L'annotation des séquences en immunogénétique: la stratégie d'IMGT basée sur IMGT-ONTOLOGY.

Véronique Giudicelli[1], Joumana Jabado-Michaloud[1], Denys Chaume[1] et Marie-Paule Lefranc[1,2]

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

L'annotation des séquences en immunogénétique: la stratégie d'IMGT basée sur IMGT-ONTOLOGY.

Véronique Giudicelli[1], Joumana Jabado-Michaloud[1], Denys Chaume[1] et Marie-Paule Lefranc[1,2]

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

1 Introduction

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [1], créé à Montpellier en 1989 (Université Montpellier II et CNRS) est spécialisé dans les séquences, structures et données génétiques des immunoglobulines (IG), récepteurs T (TR), complexe majeur d'histocompatibilité et protéines des superfamilles IgSF et MhcSF. IMGT® est la référence internationale en immunogénétique et immunoinformatique. C'est un système intégré qui comprend plusieurs bases de données, des ressources Web et des outils en ligne [2]. IMGT/LIGM-DB est la base de données de séquences nucléotidiques d'IMGT, dans laquelle sont gérées, analysées et annotées plus de 89.000 séquences d'IG et de TR (en janvier 2005) issues de l'homme et de 104 autres espèces de vertébrés. L'annotation des séquences représente la plus-value indispensable à leur compréhension et à leur réutilisation. Face à la croissance exponentielle du nombre des séquences d'IG et de TR publiées, IMGT a adopté une stratégie pour les annoter qui prend en compte à la fois la complexité de la génétique des IG et des TR [3,4] et la nécessité d'une automatisation de plus en plus importante. Cette stratégie s'appuie sur IMGT-ONTOLOGY [5], première ontologie dans le domaine de l'immunogénétique. L'annotation des séquences des IG et des TR s'effectue selon deux approches différentes selon qu'il s'agisse d'ADN génomique ou de séquences exprimées: (i) l'annotation des séquences d'ADN génomique (« germline » et réarrangé) est basée sur la recherche de motifs spécifiques et l'expertise manuelle nécessaires à la caractérisation de nouveaux gènes dans de grandes séquences. Les connaissances relatives aux gènes et allèles (déduites de l'expertise des séquences génomiques et de leur annotation) sont répertoriées et gérées en accord avec le concept de CLASSIFICATION d'IMGT-ONTOLOGY dans la base de gènes IMGT/GENE-DB [6] pour l'homme et la souris, et au niveau des ressources Web d'« IMGT Repertoire » [7] pour les autres espèces de vertébrés; (ii) l'annotation des séquences d'ADN complémentaires (ADNc) est automatisée et ce, malgré l'origine génétique complexe de ces séquences. En effet, les règles d'identification des séquences, de description de leurs motifs constitutifs et de numérotation des acides aminés des régions codantes, correspondent respectivement aux concepts d'IDENTIFICATION, de DESCRIPTION et de NUMEROTATION d'IMGT-ONTOLOGY [5]. Ces règles ont pu être codées dans IMGT/Automat [8], programme Java développé par IMGT, qui annote automatiquement les séquences d'ADNc des IG et des TR. IMGT/Automat met en œuvre IMGT/V-QUEST [9] pour l'identification des séquences, la classification des gènes et allèles du domaine

variable et la numérotation des acides aminés et IMGT/JunctionAnalysis [10] pour l'analyse précise et détaillée de leur jonction. 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. Ainsi, la stratégie d'annotation mise en place par IMGT comporte deux approches différentes dépendant de la nature, ADN génomique ou ADNc, des séquences. Par leur complémentarité, ces deux approches garantissent la validité, la précision et la cohérence des annotations des séquences nucléotidiques de IMGT/LIGM-DB. La précision et la qualité des annotations sont les facteurs déterminants pour l'exploitation de ces séquences immunogénétiques dans des secteurs aussi exigeants que la recherche fondamentale, l'industrie pharmaceutique et la recherche clinique, la recherche vétérinaire et l'ingénierie des anticorps.

