
atggaaagg	1	acacaacct	1	ccaaaactc	1	acaaacctc	2	acataaact	3	acaaaaaca	16
atatatatt	1	atgtaaacc	1	accaaacct	1	cacaaactc	2	acataaatc	3	acccaaacc	26
tcacatttg	1	acacaagcc	1	ctccaaact	1	gcaataaca	2	acaaaaatt	3	acagaaacc	30
cttcctgac	1	gtgcaaacc	1	ctcaggagc	1	ccaaaattc	2	acagaaaga	3	tcagaaacc	35
gccagaact	1	tcagaaagc	1	tacactgac	1	ctccaaacc	2	acaaaagcc	3	acataaacc	81
agacacaga	1	tcaagaacc	1	acaggaacc	1	acaaaatgg	2	acagaaact	3	acaaaaacc	88
ccacaaaca	1	accctctaa	1	aaaaaaaaa	1	ctgaatatc	2	gcaaaaatc	3	acacaaacc	115
tggaaatgg	1	cacaaatta	1	cgtaaacag	1	ggaaatgga	2	acccaaact	3		
tcaggaacc	1	cagtatcct	1	acaagaagg	1	acaaaaatg	2	ccccaaact	3		
cacaaacct	1	acaacctct	1	accataacc	1	actgaaaga	2	acataatcc	3		
aaagaaacc	1	agccatgcc	1	gcgcaaacc	1	gcataaagt	2	tcataaacc	4		

J-NONAMER					
séquence	nombre				
tgctattga	1	ggtttgtgc	1	agtttatgt	2
gattttcac	1	ggttatctc	1	gggttttat	2
gaggttgtt	1	ctctaaggc	1	gttttttgc	3
gtgttttgg	1	agttttgtc	1	ggtttctgt	3
tgtttttgc	1	ggtttctct	1	agtttttgc	3
ggttcatgt	1	gagattctt	1	gagtttgtc	3
tgattttgc	1	ggttagtgt	1	catttttgt	4
ggcttcttc	1	ggtttaggg	1	ggtttatgt	4
tgactttgc	1	ccttttaga	1	ggtttttgg	4
ggttttcct	1	tgttcatgt	1	tgtttttga	5
ggttttacc	1	ggtttttga	1	gatttttgt	5
gaattcttg	1	tttgggatt	1	ggtttgtgt	6
agtcgtcgt	1	cgttttggg	1	tgtttttgt	7
ggggatgag	1	ggttttggg	1	ccattttgt	7
gatttgtgt	1	ggtttgtac	1	ggtttttgc	9
agtttattg	1	gttcttgcc	1	agtttttgt	17
atttttctc	1	agttttagt	1	ggtttttgt	34
gttctttgt	1	cgtttttgt	1		
aacctggct	1	gctttctgt	1		
tggttttgt	1	gatttttct	1		
ggtttcatc	1	attttttgt	1		
gggtttgcc	1	ggattttgc	1		
agaatttgc	1	ggttattgc	2		
ggttattgt	1	atttattgt	2		
ggtttgcgc	1	gggttttgt	2		
gcttttggt	1	agttttacc	2		
ggtatttgt	1	gttttttgt	2		
ggttttatc	1	agtatttgt	2		
ggtttttcg	1	tggttttga	2		
catattcga	1	ggcttctgt	2		
aattcttgt	1	agtttctgt	2		
tgttactgt	1	ggattttgt	2		
agtttgtgc	1	ggttttggt	2		
gattattgt	1	ggtatttgc	2		
gttttctgt	1	tgtttctgt	2		
gtctgttgt	1	tcattttgt	2		
agttcttgt	1	ggttcttgt	2		
gttacctgt	1	ggtttcagt	2		
gctccattt	1	agtttttga	2		
tgatttgga	1	ggtttgcat	2		
ggtttctga	1	ggttttgaa	2		
gaattctgg	1	agttattgt	2		
catttctgt	1	ggttatctg	2		

5'D-NONAMER	
séquence	nombre
atttttca	1
agtttttgt	1
ggtttttgc	1
ggtttttgg	1
cgtttctga	1
ggtttttat	1
ggattccga	1
gaattcagt	1
ggattttga	1
ggtttagaa	1
catttttgt	2
cttttttgt	2
gattttgaa	2
ggttttgac	2
tgtttttgt	2
ggtttctga	3
agattctga	3
ggattctga	3
ggtttgaag	3
ggattctgt	5
ggtttgggg	6
ggttattgt	6
gctttttgt	9
ggattttgt	10
gatttttgt	11

3'D-NONAMER	
séquence	nombre
acaaaaatc	1
tcagaaacc	1
tcagaaaca	1
tcagaaaac	1
attaaccaa	1
acaaaaagc	1
ctcaaatcc	1
catggaaga	1
ccaaaaaca	1
tcctaaagc	1
acaacaaag	1
acagaaacc	1
tccaaaacg	1
accaaaact	2
acaaaaaac	2
tcaaaaact	2
tccaaaacc	2
gcaaaaacc	2
acaaaaact	2
tccaaaact	3
ccagaaacc	3
acaagaaag	3
ctcaaattc	4
gcagcaacc	6
gcaaaaact	6
tcccaaagc	6
tcaaaaacc	7
acaaaaacc	15

V-HEPTAMER			
Séquence	Nombre		
cataata	1	cacagga	2
gccacac	1	catactg	2
gtatgta	1	cacagac	2
ggccacc	1	cacagct	2
tacggta	1	acacaca	2
gagggtg	1	cacagaa	2
gcacgtg	1	ggtttgt	2
aggtttg	1	cacggta	2
catagca	1	gacagtg	2
atgtaac	1	catggtg	2
tcgtaag	1	tagagtg	2
ggaaaga	1	cacagtc	2
cacgttg	1	gtagttt	2
caacgtg	1	cacagtt	2
caccata	1	cactcta	2
cacaacg	1	cccagtg	3
gtttgtg	1	tacagtg	3
tcaggac	1	caccgtg	3
gaagttt	1	cgcagcc	3
cacatac	1	ccaagtg	3
cacaccc	1	cgcagtg	3
agtagtg	1	tcctgtg	3
cactatg	1	cacaaag	4
gaggttt	1	caccctg	4
cactagg	1	caccatg	5
ctcagtt	1	cacattg	5
cactgag	1	cacaaca	5
tactgtg	1	cacaggg	6
cagtgat	1	cagagtg	7
cacatgt	1	catagtg	7
ctcagtg	1	cacggtg	11
ctcattg	1	cactgtg	13
cagctct	1	cacagta	13
cacatca	1	cacactg	17
caccagg	1	cacagag	19
aatagtg	1	cacagcc	26
acctatg	1	cacagca	26
gactgtg	1	cacagcg	30
gtggttt	1	cacaatg	56
tttgtgg	2	cacagtg	601
atatata	2		

J-HEPTAMER			
Séquence	Nombre		
gggggtg	1	caggtgg	1
caccgca	1	aggactg	1
tacggta	1	tagagtg	1
tggactg	1	cactctg	1
cactggc	1	gttagca	1
taatgtg	1	cgctgtg	1
tacggtg	1	cagggtg	2
gcccatc	1	ccctgtg	2
tgcagag	1	ctctgtg	2
cggtgtg	1	gtctgtg	2
tattgtg	1	ctccgtg	2
catggtg	1	taacgtg	2
ggtagtg	1	ccgtctg	2
tgtggtg	1	tagtgtg	2
gattgtg	1	cacagca	2
gttagtg	1	tgatgtg	2
ggttgtg	1	ttctgtg	2
gaccacg	1	tcctgtg	2
caccgtg	1	tgttgtg	3
tagggtg	1	tacagtg	3
tggggtg	1	cactatg	3
ggccgtg	1	cactgta	3
agatgtg	1	cagagtg	3
cagagag	1	agctgtg	4
gggtgtg	1	tgcagtg	4
aactgtg	1	cagtgtg	4
aaatctg	1	tactgtg	6
cgttgcc	1	tgctgtg	6
catagtg	1	caatgtg	8
ttatgtg	1	gactgtg	10
gacagtg	1	cattgtg	12
agccgtg	1	ggctgtg	16
tatagtg	1	cacagtg	22
agcaggg	1	cactgtg	54

5'D-HEPTAMER			
Séquence	Nombre		
cacagcg	1	tgctatg	2
ggccgtg	1	cacggtg	2
caccatg	1	caccgtg	4
gattgtg	1	gactgtg	4
cagtgtg	1	cattgtg	6
cacagtc	1	cacagtg	12
caaagtg	1	tactgtg	16
cactctc	1	cactgtg	23
tgctgtg	2		

3'D-HEPTAMER			
Séquence	Nombre		
cactcaa	1	cactgtg	4
cactgtc	1	cacattg	5
catagtg	1	cacaatg	7
cacagca	2	cacagtg	50
cactgta	2		
cacacag	2		
cacgatg	2		
caccgtg	2		
cacggtg	3		

PUBLICATIONS

PUBLICATION 1

IMGT/LIGMotif: a tool for immunoglobulin and T cell receptor gene identification and description in large genomic sequences

Jérôme Lane[1], Patrice Duroux[1] and Marie-Paule Lefranc[1, 2, §]

[1] IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier cedex 5, France.

[2] Institut Universitaire de France, 103 Bd St Michel, 75005 Paris, France

[§]Corresponding author

Email addresses:

JL: jerome.lane@igh.cnrs.fr

PD: patrice.duroux@igh.cnrs.fr

MPL: marie-paule.lefranc@igh.cnrs.fr

Abstract

Background

The antigen receptors, immunoglobulins (IG) and T cell receptors (TR), are specific molecular components of the adaptive immune response of vertebrates. Their genes are organized in the genome in several loci (7 in humans) that comprise different gene types: variable (V), diversity (D), joining (J) and constant (C) genes. Synthesis of the IG and TR proteins requires rearrangements of V and J, or V, D and J genes at the DNA level, followed by the splicing at the RNA level of the rearranged V-J and V-D-J genes to C genes. Owing to the particularities of IG and TR gene structures related to these molecular mechanisms, conventional bioinformatic software and tools are not adapted to the identification and description of IG and TR genes in large genomic sequences. In order to answer that need, IMGT®, the international ImMunoGeneTics information system®, has developed IMGT/LIGMotif, a tool for IG and TR gene annotation. This tool is based on standardized rules defined in IMGT-ONTOLOGY, the first ontology in immunogenetics and immunoinformatics.

Results

IMGT/LIGMotif currently annotates human and mouse IG and TR loci in large genomic sequences. The annotation includes gene identification and orientation on DNA strand, description of the V, D and J genes by assigning IMGT® labels, gene functionality, and finally, gene delimitation and cluster assembly. IMGT/LIGMotif analyses sequences up to 2.5 megabase pairs and can analyse them in batch files.

Conclusions

IMGT/LIGMotif is currently used by the IMGT® biocurators to annotate, in a first step, IG and TR genomic sequences of human and mouse in new haplotypes and those of closely related species, nonhuman primates and rat, respectively. In a next step, and

following enrichment of its reference databases, IMGT/LIGMotif will be used to annotate IG and TR of more distantly related vertebrate species.

Background

The immune adaptive system defends multicellular organisms from pathogens (i.e. bacteria, parasites, viruses) and tumor cells which are specifically recognized by antigen receptors. These antigen receptors, immunoglobulins (IG) or antibodies [1] and T cell receptors (TR) [2], present a huge diversity (2.10^{12} IG and 2.10^{12} TR per individual) that is crucial for specific antigen recognition. These huge numbers of different proteins are encoded by a relatively limited number of genes organized in the genome in different loci (7 in humans) that comprise different types of gene: variable (V), diversity (D), joining (J) and constant (C) genes. Synthesis of the IG and TR proteins requires complex mechanisms that include, at the DNA level, rearrangements of V and J, or of V, D and J genes [3], N-Diversity at the resulting V-J and V-D-J junctions [4, 5] and, for the IG, somatic hypermutations [6, 7]. These rearrangements are followed, at the RNA level, by the splicing of rearranged V-J and V-D-J genes to a C gene. In order to manage IG and TR data, IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org/ [8] was created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS). One of the first goals of IMGT® was to identify and to describe all the human IG and TR genes present in the human genome, an indispensable requisite before analysing the immune repertoire. Owing to the particularities of the IG and TR gene structures, IMGT-ONTOLOGY [9-11], the first ontology for immunogenetics and immunoinformatics, has been built to ensure the accuracy and the consistency of the IMGT® data, as well as the coherence between the IMGT® databases, tools and Web resources [12]. Several years of expert and time

consuming manual curation led to the IMGT® gene nomenclature for IG and TR genes [1, 2] which was approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 [13] and by the World Health Organization-International Union of Immunological Societies (WHO-IUIS) [14, 15]. IMGT® IG and TR genes have been entered in IMGT/GENE-DB [16], the IMGT® gene database, in the Human Genome Database (GDB) [17], in LocusLink [18] at the National Center for Biotechnology Information (NCBI), in Entrez Gene [19] when this database superseded LocusLink, in Ensembl [20] at the European Bioinformatics Institute (EBI), and in the Vega Genome Browser [21] at the Wellcome Trust Sanger Institute.

Interestingly, the human IG and TR genes data were annotated in IMGT® [1, 2] before the release of the human genome sequence [22, 23], however most of the corresponding genomic sequences were short (1-2 kb) and large contigs still remain to be precisely annotated. Conventional software and tools such as GeneMark [24], Genescan [25] and N-SCAN [26], are not adapted to the annotation of IG and TR genes owing to the particularities of their structure. Prediction of immunoglobulin superfamily protein genes has been improved with the Exegesis [27] procedure which uses GeneWise [28] and experimental maps, by comparison with the Ensembl method. However, this procedure has not been developed for detailed and standardized annotation. To answer the need of a tool for an automated annotation of antigen receptors in genomic DNA and, thus, to avoid several manual and time consuming steps, IMGT/LIGMotif, a Java on-line software has been developed, that allows the identification, standardized description and functionality assignment of IG and TR V, D and J genes in large genomic sequences.

Methods

IG and TR gene characteristics to consider for gene identification

IG and TR V, D and J genes belong to multigene subgroups and, therefore, share a high percentage of sequence identity (>75%) as a result of gene duplications inside a given subgroup [1, 2]. As a consequence it is often difficult to assign nearly identical sequences either to different genes or to different alleles of a same gene. This can lead to possible errors in the genome assembly that need to be detected.

IG and TR loci contain several hundreds of genes, 608-665 IG and TR genes per haploid genome in human, depending on the haplotypes, and more than 800 IG and TR genes in mouse [16], these numbers including many pseudogenes (227-253 in human and 212-240 in mouse). Many of these pseudogenes are degenerated and/or partial genes and, therefore difficult to annotate. Another level of complexity results from gene insertion and/or deletion polymorphisms that are frequent in multigene families. As the genome assembly results from joined DNA fragments from different haplotypes, and from one or the other chromosome, it does not reflect a 'true' haplotype and a careful analysis is required for the gene and allele assignment. It should also be noted that some IG and TR genes, designated as orphons are localized outside the main loci [1, 2]. Although these orphons are not functional, they have a high percentage of identity with genes of the major loci and represent another source of possible confusion and errors in gene and allele identification. At last, D and J genes have very small coding regions (8-37 base pairs (bp) and 37-69 bp, respectively) making their identification difficult.

Interestingly, and despite the difficulties mentioned above, functional IG and TR genes have characteristics that, if present, allow their unambiguous identification. Thus, for example, IG and TR V genes comprise two exons, a L-PART1 (L for leader) exon and a V-EXON with a splicing frame of type 1 (sf1) (IMGT Aide-

mémoire, http://www.imgt.org/). The IG and TR V, D and J genes have recombination signals (RS) which allow them to rearrange at the DNA level in B cells (for the IG) and T cells (for the TR) [1, 2] and that constitute one of the major differences with conventional genes. RS are localized in 3' of the V genes (V-RS), in 5' of the J genes (J-RS) and on both sides of the D genes (5'D-RS and 3'D-RS) [1, 2]. They consist of conserved heptamers and nonamers separated by less conserved spacers of 12 ± 1 or 23 ± 1 bp which vary between loci and species (IMGT Repertoire, http://www.imgt.org/). An efficient recombination only occurs between a RS with a 12 bp spacer and a RS with a 23 bp spacer (12/23 rule) [29].

In IMGT®, the IG and TR gene characteristics used for gene identification are defined by concepts generated from the IMGT-ONTOLOGY 'IDENTIFICATION' axiom [9-11]. Three concept instances of the 'Molecule_EntityType' concept of identification are used in the IMGT/LIGMotif model: V-gene, D-gene and J-gene. These concept instances are defined by the gene type (variable (V), diversity (D), joining (J)), the molecule type (gDNA) and the configuration type (germline) [11].

IG and TR gene characteristics for standardized gene description

Prototypes and labels

The IG and TR gene features are described, in IMGT®, according to the standardized concepts of description generated from the IMGT-ONTOLOGY 'DESCRIPTION' axiom [9-11]. Thus, the V-gene, D-gene and J-gene are described by three concept instances of the 'Molecule_EntityPrototype' concept of description: V-GENE, D-GENE and J-GENE [11], respectively. Their graphical representation, or prototype, and the labels that describe them are shown in Figure 1 (A and B, respectively). Among the 242 IMGT® labels defined for the nucleotide sequences, 47 are used in the IMGT/LIGMotif model (Figure 1B) of which 43 are specific of one prototype (23 for a V-GENE, 11 for a D-GENE and 9 for J-GENE). Two labels (5'UTR and

3’UTR) are common to all prototypes, whereas 2 labels (ACCEPTOR-SPLICE and DONOR-SPLICE) are shared by several IMGT-ONTOLOGY prototypes (shown with a black circle in Figure 1B).

The organization of a prototype is based on the relations that order two labels [11]. Interestingly, a set of twelve relations is necessary and sufficient to describe the relations between labels in a prototype (Table 1). Ten of these relations were defined previously [9-11]. Two reciprocal relations, ‘is_in_5_prime_of’ and ‘is_in_3_prime_of’ have been added in the IMGT/LIGMotif model to indicate the relative position of labels on a 5’-3’ DNA strand when there is no intersection between labels (Table 1).

Patterns

For the purpose of gene description, IMGT/LIGMotif uses the ‘gene unit’ labels L-V-GENE-UNIT, D-GENE-UNIT and J-GENE-UNIT (Figure 1A). Indeed, these labels, in contrast to the ‘gene’ labels (V-GENE, D-GENE and J-GENE) have the advantage to be precisely delimitated in 5’ and 3’, respectively, by the 5’ end and 3’ end of constitutive labels (L-PART1 and V-RS for V, 5’D-RS and 3’D-RS for D, and J-RS and J-REGION for J, respectively). Moreover, the part of the prototype they encompass can be defined by conserved motifs that constitute a pattern (Figure 1A). In a pattern, the conserved motifs are separated from each other by a distance in base pairs (bp) comprised between a minimal and a maximal length (between braces in Figure 1A). Motifs are ordered from 5’ to 3’ with a rank (in a circle in Figure 1A) that corresponds to their relative localization in the pattern, the motif the most in 3’ having a rank that corresponds to the number of motifs in the pattern (that is 8 for V, 4 for D and 4 for J). In the J pattern, the motifs J-TRP and J-PHE are shown between brackets and separated with a coma to indicate that the two motifs are possible for the same rank.

These conserved amino acids J-TRP and J-PHE are part of a conserved motif '[W,F]-[G,A]-X-G' where W=tryptophan (J-TRP), F=phenylalanine (J-PHE), G=glycine, A=alanine and X=any amino acid except proline.

IG and TR gene characteristics for functionality identification

The gene functionality identification can only be assigned to precisely described IG and TR genes. In IMGT-ONGOLOGY, an unrearranged genomic V, D or J gene can be functional (F), open reading frame (ORF) or pseudogene (P) [9]. A gene is qualified as 'functional' if the coding region has an open reading frame without stop codon, and if there is no described defect in the splicing sites, recombination signals and/or regulatory elements. A gene is qualified as 'ORF' if the coding region has an open reading frame, but alterations have been described in the splicing sites, recombination signals and/or regulatory elements and/or changes of conserved amino acids have been suggested by the authors to lead to uncorrect folding, and/or the entity is an orphon. A gene is qualified as 'pseudogene' if the coding region has stop codon(s) and/or frameshift mutation(s). In particular, a V-GENE (or V-GENE-UNIT) is considered as 'pseudogene' if these defects occur in the L-PART1 and/or V-EXON, or if there is a mutation in the L-PART1 INIT-CODON atg. A J-GENE (or J-GENE-UNIT) is considered as 'pseudogene' if it has been identified by the presence of a RS upstream of an open reading frame, but it has no donor splice site in 5' or the donor splice is not in the expected splicing frame sf1 or if it has no conserved '[W,F]-[G,A]-X-G' motif.

Characteristics for gene delimitation and cluster assembly

The IMGT® rule to delimit V-GENE, D-GENE and J-GENE instances is to equally distribute the distance between the two genes. The IMGT-ONTOLOGY 'GeneCluster' concept, allows describing genomic sequences containing several

genes. The gene instances in a cluster can be of the same prototype (for example, a V-CLUSTER only contains V genes), or of different prototypes (for example, a V-D-J-CLUSTER contains at least one V gene, one D gene and one J gene). Seven instances of the 'GeneCluster' concept are used in the IMGT/LIGMotif model (Table 2). The IMGT-ONTOLOGY 'GeneCluster' instances are particularly useful for the annotation of the large scale genomic IG and TR loci is also used by the Sequence Ontology (SO) [30] (Table 2).

