Rapport plan quadriennal IGH
Septembre 2005

IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM

Mots clefs: Immunogénétique, immunoinformatique, immunologie, anticorps, récepteurs T.

Personnel en septembre 2005:
- Marie-Paule Lefranc, Professeur CE2 Université Montpellier II, Membre Senior de l'Institut Universitaire de France
- Gérard Lefranc, Professeur CE2 Université Montpellier II, 10%
- Denys Chaume IE CNRS 100%, responsable de la maintenance du système d'information
- Véronique Giudicelli IE Université Montpellier II 50%, bioinformatique double compétence
- Chantal Ginestoux T3 CNRS 80%, interface Web
- Joumana Jabado-Michaloud IE CDD CNRS 100%, candidate au poste d'IE mis au concours cette année, Annotateur double compétence
- Vijay Garapati IE CDD CNRS 100%, sur fonds FNS ACI IMPBio (juillet 2005-janvier 2007
- Oliver Clément IE CDD CNRS 100%, sur fonds BioSTIC 2004-2005 (fin de contrat septembre 2005)
- Géraldine Folch AI CDD CNRS, sur fonds BioSTIC 2004-2005, puis ADER, Annotateur (fin de contrat mars 2006)
- Elodie Duprat Thèse Allocataire de recherche du Ministère MENESR, soutenance prévue décembre 2005
- Quentin Kaas Thèse Allocataire de recherche du Ministère MENESR puis bourse ARC, soutenance prévue décembre 2005
- Wafae El Alaoui Thèse (cotutelle Université Montréal et Université Montpellier II), soutenance prévue 2008
- Kevin Bleakley Thèse sous la direction de Gérard Biau (thèse à l'Interface Biologie-Mathématiques, Ecole doctorale ED166 "Information, Structures et Systèmes" I2S), Allocataire de recherche du Ministère MENESR, début de thèse septembre 2005
- Un nouveau doctorant IMGT (Ecole doctorale ED168, "Sciences chimiques et biologiques pour la santé" CBS2), Allocataire de recherche de la Présidence Montpellier I, Ministère MENESR, début de thèse septembre 2005.

Rapport scientifique:

Thème et questions posées

Le système immunitaire (SI), indispensable à la survie des organismes et des espèces en préservant le « soi » et en éliminant les agents pathogènes (bactéries, virus, parasites) et les cellules tumorales, comprend le SI Inné et le SI adaptatif.
Le SI inné, présent dans toutes les espèces du monde vivant, est régi par des gènes, protéines et voies de régulation « obéissant » aux règles « classiques » de la génétique et de la biologie moléculaire de la cellule que sait gérer parfaitement la bioinformatique. En revanche, le SI adaptatif, propre aux Vertébrés, est d'une extrême complexité en raison des spécificités quasi illimitées des sites anticorps des immunoglobulines (IG) membranaires des lymphocytes B et secrétées par les plasmocytes et des sites de reconnaissance des récepteurs (TR) des

lymphocytes T. Cette diversité extrême s'explique par une organisation des gènes et une complexité dans leur fonctionnement qui ne peuvent être prises en compte par les programmes et outils bioinformatiques classiques.
La première question posée en 1989 au Congrès de New Haven (Human Genome Mapping HGM10) a été la suivante : Comment prendre en compte les gènes des IG et TR dans l'organisation des génomes ?
Le Congrès de New Haven a été le point de départ d'IMGT, avec la reconnaissance officielle des principes, publiés par LIGM, qui permettaient de décrire les différents locus IG et TR. Et l'introduction des 16 gènes du locus TRG humain, séquencés par LIGM, dans GDB (Genome DataBase).
Il restait à identifier (découvrir) les règles fondamentales qui permettraient de gérer les IG et TR, quel que soit le type de récepteur (IG ou TR), quel que soit le type de chaîne lourde (alpha, gamma, delta, epsilon et mu) et légère (kappa et lambda) pour les IG, quel que soit le type de chaîne (alpha, beta, gamma et delta) pour les TR, quel que soit le type de domaine (variable ou constant) et quelle que soit l'espèce de vertébrés (du poisson à l'homme).
Dix ans après le Congrès de New Haven, en 1999, tous les gènes IG et TR humains (664 gènes dont 421 IG and 243 TR) ont été approuvés par le Human Genome Organisation (HUGO) Nomenclature Committee HGNC, et ceci avant même que la séquence du Génome Humain soit connue. A l'heure actuelle, IMGT, the international ImMunoGeneTics information system®, http://imgt.cines.fr, est la référence internationale pour les gènes d'IG et TR. IMGT est le seul site étranger référencé par le National Center for Biotechnology Information (NCBI) aux Etats-Unis. Les résultats ont été publiés dans 2 livres qui sont devenus les références dans le domaine (*The immunoglobulin FactsBook*, et *The T cell receptor FactsBook* par M.-P. Lefranc and G. Lefranc, Academic Press, 2001).

