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

Elodie DUPRAT

le 13 décembre 2005

Titre :

Caractérisation des domaines des superfamilles IgSF et MhcSF et classification fonctionnelle dans IMGT

JURY

Mme Marie-Paule Lefranc, Professeur, Université Montpellier II, Directeur de thèse
M. Olivier Gascuel, Directeur de Recherche, LIRMM, Co-directeur de thèse
M. Gérard Lefranc, Professeur, Université Montpellier II, Président du jury
M. Olivier Poch, Directeur de recherche, IGBMC, Strasbourg, Rapporteur
M. Jean-Loup Risler, Directeur de recherche, LGI, Evry, Rapporteur
M. Vincent Schachter, Chargé de recherche, Génoscope, Evry, Examinateur

RESUME en français

La caractérisation de domaines protéiques permet de définir des superfamilles structurales; leur classification fonctionnelle contribue à la représentation des systèmes biologiques. IMGT est le système d'information international en ImMunoGénéTique®, spécialisé dans la gestion des séquences et structures 3D des protéines du système immunitaire des vertébrés. IMGT-ONTOLOGY fournit les règles de description de leurs récepteurs, chaînes et domaines; la numérotation unique de chaque type de domaine se base sur des alignements multiples et assure la description standardisée de leurs caractéristiques fonctionnelles et structurales. Des domaines de structure similaire ont été identifiés au sein de nombreuses protéines définissant ainsi les superfamilles structurales des immunoglobulines (IgSF) et du MHC (MhcSF), hétérogènes en séquences et en fonctions.
Les objectifs de cette thèse étaient l'extension de la standardisation IMGT aux protéines des IgSF et MhcSF, et la mise en place d'une méthodologie de classification de ces protéines selon leurs fonctions et leurs interactions à partir de leur séquence. Nous avons développé une procédure d'alignement de ces domaines adaptée à leur hétérogénéité de séquence; elle permet la standardisation de nouvelles protéines et la gestion des IgSF et MhcSF par les différents composants d'IMGT. Nous présentons également une approche de classification supervisée des protéines de la MhcSF selon leur liaison ou non à la beta2-microglobuline, combinant un classifieur Bayesien naïf et la numérotation unique IMGT. Cette approche est performante, fournit des informations concernant les propriétés physico-chimiques qui favorisent ou défavorisent cette interaction, et permet l'annotation automatique de protéines de vertébrés inférieurs; elle devrait s'appliquer avec succès à d'autres problématiques de classification des protéines des IgSF et MhcSF pour lesquelles on dispose de classes connues a priori, et plus généralement à la classification fonctionnelle de superfamilles protéiques à partir d'un alignement multiple.

TITRE en anglais
DOMAIN CHARACTERIZATION OF IGSF AND MHCSF SUPERFAMILIES AND FUNCTIONAL CLASSIFICATION IN IMGT

RESUME en anglais

Protein domain characterization provides the definition of structural superfamilies; their functional classification contribute to the representation of biological systems. IMGT, the international information system in ImMunoGeneTics®, is dedicated to the management of the sequences and the 3D structures of the vertebrate immune system proteins. IMGT-ONTOLOGY provides the rules of description of their receptors, chains and domains; the unique numbering of each domain type is based on multiple alignments and ensures the standardized description of their functional and structural characteristics. Domains with similar structure were identified within many proteins thus defining the structural superfamilies of the immunoglobulins (IgSF) and of the MHC (MhcSF), heterogeneous in sequences and functions.
The goals of this thesis were the extension of the IMGT standardization to proteins of IgSF and MhcSF, and the development of a method to classify these proteins according to their functions and interactions, only from their sequence. We developed a procedure for the alignement of these domains, taken into account their sequence heterogeneity; it allows the standardization of new proteins and the management of IgSF and MhcSF by the IMGT components. We also present an approach of supervised classification of the MhcSF proteins according to their interaction or not with the beta2-microglobulin, combining a simple Bayes classifier and the IMGT unique numbering. This approach is accurate, provides information concerning the physicochemical properties which support or prevent this interaction, and allows the automatic annotation of low vertebrate proteins; it should apply successfully to other problems of IgSF and MhcSF function or interaction classification, and to any problem of superfamily protein classification based on a multiple alignment.

DISCIPLINE
Bioinformatique

MOTS-CLES
Superfamille du MHC, MhcSF, superfamille des immunoglobulines, IgSF, domaines protéiques, numérotation unique IMGT, beta2-microglobuline, classification supervisée, classifieur Bayesien naïf.

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

REMERCIEMENTS

Je voudrais tout d'abord exprimer toute ma reconnaissance à Marie-Paule Lefranc et Olivier Gascuel, pour la compétence et la disponibilité dont ils ont fait preuve tout au long de ce travail de thèse. Je remercie particulièrement Marie-Paule Lefranc pour son accueil chaleureux et son soutien quotidien.

Je tiens à remercier les rapporteurs de cette thèse, Jean-Loup Risler et Olivier Poch, ainsi que les membres du jury, Vincent Schachter et Gérard Lefranc, qui ont accepté d'évaluer et de commenter ce travail.

Je remercie tous les membres de l'équipe IMGT pour leur accueil, leur soutien, leur humour, leur bonne humeur quotidienne, et particulièrement les « wonder women » Chantal, Géraldine, Joumana et Véronique. Je n'oublierai pas ces trois années passées en leur compagnie.

Mes pensées vont également à l'équipe pédagogique du DEA AGM2, qui m'a initiée à la recherche en bioinformatique avec un enthousiasme sans faille ; je remercie particulièrement Alexandre de Brevern pour son soutien et ses conseils judicieux.

Je remercie également les institutions qui m'ont apporté leur soutien financier durant ces trois années : le Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR) pour le programme de bourses « Génomes », et le CNRS.

Mes derniers remerciements vont à mes parents, pour leur soutien inégalable tout au long de mes études, à Caro, Manue et Jérôme pour la joie qu'ils me procurent depuis des années (10 ans !), à Marie-Jo et Henri pour les après-midi ensoleillées que j'ai passé en leur compagnie, et à Cyril pour tout le bonheur présent et à venir ; leurs encouragements et leur affection m'ont permis de mener à bien cette thèse.

PUBLICATIONS

Publications majeures

DUPRAT, E., LEFRANC, M.-P. and GASCUEL, O. (2005) A simple method to predict protein binding from aligned sequences – application to MHC superfamily and beta2-microglobulin. *Bioinformatics* (in press).

DUPRAT, E., LEFRANC, M.-P. and GASCUEL, O. (2005) Prédire l'interaction des protéines de la superfamille du MHC avec la beta2-microglobuline en combinant classifieur Bayesien « naïf » et alignement multiple IMGT. *Actes des Journées Ouvertes Biologie Informatique Mathématiques 2005*.

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. *Developmental and Comparative Immunology*, 29, 917-938.

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. (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Developmental and Comparative Immunology*, 29, 185-203.

KAAS, Q., **DUPRAT, E.**, TOURNEUR, G. and LEFRANC, M.-P. (2005) IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: *Immunoinformatics* (Brusic V, Schoenbach C eds.) Springer, The Netherlands (in press).