IMGT/LIGMotif model

The IMGT/LIGMotif model comprises 4 modules (Figure 2), 3 of them ('Gene identification', 'Gene description' and 'Functionality identification') take into account the IG and TR gene characteristics as defined above and deal with individual gene units, whereas the fourth module ('Gene delimitation and cluster assembly') deals with gene delimitation and assembly of genes in a cluster, providing an annotated genomic sequence.

Gene identification

The 'Gene identification' module identifies potential V, D and J genes along the genomic sequence to analyse. First, a heuristic search for local alignments is performed against IMGT/LIGMotif reference motif databases (Table 3). These databases comprise nucleotide sequences that correspond to IG and TR gene unit labels (L-V-GENE-UNIT, D-GENE-UNIT, J-GENE-UNIT) and to motifs that compose them. These databases are created dynamically from IMGT/LIGM-DB [31] sequences, using the IMGT/GENE-DB [16] interface that allows queries on labels. Thirty-four labels were queried corresponding to IG and TR sequences from human (*Homo sapiens*) and mouse (*Mus musculus*; few sequences were also included from *Mus pahari*, *Mus saxicola*, *Mus spretus*). Pseudogene genes too poorly conserved to be assigned to subgroups were excluded. The alignments obtained in this first step

provide labelled high-scoring segment pairs (or HSPs) on both DNA strands of the sequence to analyse. Then, there is a selection of the labelled HSPs and a grouping of these selected HSPs given their topology and their gene type (V, D or J). Thus, the 'Gene identification' module provide the potential V genes, D genes and J genes identified as grouped and labelled HSPs along the sequence to analyse.

Gene description

The 'Gene description' module provides the description of each potential gene identified in the first module. It comprises a search of conserved motifs based on prototypes and patterns. Codons of conserved amino acids in the patterns ('tgg' for CONSERVED-TRP and J-TRP, 'tgc' and 'tgt' for 1st-CYS and 2nd-CYS, and 'ttt' and 'ttc' for J-PHE) are difficult to identify by conventional algorithms as the motifs (triplets 'tgg', 'tgc', 'tgt', 'ttt' and 'ttc') are very frequent in sequences. For that reason, the codons of conserved amino acids of the V-REGION and V-EXON (that comprise 1st-CYS, CONSERVED-TRP and 2nd-CYS) are identified by the software IMGT/V-QUEST [32]. The expected outputs of the 'Gene description' module are described gene units (GENE-UNIT), although, as discussed in the algorithm section, partially described and undescribed outputs can also be obtained.

Functionality identification

The 'Functionality identification' module includes the control of features needed for the functionality assignment and allows to obtain annotated gene units.

Gene delimitation and cluster assembly

In this final module, genes (V-GENE, D-GENE and J-GENE) are delimited and assembled in a cluster if the analysed genomic sequence contains several genes. The final outcome of IMGT/LIGMotif is the annotated genomic sequence.

Algorithm

Gene identification of V, D and J genes

Search of labelled alignments

The algorithm starts by aligning the genomic sequence using BLASTN [33] against IMGT/LIGMotif reference motif databases (Table 3, Figure 3). The possibility is given to the biocurator to select databases on species (human and/or mouse), locus (IGH, IGK, IGL, TRA, TRB, TRG and/or TRD), gene type (V, D and/or J), functionality (F, ORF and/or P), and to choose any combination in this selection. IMGT/BLAST provides HSPs that inform on the similarity of the analysed sequence (query) with labelled motifs from the reference database (subject). These labelled HSPs are obtained on both DNA strands of the sequence to analyse. Practically, NCBI-BLASTN (version 2.2.18) is used with the minimum hit word size possible (i.e. 4), an E-value threshold of 0.01 (except for D-GENE-UNIT and its motifs where an E-value threshold of 5 is selected owing to their very short length). The BLAST was preferred to Hidden Markov Model (HMM) methods and software such as HMMER [34, 35] for practical uses.

Selection of labelled HSPs

IMGT/BLAST produces a huge quantity of HSPs but these HSPs do not have the same importance. HSPs obtained with different reference motif databases may overlap at a same location as they are expected components of a same prototype (e.g. V-EXON overlaps V-REGION). These overlapping HSPs do not need to be filtered as they delimit different and expected labels. In contrast, and owing to gene duplications in IG and TR loci, different HSPs obtained with a same reference motif database may overlap at a same location although they might belong to different genes. In consequence, the best HSPs are selected at a given location (Figure 3) on score, E-value, length and identity BLAST parameters. The method to filter overlapping HSPs obtained with the same motif database is described as follows: score(), length(),

identity(), evalue() are the functions that return 1 if a given HSP (hsp1) performs better than the other (hsp2) for the tested parameter (higher score, length, identity, and lower E-value), 0 if the two HSPs are equal and -1 if hsp1 performs less well than hsp2. Parameter g1 (sum of score() and evalue(), with $-2 \leq g1 \leq 2$) and parameter g2 (sum of length() and identity(), with $-2 \leq g2 \leq 2$) are calculated between each two overlapping HSPs obtained from the same motif database. If $g1 > 0$, or if $g1 = 0$ and $g2 > 0$, hsp2 is removed. In other cases ($g1 < 0$, or $g1 = 0$ and $g2 < 0$), hsp1 is removed.

Grouping of selected HSPs in V, D or J gene

The objective of this step is to group selected HSPs that may belong to a same gene (Figure 3). For that purpose, the positions of selected HSPs from the same DNA strand and with labels of the same gene type (V, D or J) are, in this step, compared to each other (Figure 4). If the topological relation of two compared HSPs is not coherent they are considered as belonging to distinct genes. If the topological relation is coherent, an arbitrary length, specific for each label, is added to both extremities of each compared HSPs (Figure 4) and these new regions are looked for a potential position overlap. If a position overlap is found, the two HSPs are considered as belonging to the same gene. If not, the two HSPs are considered as belonging to different genes. For a given HSP group, the position most in 5' and the position most in 3' define the area that contains a potential gene (shown as arrows in Figure 4). The labelled HSPs that constitute that group provide the gene type (V, D or J) and, using their respective localization on the DNA strands, the gene orientation. Each group of HSPs corresponds to a potential gene identified along the sequence (indicated with V_1 to V_n, D_1 to D_n and J_1 to J_n in Figure 3).

Gene description of L-V-GENE-UNIT, D-GENE-UNIT and J-GENE-UNIT

The second module of IMGT/LIGMotif, 'Gene description', describes in detail each identified gene, individually (Figure 5). This analysis is performed from 5' to 3', exploring both strands. IMGT/LIGMotif starts by searching conserved motifs of the patterns described in Figure 1A. The gene description is performed by searching and delimiting conserved motifs that are characteristic of each gene type (V, D and J).

Delimitation of conserved motif searching areas (CMSA)

In order to reduce the algorithm execution time the search of conserved motifs is limited to conserved motif searching areas (or CMSA) which are delimited by the positions of the most informative combination in each grouped HSP. The best combination of HSPs depends on the gene type (Table 4): for instance, for a V gene the best combination is L-PART1+V-EXON+V-RS, for a D gene, it is 5'D-RS + 3'D-RS, whereas for a J gene, it is J-RS + J-REGION. If the best combination is not present, other combinations are explored in the order shown in Table 4. An arbitrary length is added to both extremities (5' and 3') of the HSP combination to delimit the CMSA. For example, a length of 40 nt (the maximum length of a RS) is added to the 5' and 3' position ends of a RS HSP.

Search of conserved motifs in CMSA

Conserved motifs that include conserved amino acids (INIT-CODON for V, J-TRP and J-PHE for J), splicing sites (DONOR-SPLICE, ACCEPTOR-SPLICE), heptamers and nonamers are searched in the CMSA which are known to include them. Heptamers and nonamers are searched by alignment with the reference motif databases. If no exact match is found, an approximate form of the motifs is searched using non gapped position-specific scoring matrices (PSSM) [36]. Following the search of conserved motifs in CMSA, matches are grouped and retained as a set named 'solution' if the distances between the motifs are in the intervals defined for the pattern. A minimum of two conserved motifs is required to retain the solution. If

this condition is not fulfilled, genes cannot be described using IMGT/LIGMotif and, thus, are defined as ‘V undescribed’, ‘D undescribed’ and ‘J undescribed’ (Figure 5).

Delimitation of motifs using conserved motifs as anchors

In the next step, other motifs of the patterns are then delimited precisely using conserved motifs as anchors (or seeds) (Figure 6). Arrows taking root from anchors delimit precisely new labelled motifs. For instance, ACCEPTOR-SPLICE and V-HEPTAMER (conserved motifs) allow delimiting V-EXON. An arrow arriving from the left delimits the 5’ end of a motif whereas an arrow arriving from the right delimits the 3’ end. The number associated to an arrow indicates the number of nucleotides which must be added (+) or subtracted (-) to a conserved motif position to delimit precisely the new labelled motif. The J-PHE and J-TRP are the only conserved motifs that do not delimit labelled motifs, and consequently, no arrow takes root from them.

Additional steps for the description of a L-V-GENE-UNIT

The description of a L-V-GENE-UNIT requires two additional steps. The first one is the delimitation of the V-REGION with its constitutive regions (FR-IMGT and CDR-IMGT) and conserved motifs (1st-CYS, CONSERVED-TRP and 2nd-CYS) using IMGT/V-QUEST [32]. This step is only be performed if a V-EXON has been identified. If V-EXON is missing the gene is defined as ‘V partially described’. The final step for the description of a L-V-GENE-UNIT is the delimitation of L-PART2, a region delimited by the V-EXON acceptor splice and the 5’ end of the V-REGION determined by IMGT/V-QUEST.

Functionality identification

The third module of IMGT/LIGMotif ‘Functionality identification’ identifies the functionality of each described gene unit (Figure 7). A L-V-GENE-UNIT is identified as functional if it has all the 22 specific labels and expected splicing sites (2 labels),

no stop codons in L-PART1 and V-EXON, a splicing frame sf1 between L-PART1 and V-EXON, no frameshift (same reading frame for 1st-CYS, CONSERVED-TRP and 2nd-CYS), an expected V-SPACER length, and V-HEPTAMER and V-NONAMER identical to ones found in functional genes (Figure 7A). A J-GENE-UNIT is identified as functional if it has all the 7 specific labels and expected splicing site (1 label), no stop codons in J-REGION, a sf1 donor splice, the conserved '[W,F]-[G,A]-X-G' motif as one indicator of the absence of frameshift, an expected J-SPACER length, and J-NONAMER and J-HEPTAMER identical to ones found in functional genes (Figure 7B). A D-GENE-UNIT is identified as functional if it has all the 10 specific labels, at least one open reading frame without stop codon(s), expected 5'D-SPACER and 3'D-SPACER lengths, and the heptamers and nonamers are identical to ones found in functional genes (Figure 7C). A gene unit is identified as an open reading frame (ORF) if the last 3 criteria for V and J and the last 2 for D are not fulfilled. In other cases the gene unit is identified as a pseudogene (P). Note that if the first criterion (number of specific labels and splicing sites) is not fulfilled, the functionality is 'Unknown' as it cannot be determined automatically and its identification therefore requires a manual expertise.

Gene delimitation and cluster assembly

For each gene unit, and whatever its description status (that is either annotated (L-V-GENE-UNIT, D-GENE-UNIT, J-GENE-UNIT), partially described ('V partially described') or undescribed ('V undescribed', 'D undescribed', 'J undescribed')), the 5'UTR and 3'UTR delimitations of the corresponding V-GENE, D-GENE or J-GENE are determined by equally distributing the distance between two neighbouring gene units. Thus, the sequence can be considered as a succession of genes that constitute a cluster. The cluster is defined based on the 'Molecule_EntityPrototype' instances

found in the sequence and an IMGT® cluster label is assigned (Table 2), for instance V-D-J-CLUSTER, if the sequence contains at least one V-GENE, one D-GENE and one J-GENE. Finally, the number of genes in the analysed sequence is computed per DNA strand, per gene type and per functionality.

Results

IMGT/LIGMotif algorithm is implemented in JAVA (http://www.java.com/fr/). A web application of IMGT/LIGMotif is running on a Tomcat server (http://tomcat.apache.org/). The genomic sequence to analyse can be copied/pasted by the biocurators, or uploaded in Fasta or EMBL format. Query parameters can be modified to optimise the analysis efficiency. For instance, reference motif databases can be selected on their gene type, locus, functionality and organism. The execution time depends on the gene types and the number of genes existing in the sequence. The analysis of a single gene takes a few seconds whereas a complete locus containing more than 100 genes takes between 30 minutes to 1 hour, using standard parameters. The IMGT/LIGMotif results page (Figure 8) displays, at the top of the page, statistics that include the execution time the length of the analysed sequence, the total number of genes per DNA strand (plus or minus), and two tables: the first one indicates the number of genes per description status (GENE-UNIT, Partially described, Undescribed) and per gene type, the second table indicates the number of annotated GENE UNIT per functionality status (Functional, ORF, Pseudogene, Unknown). Below the statistics, the main table displays the content of the analysed sequence, starting from N°1 (in column 1) for the gene the most in 5'. This table provides, for each identified genes: Description (GENE-UNIT label or description status), Positions (in the analysed sequence), DNA strand, Functionality and Number of labels. Selecting gene units in the main table (Figure 8A) allows displaying their labels with

positions, nucleotide sequences and, if coding regions, amino acid sequences (Figure 8B). In the detailed display, labels of each gene unit are ordered from 5' to 3', the label with the longest nucleotide length being displayed first, if two labels start at the same 5' position. The IMGT/LIGMotif results can be exported to a spreadsheet file that can be modified at will.

Conclusions

IMGT/LIGMotif is a user friendly tool that provides the annotation of large genomic sequences containing IG and TR genes. The web user interface provides a simple way to query, visualize and download results. The execution time is suitable for the analysis of an entire locus. The annotation includes the gene identification and orientation in the sequence, the gene delimitation (V-GENE, D-GENE and J-GENE) and the complete annotation of L-V-GENE-UNIT, D-GENE-UNIT and J-GENE-UNIT that comprises a detailed description with the IMGT® labels and the functionality identification, and finally the cluster assembly. Comparison of IMGT/LIGMotif results with expert annotation has shown that the tool performs quite well for the functional and ORF genes in human. Pseudogenes are also correctly annotated, provided that they are not too degenerated. Annotation of C-GENE has not been included in this first version of IMGT/LIGMotif as the gene identification and annotation is performed easily with conventional tools by the biocurators. IMGT/LIGMotif is particularly useful for the annotation of IG and TR loci from species that are phylogenetically close to human and mouse such as nonhuman primate species, chimpanzee and rat, respectively. More distant species will still require manual expertise in the control of the annotations. However, it is expected that the progressive enrichment of the IMGT/LIGMotif reference motif databases with

data IG and TR annotated by IMGT® will save a considerable amount of time in the process of the genomic annotation of vertebrate antigen receptor loci.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JL conceived the algorithm and its implementation. PD and MPL coordinated the project. All authors read and agree to publish the manuscript.

Acknowledgements

We are deeply grateful to Véronique Giudicelli, Géraldine Folch, Joumana Michaloud and Fatena Bellahcene for helpful comments. IMGT® received funding from Centre National de la Recherche Scientifique (CNRS), Ministère de l'Enseignement Supérieur et de la Recherche MSER (Université Montpellier 2), Agence Nationale de la Recherche (ANR-06-BYOS-0005-01) and European Community ImmunoGrid (FP6-2004-IST-4).

References

1. Lefranc M-P, Lefranc G: *The Immunoglobulin FactsBook*, Academic Press 2001, pp1-458.
2. Lefranc M-P, Lefranc G: *The T cell receptor FactsBook*, Academic Press 2001, pp1-398.
3. Sakano H, Huppi K, Heinrich G, Tonegawa S: **Sequences at the somatic recombination sites of immunoglobulin light-chain genes.** *Nature* 1979, **280**:288-294.

4. Alt FW, Baltimore D: **Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions.** *Proc Natl Acad Sci U S A* 1982, **79**:4118-4122.
5. Bleakley K, Lefranc M-P, Biau G: **Recovering probabilities for nucleotide trimming processes for T cell receptor TRA and TRG V-J junctions analyzed with IMGT tools.** *BMC Bioinformatics* 2008, **9**:408.
6. Gearhart PJ, Johnson ND, Douglas R, Hood L: **IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts.** *Nature* 1981, **291**:29-34.
7. Neuberger MS, Rada C: **Somatic hypermutation: activation-induced deaminase for C/G followed by polymerase eta for A/T.** *J Exp Med* 2007, **204**:7-10.
8. Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, *et al*: **IMGT®, the international ImMunoGeneTics information system®.** *Nucleic Acids Res* 2009, **37**:D1006-1012.
9. Giudicelli V, Lefranc M-P: **Ontology for immunogenetics: the IMGT-ONTOLOGY.** *Bioinformatics* 1999, **15**:1047-1054.
10. Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, *et al*: **IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics.** *In Silico Biol* 2004, **4**:17-29. [http://www.bioinfo.de/isb/2003040004/]
11. Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V: **IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm.** *Biochimie* 2008, **90**:570-583.

12. Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, *et al*: **IMGT-Choreography for immunogenetics and immunoinformatics.** *In Silico Biol* 2005, **5**:45-60. [http://www.bioinfo.de/isb/2004050006/]
13. Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S: **Guidelines for human gene nomenclature.** *Genomics* 2002, **79**:464-470.
14. Lefranc M-P: **WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report.** *Immunogenetics* 2007, **59**:899-902.
15. 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.
16. Giudicelli V, Chaume D, Lefranc M-P: **IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes.** *Nucleic Acids Res* 2005, **33**:D256-261.
17. Letovsky SI, Cottingham RW, Porter CJ, Li PW: **GDB: the Human Genome Database.** *Nucleic Acids Res* 1998, **26**:94-99.
18. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29**:137-140.
19. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2005, **33**:D54-58.
20. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, *et al*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**:D690-697.

21. Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T, Harrow JL: **The vertebrate genome annotation (Vega) database.** *Nucleic Acids Res* 2008, **36**:D753-760.

22. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.

23. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.

24. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucleic Acids Res* 2005, **33**:6494-6506.

25. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.

26. Gross SS, Brent MR: **Using multiple alignments to improve gene prediction.** *J Comput Biol* 2006, **13**:379-393.

27. De Bono B, Chothia C: **Exegesis a procedure to improve gene predictions and its use to find immunoglobulin superfamily proteins in the human and mouse genomes.** *Nucleic Acids Res.* 2003, **31**:6096-6103.

28. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Research* 2004, **14**:988-995. doi:10.1101/gr.1865504

29. Early P, Huang H, Davis M, Calame K, Hood L: **An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH.** *Cell* 1980, **19**:981-992.

30. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biol* 2005, **6**:R44.
31. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc M-P: **IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences.** *Nucleic Acids Res* 2006, **34**:D781-784.
32. Brochet X, Lefranc M-P, Giudicelli V: **IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis.** *Nucleic Acids Res* 2008, **36**:W503-508.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
34. Eddy S: *HMMER - Profile Hidden Markov Models for Biological Sequence Analysis.* Washington University School of Medicine; 1992.
35. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press; 1998.
36. Mitrophanov AY, Borodovsky M: **Statistical significance in biological sequence analysis.** *Brief Bioinform* 2006, **7**:2-24.

Figures

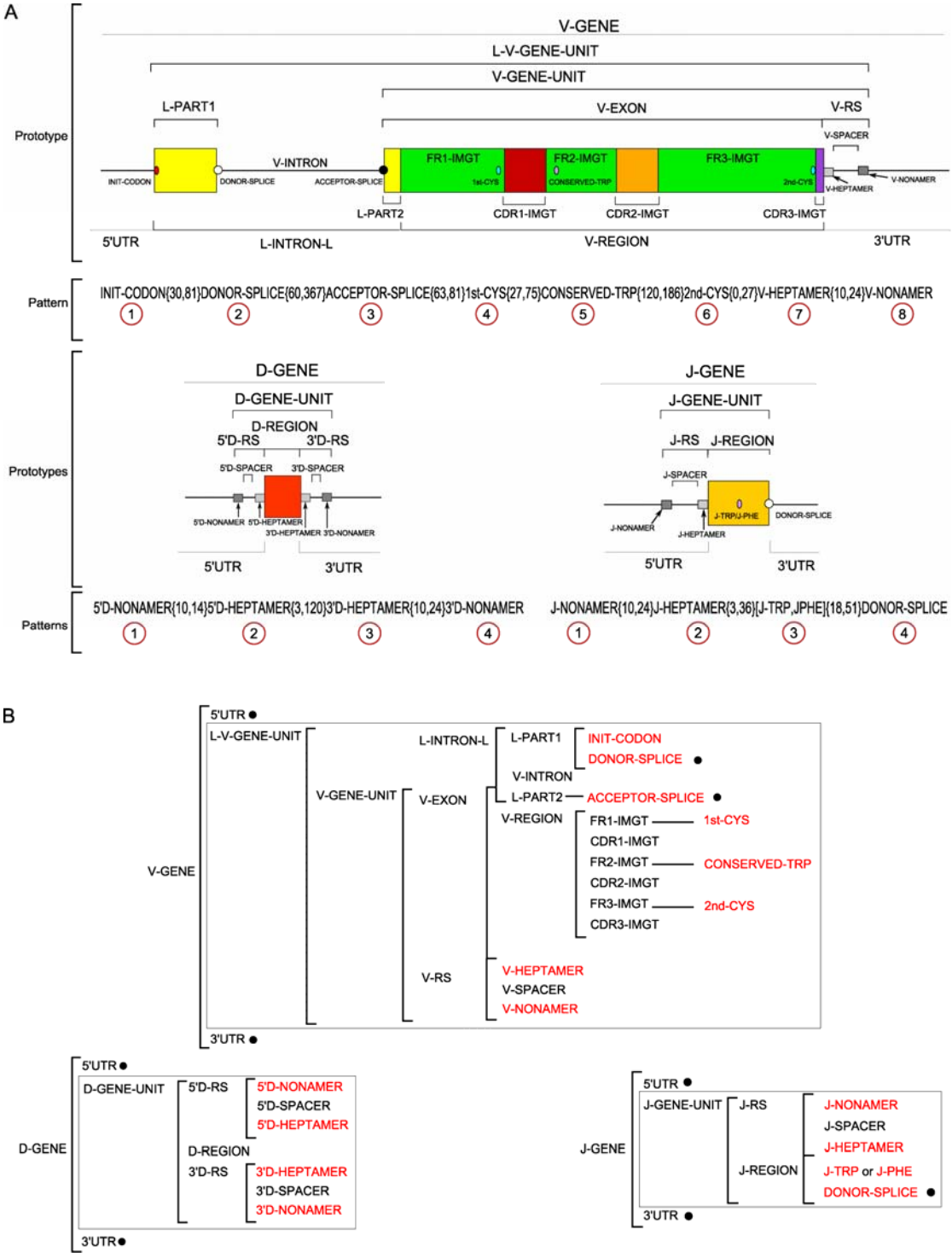


Figure 1 - Prototypes, labels and patterns.