Les thèmes de recherche abordés selon des approches génomique, génétique et structurale sont :

- les locus des gènes d'immunoglobulines et de récepteurs T des génomes de vertébrés nouvellement séquencés et qui servent de modèles animaux pour l'analyse de la réponse immunitaire adaptative (ex: souris, chimpanzé),

- les répertoires des anticorps et des sites de reconnaissance des récepteurs T dans les situations normales et pathologiques (maladies autoimmunes et infectieuses, Sida, leucémies, lymphomes, myélomes),

- les répertoire des anticorps et des sites de reconnaissance des récepteurs T dans les espèces domestiques et sauvages,

- l'étude de la diversité et de l'évolution des gènes des réponses immunitaires adaptatives.

- l'étude structurale des domaines des protéines de la superfamille des immunoglobulines (IgSF) et de la superfamille du MHC (MhcSF).

Ces recherches fondamentales répondent également aux objectifs suivants:

- en clinique pour les diagnostics des leucémies, lymphomes et myélomes (clonalités, détection et suivi des maladies résiduelles) et les approches thérapeutiques (greffes, immunothérapie, vaccinologie),

- en biotechnologie pour l'ingénierie des anticorps (construction et analyse des "single chain Fragment variable" scFv, banques combinatoires, phage displays, anticorps chimériques, humanisés et humains).

b) Résultats principaux

Les avancées d'IMGT :

- ont rendu caduque la base de données RETREMBL (*REmaining TREMBL*), maintenue par l'European Bioinformatics Institute (EBI) à Hinxton et qui contenait les séquences non gérées par les logiciels classiques (à 95 % des séquences d'IG et de TR).
- ont permis de gérer les polymorphismes génétiques de manière standardisée. L'approche IMGT est à la fois la plus rigoureuse et la plus simple en ce qui concerne les polymorphismes des régions codantes, la définition des allèles et la définition de la fonctionalité des gènes.
- ont permis une approche intégrée de familles multigéniques.
- ont permis une approche intégrée, au sein du même sytème d'information, de données hétérogènes (séquences nucléotidiques, données protéiques, gènes et données structurales). Ces types de données sont gérées dans des bases de données différentes (respectivement IMGT/LIGM-DB, IMGT/PROTEIN-DB, IMGT/GENE-DB et IMGT/3Dstructure-DB) comme cela est le cas pour les bases généralistes (EMBL/GenBank/DDBJ, Swiss-Prot, Entrez Gene NCBI et PDB). La différence est que les données d'IMGT sont décrites quel que soit le type de données et, donc, quelle que soit la base de données, avec les mêmes concepts d'identification, de classification, de description, de numérotation, d'orientation et d'obtention. Ces concepts font partie d'IMGT-ONTOLOGY, l'ontologie de référence dans le domaine de l'immunogénétique et de l'immunoinformatique. Les concepts d'IMGT-ONTOLOGY sont accessibles aux biologistes dans la charte scientifique IMGT (IMGT Scientific chart) qui fournit les règles standardisées et le vocabulaire contrôlé nécessaires à la gestion des données. Les concepts ont été formalisés, en utilisant le Schéma XML (*Extensible Markup Language*), en IMGT-ML.