DUPRAT, E., KAAS, Q., GARELLE, V., GIUDICELLI, V., LEFRANC, G. and LEFRANC, M-P. (2004) IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In: *Recent Research Developments in Human Genetics*. Trivandium, India, Research Signpost, 2, 111-136.

BERTRAND, G.[*], **DUPRAT, E.**[*], LEFRANC, M.-P., MARTI, J. and COSTE, J. (2004) Human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. *Tissue Antigens*, 64, 2, 119-131.

Autres publications

LEFRANC, M.P., GIUDICELLI, V., KAAS, Q., **DUPRAT, E.**, JABADO-MICHALOUD, J., SCAVINER, D., GINESTOUX, C., CLEMENT, O., CHAUME, D. and LEFRANC, G. (2005) IMGT, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, 33, D593-597.

LEFRANC, M.P., CLEMENT, O., KAAS, Q., **DUPRAT, E.**, CHASTELLAN, P., COELHO, I., COMBRES, K., GINESTOUX, C., GIUDICELLI, V., CHAUME, D. and LEFRANC, G. (2005) IMGT-Choreography for immunogenetics and immunoinformatics. *In Silico Biology*, 5, 45-60.

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., POMMIE, C., CHAUME, D. and LEFRANC, G. (2004) IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biology*, 4, 17-29.

[*] contribution équivalente

TABLE DES MATIERES

INTRODUCTION 1

CHAPITRE 1 – Les protéines du système immunitaire adaptatif 5

1.1. Système immunitaire adaptatif et reconnaissance antigénique 6

1.1.1. Immunité humorale : lymphocytes B et immunoglobulines 6
1.1.2. Immunité cellulaire : lymphocytes T et récepteurs T 7

1.2. IG et TR : DESCRIPTION et NUMEROTATION 8

1.2.1. DESCRIPTION des récepteurs, chaînes et domaines protéiques 8
1.2.2. NUMEROTATION des V-DOMAINs et C-DOMAINs 13

1.3. MHC-I et MHC-II : DESCRIPTION et NUMEROTATION 17

1.3.1. DESCRIPTION des récepteurs, chaînes et domaines protéiques 17
1.3.2. NUMEROTATION des G-DOMAINs 20

CHAPITRE 2 – Les superfamilles IgSF et MhcSF 23

2.1. La superfamille des immunoglobulines (IgSF) 23

2.2. La superfamille du MHC (MhcSF) 36

CHAPITRE 3 – Standardisation et caractérisation des protéines de la MhcSF 32

3.1. Procédure d'alignement des G-DOMAINs et G-LIKE-DOMAINs 32

3.1.1. Extraction des données 33
3.1.2. Description des domaines protéiques 34
3.1.3. Alignement des G-DOMAINs et numérotation 34
3.1.4. Numérotation des G-LIKE-DOMAINs 35
3.1.5. Mise à jour de l'alignement et numérotation de nouvelles séquences 37

3.2. Analyse position-dépendante des G-DOMAINs et G-LIKE-DOMAINs 37

3.3. Caractéristiques évolutives 40

3.4. Les interactions protéine-ligand au sein de la MhcSF 42

3.4.1. Diversité fonctionnelle 42
3.4.2. B2M 44

CHAPITRE 4 – Classification fonctionnelle de familles protéiques 49

4.1. Classification non supervisée 50

4.2. Classification supervisée 53

4.2.1. Analyse discriminante de Fisher 54
4.2.2. *k* plus proches voisins (*k*-NN) 54
4.2.3. Bayes naïf 55

4.2.4. Segmentation par arbre binaire (CART) 56
4.2.5. Réseaux de neurones multicouches (ANN) 57
4.2.6. Machines à vecteurs supports (SVM) 58
4.3. Classifieur Bayesien naïf et numérotation unique IMGT 59

DISCUSSION ET CONCLUSION 63

BIBLIOGRAPHIE 66

ANNEXES 73

Annexe 1. Les concepts d'IMGT-ONTOLOGY 74
Annexe 2. DESCRIPTION des récepteurs, chaînes et domaines des fragments protéolytiques d'IG 75
Annexe.3. Organisation exon/intron des gènes MHC-I, MHC-II et MHC-I-like 77
Annexe 4. Modèle de l'origine évolutive des gènes MHC-I et MHC-II 82
Annexe 5. Données de séquence et de structure 3D des protéines MHC-I, MHC-II et MHC-I-like 83
Annexe.6. Regroupements des acides aminés selon leurs propriétés physico-chimiques 86
Annexe 7. Bootstrap et validation croisée 87
Annexe 8. Les mesures d'entropie relative, d'information mutuelle et du χ^2 88

PUBLICATIONS 90

INTRODUCTION

Le développement intensif du séquençage, des techniques de biologie structurale et de la biologie moléculaire a abouti à l'accumulation de nombreuses données concernant la séquence, la structure tridimensionnelle et la fonction des protéines. La mise en place d'approches *in silico*, destinées à classer automatiquement les protéines nouvellement identifiées d'après leurs caractéristiques structurales ou fonctionnelles, constitue un axe important de la bioinformatique actuelle. L'enjeu de ces approches est la compréhension de ce qui détermine la structure et la fonction des protéines ; les informations fournies permettent d'optimiser l'utilisation des protocoles expérimentaux longs et coûteux.

Les domaines protéiques sont des entités structurales autonomes ; de nombreuses bases de données de domaines protéiques ont été créées, regroupant les domaines en familles structurales (Murzin et al. 1995, Orengo et al. 1997) ou fonctionnelles (Webb 1992, Sonnhammer et al. 1997). Les séquences protéiques d'une même famille peuvent correspondre à différentes spécificités de fonctions ou d'interactions, connues ou non ; par exemple, la famille fonctionnelle des protéines kinases comprend des protéines qui présentent une activité similaire (la phosphorylation d'acides aminés) mais des substrats différents, correspondant aux différents types d'acides aminés phosphorylés. La classification fonctionnelle des familles protéiques consiste à découvrir les différentes fonctions qui les composent, ou à prédire la fonction de nouvelles protéines. Ces problématiques correspondent respectivement aux méthodes de classification dite supervisée et non supervisée. La classification fonctionnelle des familles protéiques obtenue à partir de ces approches est de plus en plus fine, et favorise ainsi une représentation de plus en plus réaliste des systèmes biologiques.

Parmi les systèmes biologiques, le système immunitaire permet la protection de l'hôte des maladies infectieuses et du cancer. Alors qu'une immunité naturelle est observée chez de nombreux organismes multicellulaires, les acteurs cellulaires et protéiques qui permettent une reconnaissance spécifique et une mémoire des agents pathogènes ne sont rencontrés que chez les vertébrés ; cette spécificité et la reconnaissance du soi et du non soi caractérisent l'immunité acquise, basée sur des interactions de type protéine-ligand impliquant principalement les immunoglobulines (IG), les récepteurs des cellules T (TR) et les protéines du complexe majeur d'histocompatibilité (MHC). Le système immunitaire est complexe, au niveau génomique (les IG, TR et MHC correspondent à des familles multigéniques, et l'ADN de la lignée germinale des cellules matures exprimant les IG et TR est soumis à des

réarrangements) (Tonegawa 1983) comme à l'échelle moléculaire (par la spécificité élevée des interactions impliquant les IG et TR) (Lefranc and Lefranc 2001a, 2001b). La bioinformatique est essentielle pour modéliser cette complexité, comprendre les dysfonctionnements du système immunitaire et développer des stratégies thérapeutiques adaptées.