A. V-GENE, D-GENE and J-GENE prototypes with labels and corresponding patterns. B. Labels for each prototype and gene unit. A L-V-GENE-UNIT is described with 24 labels (22 specific and 2 shared ones). A D-GENE-UNIT is described with 10 labels (all specific). A J-GENE-UNIT is described with 8 labels (7 specific labels, J-PHE and J-TRP being mutually exclusive and 1 shared one). Three additional labels, 1 specific and 2 common ones (5'UTR and 3'UTR), allow to describe the V-GENE (27 labels), D-GENE (13 labels) and J-GENE (11 labels). Shared and common labels are shown in black circle. Labels in red are conserved motifs.

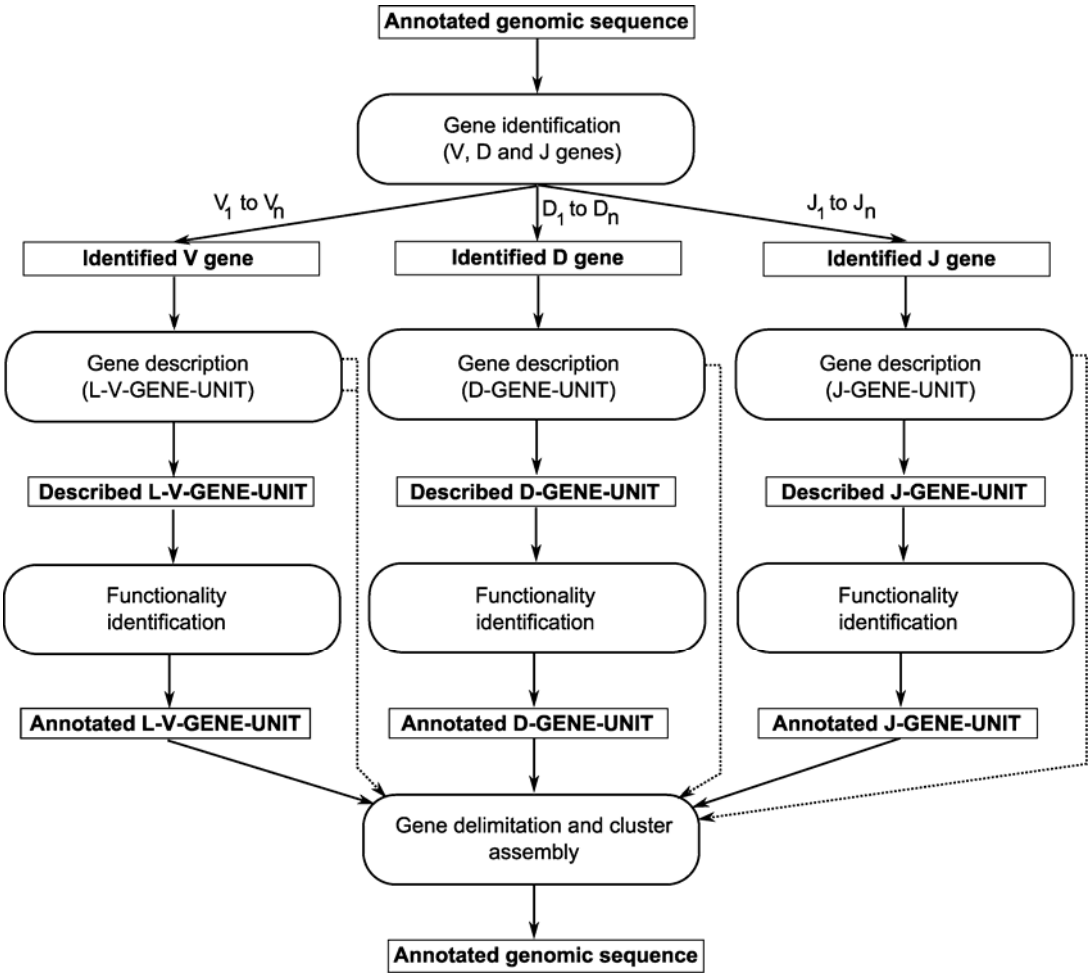


Figure 2 - IMGT/LIGMotif model overview.

The 4 modules IMGT/LIGMotif comprise 'Gene identification', 'Gene description', 'Functionality identification' and 'Gene delimitation and cluster assembly'. Dots indicated partially described and undescribed outputs (for V, D and J) of the 'gene description' module that are entered, for complete analysis of the sequence to analyse, in the last module.

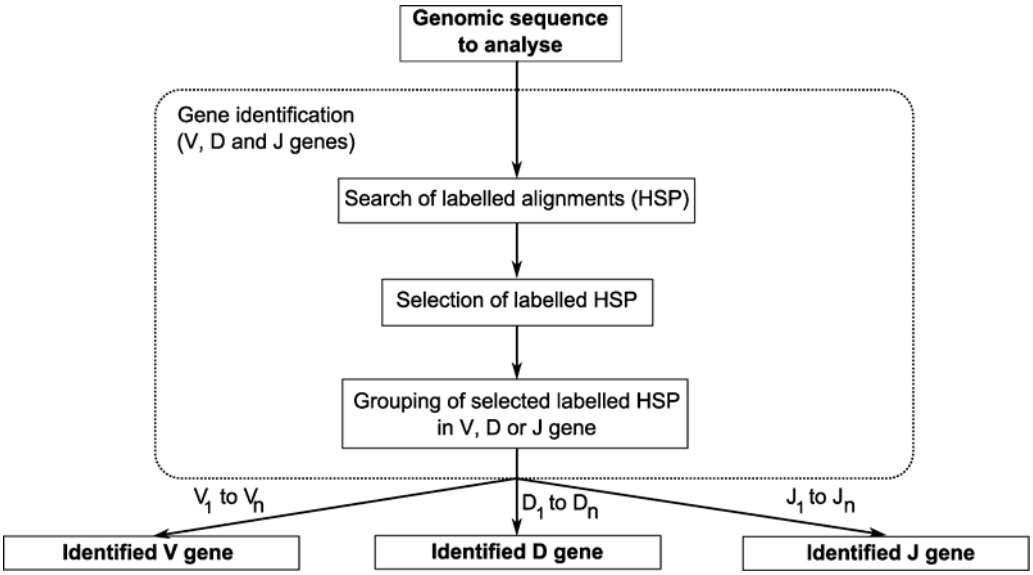


Figure 3 - IMGT/LIGMotif 'Gene identification' module.

The output of the 'Gene identification' module is an 'Identified V gene', 'Identified D gene' and 'Identified J gene'. Each gene are identified by group of labelled alignment (HSP) selected on their BLAST parameters.

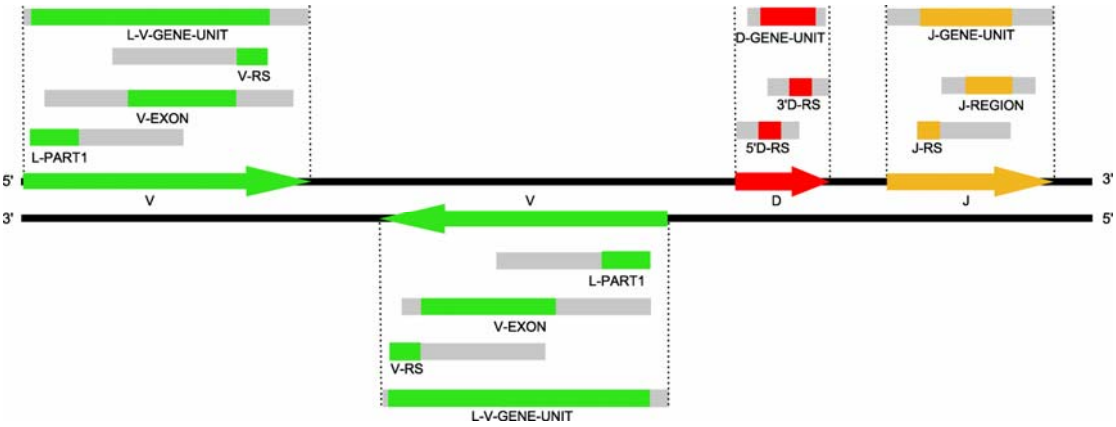


Figure 4 - Grouping of selected labelled HSP into V, D and J genes.

Selected labelled HSP are grouped into V, D and J genes, respectively. The rectangles in grey represent the length added in 5' and or 3' of the selected HSP. A search of position overlaps with the other selected and extended HSP has allowed the grouping. Arrows indicates the gene orientation relative to the DNA strand, as deduced by the algorithm by the respective positions of the labelled HSP. Note that the labels at this stage refer to labelled HSP and gene identification, and not to a detailed description of the genes.

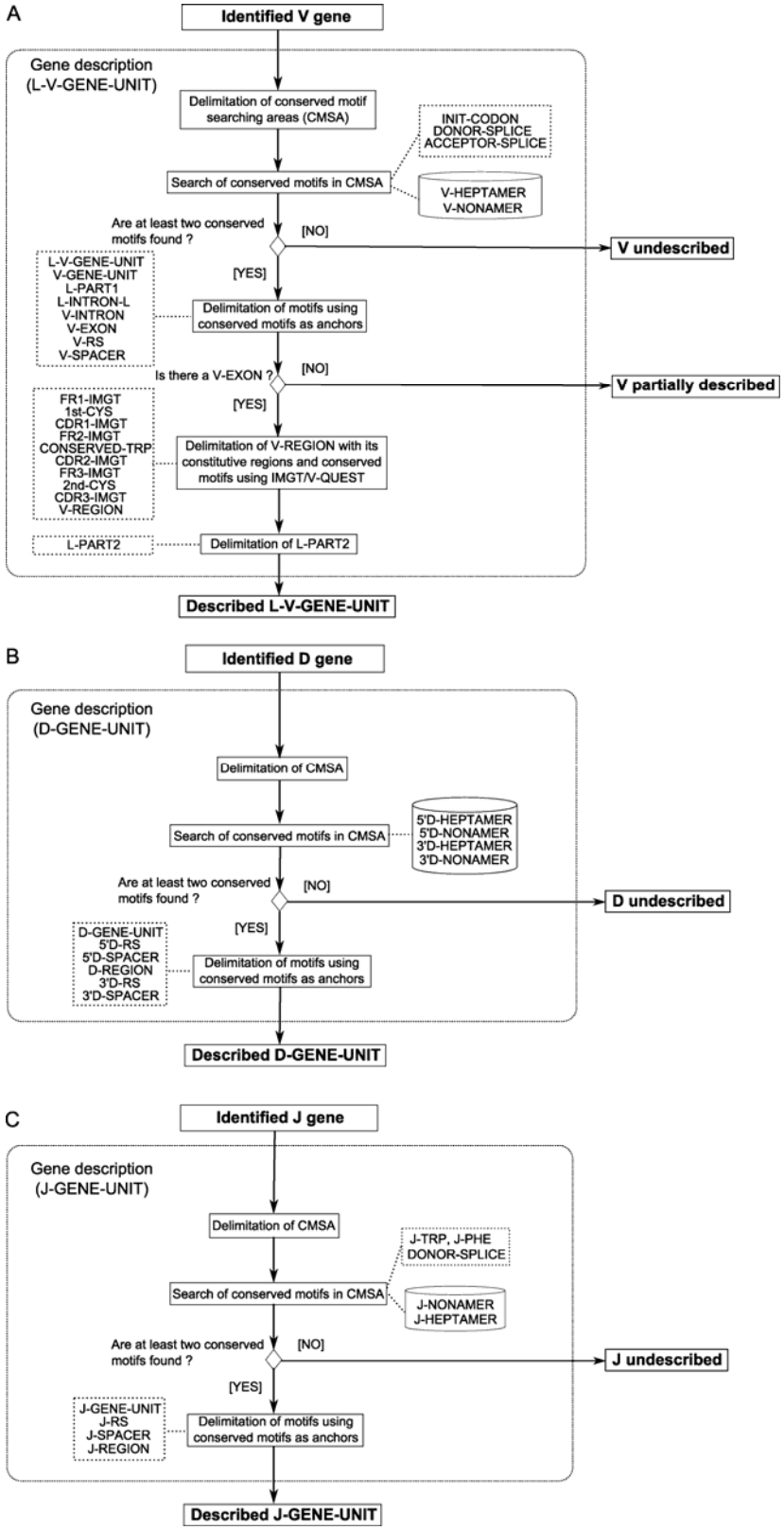


Figure 5 - IMGT/LIGMotif 'Gene description' module.

The output of the 'Gene description' module is a 'Described V-GENE-UNIT' (A), 'Described D-GENE-UNIT' (B) or 'Described J-GENE-UNIT' (C). The delimitation of conserved motif searching areas (CMSA) is described in the text. At least two conserved motifs need to be found. If not, the output is 'V undescribed', 'D undescribed' and 'J undescribed'. The absence of V-EXON for a V gene leads to a 'V partially described'. The delimitation of motifs using conserved motifs as anchors is shown in Figure 6.

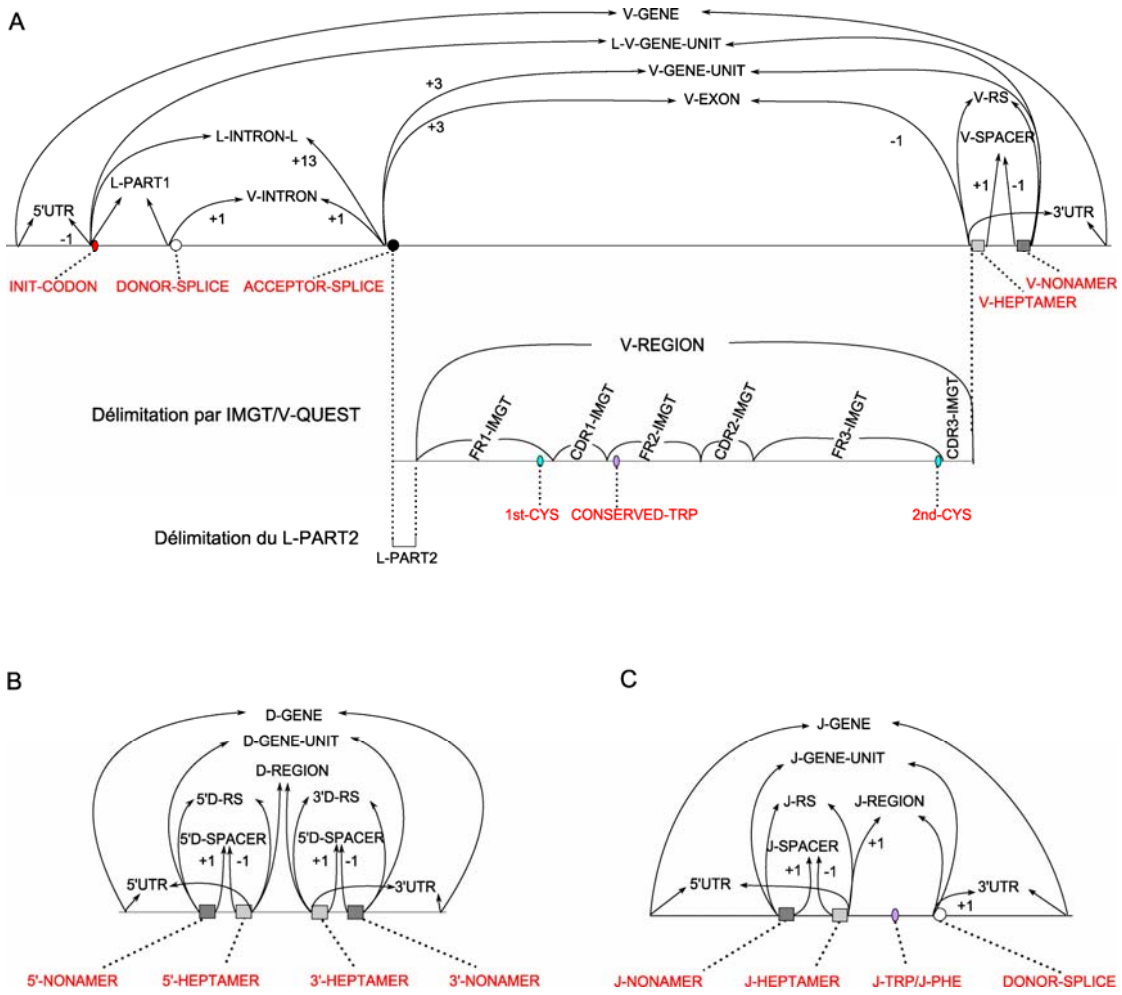


Figure 6 - Delimitation of motifs using conserved motifs as anchors.

The delimitation of motifs using conserved motifs as anchors is used for the description of a V-GENE-UNIT (A), D-GENE-UNIT (B) and J-GENE-UNIT (C). This approach is necessary and sufficient for the description of a D-GENE-UNIT or that of a J-GENE-UNIT. The description of a V-GENE-UNIT requires two additional steps: the delimitation of the V-REGION and of its constitutive regions and conserved motifs by IMGT/V-QUEST [30], and the delimitation of L-PART2.

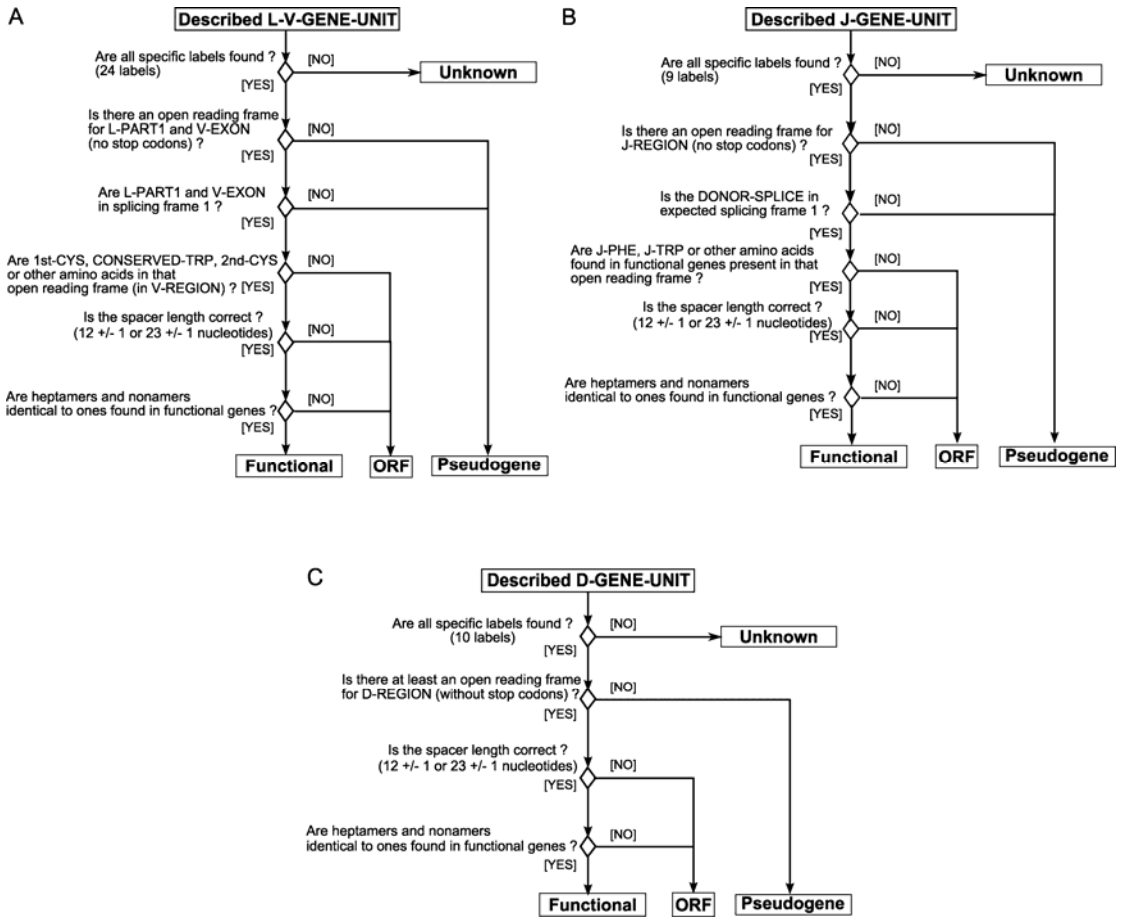


Figure 7 - Functionality identification.