Résultats majeurs des 4 dernières années:

Les travaux de recherche des 4 dernières années ont permis d'intégrer des données génétiques et des données structurales. Nous avons pris en compte la remarquable conservation structurale des domaines (domaines variables et domaines constants des IG et TR) afin de décrire les mutations, les caractéristiques physico-structurales et contacts entre acides aminés et ceci, malgré les différences importantes au niveau des séquences.
Une numérotation unique IMGT a été mise en place pour les domaines variables (V-DOMAIN) qui résultent de la jonction de deux ou trois gènes (selon les locus) et correspond au niveau protéique soit à une V-J-REGION (cas des chaînes légères des IG et des chaînes alpha et gamma des TR), soit à une V-D-J-REGION (cas des chaînes lourdes des IG et des chaînes beta et delta des TR). Grâce à cette numérotation unique IMGT, des représentations graphiques en 2 dimensions, ou « IMGT Colliers de Perles », ont été créées. Les IMGT Colliers de Perles permettent une comparaison aisée entre séquences et structures tridimensionnelles, et la prédiction des caractéristiques structurales de domaines pour lesquels la structure 3D n'est pas encore disponible. Le succès de cette standardisation a introduit l'expression IMGT Colliers de Perles aux Etats-Unis et en Angleterre.

La démonstration la plus inattendue de la validité de cette démarche standardisée est venue de l'identification récente, chez un vertébré inférieur (truite), d'une protéine avec un domaine V-LIKE (codé par un gène qui ne réarrange pas) et dont l'IMGT Collier de perles était l'illustration parfaite du Collier de Perles « virtuel », conçu par IMGT, pour la première fois en 1997. Toutes les positions « gaps » de ce collier sont occupées et, de plus, le motif

caractéristique des J-REGIONs des IG (Phe-Gly- X-Gly) est retrouvé aux positions (118-121) où il serait présent si ce domaine correspondait à un gène réarrangé d'IG.
Ceci conforte notre approche selon laquelle la standardisation des domaines des IG et des TR peut être extrapolée à tous les domaines V-LIKE et C-LIKE des protéines de la superfamille des immunoglobulines (IgSF) autres que les IG et TR.

Une numérotation unique IMGT a également été mise en place pour les G-DOMAINs (« groove ») des protéines du complexe majeur d'histocompatibilité (CMH, HLA *human leucocyte antigen* chez l'homme). Ceci nous a permis de réaliser, pour la première fois, des alignements des domaines G-ALPHA1 et G-ALPHA2 des CMH de classe I et les domaines G-ALPHA et G-BETA des CMH de classe II, et ceci quelle que soit l'espèce. La standardisation des domaines du MHC a été extrapolée à tous les domaines G-LIKE des protéines de la superfamille du MHC (MhcSF) autres que le MHC.

De 2002 à 2005, 40 publications dans des revues internationales à comité de lecture (28 dans des journaux et 12 ouvrages ou chapîtres de livres, auxquelles s'ajoutent 51 communications dans des congrès et 43 conférences invitées, l'obtention du prix ADER Recherche-Innovation-Entreprise Languedoc-Roussillon en 2003 témoignent de l'activité scientifique de l'équipe IMGT.

Au niveau national, IMGT est Plate-Forme Bioinformatique RIO (CNRS, INSERM, CEA, INRA) depuis 2001. IMGT est Plate-Forme Bioinformatique du Réseau National des Génopoles. IMGT participe à l'action concertée incitative Informatique, Mathématiques, Physique en Biologie (ACI IMPBio 2004-2007) et à l'activité GIS AGENAE 2004-2007.

Au niveau européen, IMGT a assuré la coordination de trois précédents contrats européens: BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037) et Quality of Life and Management of Living Resources (QLG2-2000-01287) du 5ème PCRDT. IMGT est actuellement participant d' un STREPS ImmunoGrid FP6-2004-IST-4 "The European Virtual Human Immune System Project" présélectionné dans le cadre du 6ème PCRDT et en phase de négociation.

Les autres étapes marquantes sont:
- la page de couverture de *Nucleic Acids Research* pour les 10 ans d'IMGT,
- NetWatch of Science "Blueprints of Immunity" a NetWatch report on IMGT, *Science*, 296,1207 (2002), et le WebWatch BioTechniques (2004)
- l'atelier organisé à Montréal en juillet 2004 au Congrès International d'Immunologie pour les 15 ans d'IMGT et l'atelier organisé à San Diego (Californie) en Décembre 2005

IMGT est une marque déposée du CNRS (France, EU, Canada et Etats-Unis). IMGT reçoit une moyenne de 140.000 requêtes par mois, les utilisateurs étant répartis à parts égales entre l'Europe (1/3), les Etats-Unis (1/3) et le reste du monde (1/3).