IMGT, the international ImMunoGeneTics information system® créé par Marie-Paule Lefranc en 1989 à Montpellier, est spécialisé dans la gestion des données de séquence et de structure 3D des IG, TR et MHC de vertébrés (Lefranc et al. 2005a). Les données sont annotées à différentes échelles (chromosomes, locus, gènes, domaines protéiques), d'après les concepts d'IMGT-ONTOLOGY (Giudicelli and Lefranc 1999, Lefranc et al. 2004). Cette standardisation a permis de décrire l'organisation modulaire des récepteurs et chaînes des IG, TR et MHC, et de définir trois familles structurales de domaines protéiques : V-DOMAIN (Lefranc et al. 2003), C-DOMAIN (Lefranc et al. 2005b) et G-DOMAIN (Lefranc et al. 2005c). La numérotation unique IMGT de chaque type de domaine se base sur des alignements multiples et assure la description standardisée de leurs caractéristiques fonctionnelles et structurales. Des domaines de structure 3D similaire ont été identifiés au sein de protéines impliquées dans une grande variété de processus biologiques (tels que l'adhésion cellulaire et la régulation métabolique), localisés dans des compartiments cellulaires différents et correspondant à différents sites d'interaction protéine-ligand. Ces domaines protéiques sont nommés respectivement V-LIKE-DOMAIN, C-LIKE-DOMAIN et G-LIKE-DOMAIN. Les protéines composées d'au moins un V-DOMAIN, C-DOMAIN, V-LIKE-DOMAIN ou C-LIKE-DOMAIN appartiennent à la superfamille des immunoglobulines (IgSF), tandis que celles composées d'au moins un G-DOMAIN ou G-LIKE-DOMAIN appartiennent à la superfamille du MHC (MhcSF). Chacune de ces superfamilles structurales est hétérogène en terme de séquences, et comprend plusieurs familles fonctionnelles.

Les objectifs de cette thèse étaient l'extension de la standardisation IMGT aux protéines des IgSF et MhcSF, et la mise en place d'une méthodologie de classification de ces protéines selon leurs fonctions et leurs interactions à partir de leur séquence. Nous nous sommes particulièrement intéressés à la classification des protéines de la superfamille du MHC (MhcSF), selon qu'elles se lient ou non à la beta2-microglobuline (B2M). Pour les protéines de la MhcSF qui se lient constitutivement à la B2M, cette liaison non covalente est nécessaire à la stabilité de leur structure, à leur expression à la surface cellulaire et à leur fonction ; une interaction défectueuse pour ces protéines aboutit à de nombreuses pathologies décrites dans la littérature (Rose et al. 1983, Wang et al. 1993). Néanmoins, certaines protéines de la

MhcSF ne se lient pas à la B2M, sans conséquence fonctionnelle. Notre objectif était de prédire si une nouvelle protéine de la MhcSF se lie ou non à la B2M, et de comprendre pourquoi. Les enjeux biologiques de cette approche de classification supervisée sont la détection de mutants pathologiques dont l'expression cellulaire défecteuse est liée à une absence d'interaction avec la B2M, l'annotation automatique des nouvelles protéines de la MhcSF issues du séquençage des génomes de vertébrés inférieurs, et la prédiction de ligands. Certains récepteurs des cellules NK reconnaissent en effet spécifiquement la B2M (Michaëlsson et al. 2001) ; en indiquant qu'une protéine se lie à la B2M, nous pourrons donc prédire ce type d'interaction.

Le premier objectif était de standardiser les protéines des IgSF et MhcSF à partir des règles existantes, ou en établissant des règles adaptées. Nous avons d'une part démontré que les numérotations des V-DOMAINs et des C-DOMAINs étaient applicables aux V-LIKE-DOMAINs et C-LIKE-DOMAINs, et d'autre part mis en place la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs ; nous avons pour cela développé une procédure d'alignement de ces domaines protéiques basée sur la combinaison d'alignements de séquences et structures 3D, particulièrement adaptée à leur hétérogénéité de séquences.

A partir de ces séquences standardisées, nous avons établi la phylogénie des protéines de la MhcSF et caractérisé leurs contacts protéine-ligand. Ces analyses nous ont permis d'une part de mettre en évidence l'antériorité évolutive de la spécialisation des protéines de la MhcSF sur la spéciation et la perte partielle de la capacité de liaison à la B2M au cours de l'évolution, et d'autre part de définir la zone de contact potentiel des protéines de la MhcSF avec la B2M. Nous avons ainsi déterminé les contraintes inhérentes aux données et à notre problématique de classification des protéines de la MhcSF selon leur interaction ou non avec la B2M.

Un état de l'art des principales méthodes de classification supervisée et non supervisée appliquées à la classification fonctionnelle de familles protéiques nous a alors permis de développer une approche efficace, adaptée à nos données. Cette approche supervisée combine un classifieur Bayesien naïf et la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs. Elle comprend deux étapes : la sélection d'un ensemble de descripteurs binaires discriminants (qui associent une position dans l'alignement et un groupe d'acides aminés), et la construction du classifieur par estimation des fréquences de ces descripteurs conditionnellement aux classes que l'on cherche à séparer. Cette approche est performante quelle que soit la similarité de la séquence à classer avec les séquences du jeu de données. Le caractère explicatif de ce classifieur nous permet d'identifier les propriétés physico-chimiques

qui, observées à une position donnée des G-DOMAINs ou G-LIKE-DOMAINs, favorisent ou défavorisent l'interaction des protéines de la MhcSF à la B2M, et d'en donner une interprétation structurale. Ces informations pourraient s'avérer cruciales pour de futures expériences de mutagenèse dirigée. Enfin, l'utilisation de notre classifieur comme outil prédictif nous permet d'émettre des hypothèses concernant l'interaction des protéines du MHC de vertébrés inférieurs avec la B2M. Cette méthode devrait s'appliquer avec succès à d'autres problématiques de classification des protéines des IgSF et MhcSF pour lesquelles on dispose de classes connues a priori, et plus généralement à la classification fonctionnelle de familles protéiques à partir d'un alignement multiple.

Certains chapitres du manuscrit sont associés à un ou plusieurs articles publiés ou en cours de publication. Les différents chapitres sont les suivants : Le **chapitre 1** présente le système immunitaire adaptatif, les règles de DESCRIPTION des IG, TR, MHC et les règles de NUMEROTATION de leurs domaines protéiques. Le **chapitre 2** met en évidence la diversité fonctionnelle des protéines des superfamilles IgSF et MhcSF, et présente la standardisation IMGT de leurs récepteurs, chaînes et domaines que nous avons obtenue à partir des règles existantes pour les IG, TR et MHC. Le **chapitre 3** détaille la procédure que nous avons mise en place pour l'alignement des G-DOMAINs et G-LIKE-DOMAINs et la standardisation des protéines de la MhcSF, et l'analyse évolutive et structurale de ces données standardisées. Le **chapitre 4** présente d'une part les principales méthodes de classification supervisée et non supervisée appliquées à la classification fonctionnelle de familles protéiques, et d'autre part la méthodologie que nous avons développée et nos résultats concernant la classification des protéines de la MhcSF selon leur interaction ou non à la B2M.