The output of the 'Functionality identification' module is 'Annotated L-V-GENE-UNIT' (A), 'Annotated D-GENE-UNIT' (B) and 'Annotated J-GENE-UNIT' (C). The identification of the functionality is 'Functional', 'ORF' and 'Pseudogene'. If the first criterion is not fulfilled, the functionality is 'Unknown' as it cannot be identified automatically. In the figure, heptamers and nonamers are quoted from 5' to 3' relative to gene units.

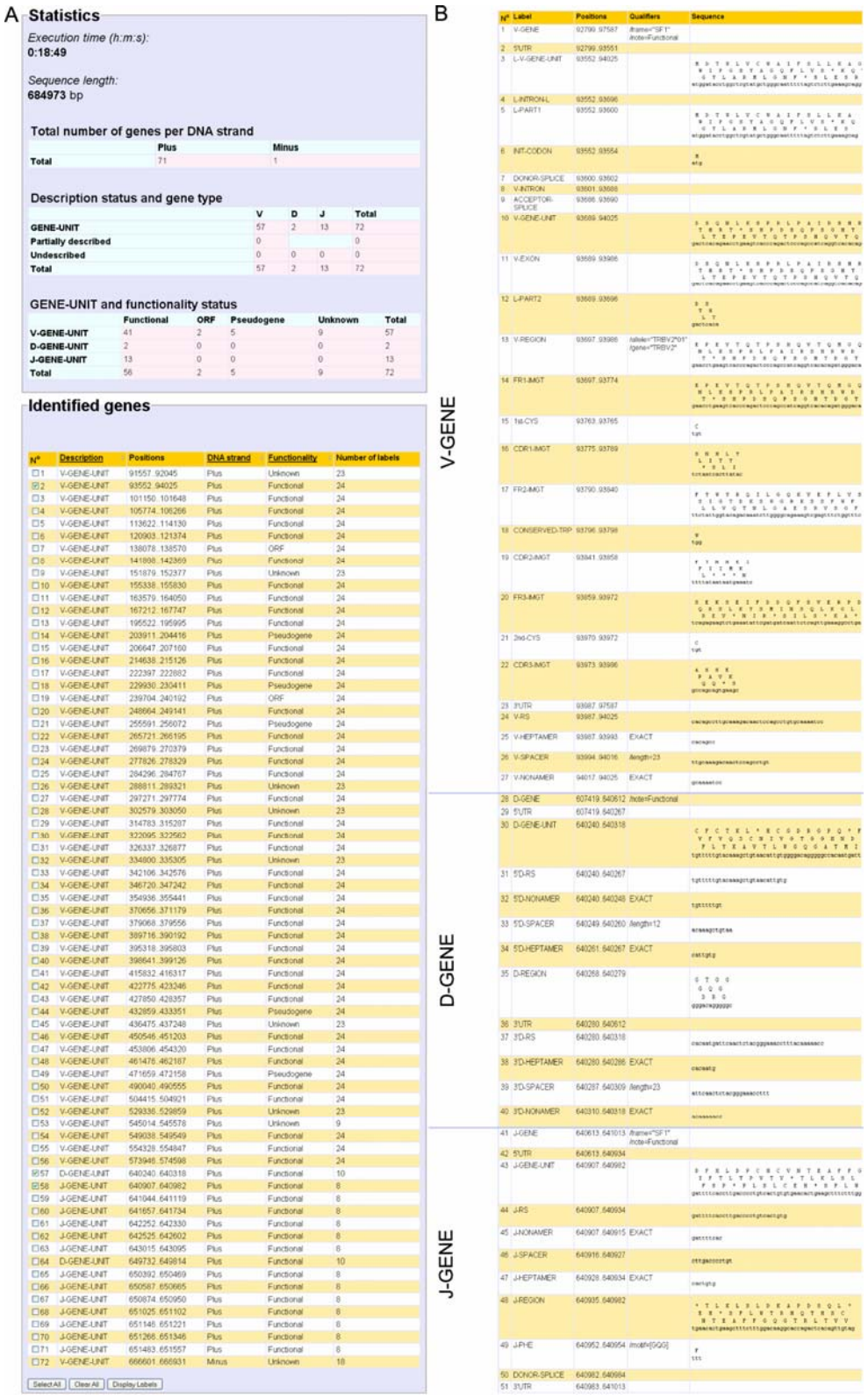


Figure 8 - IMGT/LIGMotif result page.

A. Statistics and Identified genes. The main table provides the identified genes from the 5' end to the 3' end of the DNA sequence (from top to bottom of the table). Three genes were selected to display their labels: a V-GENE-UNIT (N° 2, a D-GENE-UNIT (N° 58) and a J-GENE-UNIT (N° 59).
B. Display of labels for the genes selected in the main table.

Tables

Table 1 - IMGT-ONTOLOGY relations between labels.

Relations between labels used for the description of prototypes (graphical representation of instances of the 'Molecule_EntityPrototype' concept of description) in IMGT-ONTOLOGY [9-11].

Relation	Reciprocal relation
'adjacent_at_its_5_prime_to'	'adjacent_at_its_3_prime_to'
'included_with_same_5_prime_in'	'includes_with_same_5_prime'
'included_with_same_3_prime_in'	'includes_with_same_3_prime'
'overlaps_at_its_5_prime_with'	'overlaps_at_its_3_prime_with'
'included_in'	'includes'
'is_in_5_prime_of'	'is_in_3_prime_of'

Table 2 - 'GeneCluster' concept instances used in IMGT/LIGMotif.

Seven 'GeneCluster' concept instances of IMGT-ONTOLOGY are used in IMGT/LIGMotif. Six of them are also used by Sequence Ontology (SO) [28]. Relations with the 'Molecule_EntityPrototype' concept instances comprise the minimal number of different instances for each 'GeneCluster' concept instance and the name of the different instances, as defined in the IMGT/LIGM-DB list of labels [29].

IMGT-ONTOLOGY "GeneCluster" concept instance	Sequence Ontology	'Molecule_EntityPrototype' concept instance	
		Minimal number of different instances	Name of the different instances
V-CLUSTER	SO:0000526	1	V-GENE
J-CLUSTER	SO:0000513	1	J-GENE
D-CLUSTER	SO:0000559	1	D-GENE
D-J-CLUSTER	SO:0000560	2	D-GENE
			J-GENE
V-D-CLUSTER		2	V-GENE
			D-GENE
V-J-CLUSTER	SO:0000534	2	V-GENE
			J-GENE
V-D-J-CLUSTER	SO:0000532	3	V-GENE
			D-GENE
			J-GENE

Table 3 - IMGT/LIGMotif reference motif databases.

Prototype	Number of databases	Reference motif databases	Number of sequences		Gene identification (Blast HSPs)	Gene description and functionality identification
			Human	Mouse		
V-GENE	16	L-V-GENE-UNIT	368	204	+	
		V-GENE-UNIT	385	221	+	
		L-PART1	534	550	+	
		V-INTRON	575	518	+	
		L-INTRON-L	470	288	+	
		V-EXON	660	550	+	+
		L-PART2	591	580	+	
		V-REGION	887	1036	+	
		FR1-IMGT	785	730	+	
		CDR1-IMGT	790	731	+	
		FR2-IMGT	790	739	+	
		CDR2-IMGT	788	737	+	
		FR3-IMGT	788	737	+	
		CDR3-IMGT	674	607	+	
		V-RS	378	222	+	
		V-SPACER	485	449	+	
	2	V-HEPTAMER	326	317		+
		V-NONAMER	262	298		+
D-GENE	6	D-GENE-UNIT	36	28	+	
		5'D-RS	40	29	+	
		5'D-SPACER	50	29	+	
		D-REGION	50	38	+	
		3'D-RS	36	32	+	
		3'D-SPACER	47	32	+	
	4	5'D-NONAMER	36	22		+
		5'D-HEPTAMER	36	22		+
		3'D-HEPTAMER	36	21		+
		3'D-NONAMER	33	21		+
J-GENE	4	J-GENE-UNIT	120	107	+	
		J-RS	120	107	+	
		J-SPACER	120	108	+	
		J-REGION	130	121	+	
	2	J-NONAMER	101	76		+
		J-HEPTAMER	101	77		+

Table 4 - Combinations of labelled HSPs used for the delimitation of conserved motif searching areas (CMSA).

The underlined combinations are those used in priority. If not present, the combinations below in the columns are chosen in the order from top to bottom.

Prototype	L-V-GENE-UNIT	D-GENE-UNIT	J-GENE-UNIT
Combination of labelled HSPs	L-PART1 + V-EXON + V-RS	5'D-RS + 3'D-RS[2]	J-RS + J-REGION
	L-PART1 + V-EXON	5'D-RS + D-REGION	J-REGION
	V-EXON[1] + V-RS	D-REGION + 3'D-RS	J-RS
	V-EXON[1]	5'D-RS	J-GENE-UNIT
	L-V-GENE-UNIT	3'D-RS	
		D-REGION	
		D-GENE-UNIT	

(1) If L-PART1 is missing, its motifs (INIT-CODON and DONOR-SPLICE) are not delimited.

(2) D-REGION is not used even if it is identified because 5'D-RS and 3'D-RS are sufficient for its precise delimitation, as well as those of the heptamers and nonamers.

PUBLICATION 2

D1006–D1012 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 31 October 2008*
doi:10.1093/nar/gkn838

IMGT®, the international ImMunoGeneTics information system®

Marie-Paule Lefranc*, Véronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Géraldine Folch, Fatena Bellahcene, Yan Wu, Elodie Gemrot, Xavier Brochet, Jérôme Lane, Laetitia Regnier, François Ehrenmann, Gérard Lefranc and Patrice Duroux

IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Received September 13, 2008; Accepted October 14, 2008

ABSTRACT

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org), was created in 1989 by Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and manage the complexity of immunogenetics data. The building of a unique ontology, IMGT-ONTOLOGY, has made IMGT® the global reference in immunogenetics and immunoinformatics. IMGT® is a high-quality integrated knowledge resource specialized in the immunoglobulins or antibodies, T cell receptors, major histocompatibility complex, of human and other vertebrate species, proteins of the IgSF and MhcSF, and related proteins of the immune systems of any species. IMGT® provides a common access to standardized data from genome, proteome, genetics and 3D structures. IMGT® consists of five databases (IMGT/LIGM-DB, IMGT/GENE-DB, IMGT/3Dstructure-DB, etc.), fifteen interactive online tools for sequence, genome and 3D structure analysis, and more than 10 000 HTML pages of synthesis and knowledge. IMGT® is used in medical research (autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas and myelomas), veterinary research, biotechnology related to antibody engineering (phage displays, combinatorial libraries, chimeric, humanized and human antibodies), diagnostics (clonalities, detection and follow-up of residual diseases) and therapeutical approaches (graft, immunotherapy, vaccinology). IMGT is freely available at http://www.imgt.org.

INTRODUCTION

The number of genomics, genetics, 3D and functional data published in the immunogenetics field is growing exponentially and involves fundamental, clinical, veterinary, and pharmaceutical research. The number of potential protein forms of the antigen receptors, immunoglobulins (IG) and T cell receptors (TR) is almost unlimited. The potential repertoire of each individual is estimated to comprise about 10^{12} different IG (or antibodies) and TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce. This huge diversity is inherent to the particularly complex and unique molecular synthesis and genetics of the antigen receptor chains. This 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 (1,2).

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) (3), was created in 1989 by Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and manage the complexity of immunogenetics data. IMGT® has reached that goal through the building of a unique ontology, IMGT-ONTOLOGY (4), the first ontology in immunogenetics and immunoinformatics. IMGT-ONTOLOGY has allowed the setting up of the official nomenclature of the IG and TR genes and alleles (5,6), the definition of IMGT standardized labels, and the IMGT unique numbering that bridges the gap between sequences and 3D structures for the variable (V) and constant (C) domains of the IG and TR (7–10) and for the groove (G) domains of the major histocompatibility

*To whom correspondence should be addressed. Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: marie-paule.lefranc@igh.cnrs.fr

complex (MHC) (11). IMGT® is recognized as the global reference that provides the standards in immunogenetics and immunoinformatics. IMGT® is a high-quality integrated knowledge resource, specialized in the IG, TR, MHC of human and other vertebrates, the proteins that belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF), and the related proteins of the immune systems (RPI) of any species. IMGT® provides a common access to standardized data from genome, proteome, genetics and 3D structures.

The IMGT® information system consists of databases, tools and Web resources (3). IMGT® databases include one genome database, several sequence databases and one 3D structure database. Fifteen IMGT® interactive online tools are provided for genome, sequence and 3D structure analysis. IMGT® Web resources comprise more than 10 000 HTML pages of synthesis and knowledge (IMGT Scientific chart, IMGT Repertoire, The IMGT Medical page, The IMGT Veterinary page, The IMGT Biotechnology page, IMGT Education, IMGT Lexique, IMGT Aide-Mémoire, Tutorials, IMGT Index), external links (IMGT Bloc-notes, The IMGT Immunoinformatics page) and IMGT other accesses (SRS, MRS). Despite the heterogeneity of these different components, all data in IMGT® are expertly annotated. The accuracy, the consistency and the integration of the IMGT® data, as well as the coherence between the different IMGT® components (databases, tools and Web resources) are based on the IMGT-ONTOLOGY axioms and concepts (4,12).

IMGT-ONTOLOGY

Formal IMGT-ONTOLOGY

The Formal IMGT-ONTOLOGY, also designated as IMGT Kaleidoscope (12), comprises seven axioms: IDENTIFICATION, DESCRIPTION, CLASSIFICATION, NUMEROTATION, ORIENTATION, LOCALIZATION and OBTENTION that postulate that objects, processes and relations have to be identified, described, classified, numerotated, localized, orientated, and that the way they are obtained has to be determined. IMGT-ONTOLOGY concepts derived from these axioms are available, for the biologists and IMGT® users, in the IMGT Scientific chart, and have been formalized, for the computing scientists, in IMGT-ML which is an XML Schema (http://www.w3.org/TR/xmlschema-0/). In order to formalize the semantic relations between concepts and instances that are essential for high-quality data processing and coherence control, IMGT-ONTOLOGY is currently designed with Protégé (13) and OBO-Edit (http://oboedit.org/).

IMGT Scientific chart

The IMGT Scientific chart is constituted by controlled vocabulary and annotation rules for data and knowledge management of the IG, TR, MHC, IgSF, MhcSF and RPI. All IMGT® data are expertly annotated according to the IMGT Scientific chart rules.

Keywords and labels. IMGT standardized keywords (concepts of identification) are assigned to all entries in the IMGT® databases. More than 500 IMGT standardized labels (concepts of description) were necessary to describe all structural and functional subregions that compose IG and TR (221 labels for sequences and 285 for 3D structures). Interestingly, 64 IMGT specific labels defined for nucleotide sequences have been entered in the newly created Sequence Ontology (SO) (14).

Gene and allele nomenclature. All the human IMGT standardized gene names (5,6) (concepts of classification) were approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 (15), and entered in IMGT/GENE-DB (16), and in Entrez Gene NCBI (17), and more recently on the Ensembl server (18) at the European Bioinformatics Institute (EBI) in 2006, and in the Vega (19) database at the Wellcome Trust Sanger Institute in 2008. All the mouse IMGT® gene and allele names and the corresponding IMGT reference sequences were provided to HGNC and to the Mouse Genome Informatics Mouse Genome Database (20) in July 2002 and were presented by IMGT® at the 19th International Mouse Genome Conference, IMGC 2005, in Strasbourg, France, and entered in 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.

IMGT unique numbering. The IMGT unique numbering (concepts of numerotation) (7–11) is, with its 2D graphical representation or IMGT Collier de Perles (21,22), the flagship of IMGT® that allows to bridge the gap between sequences, genes and 3D structures in the IMGT® databases, tools and Web resources (23). Structural and functional domains of the IG and TR chains comprise the V-DOMAIN (9-strand β-sandwich) which corresponds to the V-J-REGION or V-D-J-REGION and is encoded by two or three genes (1,2), and the constant domain or C-DOMAIN (7-strand β-sandwich). The IMGT unique numbering initially defined for the IG and TR domains has been extended to the V-LIKE-DOMAIN and C-LIKE-DOMAIN of IgSF proteins other than IG and TR (9,10,22). The IMGT unique numbering for the MHC G-DOMAIN (four β-strand and one α-helix) has been extended to the G-LIKE-DOMAIN of MhcSF proteins other than MHC (11,22).

IMGT Choreography

In order to extract knowledge from IMGT® standardized immunogenetics data, three main IMGT® biological approaches have been developed: genomic, genetic and structural approaches (Figure 1). The IMGT® genomic approach is gene-centered and mainly orientated towards the study of the genes within their loci and on the chromosomes. The IMGT® genetic approach refers to the study of the genes in relation with their sequence polymorphisms and mutations, their expression, their specificity and their evolution. The IMGT® structural

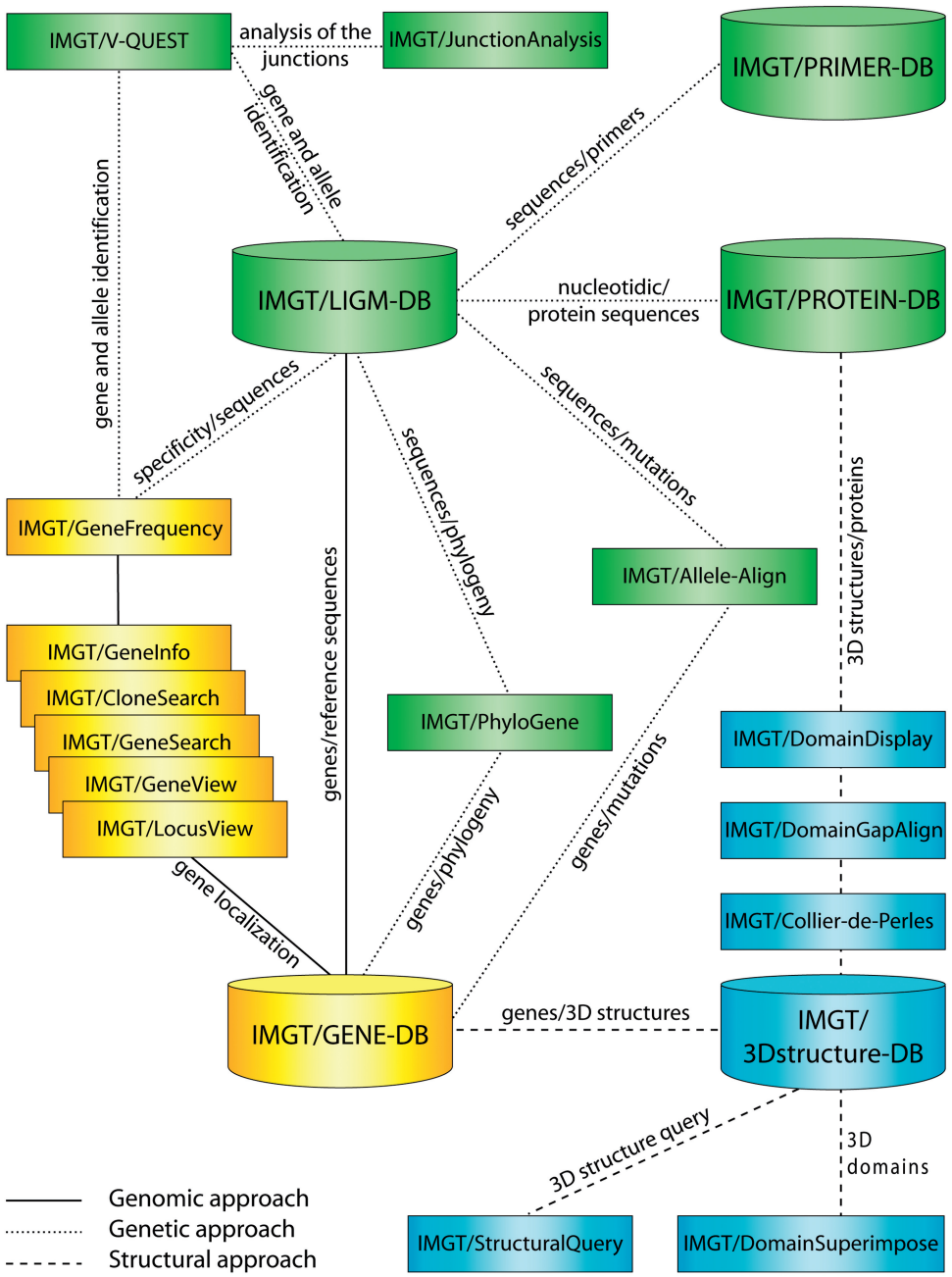


Figure 1. IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org). Genomic, genetic and structural components (databases and tools) are in yellow, green and blue, respectively. The IMGT Repertoire and other Web resources are not shown. Interactions in the genomic, genetic and structural approaches are represented with continuous, dotted and broken lines, respectively.

approach refers to the study of the 2D and 3D structures of the IG, TR, MHC, IgSF, MhcSF and RPI, and to the antigen- or ligand-binding characteristics in relationship with the protein functions, polymorphisms and evolution. For each approach, IMGT® provides databases, tools and Web resources (Figure 1 and Table 1). IMGT-Choreography (33), based on the Web service architecture paradigm, has been developed with the goal to enable significant biological and clinical requests involving every part of the IMGT® information system.