Tableau synoptique des avancées scientifiques et applications

Année	Avancées scientifiques	Applications
2002	• **IMGT®, système d'information international, la référence en Immunogénétique et en Immunoinformatique** *(NAR 2002, Dev. Comp. Immunol., 2002)* • **Création de IMGT/PRIMER-DB**	• IMGT/LIGM-DB: Accès à plus de 54 000 séquences de 105 espèces en janvier 2002. • Nouvelle base de données d'IMGT pour les oligonucléotides d'IG et de TR.
2003	• **IMGT®** *(NAR 2003, Dev. Comp. Immunol., 2003, Leukemia 2003)* • **Création de IMGT/GENE-DB**	• IMGT/LIGM-DB: Accès à plus de 70 500 séquences de 105 espèces en mai 2003. • Nouvelle base de données d'IMGT pour les gènes des IG et TR.
2004	• **IMGT®***(In Silico Biology 2004, NAR 2004, J. Mol. Recognit. 2004)* • **IMGT/V-QUEST** *(NAR 2004)* • **IMGT/JunctionAnalysis** *(NAR 2004)* • **IMGT/GeneInfo** *(Bioinformatics 2004)*	• IMGT/LIGM-DB: Accès à plus de 86 000 séquences de 120 espèces en octobre 2004. • Outils d'analyse des séquences d'IG et TR. • Outils d'analyse des jonctions d'IG et TR. • Outil d'analyse des réarrangements des TR
2005	• **IMGT®** *(NAR 2005, Dev. Comp. Immunol. 2005, In Silico Biology 2005)*	• IMGT/LIGM-DB: Accès à plus de 96 500 séquences de 150 espèces en septembre 2005.

c) Objectifs, projet

IMGT fournit un accès commun à des données standardisées qui comprennent des séquences nucléotidiques, des séquences protéiques, des cartes de locus, des polymorphismes génétiques et des structures tridimensionnelles (3D). Ceci forme un ensemble riche et complexe de données. IMGT est donc un système d'information complet et hétérogène à la fois par ses composantes (bases de données, outils, ressources Web) et ses données (génétiques, génomiques, structurales, etc.).

Pour maintenir IMGT en tant que référence internationale en immunogénétique et immunoinformatique, il est capital de lui garder ses compétences innovantes et, pour cela, de créer un système d'interactions dynamiques entre les bases de données, les outils et les ressources Web du système d'information.

1. La méthodologie informatique: Web services

Dans l'optique de création d'un " continuum " par des interactions dynamiques entre les outils et les bases de données, nous avons choisi d'utiliser les Web Services. Ces services (créés en janvier 2002) permettent à un programme client externe (comme une base de donnée ou un outil) d'accéder à des informations ou de mettre en œuvre des traitements de données disponibles sur le serveur IMGT via des protocoles standards. Dans notre cas la requête sera exprimée avec un formalisme XML d'IMGT-ONTOLOGY: IMGT-ML, qui est en cours d'élaboration.

L'ensemble des interactions entre composantes, qui nécessite la capacité de composer et de décrire les relations entre les Web Services (Web Services Choreography), est désigné sous le nom d'IMGT-Choreography.

IMGT-Choreography a démarré dans le cadre des actions BioSTIC-LR 2004-2005 et ACI-IMPBio. Nous souhaitons vivement poursuivre ce projet qui permet de gérer et d'extraire les connaissances d'une manière beaucoup plus efficace. IMGT-Choreography choisit dynamiquement, en fonction de la démarche engagée, les outils adéquats, le cheminement de recherche approprié et le type d'affichage des résultats des requêtes. Avec IMGT-Choreography, nous poursuivons et renforçons notre objectif que chaque outil, base de données ou page HTML d'IMGT ne soit pas une fin en soi, mais encourage à poursuivre le parcours vers d'autres ressources. Cette démarche a pour but d'approfondir la recherche ou le traitement des données et ainsi de générer de nouvelles connaissances.

2. Aspects innovants du projet

La réalisation d'un tel projet informatique, dans le domaine biologique qui est complexe, hétérogène et évolutif, est inédite. L'approche d'IMGT-Choreography sera certainement suivie avec intérêt par d'autres laboratoires car, de plus en plus, l' approche standardisée et intégrée d'IMGT est utilisée dans des projets en Bioinformatique. IMGT-Choreography renforcera la position d'IMGT en tant que système de référence international en immunogénétique et immunoinformatique.