CHAPITRE 1

Les protéines du système immunitaire adaptatif

Le système immunitaire assure l'intégrité des organismes multicellulaires en les protégeant des agents pathogènes. L'immunité non spécifique (ou innée) est constituée des barrières naturelles anatomiques, physiologiques, phagocytaires et inflammatoires. Le système immunitaire spécifique (ou adaptatif, observé à ce jour uniquement chez les vertébrés) est impliqué dans la reconnaissance et l'élimination spécifiques des agents pathogènes, et comprend l'immunité humorale et cellulaire ; ce système est basé sur des interactions de type protéine-ligand impliquant principalement les immunoglobulines (IG), les récepteurs T (TR) et les protéines du complexe majeur d'histocompatibilité (MHC) (Fig. 1.1).

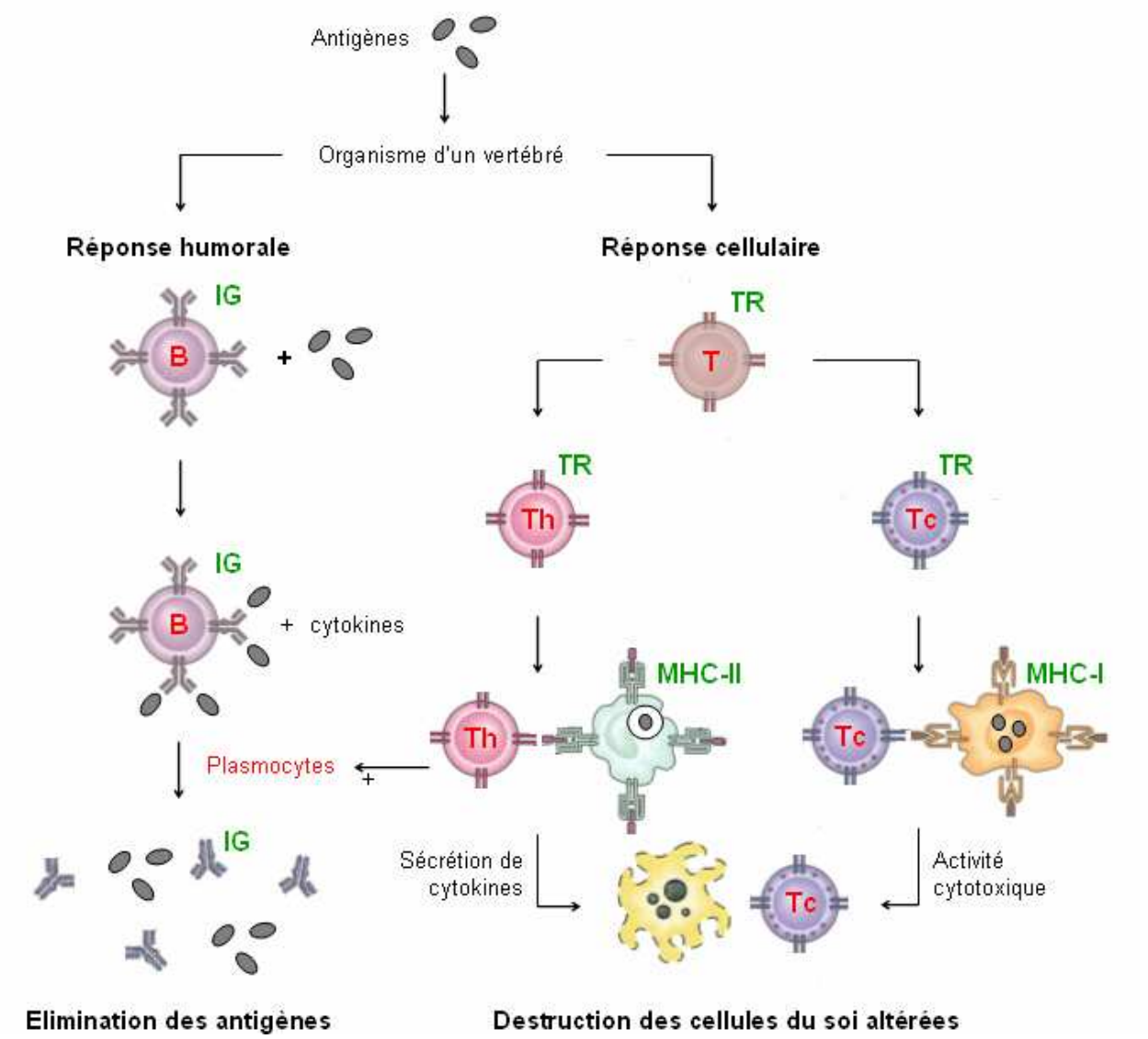


Figure 1.1. Les protéines IG, TR et MHC et leurs interactions au sein du système immunitaire (modifié d'après Abbas AK. et al. Cellular and Molecular Immunology, Fourth Edition)**.** Les cellules : B (lymphocytes ou cellules B) et plasmocytes, T (lymphocytes ou cellules T), Tc (lymphocytes T cytotoxiques), Th (lymphocytes T auxiliaires) ; les protéines : IG (immunoglobulines), TR (récepteurs T), MHC-I et MHC-II (complexe majeur d'histocompatibilité de classes I et II).

Ces protéines sont constituées d'un ou plusieurs récepteurs aux antigènes ; chaque récepteur est un hétérodimère de chaînes polypeptidiques comprenant différents domaines protéiques. Les concepts d'IMGT-ONTOLOGY (Annexe 1, Giudicelli and Lefranc 1999, Lefranc et al. 2004) fournissent une classification sémantique des connaissances en immunogénétique ; plus particulièrement, les concepts de DESCRIPTION et de NUMEROTATION fournissent les termes et règles nécessaires respectivement à la description des récepteurs, chaînes et domaines des IG, TR et MHC des vertébrés, et à la numérotation des acides aminés des séquences et structures 3D de leurs domaines (Lefranc 1999).

Ce chapitre présente tout d'abord le système immunitaire adaptatif et le rôle des IG, TR et MHC dans la reconnaissance antigénique. Les IG et TR sont constitués de domaines protéiques de structure similaire, et nous détaillons donc par la suite les règles de DESCRIPTION de ces protéines et les règles de NUMEROTATION de leurs domaines. Enfin, nous détaillons les règles de DESCRIPTION des protéines du MHC et les règles de NUMEROTATION de leurs domaines. Comme nous le verrons dans la suite de cette thèse, cette standardisation des acteurs moléculaires du système immunitaire adaptatif est essentielle à la comparaison de leurs caractéristiques de séquence, de structure et de fonction.