IMGT® DATABASES

Gene database

IMGT/GENE-DB (16) is the comprehensive IMGT® genome database. IMGT/GENE-DB is the official repository of all the IG and TR genes and alleles approved by the World Health Organization (WHO)/International Union of Immunological Societies (IUIS) Nomenclature Subcommittee for IG and TR (34,35). In September 2008, IMGT/GENE-DB contained 1911 IG and TR genes from human, mouse and rat and 2909 alleles. Reciprocal links

Table 1. IMGT® databases, tools and Web resources for genomic, genetic and structural approaches

Approaches	Databases	Tools	Web resources[a]
Genomic	IMGT/GENE-DB (16)	IMGT/GeneView IMGT/LocusView IMGT/CloneSearch IMGT/GeneSearch IMGT/GeneInfo (28) IMGT/GeneFrequency	IMGT Repertoire 'Locus and genes' section: –chromosomal localizations (1,2) –locus representations (1,2) –locus description –gene tables, etc. –potential germline repertoires –lists of genes –correspondence between nomenclatures (1,2)
Genetic	IMGT/LIGM-DB (24) IMGT/PRIMER-DB (25) IMGT/MHC-DB (26)	IMGT/V-QUEST (29) IMGT/JunctionAnalysis (30) IMGT/Allele-Align IMGT/PhyloGene (31) IMGT/DomainDisplay	IMGT Repertoire 'Proteins and alleles' section: –alignments of alleles –protein displays –tables of alleles etc.
Structural	IMGT/3Dstructure-DB (27)	IMGT/DomainGapAlign IMGT/Collier-de-Perles IMGT/DomainSuperimpose IMGT/StructuralQuery (27)	IMGT Repertoire '2D and 3D structures' section: –IMGT Colliers de Perles (2D representations on one layer or two layers) (21,22) –IMGT classes for amino acid characteristics (32) –IMGT Colliers de Perles reference profiles (32) –3D representations

[a]Only Web resources examples from the IMGT Repertoire section are shown.

exist between IMGT/GENE-DB and the HGNC database (36) and Entrez Gene (17). IMGT-GENE-DB allows a query per gene and allele name. IMGT/GENE-DB interacts dynamically with IMGT/LIGM-DB (24) to download and display human, mouse and rat gene-related sequence data. This is the first example of an interaction between IMGT® databases using the concepts of classification.

Sequence databases

IMGT/LIGM-DB. IMGT/LIGM-DB (24) is the comprehensive IMGT® database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, created in 1989 by LIGM, Montpellier, France, on the Web since July 1995. IMGT/LIGM-DB is the first and the largest IMGT® database. In September 2008, IMGT/LIGM-DB contained 126 667 nucleotide sequences of IG and TR from 223 vertebrate species. The unique source of data for IMGT/LIGM-DB is EMBL-Bank (18) which shares data with the other two generalist databases GenBank (37) and DDBJ (38). IMGT/LIGM-DB sequence data are identified by the EMBL/GenBank/DDBJ accession number. Based on expert analysis, specific detailed annotations are added to IMGT flat files. The Web interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge in a computing language. Selection is displayed at the top of the resulting sequences page, so the users can check their own queries. Users have the possibility to modify their request or consult the results with a choice of nine possibilities. The IMGT/LIGM-DB annotations (gene and allele name assignment, labels) allow data retrieval not only from IMGT/LIGM-DB, but also from other IMGT® databases. For example, the IMGT/GENE-DB entries provide the IMGT/LIGM-DB accession numbers of the IG and TR cDNA sequences that contain a given V, D, J or C gene. The automatic annotation of rearranged human and mouse cDNA sequences in IMGT/LIGM-DB is performed by IMGT/Automat (39), an internal Java tool that implements IMGT/V-QUEST (29) and IMGT/JunctionAnalysis (30). IMGT/LIGM-DB data are also distributed by anonymous FTP servers at CINES (ftp://ftp.cines.fr/IMGT/) and EBI (ftp://ftp.ebi.ac.uk/pub/databases/imgt/) and from several Sequence Retrieval System (SRS) sites. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI, IGH, Institut Pasteur Paris).

Other IMGT sequence databases. IMGT/PRIMER-DB (25) is the IMGT® oligonucleotide database on the Web since February 2002. In September 2008, IMGT/PRIMER-DB contained 1864 entries. The database manages standardized information on oligonucleotides (or Primers) and combinations of primers (Sets and Couples) for IG and TR. These primers are useful for combinatorial library constructions, scFv, phage display or microarray technologies. IMGT/PROTEIN-DB, in developement, will contain the translations of the IMGT/LIGM-DB and IMGT/GENE-DB sequences. IMGT/MHC-DB hosted at EBI comprises IMGT/HLA for human MHC (or HLA) and IMGT/MHC-NHP, for MHC of non-human primates (26).

Structure database

IMGT/3Dstructure-DB is the IMGT® 3D structure database, created by LIGM, on the Web since November 2001 (27). IMGT/3Dstructure-DB comprises IG, TR, MHC, IgSF, MhcSF and RPI with known 3D structures. In September 2008, IMGT/3Dstructure-DB contained 1461 atomic coordinate files. These coordinate files extracted from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/) (40) are renumbered according to the standardized IMGT unique numbering (9–11). The IMGT/3Dstructure-DB cards provide chain details with IMGT

annotations (receptor, chain and domain description with IMGT labels, assignment of IMGT gene and allele names, domain delimitations and amino acid positions according to the IMGT unique numbering, and IMGT Colliers de Perles on one layer and two layers), contact analysis, downloadable renumbered IMGT/3Dstructure-DB flat files, visualization tools (Jmol and QuickPDB), and external links. IMGT Residue@Position cards provide detailed information on the inter- and intra-domain contacts at each residue position, based on the IMGT unique numbering. The contacts are described per domain (intra- and inter-domain contacts) and annotated in terms of IMGT® labels (chain and domain), positions (IMGT unique numbering), backbone or side-chain implication.

IMGT® TOOLS

Gene tools

The IMGT® gene tools (genomic approach) manage the locus organization and gene location and provide the display of physical maps for the human and mouse IG, TR and MHC loci. They allow to view genes in a locus (IMGT/GeneView, IMGT/LocusView), to search for clones (IMGT/CloneSearch), or to search for genes in a locus (IMGT/GeneSearch, IMGT/GeneInfo) based on IMGT® gene names, functionality or localization on the chromosome. IMGT/GeneFrequency provides a graphical representation of the numbers of cDNA and gDNA IMGT/LIGM-DB sequences containing rearranged IG and TR genes.

Sequence tools

The IMGT® sequence analysis tools (genetic approach) comprise IMGT/V-QUEST (29) for the identification of the V, D and J genes and of their mutations, IMGT/JunctionAnalysis (30) for the analysis of the V-J and V-D-J junctions that confer the antigen receptor specificity, IMGT/Allele-Align for the detection of polymorphisms, IMGT/Phylogene (31) for gene evolution analyses, and IMGT/DomainDisplay for the display of amino acid sequences from the IMGT domain directory. IMGT/V-QUEST (V-QUEry and STandardization) (29) is an integrated software for IG and TR. This tool, which is easy to use, analyses an input of up to fifty IG or TR germline or rearranged variable nucleotide sequences. IMGT/V-QUEST results comprise, for rearranged sequences, the identification of the V, D and J genes and alleles, nucleotide alignments by comparison with the IMGT reference directory, the delimitations of the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) based on the IMGT unique numbering, the protein translation of the input sequences, the result of IMGT/JunctionAnalysis, the description of the mutations and amino acid changes of the V-REGION and the IMGT Collier de Perles representation of the V-DOMAIN.

Structure tools

The IMGT® structure tools bridge the gap between sequences and 3D structures: IMGT/DomainGapAlign analyses amino acid sequences per domain, IMGT/Collier-de-Perles allows to make your own IMGT Collier de Perles, and IMGT/DomainSuperimpose allows to superimpose two domain 3D structures from IMGT/3Dstructure-DB. IMGT/StructuralQuery (27) allows to retrieve the IMGT/3Dstructure-DB entries containing a V-DOMAIN, based on specific structural characteristics of the intramolecular interactions: phi and psi angles, accessible surface area, amino acid type, distance in angstrom between amino acids, and CDR-IMGT lengths.

IMGT® WEB RESOURCES

IMGT Repertoire

The IMGT® Web resources for genomic, genetic and structural approaches are compiled in the sections of the IMGT Repertoire and provide a synthetic view of data managed in the databases and tools.

Genomics Web resources. The IMGT® genomics resources are compiled in the 'Locus and genes' section which includes 'Chromosomal localizations', 'Locus representations', 'Locus description', 'Gene exon/intron organization', 'Gene exon/intron splicing sites', 'Gene tables', 'Potential germline repertoires', lists of IG and TR genes and links between IMGT®, HGNC, Entrez Gene and OMIM, and correspondence between nomenclatures (1,2). The IMGT Repertoire 'Probes and RFLP' section provides data on gene insertion/deletion.

Genetics Web resources. The IMGT® genetics resources are compiled in the 'Proteins and alleles' section which includes 'Alignments of alleles', 'Tables of alleles', 'Allotypes', 'Isotypes', 'Protein displays', etc.

Structural Web resources. The IMGT® structural resources are compiled in the '2D and 3D structures' section which includes IMGT Colliers de Perles (21,22), FR-IMGT and CDR-IMGT lengths, amino acid chemical characteristics profiles (32). To appropriately analyse the amino acid resemblances and differences between IG, TR, MHC and RPI chains, eleven IMGT classes were defined for the amino acid 'chemical characteristics' properties and used to set up IMGT Colliers de Perles references profiles. IMGT Colliers de Perles reference profiles allow to easily compare amino acid properties at each position whatever the domain, the chain, the receptor or the species. The visualization of 3D representations of IG and TR variable domains allows rapid correlation between protein sequences and 3D data.

Other Web resources

In addition to the IMGT Scientific chart and IMGT Repertoire, other major components of the IMGT® Web resources comprise The IMGT Medical page, The IMGT Veterinary page, The IMGT Biotechnology page,

IMGT Education, IMGT Lexique, IMGT Aide-Mémoire, Tutorials, IMGT Index, and external links (IMGT Bloc-notes, The IMGT Immunoinformatics page, Interesting links) and IMGT other accesses (SRS,MRS).

CONCLUSION

Since July 1995, IMGT® has been available on the Web at the IMGT Home page http://www.imgt.org (Montpellier, France). IMGT® has an exceptional response with more than 150 000 requests a month. The information is of much value to clinicians and biological scientists in general. IMGT® databases, tools, and Web resources are extensively queried and used by scientists from both academic and industrial laboratories, who are equally distributed between the United States, Europe and the remaining world. IMGT® is used in very diverse domains: (i) fundamental and medical research (repertoire analysis of the IG antibody recognition sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas and 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 (single chain Fragment variable (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 and vaccinology). The creation of dynamic interactions between the IMGT® databases and tools, using Web services and IMGT-ML, and the design of IMGT-Choreography, represent novel and major developments of IMGT®, the international reference in immunogenetics and immunoinformatics.

CITING IMGT

Users are requested to cite this article and quote the IMGT home page URL, http://www.imgt.org.

ACKNOWLEDGEMENTS

We thank all the IMGT® users from academic and industrial laboratories and the clinicians and scientists from the European Research Initiative on CLL who help promoting standardization. IMGT® has received the National Bioinformatics Platform RIO label since the RIO creation in 2001 (CNRS, INSERM, CEA, INRA) and the National Bioinformatics Platform IBiSA label since the IBiSA creation in 2007. IMGT® is an Institutional Academic Member of the International Medical Informatics Association. IMGT® is a registered mark of the Centre National de la Recherche Scientifique.

FUNDING

The BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources programmes (QLG2-2000-01287) programmes of the European Union (EU). IMGT® is currently supported by the CNRS, the Ministère de l'Enseignement Supérieur et de la Recherche (MESR) (Université Montpellier 2 Plan Pluri-Formation, Institut Universitaire de France), Réseau National des Génopoles, the Région Languedoc-Roussillon, the Agence Nationale de la recherche ANR (BIOSYS06_135457, FLAVORES), and the EU ImmunoGrid (IST-028069). Funding for open access charge: CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Lefranc,M.-P. and Lefranc,G. (2001) *The Immunoglobulin FactsBook*, Academic Press, London, UK, pp. 1–458.
2. Lefranc,M.-P. and Lefranc,G. (2001) *The T cell receptor FactsBook*, Academic Press, London, UK, pp. 1–398.
3. Lefranc,M.-P., Giudicelli,V., Kaas,Q., Duprat,E., Jabado-Michaloud,J., Scaviner,D., Ginestoux,C., Clément,O., Chaume,D. and Lefranc,G. (2005) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, **33**, D593–D597.
4. Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for Immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics*, **12**, 1047–1054.
5. Lefranc,M.-P. (2000) Nomenclature of the human immunoglobulin genes. In Coligan,J.E., Bierer,B.E., Margulies,D.E., Shevach,E.M. and Strober,W. (eds), *Current Protocols in Immunology*, John Wiley & Sons, Inc., NJ, Hoboken, pp. A.1P.1–A.1P.37.
6. Lefranc,M.-P. (2000) Nomenclature of the human T cell receptor genes. In Coligan,J.E., Bierer,B.E., Margulies,D.E., Shevach,E.M. and Strober,W. (eds), *Current Protocols in Immunology*, John Wiley & Sons, Inc., NJ, Hoboken, pp. A.1O.1–A.1O.23.
7. Lefranc,M.-P. (1997) Unique database numbering system for immunogenetic analysis. *Immunol. Today*, **18**, 509.
8. Lefranc,M.-P. (1999) The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist*, **7**, 132–136.
9. Lefranc,M.-P., Pommié,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
10. Lefranc,M.-P., Pommié,C., Kaas,Q., Duprat,E., Bosc,N., Guiraudou,D., Jean,C., Ruiz,M., Da Piedade,I., Rouard,M. *et al.* (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.*, **29**, 185–203.
11. Lefranc,M.-P., Duprat,E., Kaas,Q., Tranne,M., Thiriot,A. and Lefranc,G. (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev. Comp. Immunol.*, **29**, 917–938.
12. Duroux,P., Kaas,Q., Brochet,X., Lane,J., Ginestoux,C., Lefranc,M.-P. and Giudicelli,V. (2008) IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. *Biochimie*, **90**, 570–583.
13. Noy,N.F., Crubezy,M., Fergerson,R.W., Knublauch,H., Tu,S.W., Vendetti,J. and Musen,M.A. (2003) Protege-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu. Symp. Proc.*, **2003**, 953.
14. Eilbeck,K and Lewis,S.E (2004) Sequence Ontology Annotation Guide. *Com. Funct. Genomics*, **5**, 642–647.
15. 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.

16. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*, **33**, D256–D261.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
18. Cochrane,G., Akhtar,R., Aldebert,P., Althorpe,N., Baldwin,A., Bates,K., Bhattacharyya,S., Bonfield,J., Bower,L., Browne,P. *et al.* (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **36**, D5–D12.
19. Wilming,L.G., Gilbert,J.G.R., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
20. Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Blake,J.A. and the Mouse Genome Database Group (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–D728.
21. Ruiz,M. and Lefranc,M.-P. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, **53**, 857–883.
22. Kaas,Q., Ehrenmann,F. and Lefranc,M.-P. (2007) IG, TR, MHC, IgSf and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief. *Funct. Genomic Proteomic*, **6**, 253–264.
23. Lefranc,M.-P., Giudicelli,V., Regnier,L. and Duroux,P. (2008) IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief. Bioinform.*, **9**, 263–275.
24. Giudicelli,V., Duroux,P., Ginestoux,C., Folch,G., Jabado-Michaloud,J., Chaume,D. and Lefranc,M.-P. (2006) IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.*, **34**, D781–D784.
25. Folch,G., Bertrand,J., Lemaitre,M. and Lefranc,M.-P. (2004) IMGT/PRIMER-DB. The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res.*, **32**, 3–22.
26. 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.
27. Kaas,Q., Ruiz,M. and Lefranc,M.-P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res.*, **32**, D208–D210.
28. Baum,T.P., Hierle,V., Pasqual,N., Bellahcene,F., Chaume,D., Lefranc,M.-P., Jouvin-Marche,E., Marche,P.N. and Demongeot,J. (2006) IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. *BMC Bioinformatics*, **7**, 224.
29. Brochet,X., Lefranc,M.-P. and Giudicelli,V. (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*, **36**, W503–W508.
30. Yousfi Monod,M., Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics*, **20**, i379–i385.
31. Elemento,O. and Lefranc,M.-P. (2003) IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. *Dev. Comp. Immunol.*, **27**, 763–779.
32. Pommié,C., Sabatier,S., Lefranc,G. and Lefranc,M.-P. (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.*, **17**, 17–32.
33. Lefranc,M.-P., Clément,O., Kaas,Q., Duprat,E., Chastellan,P., Coelho,I., Combres,K., Ginestoux,C., Giudicelli,V., Chaume,D. *et al.* (2004) IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol.*, **5**, 45–60.
34. Lefranc,M.-P. (2008) WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report Immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev. Comp. Immunol.*, **32**, 461–463.
35. Lefranc,M.-P. (2007) WHO-IUIS Nomenclature Subcommittee for Immunoglobulins and T cell receptors report. *Immunogenetics*, **59**, 899–902.
36. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.
37. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
38. Sugawara,H., Ogasawara,O., Okubo,K., Gojobori,T. and Tateno,Y. (2008) DDBJ with new system and face. *Nucleic Acids Res.*, **36**, D22–D24.
39. Giudicelli,V., Chaume,D., Jabado-Michaloud,J. and Lefranc,M.-P. (2005) Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. *Stud. Health Technol. Inform.*, **116**, 3–8.
40. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* (2008) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.

PUBLICATION 3

Available online at www.sciencedirect.com
ScienceDirect
Biochimie 90 (2008) 570–583
BIOCHIMIE
www.elsevier.com/locate/biochi

Research paper

IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm

Patrice Duroux [a], Quentin Kaas [a], Xavier Brochet [a], Jérôme Lane [a], Chantal Ginestoux [a], Marie-Paule Lefranc [a,b,*], Véronique Giudicelli [a]

[a] *IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France*
[b] *Institut Universitaire de France, 103 Bd Saint-Michel, 75005 Paris, France*

Received 8 June 2007; accepted 4 September 2007
Available online 11 September 2007

Abstract

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

Keywords: IMGT; Ontology; System biology; Immunogenetics; Immunoinformatics

1. Introduction

Immunogenetics, the science that studies the genetics of the immune responses, has shown a considerable expansion in biomedical fields since the last decades. It has highlighted the complex mechanisms by which B cells and T cells are at the origin of the extreme diversity of antigen receptors that comprise the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (10^{12} different immunoglobulins and 10^{12} different T cell receptors per individual, in humans) [1,2]. These mechanisms include in particular DNA rearrangements [3] and, for the IG, somatic hypermutations [1,2]. In addition, there is a considerable polymorphism of the major histocompatibility complex (MHC), human leucocyte antigens (HLA) in humans. These particularities of the adaptive immune system of the vertebrates, and a better knowledge of the innate immune response found in any species, allow the immune system to be an excellent model for system biology. The huge amount of immunological experimental data

* Corresponding author. IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2, 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.
E-mail address: marie-paule.lefranc@igh.cnrs.fr (M.-P. Lefranc).

0300-9084/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2007.09.003

continues to grow exponentially and necessitates to be managed and analysed computationally. This is the goal of immunoinformatics, a new emerging science that implements the bioinformatics methodologies to answer these needs. At the same time, standardized representation of genomics, genetics, proteomics and three-dimensional (3D) structures data is required to organize immunogenetics knowledge towards system biology and for the modelling and a better understanding of the immune system.

IMGT®, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [4], is the international reference in immunogenetics and immunoinformatics. Created in 1989 at the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) by Marie-Paule Lefranc (Université Montpellier 2 and CNRS) in Montpellier, France, IMGT provides a high-quality integrated knowledge resource, specialized in the IG, TR, MHC of human and other vertebrates, and related proteins of the immune system (RPI), which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF) of any species. The IMGT information system consists of databases (three of sequences, one of genes and one of 3D structures) and interactive tools for sequence, genome and 3D structure analysis, which interact together according to genomic, genetic and structural approaches [5] (Fig. 1).