2 Matériel et méthodes

Matériel

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 [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.

Méthodes

Annotation des séquences d'ADN génomique

L'annotation des séquences génomiques consiste à localiser les gènes, à prédire les exons, à déterminer les signaux de régulation tels que les promoteurs et les sites d'épissage, et à déterminer les homologies avec les gènes connus. Les programmes classiques développés pour accomplir ces tâches s'avèrent absolument inefficaces pour la recherche et l'analyse des gènes IG et TR. En effet, à titre d'exemple la figure 1 montre que le deuxième exon des gènes variables en configuration germline (V-GENE) ne comporte pas de site d'épissage, ni de codon de terminaison en 3', mais des signaux de recombinaison caractéristiques des IG et des TR, impliqués dans le réarrangement de l'ADN. Les gènes de diversité (D-GENE) et de jonction (J-GENE) ne comprennent ni codon d'initiation en 5', ni site d'épissage [3,4]. Leurs régions codantes sont très courtes (inférieures à 20 acides aminés) et les outils classiques de recherche d'exons ne sauront ni les détecter, ni les délimiter correctement. IMGT a donc développé un programme de recherche de motifs spécifiques aux IG et aux TR, LIGMotif, basé sur la reconnaissance des signaux de recombinaisons, des sites d'épissage et la numérotation standardisée IMGT des acides aminés des parties codantes qui permet en particulier de localiser les acides aminés conservés (Figure 1). IMGT a défini, pour chaque gène et chaque allèle caractérisé, une séquence de référence. L'ensemble de ces séquences de référence constitue l'« IMGT reference directory ». Les séquences codantes des nouveaux gènes sont comparées, par BLAST, FASTA et IMGT/V-QUEST, aux séquences de l'IMGT reference directory. L'expertise manuelle, effectuée après cette étape d'identification, est indispensable et cruciale pour la classification du gène, la caractérisation précise de ses allèles, la définition de sa structure et de sa fonctionnalité, selon les concepts de IMGT-ONTOLOGY [5] et les règles de la charte scientifique (IMGT Scientific chart) [7].