1.1 Système immunitaire adaptatif et reconnaissance antigénique

1.1.1 Immunité humorale : lymphocytes B et immunoglobulines

L'immunité humorale agit contre les bactéries et les virus circulant dans le sang et la lymphe, et est basée sur la reconnaissance spécifique d'épitopes (ou déterminants antigéniques) de ces agents pathogènes par les immunoglobulines (IG) ou anticorps.

Les IG sont sécrétées par les plasmocytes, qui représentent la forme de différenciation terminale des lymphocytes B. Les lymphocytes B se développent à partir de cellules souches dans la moelle osseuse (organe lymphoïde primaire). Des mécanismes de réarrangement de l'ADN génèrent une grande diversité de lymphocytes B (10^{12} différents chez l'homme) ; chaque lymphocyte B généré est unique et exprime à sa surface des IG identiques, ayant par conséquent la même spécificité. La maturation des lymphocytes B comprend une sélection négative : les lymphocytes B qui expriment des IG spécifiques d'antigènes du soi sont éliminés. Les lymphocytes B matures circulent alors dans la lymphe et gagnent les organes lymphoïdes secondaires (ganglions lymphatiques, rate).

Dans les ganglions lymphatiques, la reconnaissance spécifique d'un antigène par un lymphocyte B mature induit sa division cellulaire (ou expansion clonale) et la différenciation

des clones en plasmocytes (qui sécrètent des IG) et en lymphocytes B mémoires. Les IG circulantes aboutissent à la destruction spécifique de l'agent pathogène ; cette destruction est liée à la formation du complexe immun (complexe antigène-anticorps), qui favorise la phagocytose du pathogène par les macrophages. Les lymphocytes B mémoires expriment des IG membranaires et ont une durée de vie beaucoup plus longue que les plasmocytes ; ils pourront être activés et se différencier en plasmocytes lors d'une nouvelle rencontre avec l'antigène.

1.1.2 Immunité cellulaire : lymphocytes T et récepteurs T

L'immunité cellulaire agit contre les cellules infectées par des virus ou des bactéries, par l'interaction spécifique des récepteurs T (TR) des lymphocytes T avec les protéines du MHC présentant des peptides endogènes ou exogènes (Fig. 1.1).

Les protéines du MHC (identifiées par Gorer en 1936) se subdivisent en deux classes. Les protéines MHC de classe I (MHC-I) sont exprimées à la surface de la majorité des cellules de l'organisme. Elles ont pour fonction de présenter aux lymphocytes T CD8+ (pour la plupart cytotoxiques) des peptides issus de la dégradation de protéines cytosoliques par le protéasome ; ces peptides endogènes sont liés aux protéines du MHC-I dans le réticulum endoplasmique. Les protéines MHC de classe II (MHC-II) sont exprimées à la surface des cellules présentatrices d'antigènes (CPA) professionnelles (macrophages, cellules dendritiques, lymphocytes B). Elles ont pour fonction de présenter aux lymphocytes T CD4+ (pour la plupart helper ou auxiliaires) des peptides issus des vésicules d'endocytose ; ces peptides exogènes sont liés aux protéines du MHC-II dans les endosomes.

Les récepteurs T assurent la reconnaissance spécifique des cellules du soi infectées. Ils sont exprimés à la surface des lymphocytes T, qui se développent dans la moelle osseuse à partir de cellules souches et arrivent à maturation dans le thymus. Des mécanismes de réarrangement de l'ADN (analogues à ceux à l'origine de la diversité des lymphocytes B) génèrent une grande diversité de lymphocytes T (environ 10^{12} différents chez l'homme) ; chaque lymphocyte T est unique et exprime des TR identiques de même spécificité. Leur maturation consiste en une sélection négative qui élimine les lymphocytes T spécifiques des peptides du soi, et en une sélection positive qui entraîne la prolifération des lymphocytes T spécifiques de peptides du non soi. Les peptides reconnus par les lymphocytes T sont présentés liés aux protéines du MHC. Une cellule saine, qui présente à sa surface des protéines du MHC liées à des peptides du soi, ne déclenchera donc aucune réaction immunitaire, tandis qu'une cellule étrangère qui présente des protéines du MHC différentes

sera rejetée (cas des greffes allogéniques), et qu'une cellule tumorale ou infectée par un virus dont le MHC présente des peptides étrangers entraînera une réponse immunitaire. Les lymphocytes T matures qui circulent dans la lymphe portent des corécepteurs CD4 ou CD8. La majorité des lymphocytes T CD8+ sont cytotoxiques et reconnaissent les peptides présentés par les protéines du MHC-I ; ils sont restreints au MHC-I. La majorité des lymphocytes T CD4+ reconnaissent les peptides présentés par les protéines du MHC-II et sont auxiliaires ou helper ; ils sont restreints au MHC-II.

L'interaction TR/peptide/MHC 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 ou auxiliaires) et en lymphocytes T mémoires. Les lymphocytes T auxiliaires sécrètent des cytokines qui stimulent la transformation des lymphocytes B en plasmocytes (immunité humorale) et l'activité des lymphocytes T cytotoxiques (immunité cellulaire), et aboutissent ainsi à la destruction de la cellule infectée.

Les lymphocytes B et T sont les principaux acteurs cellulaires de l'immunité à médiation humorale et cellulaire. L'ADN de ces cellules matures diffère de l'ADN de la lignée germinale ; le réarrangement génique est une caractéristique essentielle de la maturation des lymphocytes, et n'a jamais été observé dans d'autres types de cellules de vertébrés. Les acteurs moléculaires de la reconnaissance antigénique sont les IG, les TR, et les protéines du MHC (MHC-I et MHC-II). Ces protéines sont les garants de l'intégrité des organismes de vertébrés : elles permettent la reconnaissance spécifique des agents pathogènes, l'induction de leur destruction et la préservation des cellules et molécules du soi.

1.2 IG et TR : DESCRIPTION et NUMEROTATION

Chez l'homme, les IG correspondent à 5 classes ou isotypes IgM, IgD, IgG, IgA et IgE, les IgG et IgA étant elles-même divisées en sous-classes IgG1, IgG2, IgG3, IgG4, IgA1 et IgA2 (Lefranc and Lefranc 2001a). Ces classes et sous-classes diffèrent par leurs propriétés physico-chimiques et leurs fonctions biologiques (Burton 1987, Lefranc and Lefranc 1990). Il existe deux types de TR : alpha-beta (les plus représentés) et gamma-delta (Lefranc and Lefranc 2001b). Les TR gamma-delta ne sont restreints ni par les protéines du MHC ni par les peptides qu'elles présentent (Porcelli et al. 1991, Allison and Garboczi 2001).