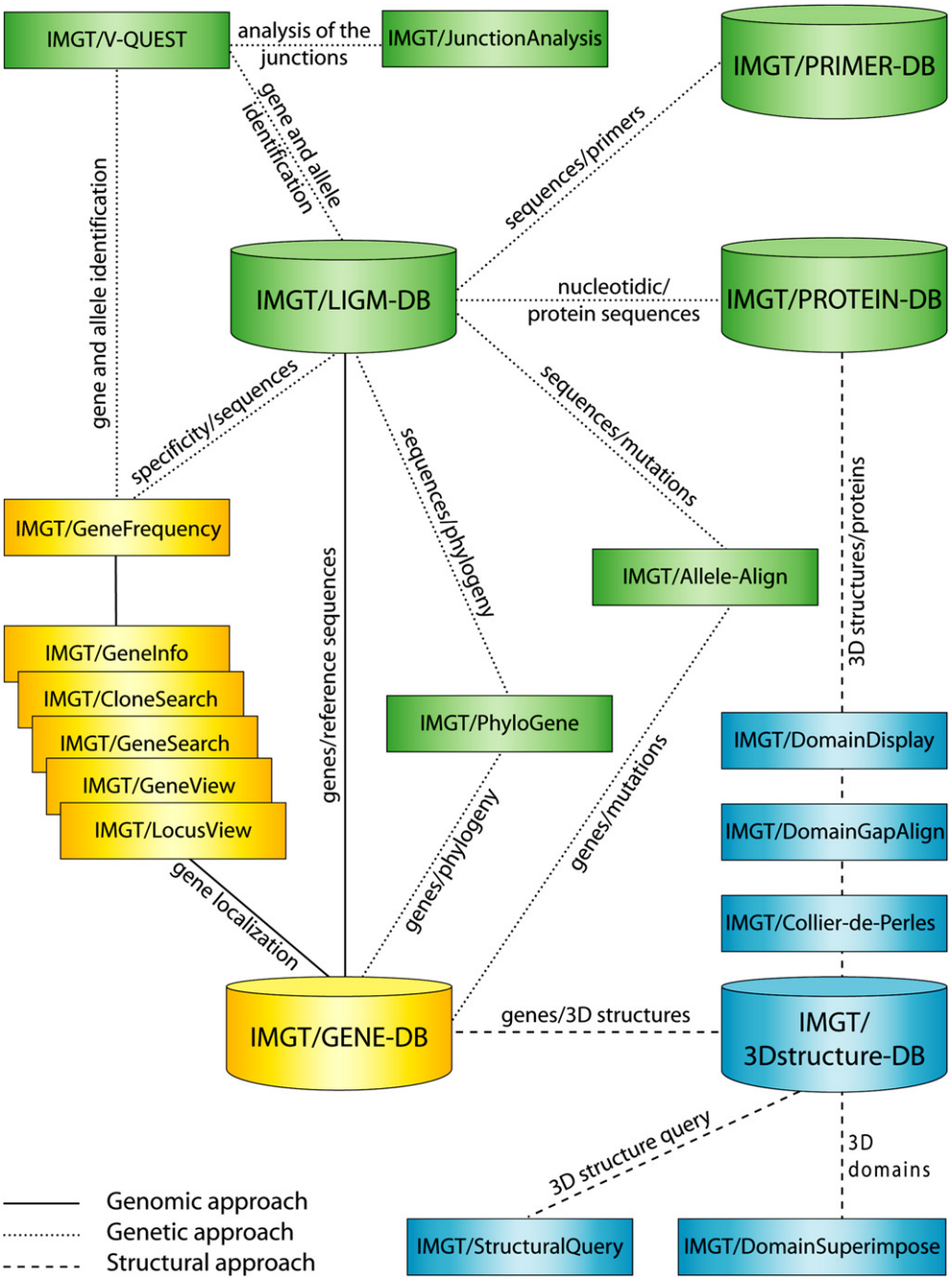


Fig. 1. IMGT® databases and tools with their interactions according to the genomic, genetic and/or structural approaches.

Moreover, IMGT provides Web resources comprising more than 10,000 HTML pages of synthesis (IMGT Repertoire), knowledge (IMGT Scientific chart, IMGT Education, IMGT Index) and external links (IMGT Bloc-notes and IMGT other accesses) [4].

The accuracy and the consistency of the IMGT data, as well as the coherence between the different IMGT components (databases, tools and Web resources), are based on IMGT-ONTOLOGY, the first ontology for immunogenetics and immunoinformatics [6]. 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 3D structures [7–9]. IMGT-ONTOLOGY results from a deep expertise in the domain and an extensive effort of conceptualization. The first standardization step was the identification of the IG and TR nucleotide sequences and the second step their description which led to the creation of IMGT/LIGM-DB [10], the first on-line IMGT database. The resulting controlled vocabulary comprises a thesaurus of keywords for the sequence identification and a set of labels for the description of the constitutive motifs. The third standardization step was the classification of the IG and TR genes which gave rise to the IMGT gene nomenclature for IG and TR of human and other vertebrates [1,2], approved by the Human Genome Organisation (HUGO) Nomenclature Committee HGNC in 1999 [11] and currently used in the generalist genome databases. The fourth standardization step was the setting up of the principles for the unique numbering of antigen receptor sequences and structures [12–16].

The standardization rules, defined in the IMGT Scientific chart [4], are based on the concepts of identification, description, classification and numerotation which characterize IMGT-ONTOLOGY [6] and which, interestingly, were defined before the term ''ontology'' became commonly used in biology and bioinformatics. IMGT-ONTOLOGY manages the immunogenetics knowledge through diverse facets that rely on the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope. Four of these axioms, ''IDENTIFICATION'', ''DESCRIPTION'', ''CLASSIFICATION'' and ''NUMEROTATION'' are presented in this paper, with the concepts that have been essential for the conceptualization of the molecular immunogenetics knowledge. As the same axioms can be used to generate concepts for multi-scale level approaches, the Formal IMGT-ONTOLOGY represents a paradigm for system biology ontologies, which need to identify, to describe, to classify and to numerotate objects, processes and relations at the molecule, cell, tissue, organ, organism or population levels.

2. Methods

2.1. Terminology

An ontology is a formal representation of a knowledge domain [6,17–19]. 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. 2). The axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope.

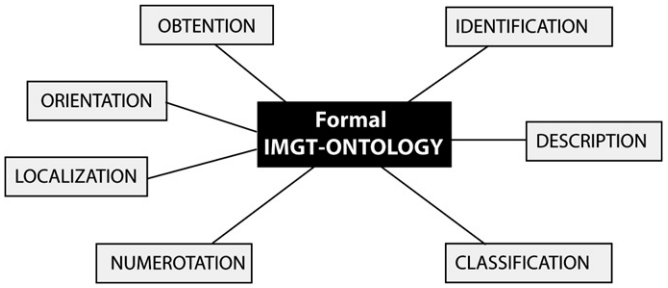


Fig. 2. The axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope.

Each axiom gives rise to a set of concepts. Concepts are general in the reality [6,20–23]. Concept instances correspond to all possible examples of representation of a concept at a given granularity. A concept may be exemplified by one or several concept instances. New concept instances may be defined with the advancement of science. Concepts are linked by relations, the simplest being ''is_a'' which represents the edge between concepts at different levels of granularity and organizes the main hierarchy of IMGT-ONTOLOGY. Properties are semantic characteristics of a concept or of a concept instance: they may be simple attributes as a name alias, or they may be specific relations between concepts and instances across the main hierarchy. These relations are fundamental since they reveal strong semantic constraints and dependencies on which relies the coherence within or between IMGT components.

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

The immunoglobulin synthesis, an example of knowledge at the molecular level, will be used to define the concepts generated by four of the axioms of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope. The concepts of identification (IDENTIFICATION axiom) identify the nucleotide and protein sequences and the 3D structures according to a structured terminology, the concepts of description (DESCRIPTION axiom) describe the composition of the sequences and structures with standardized labels, the concepts of classification (CLASSIFICATION axiom) classify the genes and alleles with a standardized nomenclature, and the concepts of numerotation (NUMEROTATION axiom) numerotate the nucleotide and amino acid numbering within sequences and structures.

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

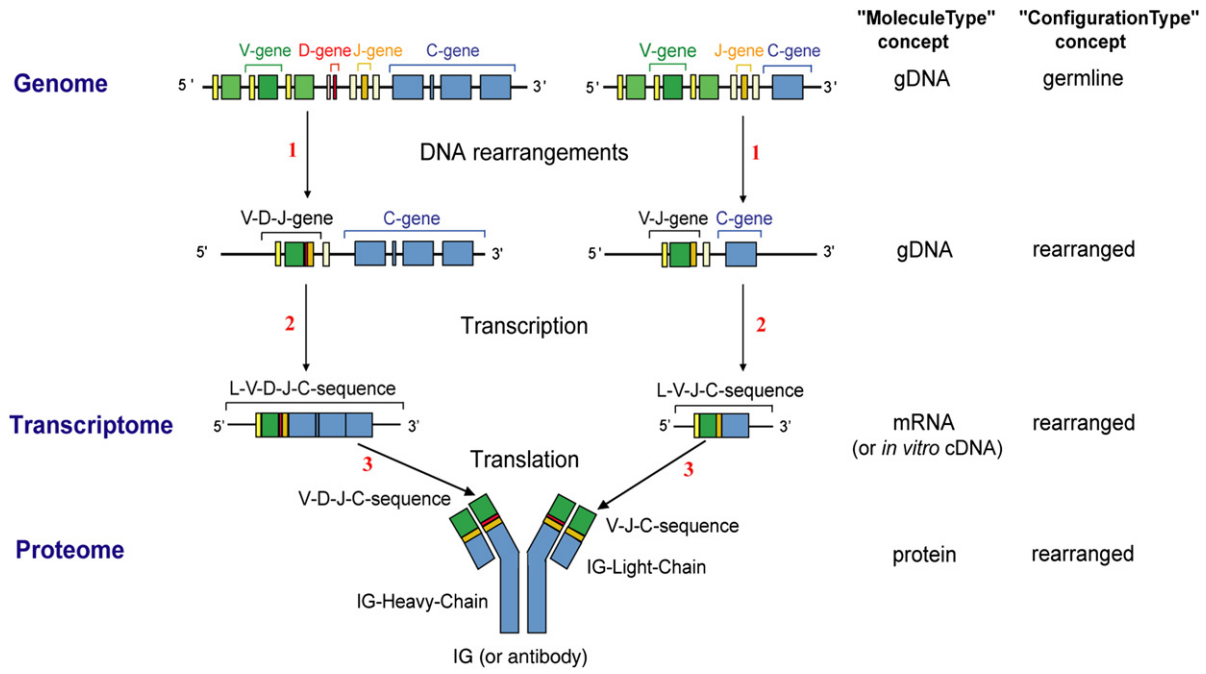


Fig. 3. An example of knowledge at the molecular level: the synthesis of an IG or antibody in humans. A human being may potentially synthesize 10^{12} different antibodies [1]. 1: DNA rearrangements (is_rearranged_into), 2: Transcription (is_transcribed_into), 3: Translation (is_translated_into). The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

requires rearrangements of the IGH, IGK and IGL genes during the differentiation of the B lymphocytes.

In the human genome (genomic DNA or gDNA), four types of genes code the IG (and TR): variable (V), diversity (D), joining (J) and constant (C). The configuration of the V-gene, D-gene and J-gene is identified as ''germline'' (Fig. 3), the configuration of the C-gene is ''undefined''. During the differentiation of the B lymphocytes in the bone marrow, the genomic DNA is rearranged first in the IGH locus, and then in the IGK and IGL loci. The rearrangements in the IGH locus lead to the junction of a D-gene and a J-gene to form a D-J-gene, and then to the junction of a V-gene to the D-J-gene to form a V-D-J-gene. The rearrangements in the IGK or IGL loci lead to the junction of a V-gene and a J-gene to form a V-J-gene. The configuration of these genes is identified as ''rearranged''. After transcription and maturation of the pre-messenger by splicing, the messenger RNA (mRNA) L-V-D-J-C-sequence and L-V-J-C-sequence (L for leader) are obtained and then translated into the heavy chain (IG-Heavy-Chain) and the light chain (IG-Light-Chain) of an IG (or antibody) (Fig. 3).

The variable domains VH and VL are coded by the V-D-J-REGION and the V-J-REGION (Fig. 4). Each domain includes four framework regions (FR) (in pale grey in Fig. 4) and three hypervariable loops or complementarity determining regions (CDR). The CDR, and more particularly the CDR3 that result from the junction of the V-D-J genes (in the VH domain) and V-J genes (in the VL domain), are involved in the antigen recognition. The VH and VL amino acids in contact with the antigen constitute the paratope. The part of the antigen recognized by the antibody is the epitope. The number of potential V-D-J and V-J rearrangements depends on the number of functional V, D and J genes in the genome. Additional mechanisms (N diversity at the V-D-J and V-J junctions and somatic hypermutations) allow to reach 10^{12} different antibodies per individual [1] (IMGT®, http://imgt.cines.fr).

2.3. Implementation

The main hierarchy of the IMGT-ONTOLOGY concepts has previously been described [6]. IMGT-ONTOLOGY

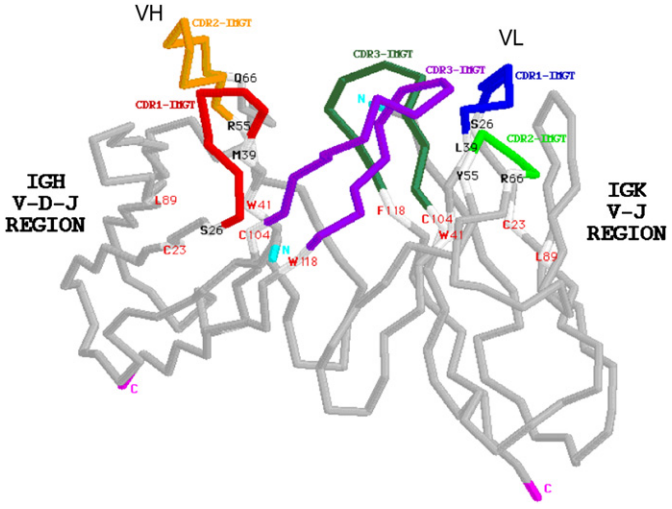


Fig. 4. The variable domains VH and VL of the heavy and light chains of an IG or antibody. VH CDR1-IMGT is in red, CDR2-IMGT in orange and CDR3-IMGT in purple. VL CDR1-IMGT is in blue, CDR2-IMGT in green and CDR3-IMGT in dark green.

concepts are available for the biologists and IMGT users in natural language in the IMGT Scientific chart [4], and have been formalized for programming purpose in IMGT-ML [24,25] which is an XML Schema (http://www.w3.org/TR/xmlschema-0/). In order to formalize the semantic relations between concepts and instances that are essential for high-quality data processing and coherence control, IMGT-ONTOLOGY is currently designed with Protégé [26] and OBO-Edit (http://oboedit.org/), that are frequently used ontology editors for biological ontologies. Protégé and OBO-Edit ontologies can be exported into RDF (http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/) and OWL (http://www.w3.org/2004/OWL/) which allow interoperability with other ontologies.

3. Results

3.1. The necessity of identification: the IDENTIFICATION axiom

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

3.1.1. Identification of an organism: the ''Taxon'' concept

The ''Taxon'' concept allows to identify the type of taxon in which an object, process or relation is found. The ''Taxon'' concept manages a hierarchy of concepts at various levels of granularity. The corresponding hierarchical taxonomy is that provided by the National Center for Biotechnology Information NCBI (http://www.ncbi.nlm.nih.gov/) up to the rank of species (''Species'' concept) and subspecies (''Subspecies'' concept) in order to establish complete interoperability with generalist databases. Since IG, TR and MHC genes are only present in jawed vertebrates (gnathostoma), only vertebrate species were originally represented in IMGT-ONTOLOGY. However, with the extension of IMGT-ONTOLOGY to the IgSF and MhcSF, invertebrate species are incorporated whenever necessary. The ''EthnicGroup'', ''Breed'' and ''Strain'' concepts have been added to IMGT-ONTOLOGY to allow the identification of data specific to ethnic groups for humans (http://www.ebi.ac.uk/imgt/hla/help/ethnic_help.html), breeds for domestic animals or strains for laboratory [27] and wild animals.

3.1.2. Identification of an entity: the ''EntityType'' concept

The ''EntityType'' concept identifies the type of entity. An entity can be a molecule, a cell, a tissue, an organ, an organism or a population. If the object is a molecule, the ''EntityType'' concept is designated as ''Molecule_EntityType'', which is defined by the ''MoleculeType'', ''GeneType'' and ''ConfigurationType'' concepts of identification and has properties identified in the ''Functionality'' and ''StructureType'' concepts (Fig. 5).

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

3.1.2.2. The ''GeneType'' concept. The ''GeneType'' concept identifies the type of gene and comprises five instances (Fig. 5). The first instance, 'conventional', refers to any (coding or not coding) gene other than IG or TR genes. The other four instances are specific to immunogenetics: 'variable' (V), 'diversity' (D) and 'joining' (J) gene types that rearrange at the DNA level and code the variable domains of IG and TR, and 'constant' (C) gene type that codes the constant region of IG and TR [1,2].

3.1.2.3. The ''ConfigurationType'' concept. The ''ConfigurationType'' concept identifies the type of gene configuration and comprises three instances (Fig. 5). The instance 'undefined' identifies the configuration of the conventional and of the constant (C) genes. The instances 'germline' and 'rearranged' identify the status of the V, D and J genes, before and after DNA rearrangements, respectively [1,2].

3.1.2.4. The ''Molecule_EntityType'' concept. The ''Molecule_EntityType'' concept, defined by the ''MoleculeType'', ''GeneType'' and ''ConfigurationType'' concepts, includes 19 instances. Three instances, 'gene', 'nt-sequence' and 'AA-sequence', respectively identify the gDNA, mRNA and protein (''MoleculeType'') of a conventional gene (''GeneType'') in undefined configuration (''ConfigurationType''). The nt-sequence instance is also valid for cDNA. Sixteen instances allow to identify the IG and TR. Ten of them are represented in Fig. 3: six for the gDNA ('V-gene', 'D-gene', 'J-gene', 'C-gene', 'V-D-J-gene' and 'V-J-gene'), two for the mRNA, 'L-V-D-J-C-sequence' and 'L-V-J-C-sequence', also valid for cDNA, and two for the protein, 'V-D-J-C-sequence' and 'V-J-C-sequence'. For example, the instance 'V-gene' identifies a gDNA (''MoleculeType'') containing a gene V (''GeneType''), in germline configuration (''ConfigurationType''). The instance 'L-V-J-C-sequence' identifies a sequence of mRNA or cDNA (''MoleculeType'') corresponding to V, J and C genes (''GeneType''), in rearranged configuration

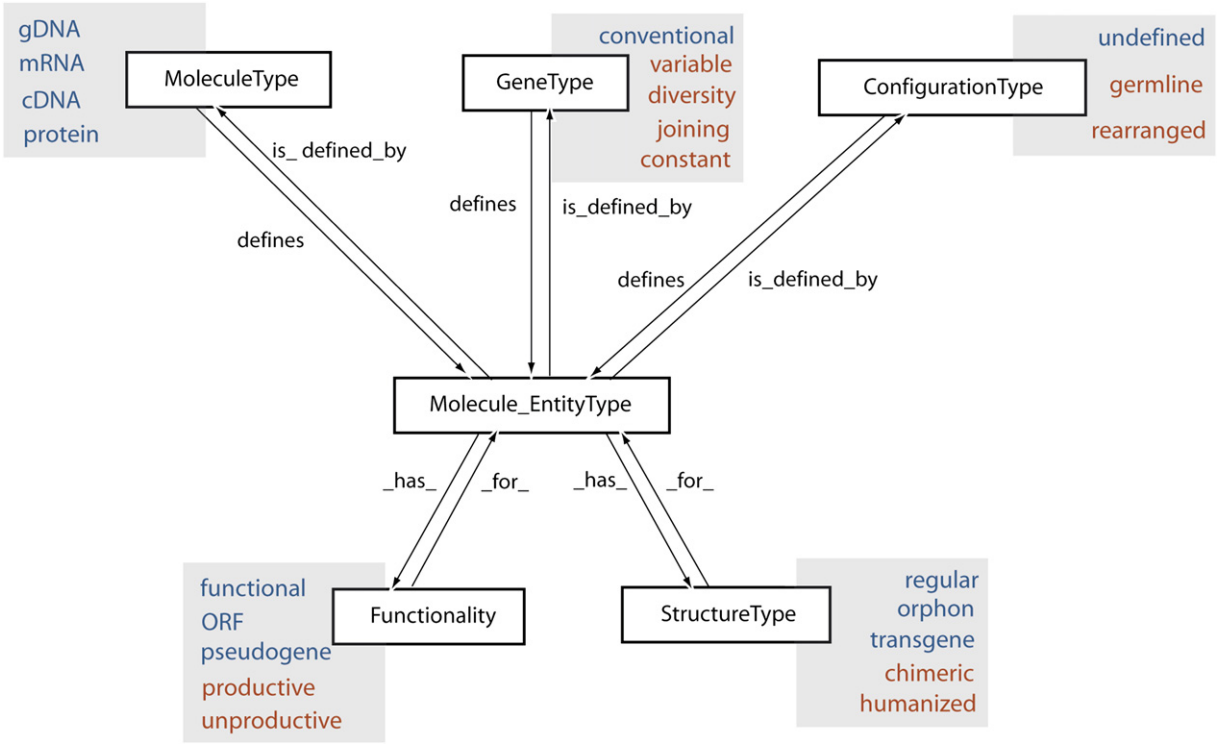


Fig. 5. The ‘‘Molecule_EntityType’’ concept. The ‘‘Molecule_EntityType’’ concept is defined by the ‘‘MoleculeType’’, ‘‘GeneType’’ and ‘‘ConfigurationType’’ concepts of identification and has properties identified in the ‘‘Functionality’’ and ‘‘StructureType’’ concepts (IDENTIFICATION axiom). Arrows indicate reciprocal relations ‘‘is_defined_by’’ and ‘‘defines’’, ‘‘_has_’’ and ‘‘_for_’’. Concept instances which are general are in blue, those which are specific of the IG and TR are in red. The ‘‘Molecule_EntityType’’ concept has 19 instances (listed in Section 3.1.2.4). Only a few examples of the ‘‘StructureType’’ concept instances are shown.

(‘‘ConfigurationType’’) (Fig. 3). The last six instances correspond to partial rearrangement (‘D-J-gene’) or to sterile transcripts (‘L-V-sequence’, ‘D-sequence’, ‘J-sequence’, ‘J-C-sequence’ and ‘C-sequence’).

3.1.2.5. The ‘‘Functionality’’ concept. The ‘‘Functionality’’ concept identifies the type of functionality for the ‘‘Molecule_EntityType’’ concept (Fig. 5). It includes five instances, divided into two categories, according to the configuration type. Three instances, ‘functional’, ‘ORF’ (open reading frame) and ‘pseudogene’ identify the functionality of a ‘‘Molecule_EntityType’’ instance in undefined or germline configuration. They allow to identify the functionality of conventional genes, that of C genes, and that of V, D and J genes before their rearrangement in the genome, and by extension the functionality of their transcripts and proteins. The two instances ‘productive’ and ‘unproductive’ identify the functionality of ‘‘Molecule_EntityType’’ instances in rearranged configuration. They allow to identify the functionality of IG and TR entities after their rearrangement in the genome, that of fusion genes resulting from translocations, and that of hybrid genes obtained by biotechnology molecular engineering, and by extension the functionality of their transcripts and proteins.