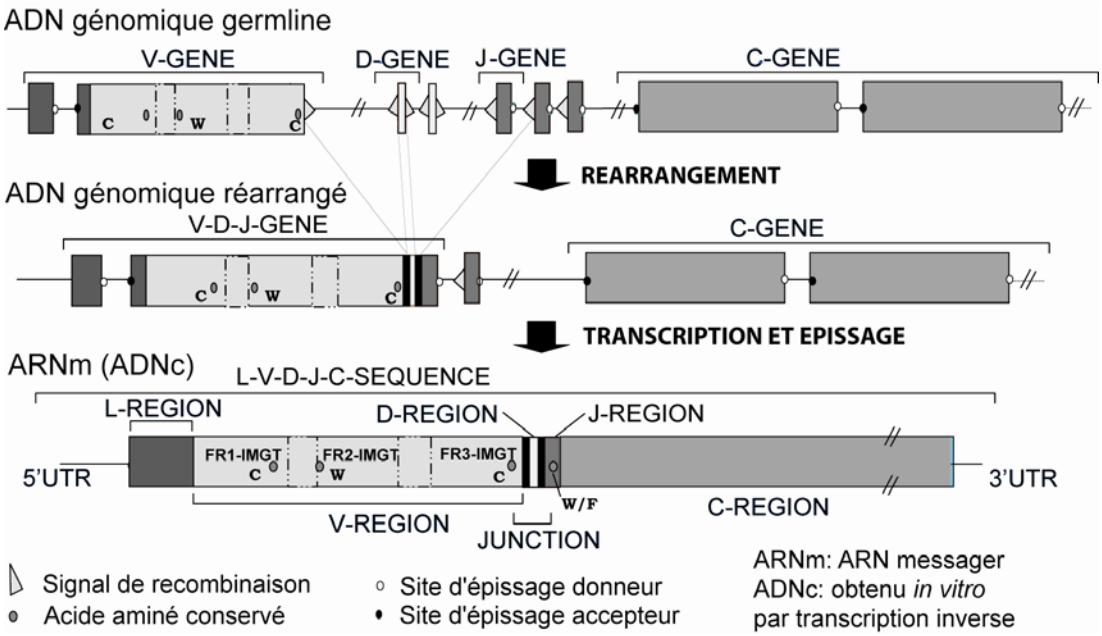

Figure 1: Représentation graphique de la synthèse d'une chaîne lourde d'IG.

Annotation des séquences d'ADNc

Lorsque les séquences d'ADNc sont issues de gènes et allèles connus et caractérisés dans IMGT (séquences répertoriées dans l'IMGT reference directory), l'annotation peut être réalisée automatiquement par IMGT/Automat [8]. IMGT/Automat, programme écrit en langage Java, analyse chaque séquence d'ADNc et en déduit une annotation complète de façon totalement automatique. IMGT/Automat s'appuie sur les principaux concepts d'IMGT-ONTOLOGY [5]. Les principales étapes de l'annotation automatique sont représentées dans la figure 2. IMGT/Automat utilise dans un premier temps le logiciel IMGT/V-QUEST [9] pour comparer et aligner la séquence ADNc avec les séquences de l' IMGT reference directory 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 et 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) [3,4]. La description précise de la zone de jonction des gènes V-D-J ou V-J est réalisée à l'aide de IMGT/JunctionAnalysis [10]. Des méthodes Java, basées sur la comparaison de motifs, permettent 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 dont un exemple est représenté dans la figure 1. Le module de cohérence de la troisième étape vérifie et valide les séquences définies comme « productive » (les séquences « unproductive » nécessitent une expertise complémentaire manuelle).

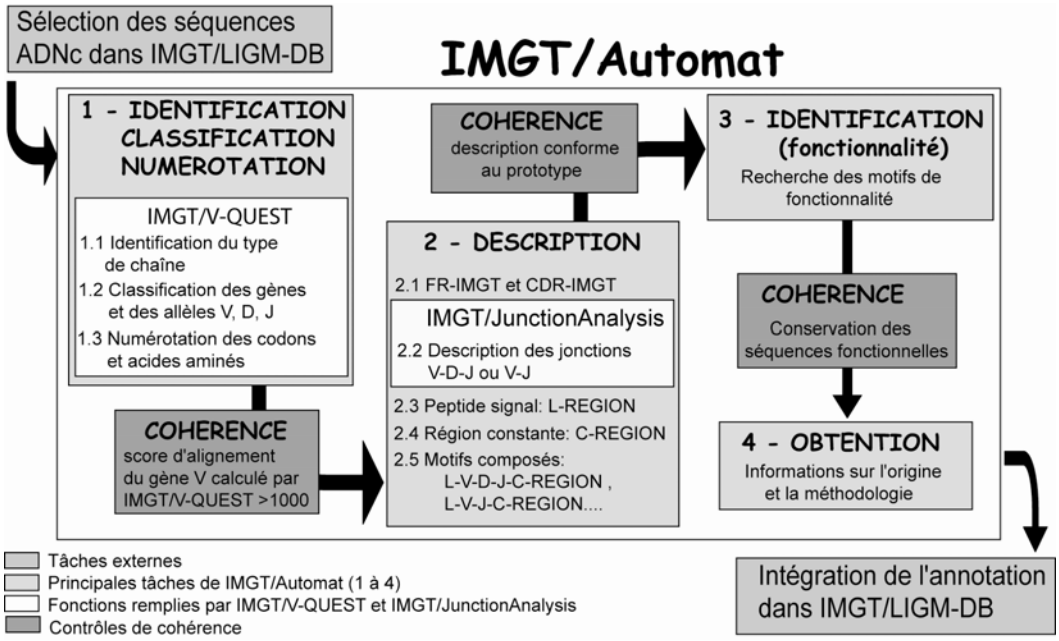

Figure 2: Principales étapes de l'annotation automatique des ADNc réalisée par IMGT/Automat.