1.2.1 DESCRIPTION des récepteurs, chaînes et domaines protéiques

Les IG comportent quatre chaînes polypeptidiques, constituant deux récepteurs aux antigènes décrits en Table 1.1. Chaque récepteur comprend une chaîne lourde transmembranaire

Récepteur	Chaîne	Domaine
IG-ALPHA_KAPPA	H-ALPHA	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-ALPHA-1_KAPPA	H-ALPHA-1	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-ALPHA-2_KAPPA	H-ALPHA-2	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-DELTA_KAPPA	H-DELTA	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-EPSILON_KAPPA	H-EPSILON	VH, CH1, CH2, CH3, CH4
	L-KAPPA	V-KAPPA, C-KAPPA
IG-GAMMA-1_KAPPA	H-GAMMA-1	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-GAMMA-2_KAPPA	H-GAMMA-2	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-GAMMA-2-A_KAPPA	H-GAMMA-2-A	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-GAMMA-2-B_KAPPA	H-GAMMA-2-B	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-GAMMA-2-C_KAPPA	H-GAMMA-2-C	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-GAMMA-3_KAPPA	H-GAMMA-3	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-GAMMA-4_KAPPA	H-GAMMA-4	VH, CH1, CH2, CH3
	L-KAPPA	V-KAPPA, C-KAPPA
IG-MU_KAPPA	H-MU	VH, CH1, CH2, CH3, CH4
	L-KAPPA	V-KAPPA, C-KAPPA
IG-ALPHA_LAMBDA	H-ALPHA	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-ALPHA-1_LAMBDA	H-ALPHA-1	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-ALPHA-2_LAMBDA	H-ALPHA-2	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-DELTA_LAMBDA	H-DELTA	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-EPSILON_LAMBDA	H-EPSILON	VH, CH1, CH2, CH3, CH4
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-GAMMA-1_LAMBDA	H-GAMMA-1	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-GAMMA-2_LAMBDA	H-GAMMA-2	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-GAMMA-2-A_LAMBDA	H-GAMMA-2-A	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-GAMMA-2-B_LAMBDA	H-GAMMA-2-B	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-GAMMA-2-C_LAMBDA	H-GAMMA-2-C	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-GAMMA-3_LAMBDA	H-GAMMA-3	VH, CH1, CH2, CH3
	L-LAMBDA	V-LAMBDA, C-LAMBDA
IG-MU_LAMBDA	H-MU	VH, CH1, CH2, CH3, CH4
	L-LAMBDA	V-LAMBDA, C-LAMBDA

Table 1.1. DESCRIPTION des récepteurs, chaînes et domaines d'IG de mammifères.

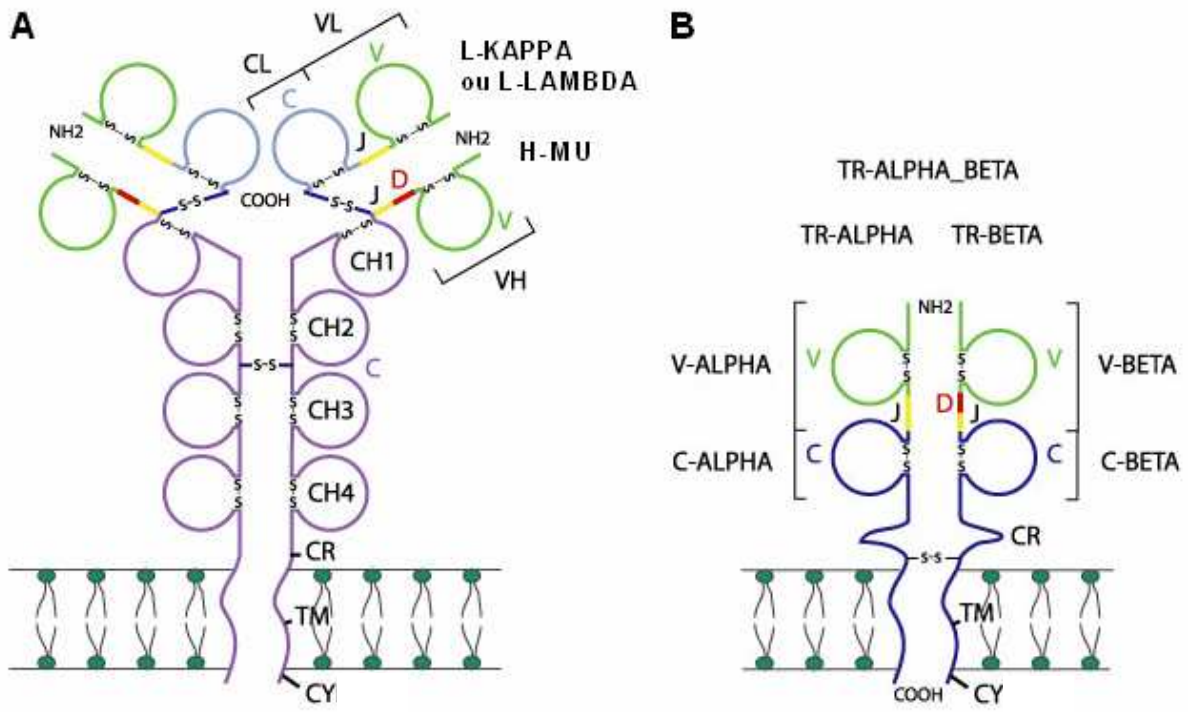


Figure 1.2. Représentation schématique d'une IG et d'un TR. (A) Homodimère de récepteurs IG-MU-KAPPA ou IG-MU_LAMBDA comportant respectivement les chaînes L-KAPPA et H-MU ou L-LAMBDA et H-MU. Chaque chaîne légère comporte deux domaines : VL (V-KAPPA ou V-LAMBDA) et CL (C-KAPPA ou L-LAMBDA) ; chaque chaîne lourde comporte les domaines VH, CH1, CH2, CH3, CH4, et les régions CR, TM et CY (respectivement CONNECTING-REGION, TRANSMEMBRANE-REGION et INTRACYTOPLASMIC-REGION). V, D et J indiquent les régions du gène réarrangé pour chaque chaîne (V-D-J pour les chaînes lourdes et V-J pour les chaînes légères), respectivement V-REGION, D-REGION et J-REGION ; C indique la C-REGION. La dégradation protéolytique d'un récepteur IG-MU-KAPPA ou IG-MU_LAMBDA génère respectivement les récepteurs FAB-MU_KAPPA (chaînes VH-CH1 et L-KAPPA) et FC-MU (chaîne CH2-CH3-CH4), et FAB-MU_LAMBDA (chaînes VH-CH1 et L-LAMBDA) et FC-MU. (B) Récepteur TR-ALPHA_BETA comportant les chaînes TR-ALPHA et TR-BETA. La chaîne TR-ALPHA comprend les domaines V-ALPHA et C-ALPHA ; la chaîne TR-BETA comprend les domaines V-BETA et C-BETA. V, D et J indiquent les régions du gène réarrangé pour chaque chaîne (V-D-J pour TR-BETA et V-J pour TR-ALPHA), respectivement V-REGION, D-REGION et J-REGION ; C indique la C-REGION.