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

3.1.3. Identification of a receptor: the ‘‘ReceptorType’’ concept

The ‘‘ReceptorType’’ concept identifies the type of receptor. A receptor can be a molecule, a cell, a tissue, an organ, an organism or a population. If the object is a molecule, the ‘‘ReceptorType’’ concept is designated as ‘‘Molecule_ReceptorType’’ which is defined by the ‘‘ChainType’’ concept of identification and has properties identified in the ‘‘StructureType’’, ‘‘Specificity’’ and ‘‘Function’’ concepts (Fig. 6). The ‘‘ChainType’’ concept is itself defined by the ‘‘Molecule_EntityType’’ and the ‘‘DomainType’’ concepts of identification and by concepts of classification (see CLASSIFICATION axiom). These latter are organized in a hierarchy which confers different levels of granularity to the ‘‘Molecule_ReceptorType’’ and ‘‘ChainType’’ concepts.

3.1.3.1. The ‘‘Molecule_ReceptorType’’ concept. The ‘‘Molecule_ReceptorType’’ concept identifies the type of protein receptor, defined by its chain composition. Thus, IG is an instance of the ‘‘Molecule_ReceptorType’’ concept, defined as comprising 4 chains, two heavy chains and two light chains,

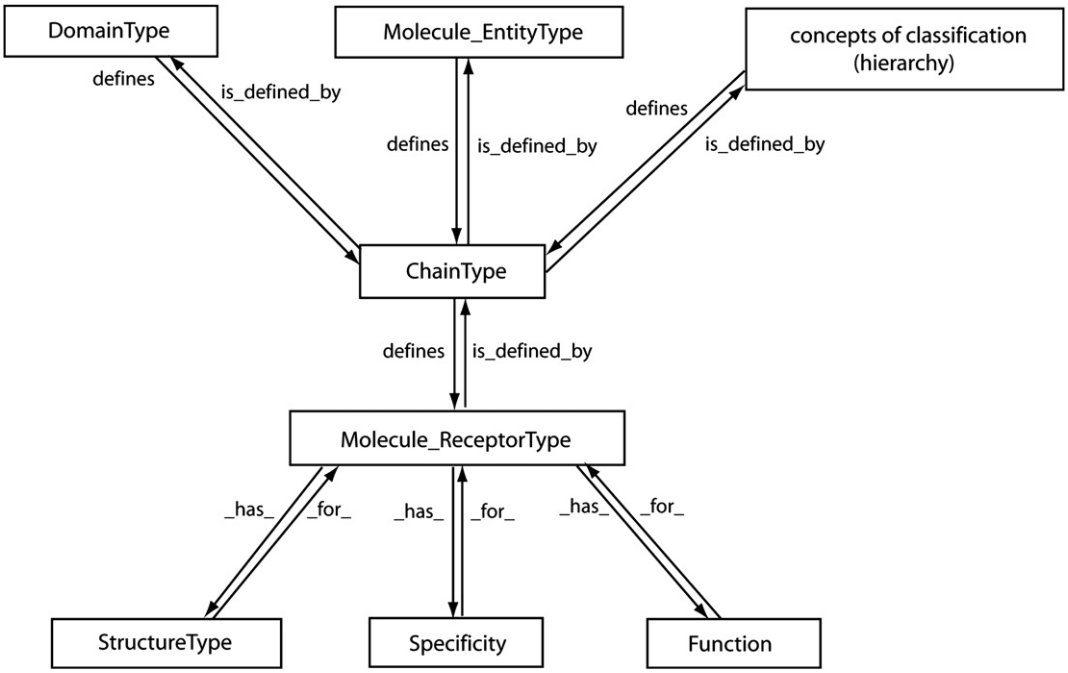


Fig. 6. The ''Molecule_ReceptorType'' concept. The ''Molecule_ReceptorType'' concept, defined by the ''ChainType'' concept of identification, has properties identified in the ''StructureType'', ''Specificity'' and ''Function'' concepts (IDENTIFICATION axiom). The ''ChainType'' concept is itself defined by the ''Molecule_EntityType'' and ''DomainType'' concepts and by concepts of classification (hierarchy). Arrows indicate reciprocal relations ''is_defined_by'' and ''defines'', ''_has_'' and ''_for_''. These concepts have different levels of granularity, up to six for ''Molecule_ReceptorType'' and ''ChainType''.

identical two by two and covalently linked (Fig. 7). A receptor may comprise one chain (monomer) or several associated chains (multimer).

3.1.3.2. The ''ChainType'' concept. The ''ChainType'' concept identifies the type of chain (Fig. 6). It is one of the most important concepts of identification for the standardization of genome, transcriptome and proteome data in system biology. Indeed, being able to identify a type of chain means that it is possible to identify the transcript and the encoding gene(s). The ''ChainType'' concept contains a hierarchy of concepts which identify the chain type at different levels of granularity. The finest level of granularity, the ''GeneLevelChainType'' concept, identifies the type of chain by reference to the gene(s) which code(s) the chain. It represents the main concept for a very precise identification because it establishes a relationship with the ''Gene'' concept which belongs to the concepts of classification (reciprocal relations ''is_coded_by'' and ''codes''). The number of instances of the ''GeneLevelChainType'' concept depends on the number of functional genes and ORF per haploid genome in a given species (in the case of the IG and TR, it is the number of functional and ORF constant genes which is taken into account). If only the functional genes are considered, the instances of this concept correspond to the isotypes.

3.1.3.3. The ''DomainType'' concept. A chain can be defined by its constitutive structural units (''DomainType'' concept) (Fig. 6). A domain is a chain subunit characterized by its three-dimensional (3D) structure, and by extension its amino acid sequence and the nucleotide sequence which encodes it. This concept may theoretically comprise many instances, but so far only the instances which have been carefully characterized by LIGM have been entered in IMGT-ONTOLOGY. The ''DomainType'' concept has currently three instances, V type domain (variable domains of the IG and TR and V-like domains of other IgSF proteins), C type domain (constant domains of the IG and TR and C-like domains of other IgSF proteins) and G type domain (groove domains of the MHC and G-like domains of other MhcSF proteins) [14–16].

3.1.3.4. The ''Specificity'' and ''Function'' concepts. The ''Specificity'' concept identifies the specificity of the ''Molecule_ReceptorType'' (Fig. 6), and by extension the specificity of the chains and domains and of the corresponding transcripts. Instances of the ''Specificity'' concept identify the antigen recognized by an antigen receptor (IG or TR). The ''Specificity'' concept is particularly important because of the unlimited number of antigens and of the complexity of the antigen/antigen receptor interactions. The conceptualization of knowledge associated with this concept is in the course of modelling. The instances of the ''Specificity'' concept (several hundreds at the present time) will be connected on the one hand, with the ''Epitope'' concept which identifies the part of the antigen recognized by the antigen receptor and on the other hand, with the ''Paratope'' concept which identifies the part of the antigen receptor (IG or TR) which recognizes and binds to the antigen. The ''Function'' concept identifies the function of the ''Molecule_ReceptorType'' (Fig. 6), and by extension the function of the chains and domains and of the corresponding

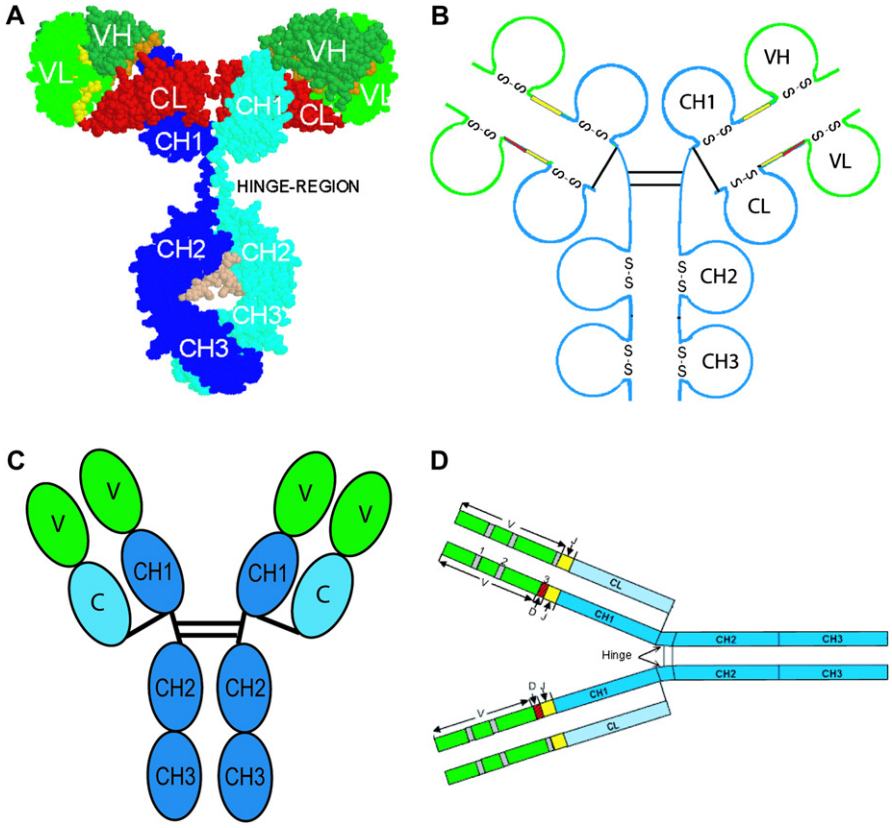


Fig. 7. Identification of an IG or antibody as an instance of the ''Molecule_ReceptorType'' concept made of four chains, two IG-Heavy-Chain and two IG-Light-Chain (''ChainType'' concept). The four representations, although different, allow to identify an IG as a receptor of four chains, themselves organized in domains (''DomainType'' concept). VH and VL are V type domains, coded by the V-D-J region and V-J region, respectively. CL, CH1, CH2 and CH3 are C type domains. (A) 3D structure, (B) organization in Ig-like domains, (C) organization in modules, (D) regions coded by the V, D, J and C gene types. The C gene type codes the constant region (CL for the IG-Light-Chain and CH1, hinge, CH2 and CH3 for the IG-Heavy-Chain). This representation, schematized as a Y shape, is frequently used to represent an IG.

transcripts. Instances of the ''Function'' concept identify the dual function of the antigen receptors [2]. Their identification and definition are still in development.

3.2. The necessity of description: the DESCRIPTION axiom

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

3.2.1. Description of an entity: the ''EntityPrototype'' concept

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

3.2.2. The ''Molecule_EntityPrototype'' concept

In molecular biology, the DESCRIPTION axiom has generated the concepts of description which provide the terms and the rules to describe motifs in the nucleotide and protein sequences and in 3D structures. These concepts gave rise to a standardized terminology and to a precise definition of the annotation rules. The ontology for sequences and 3D structures has been the focus of IMGT for many years. The instances of the concepts of description correspond to IMGT labels. More than 550 labels were defined (270 for the nucleotide sequences (http://imgt.cines.fr/cgi-bin/IMGTlect.jv?query=7) [10] and 285 for the 3D structures [28] (http://imgt.cines.fr/textes/IMGTScientificChart/SequenceDescription/IMGT3Dlabeldef.html). Interestingly, 64 IMGT labels defined for nucleotide sequences are used and cross-referenced in the recently created Sequence Ontology (SO) (http://song.sourceforge.net/) [29] to describe specific IG and TR gene organization (http://imgt.cines.fr/textes/IMGTindex/ontology.html).

The "Molecule_EntityPrototype" concept allows the description of the entity (gene, transcrit and protein) organization and of their constitutive motifs. This concept is fundamental in IMGT-ONTOLOGY because it allows the representation of the knowledge related to the complex mechanisms of IG and TR gene rearrangements (Fig. 8). The relation "is_rearranged_into" is specific to the synthesis of the IG and TR. The relations "is_transcribed_into" and "is_translated_into" are general for molecular biology. These three relations allow the organization of the various instances of the "Molecule_EntityPrototype" concept during the synthesis of the IG and the TR, and in a more general way for the expression of any protein. They allow in addition, by more specific relations, to take into account the alternative transcripts, the protein isoforms and the post-translational modifications.

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

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

3.2.3. The "Core" concept

The "Core" concept allows to describe the coding region of genes and contains five instances which are 'REGION' (for conventional gene type), 'V-REGION', 'D-REGION', 'J-REGION' and 'C-REGION' (for V, D, J and C gene types, respectively). These instances are particularly important since they can be described in all the instances of the "Molecule_EntityPrototype" concept. They allow to describe the chains of the antigen receptors in spite of the complexity of their structure and to link sequences, structures and functions. Moreover, these are the instances of the "Core" concept which allowed the definition and standardized description of the IG and TR alleles (concepts of classification), now approved at the international level [1,2].

3.3. The necessity of classification: the CLASSIFICATION axiom

The CLASSIFICATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be classified. In molecular biology, the concepts of classification generated from the CLASSIFICATION axiom allow to classify and name the genes and their alleles. The genes which code the IG and TR belong to highly polymorphic multigenic families. A major contribution of IMGT-ONTOLOGY was to set the principles of their classification and to propose a standardized nomenclature [1,2] (Fig. 10). The IMGT gene nomenclature has been approved at the international level by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC), in 1999 [11]. The IMGT IG and TR gene names are the official reference for the genome projects and, as such, have been integrated in the Genome Database (GDB), in LocusLink and in Entrez Gene at NCBI [30]. The IG and TR genes [1,2] are managed in the IMGT/GENE-DB database [31].

3.3.1. The "Group" and "Subgroup" concepts

The "Group" concept classifies a set of genes which belong to the same multigene family, within the same species

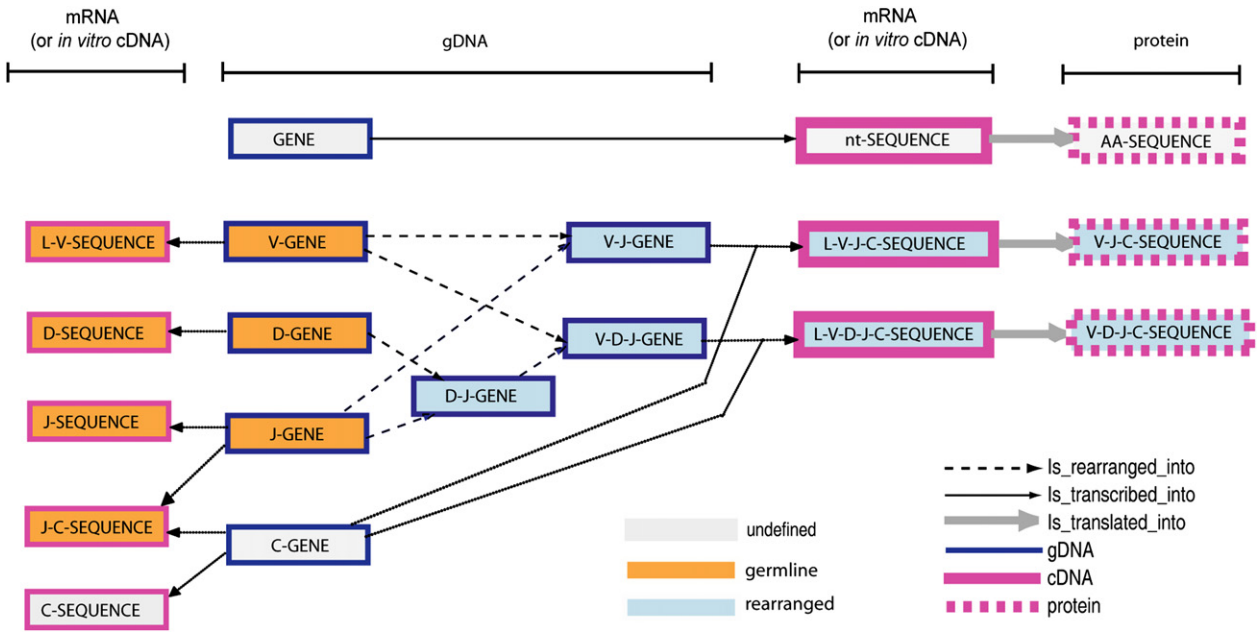


Fig. 8. Instances of the "Molecule_EntityPrototype" concept (DESCRIPTION axiom). The three instances "GENE", "nt-SEQUENCE" and "AA-SEQUENCE" correspond to conventional genes while the 16 other instances are specific of the IG and TR. The concept instances for mRNA are also valid for *in vitro* cDNA.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

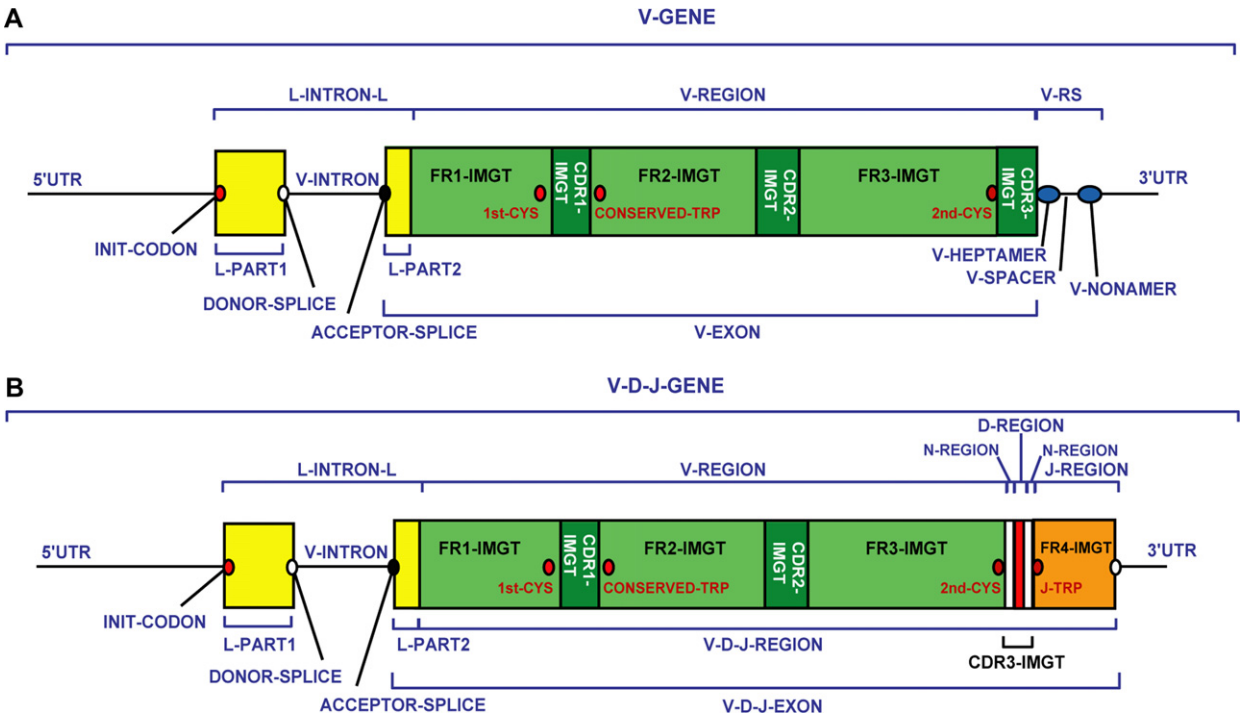


Fig. 9. Graphical representation of two instances of the "Molecule_EntityPrototype" concept (DESCRIPTION axiom). (A) V-GENE. (B) V-D-J-GENE. Twenty-five labels and ten relations are necessary and sufficient for a complete description of these instances.

or between different species. For the IG and TR, the set of genes is identified by an instance of the "GeneType" concept (V, D, J or C). The "Subgroup" concept classifies a subset of genes which belong to the same group, and which, in a given species, share at least 75% of identity at the nucleotide sequence level (and in the germline configuration for the V, D, and J genes).

3.3.2. The "Gene" and "Allele" concepts

The "Gene" concept classifies a unit of DNA sequence that can be potentially transcribed and/or translated (this definition includes the regulatory elements in 5′ and 3′, and the introns, if present). The instances of the "Gene" concept are gene names. In IMGT-ONTOLOGY, a gene name is composed of the name of the species (instance of the Taxon "Species" concept) and of the international HGNC/IMGT gene symbol, for example, *Homo sapiens* IGLV1−2. By extension, orphons and pseudogenes are also instances of the "Gene" concept. The "Allele" concept classifies a polymorphic variant of a gene. The instances of the "Allele" concept are allele names. Alleles identified by the mutations of the nucleotide sequence are classified by reference to allele *01.

Table 1
Relations for sequence description (LOCALIZATION axiom)

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

Full description of mutations and allele name designations are currently recorded for the core sequences (V-REGION, D-REGION, J-REGION, C-REGION). They are reported in Alignment tables, in IMGT Repertoire http://imgt.cines.fr and in IMGT/GENE-DB [16].

3.4. The necessity of numbering: the NUMEROTATION axiom

The NUMEROTATION axiom of the Formal IMGT-ONTOLOGY or IMGT-Kaleidoscope postulates that molecules, cells, tissues, organs, organisms or populations, their processes and their relations, have to be numerotated. So far, these concepts have essentially been defined at the molecular level. The NUMEROTATION axiom and the concepts of numerotation determine the principles of a unique numbering for a domain (sequences and 3D structures) [14−16] (Fig. 11). The "IMGT_unique_numbering" concept has three concept instances: "IMGT_unique_numbering_for_V_Type_domain", "IMGT_unique_numbering_for_C_type_domain", "IMGT_unique_numbering_for_G_type_domain" [14−16].