3 Résultats

L'annotation des séquences d'ADN génomique des IG et TR, basée sur la recherche de motifs spécifiques et l'expertise manuelle, représente l'étape indispensable et préalable à l'inventaire, la caractérisation et la classification de tous les gènes et allèles, pour une espèce donnée. Cette expertise a été complètement réalisée par IMGT, pour les gènes et allèles de l'homme et de la souris, et les données sont accessibles dans IMGT/GENE-DB [6]. La nomenclature élaborée par IMGT pour les gènes IG et TR humains [3,4] a été acceptée en 1999 par le « HUGO Nomenclature Committee » (HGNC) [12] et est devenue la référence internationale. Ainsi, des liens directs ont été mis en place sur IMGT/GENE-DB à partir de LocusLink et Entrez Gene au NCBI, de GDB, de GENATLAS et de GeneCards. L'annotation des séquences d'ADNc par IMGT/Automat permet de traiter rapidement un grand nombre de ces séquences qui représentent plus de 50% de IMGT/LIGM-DB. A l'heure actuelle IMGT/Automat, utilisé comme outil interne, annote principalement les ADNc d'homme et de souris. Ainsi sur 18.000 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.

4 Discussion

IMGT a adopté, pour l'annotation des séquences nucléotidiques de la base de données IMGT/LIGM-DB, une stratégie qui comporte deux approches. Cette stratégie tient 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. L'annotation des séquences d'ADN génomique est basée sur une expertise manuelle importante. En effet, la robustesse et la précision de cette annotation manuelle permettent, dans un deuxième temps, la mise en place d'une annotation automatique des ADNc par IMGT/Automat. De manière

remarquable, la qualité de l'annotation automatique des ADNc est équivalente à la qualité d'une annotation manuelle. Cette stratégie, qui combine expertise manuelle et automatisme est à l'heure actuelle la seule stratégie qui puisse garantir qualité et précision de l'annotation et le traitement d'un nombre toujours plus élevé des séquences nucléotidiques dans le domaine de l'immunogénétique. Cette automatisation pourra être appliquée à l'annotation des séquences génomiques réarrangées, moyennant la mise en place de contrôles et règles complémentaires. Les annotations fournies par IMGT, la référence internationale en immunogénétique et immunoinformatique, sont particulièrement utilisées en recherche médicale pour l'étude des répertoires en situation normales et pathologiques (maladies auto-immunes et infectieuses, Sida, leucémies, lymphomes, myélomes), en biotechnologie et ingénierie des anticorps (banques combinatoires, phage displays) et pour les approches thérapeutiques (greffes, immunothérapie).

Références

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

[2] Lefranc M-P. IMGT® databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis, http://imgt.cines.fr. Leukemia, 2003; 17, 260–266.

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

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

[5] Giudicelli V. and Lefranc M-P. Ontology for Immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 1999; 12, 1047–1054.

[6] 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-D261.

[7] 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, 2004; 4, 17–29.

[8] Giudicelli V, Protat C and Lefranc M-P. The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. ECCB'2003, European Conference on Computational Biology. Ed DISC/Spid DKB-31, 2003; pp103–104.

[9] 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. 2004; 32, W435–W440.

[10] 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 2004; 20, I379–I385.

[11] Kanz C, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 2005; 33, D29-D33.

[12] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW and Povey S. Guidelines for human gene nomenclature. Genomics, 200, 79, 464–470.

Adresse de correspondance

Marie-Paule Lefranc, 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 Email: lefranc@ligm.igh.cnrs.fr URL : http://imgt.cines.fr