distincte selon la classe d'IG et une chaîne légère L-KAPPA ou L-LAMBDA (Fig. 1.2). Ces chaînes sont définies par leur poids moléculaire : respectivement 50 kilodaltons (KDa) ou plus et 25 KDa (Edelman et al. 1969). La chaîne légère est liée à la chaîne lourde par un pont disulfure (liaison covalente forte), et par un ensemble de liaisons non covalentes telles que les liaisons salines, les liaisons hydrogènes et les interactions hydrophobes. Les deux récepteurs sont liés par des interactions similaires. La digestion enzymatique d'une IG par la papaïne produit 2 fragments Fab (« antigen binding ») et 1 fragment Fc (« cristallisable ») (Porter 1959) ; ces fragments sont décrits comme des récepteurs au sein d'IMGT (Annexe 2). Les chaînes lourdes sont constituées d'un V-DOMAIN VH (représenté par une V-D-J-REGION, issue du réarrangement des V-GENE, D-GENE et J-GENE de la lignée germinale, Fig. 1.3), de trois ou quatre C-DOMAINs et des CONNECTING-REGION, TRANSMEMBRANE-REGION et INTRACYTOPLASMIC-REGION (Fig. 1.2 et 1.3). Chaque C-DOMAIN est codé par un exon du C-GENE exprimé (Fig. 1.3) ; le type de chaîne lourde et de récepteur

dépend du C-GENE exprimé. Les C-DOMAINs des chaînes lourdes ont des propriétés effectrices : leur reconnaissance spécifique par les Fc récepteurs établit un pont entre les réponses humorales et cellulaires du système immunitaire (Padlan 1994). Les chaînes légères comprennent un V-DOMAIN V-KAPPA ou V-LAMBDA (codé par une V-J-REGION issue du réarrangement des V-GENE et J-GENE de la lignée germinale, Fig. 1.3) et un C-DOMAIN C-KAPPA ou C-LAMBDA (Fig. 1.2).

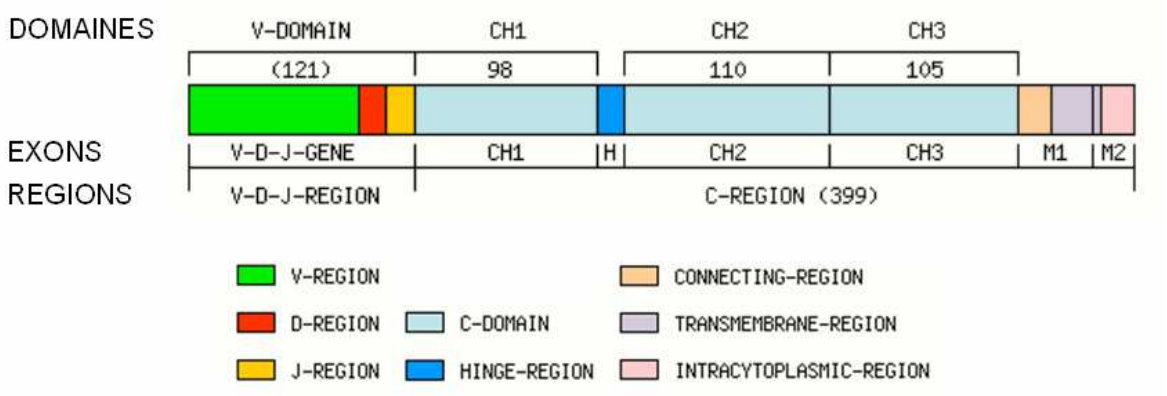


Figure 1.3. Correspondance entre les domaines protéiques d'une chaîne H-GAMMA-1 membranaire et les exons et régions du gène réarrangé. La taille des domaines et des régions est en nombre d'acides aminés.

Les TR sont des hétérodimères de chaînes transmembranaires, distinctes selon le type de TR (alpha-beta ou gamma-delta) ; chaque chaîne comprend un V-DOMAIN (représenté par une V-D-J-REGION ou par une V-J-REGION), un C-DOMAIN et les CONNECTING-REGION, TRANSMEMBRANE-REGION et INTRACYTOPLASMIC-REGION (Table 1.2, Fig. 1.2). Les deux chaînes d'un récepteur sont le plus souvent reliées entre elles par un pont disulfure et des liaisons non covalentes.

Récepteur	Chaîne	Domaine
TR-ALPHA_BETA-1	TR-ALPHA TR-BETA-1	V-ALPHA, C-ALPHA V-BETA, C-BETA-1
TR-ALPHA_BETA-2	TR-ALPHA TR-BETA-2	V-ALPHA, C-ALPHA V-BETA, C-BETA-2
TR-GAMMA-1_DELTA	TR-GAMMA-1 TR-DELTA	V-GAMMA, C-GAMMA-1 V-DELTA, C-DELTA
TR-GAMMA-2_DELTA	TR-GAMMA-2 TR-DELTA	V-GAMMA, C-GAMMA-2 V-DELTA, C-DELTA
TR-GAMMA-3_DELTA	TR-GAMMA-3 TR-DELTA	V-GAMMA, C-GAMMA-3 V-DELTA, C-DELTA
TR-GAMMA-4_DELTA	TR-GAMMA-4 TR-DELTA	V-GAMMA, C-GAMMA-4 V-DELTA, C-DELTA
TR-GAMMA-5_DELTA	TR-GAMMA-5 TR-DELTA	V-GAMMA, C-GAMMA-5 V-DELTA, C-DELTA

Table 1.2. DESCRIPTION des récepteurs, chaînes et domaines de TR des vertébrés.

Un V-DOMAIN est une unité structurale caractérisée par un repliement en sandwich beta : 9 brins beta antiparallèles (A, B, C, C', C'', D, E, F et G) sont organisés en deux feuillets (Fig. 1.4) (Lefranc and Lefranc 2001a, 2001b, Lefranc et al. 2003). Les feuillets ABED et GFCC'C'' sont maintenus rapprochés par un pont disulfure entre deux cystéines des brins B et F, et par le cœur hydrophobe du domaine (Lesk and Chothia 1982, Chothia et al. 1998, Pommié et al. 2004). La variabilité des IG et TR est essentiellement localisée au niveau de 3 boucles structurales de leurs V-DOMAINs, impliquées dans la reconnaissance des antigènes (Fig. 1.4 et 1.5) ; ces boucles sont nommées CDR (complementarity determining region), et relient respectivement les brins BC (CDR1), C'C'' (CDR2) et FG (CDR3).