The "IMGT_unique_numbering" concept determines the "FR-IMGT_length", "CDR-IMGT_length", "Strand_length", and "Helix_length" concepts [14−16]. The "IMGT_unique_numbering" concept is illustrated by the "IMGT_Collier_de_Perles" concept which allows graphical representation in two dimensions (2D) of the amino acid sequences of V, C or G type domains [32,33] and comprises three concept instances (Fig. 12). This concept is largely recognized at the international level and the expression "IMGT Collier de Perles" is now used in scientific publications.

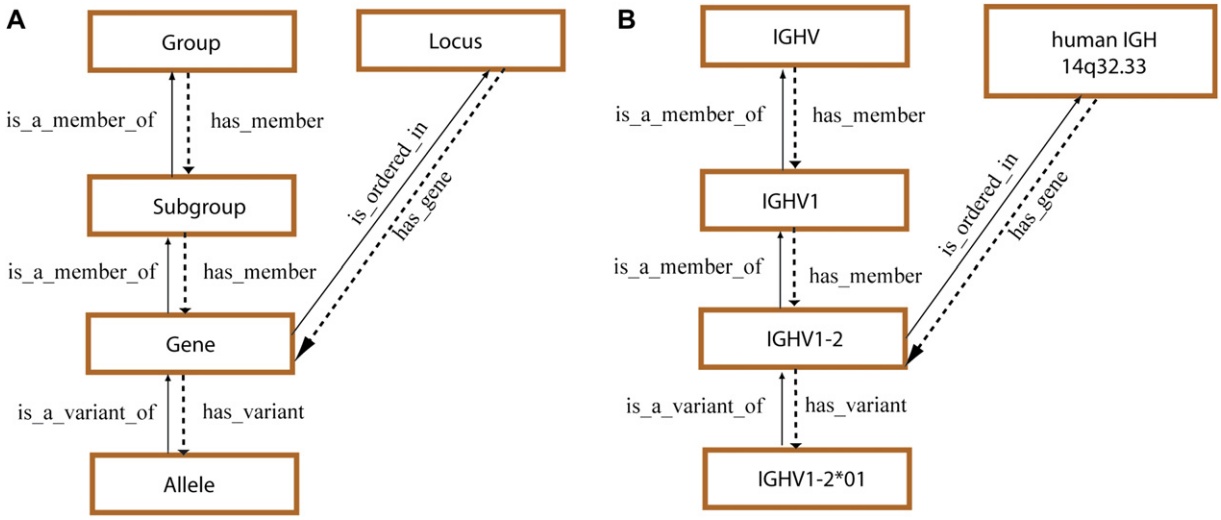


Fig. 10. Concepts of classification for gene and allele nomenclature (CLASSIFICATION axiom). (A) Hierarchy of the concepts of classification and their relations. (B) Examples of concept instances for each concept of classification. The concepts instances are associated to an instance of the "Taxon" concept, and more precisely for the "Gene" and "Allele" concepts to an instance of the "Species" concept (here, *Homo sapiens*). The "Locus" concept is a concept of localization (LOCALIZATION axiom).

The "IMGT_Collier_de_Perles" concept is particularly used in antibody engineering for the humanization of murine antibodies in which it is necessary to precisely delimit the murine CDR-IMGT to be grafted, in order to preserve the antibody specificity. The concepts of numerotation are also at the origin of the standardization of the allele description and, more generally of the mutation description (IMGT Scientific chart, http://imgt.cines.fr).

4. Conclusion

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

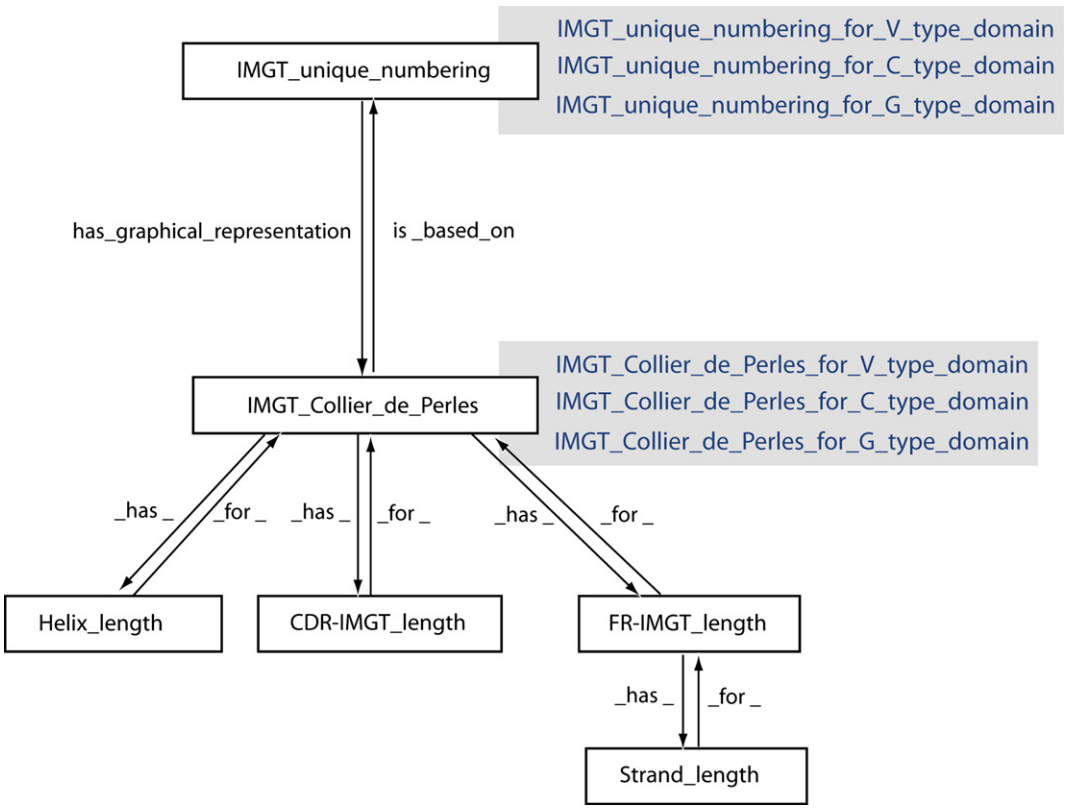


Fig. 11. The "IMGT_unique_numbering" and "IMGT_Collier_de_Perles" concepts and relations with other concepts of numerotation (NUMEROTATION Axiom). Concept instances are written in blue. Arrows indicate reciprocal relations "has_graphical_representation" and "is_based_on", "_has_" and "_for_".

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, CLASSIFICATION and NUMEROTATION axioms 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

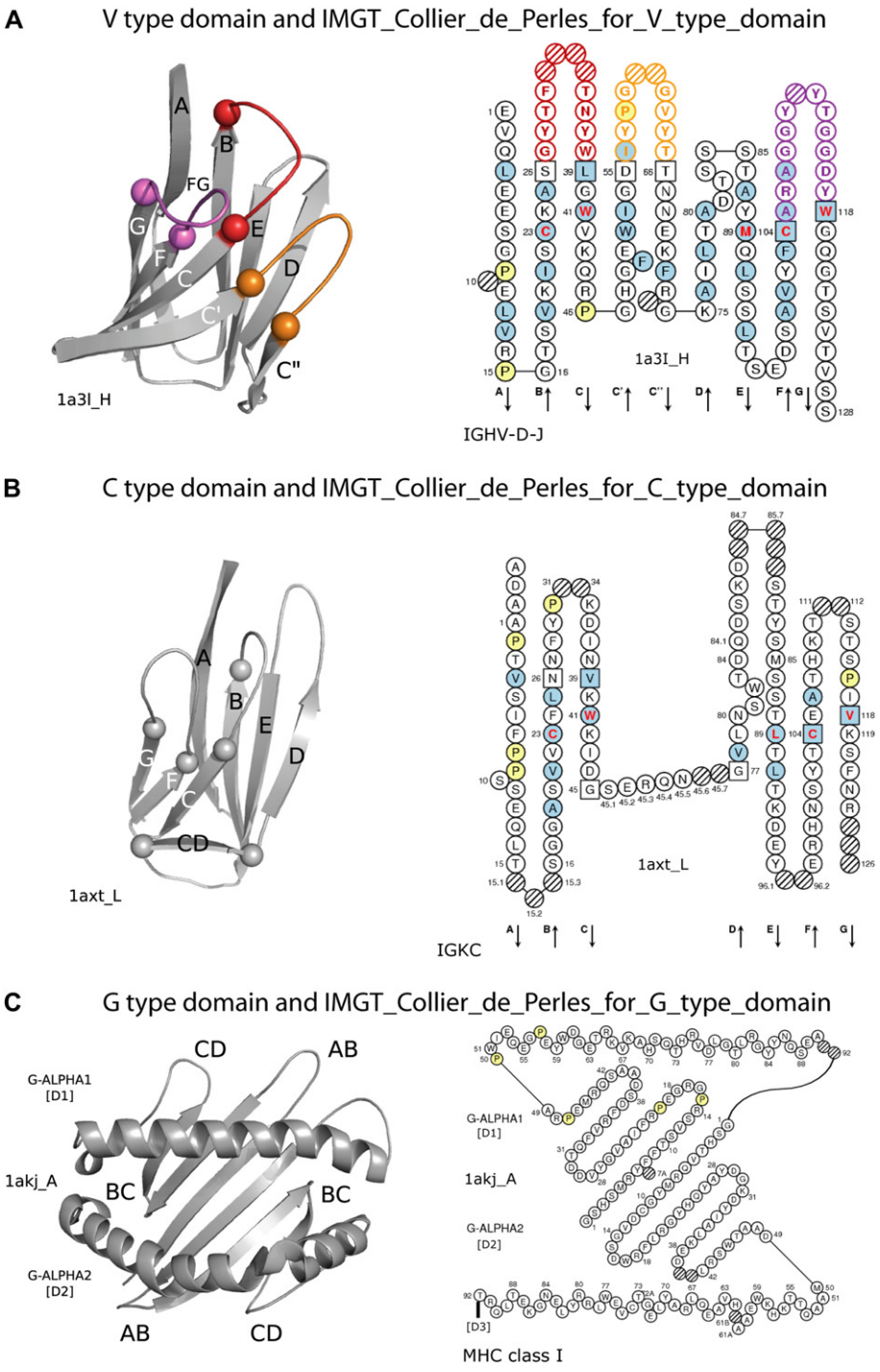


Fig. 12. "DomainType" and "IMGT_Collier_de_Perles" concept instances.

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 IMGT and 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) [34], and in immunology, such as the Immunome Epitope database and Analysis Resource (IEDB) [35] 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, leukaemias, lymphomas, myelomas), (ii) veterinary research (IG and TR repertoires in farm and wild life species), (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) biotechnology related to antibody engineering (scFv, phage displays, combinatorial libraries, chimaeric, 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.

Acknowledgements

We are grateful to Gérard Lefranc for helpful discussion and to the IMGT® team for its constant motivation and its expertise. IMGT® was funded by the Centre National de la Recherche Scientifique (CNRS), the BIOMED1 (BIOCT 930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of LivingResources (QLG2-2000-01287) programmes of the European Union. IMGT received subventions from Association pour la Recherche sur le Cancer (ARC), Génopole-Montpellier-Languedoc-Roussillon and the Réseau National des Génopoles (RNG). IMGT has been a National Bioinformatics RIO Platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT is currently supported by the CNRS, the Ministère de l'Education Nationale de l'Enseignement Supérieur et de la Recherche (MENESR), Université Montpellier 2 Plan Pluri-Formation, Région Languedoc-Roussillon, BIOSTIC-LR2004, ACI-IMPBIO (IMP82-2004), GIS-AGENAE, Agence Nationale de la Recherche ANR (BIOSYS06_135457) and the European Union ImmunoGrid project (IST-2004-0280069).

References

[1] M.-P. Lefranc, G. Lefranc, The Immunoglobulin FactsBook, Academic Press, London UK, 2001, 458 pp.
[2] M.-P. Lefranc, G. Lefranc, The T Cell Receptor FactsBook, Academic Press, London UK, 2001, 398 pp.
[3] H. Sakano, K. Huppi, G. Heinrich, S. Tonegawa, Sequences at the somatic recombination sites of immunoglobulin light-chain genes, Nature 280 (1979) 288–294.
[4] M.-P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. Clément, D. Chaume, G. Lefranc, IMGT, the international ImMunoGeneTics information system®, Nucleic Acids Res. 33 (2005) D593–D597.
[5] M.-P. Lefranc, O. Clément, Q. Kaas, E. Duprat, P. Chastellan, I. Coelho, K. Combres, C. Ginestoux, V. Giudicelli, D. Chaume, G. Lefranc, IMGT-Choreography for immunogenetics and immunoinformatics, In Silico Biol. 5 (2005) 45–60.
[6] V. Giudicelli, M.-P. Lefranc, Ontology for Immunogenetics: IMGT-ONTOLOGY, Bioinformatics 15 (1999) 1047–1054.
[7] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, N. Bosc, G. Folch, D. Guiraudou, J. Jabado-Michaloud, S. Magris, D. Scaviner, V. Thouvenin, K. Combres, D. Girod, S. Jeanjean, C. Protat, M. Yousfi Monod, E. Duprat, Q. Kaas, C. Pommié, D. Chaume, G. Lefranc, IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics http://imgt.cines.fr, In Silico Biol. 4 (2004) 17–29.
[8] V. Giudicelli, D. Chaume, J. Jabado-Michaloud, M.-P. Lefranc, Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY, Stud. Health Technol. Inform. 116 (2005) 3–8.
[9] Q. Kaas, M.-P. Lefranc, T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB, In Silico. Biol. 5 (2005) 505–528.
[10] V. Giudicelli, C. Ginestoux, G. Folch, J. Jabado-Michaloud, D. Chaume, M.-P. Lefranc, IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences, Nucleic Acids Res. 34 (2006) D781–D784.
[11] H.M. Wain, E.A. Bruford, R.C. Lovering, M.J. Lush, M.W. Wright, S. Povey, Guidelines for human gene nomenclature, Genomics 79 (2002) 464–470.
[12] M.-P. Lefranc, Unique database numbering system for immunogenetic analysis, Immunol. Today 18 (1997) 509.
[13] M.-P. Lefranc, The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains, Immunologist 7 (1999) 132–136.
[14] M.-P. Lefranc, C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, G. Lefranc, IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains, Dev. Comp. Immunol. 27 (2003) 55–77.
[15] M.-P. Lefranc, C. Pommié, Q. Kaas, E. Duprat, N. Bosc, D. Guiraudou, C. Jean, M. Ruiz, I. Da Piedade, M. Rouard, E. Foulquier, V. Thouvenin, G. Lefranc, IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains, Dev. Comp. Immunol. 29 (2005) 185–203.
[16] M.-P. Lefranc, E. Duprat, Q. Kaas, M. Tranne, A. Thiriot, G. Lefranc, IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN, Dev. Comp. Immunol. 29 (2005) 917–938.
[17] T.R. Gruber, A translation approach to portable ontologies, Knowledge Acquisit. 5 (1993) 199–220.
[18] N. Guarino, P. Giaretta, Ontologies and knowledge bases: towards a terminological clarification, in: N. Mars (Ed.), Towards Very Large Knowledge Bases, IOS Press, Amsterdam, 1995, pp. 29–45.
[19] N. Guarino, Understanding, building and using ontologies, Int. J. Human-Comput. Stud. 46 (1997) 293–310.
[20] Noy N.F., McGuinness D.L., Ontology development 101: A guide to creating your first ontology, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[21] B. Smith, Ontology, in: L. Floridi (Ed.), Blackwell Guide to the Philosophy of Computing and Information, Blackwell, Oxford, 2003, pp. 155−166.
[22] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, C. Rosse, Relations in biomedical ontologies, Genome Biol. 6 (2005) R46.
[23] L.-N. Soldatova, A. Clare, A. Sparkes, R.D. King, An ontology for a Robot Scientist, Bioinformatics 22 (2006) e464−e471.
[24] D. Chaume, V. Giudicelli, M.-P. Lefranc, IMGT-ML a XML language for IMGT-ONTOLOGY and IMGT/LIGM-DB data, in: Proceedings of NETTAB 2001, Network Tools and Applications in Biology, Genoa, Italy, May 17−18, 2001, pp. 71−75.
[25] D. Chaume, K. Combres, V. Giudicelli, M.-P. Lefranc, Retrieving factual data and documents using IMGT-ML in the IMGT information system®, in: Proceedings of NETTAB 2005, Workflows management: new abilities for the biological information overflow, Naples, Italy, Oct 5−7, 2005, pp. 47−51.
[26] N.F. Noy, M. Crubezy, R.W. Fergerson, H. Knublauch, S.W. Tu, J. Vendetti, et al., Protege-2000: an open-source ontology-development and knowledge-acquisition environment, AMIA Annu. Symp. Proc. (2003) 953.
[27] J.T. Eppig, C.J. Bult, J.A. Kadin, J.E. Richardson, J.A. Blake, the members of the Mouse Genome Database Group 2005, The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology, Nucleic Acids Res. 33 (2005) D471−D475.
[28] Q. Kaas, M. Ruiz, M.-P. Lefranc, IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data, Nucleic Acids Res. 32 (2004) D208−D210.
[29] K. Eilbeck, S.E. Lewis, Sequence Ontology annotation guide, Comp. Funct. Genomics 5 (2004) 642−647.
[30] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res. 35 (2007) D26−D31.
[31] V. Giudicelli, D. Chaume, M.-P. Lefranc, IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes, Nucleic Acids Res. 33 (2005) D256−D261.
[32] E. Duprat, Q. Kaas, V. Garelle, G. Lefranc, M.-P. Lefranc, IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily, in: S.G. Pandalai (Ed.), Recent Research Developments in Human Genetics, 2, Research Signpost, Trivandrum, India, 2004, pp. 111−136.
[33] Q. Kaas, M.-P. Lefranc, IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains, Curr. Bioinformat. 2 (2007) 21−30.
[34] The Gene Ontology Consortium, The Gene Ontology (GO) project in 2006, Nucleic Acids Res. 34 (2006) D322−D326.
[35] B. Peters, J. Sidney, P. Bourne, H.H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J.V. Ponomarenko, M. Sathiamurthy, S. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, A. Sette, The immune epitope database and analysis resource: from vision to blueprint, PLoS Biol. 3 (2005) e91.

RESUME en français

Les récepteurs d'antigènes, immunoglobulines (IG) et récepteurs T (TR), sont des composants moléculaires important de la réponse immunitaire adaptative des vertébrés. Les locus d'IG et TR sont constitués des gènes variables (V), de diversité (D), de jonction (J) et constants (C). La synthèse des chaînes IG et TR exige des réarrangements des gènes V et J, ou V, D et J, au niveau de l'ADN et l'épissage des gènes réarrangés V-J et V-D-J après leur transcription avec les gènes C. A cause des de l'originalité de ces mécanismes et des particularités structurales des gènes d'IG et TR rattachées à ces mécanismes, les logiciels bioinformatiques conventionnels ne sont pas adaptés à leur annotation dans de grandes séquences génomiques. Pour répondre au besoin des biocurateurs-annotateurs d'IMGT®, the international ImMunoGeneTics information system®, j'ai développé IMGT/LIGMotif, un outil pour l'annotation des gènes d'IG et TR. Cet outil est fondé sur les règles standardisées définies dans IMGT-ONTOLOGY, la première ontologie en immunogénétique et immunoinformatique. IMGT/LIGMotif annote actuellement les gènes V, D et J de l'homme et de la souris, dans de grandes séquences génomiques. Le processus d'annotation intègre plusieurs modules qui permettent la description des gènes en leur attribuant des labels, leur orientation dans la séquence analysée et l'identification de leur fonctionnalité. IMGT/LIGMotif analyse actuellement des séquences allant jusqu'à 2.5 mégabases ainsi que des locus par lot de fichiers de séquences. IMGT/LIGMotif est particulièrement utile pour annoter les séquences génomiques de locus d'homme ou de souris ainsi que celles d'espèces proches, respectivement, les primates non-humains et le rat.

TITRE en anglais

ANALYSE AND CONCEPTION AT IMGT® OF AN INTEGRATED BIOINFORMATIC APPROACH FOR IMMUNOGLOBULIN AND T RECEPTOR GENES IN GENOMIC LOCUS OF LARGE SIZE

RESUME en anglais

The antigen receptors, immunoglobulins (IG) and T cell receptors (TR), are important molecular components of the adaptive immune responses of vertebrates. The locus of IG and TR consist of variable (V), diversity (D), joining (J) and constant (C) genes. Chain synthesis requires IG and TR rearrangements of V and J, or V, D and J genes at the DNA level, and transcript splicing of rearranged V-J and V-D-J genes after their transcription with the C gene. Because of the characteristics of the IG and TR genes, associated with these mechanisms, conventional bioinformatic software are not adapted to their annotation in large genomic sequences. To meet the need of the biocurators working at IMGT®, the international ImMunoGeneTics information system®, I developed IMGT/LIGMotif a tool for the annotation of the IG and TR genes. This tool is based on standardized rules defined in IMGT-ONTOLOGY, the first ontology in immunogenetics and immunoinformatics. IMGT/LIGMotif currently provides annotation of V, D and J gene clusters in human and mouse large genomic sequences. The annotation process includes several modules that allow the description of genes by assigning labels, the gene orientation in the analysed sequence and the identification of their functionality. IMGT/LIGMotif performs analysis of sequences extending up to 2.5 megabase pairs and can analyse locus batch file sequences. IMGT/LIGMotif is particularly useful for annotating genomic sequences of human and mouse loci and those of closely related species, the non-human primates and rat, respectively.

DISCIPLINE

Bioinformatique

MOTS-CLES

Immunogénétique, immunoinformatique, immunoglobuline, récepteur des cellules T, annotation automatique, génomique.

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