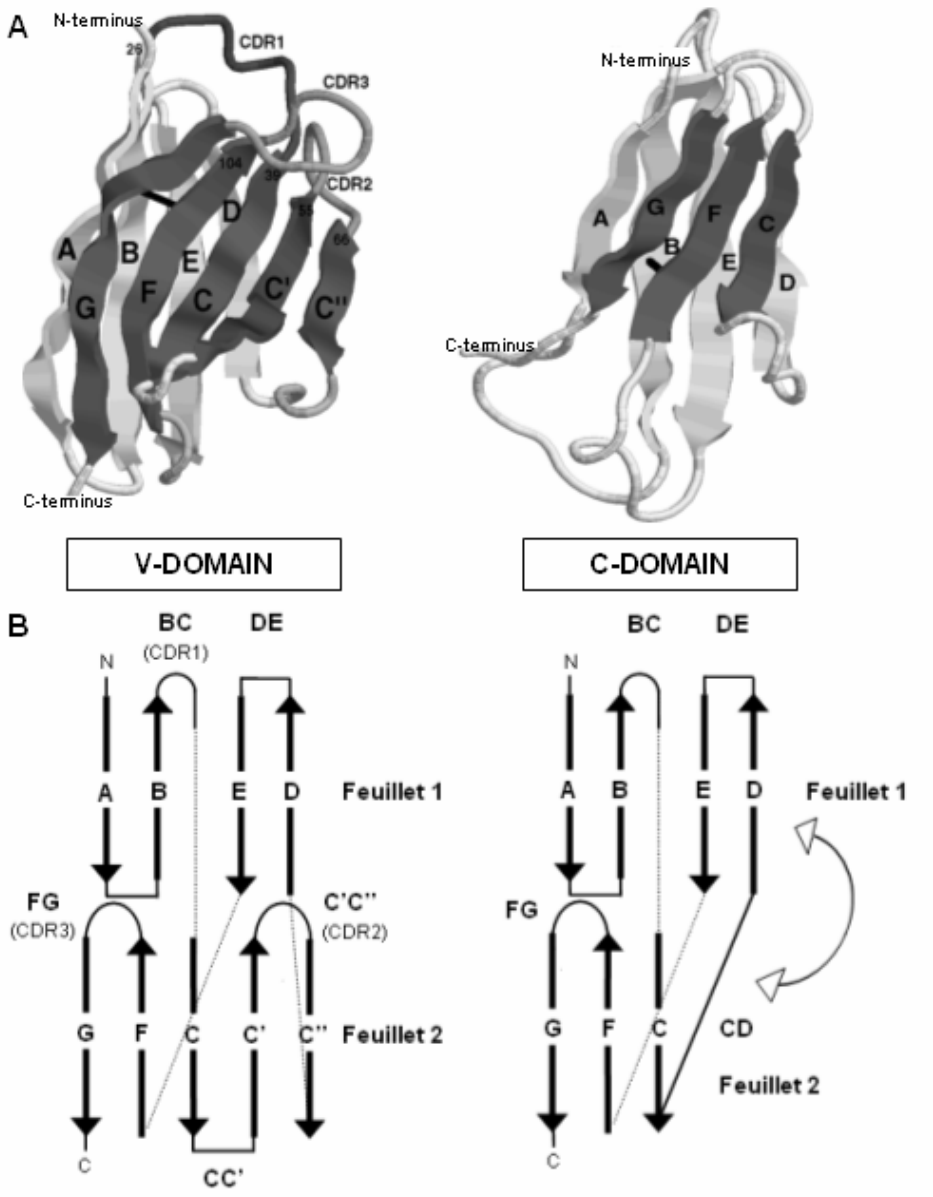


Figure 1.4. Structure 3D et représentation schématique des V-DOMAINs et C-DOMAINs des IG et TR (extrait de Lefranc and Lefranc 2001a, 2001b, Lefranc et al. 2005b, Publication 1). Les V-DOMAINs et C-DOMAINs sont constitués de 2 feuillets ; le pont disulfure entre les brins B et F est impliqué dans le maintien de la structure de ces domaines.

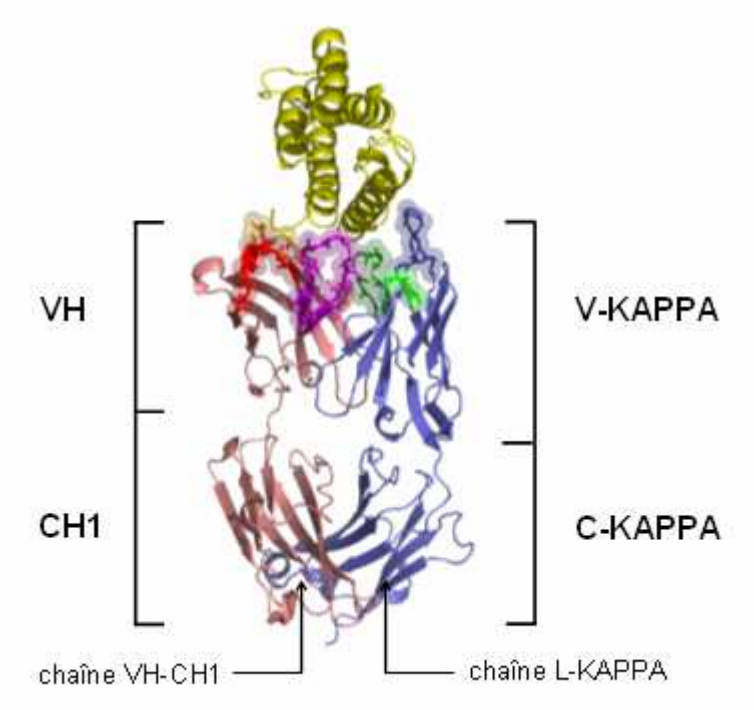


Figure 1.5. CDR-IMGT et reconnaissance de l'antigène par les IG. Structure 3D d'un homodimère de récepteurs FAB-GAMMA-1_KAPPA de *Mus musculus*, en interaction avec la protéine de la capside du virus HIV-1 (en jaune) (IMGT/3Dstructure-DB 1afv, Kaas et al. 2004). Les CDR-IMGT du domaine VH sont représentés en rouge (CDR1), orange (CDR2) et violet (CDR3) ; les CDR-IMGT du domaine V-KAPPA sont représentés en bleu (CDR1), vert clair (CDR2) et vert foncé (CDR3).

La variabilité des V-DOMAINs des IG et TR provient de la combinaison aléatoire de V-GENE, (D-GENE) et J-GENE (Lefranc and Lefranc 2001a, 2001b), de mécanismes d'ajout de N-nucléotides et de P-nucléotides lors des réarrangements de l'ADN de la lignée germinale (Alt and Baltimore 1982, Landau et al. 1984), et d'hypermutations somatiques dans le cas des IG (Gearhart et al. 1981, Lieber 2000).

Un C-DOMAIN est une unité structurale caractérisée par un repliement en sandwich beta, constitués de 7 brins beta antiparallèles (A, B, C, D, E, F et G) organisés en deux feuillets (ABE et GFCD ou ABED et GFC, le brin D pouvant appartenir à l'un ou l'autre des feuillets, Lefranc and Lefranc 2001a, 2001b, Lefranc et al. 2005b, Publication 1) (Fig. 1.4). Les C-DOMAINs ont une topologie et une structure tridimensionnelle similaire à ceux des V-DOMAINs, mais ils sont dépourvus des brins C' et C'' ; ces domaines étaient initiallement désignés sous le terme commun de domaines de type immunoglobulin (Williams and Barclay 1988, Bork et al. 1994). Les deux feuillets sont maintenus par un pont disulfure entre les brins B et F, caractéristique commune aux V-DOMAINs et C-DOMAINs.

1.2.2 NUMEROTATION des V-DOMAINs et C-DOMAINs

La numérotation unique des V-DOMAINs (Lefranc et al. 2003) a été établie par alignement multiple de plus de 5000 séquences protéiques d'IG et de TR de vertébrés ; un numéro est

attribué à chaque position des séquences alignées. Cette numérotation est unique car elle s'applique quel que soit le type de récepteur, le type de chaîne et l'espèce, et n'est donc pas modifiée lors de l'ajout de nouvelles séquences. Les caractéristiques de séquence et de structure peuvent donc être décrites par un identifiant unique et stable.

Les caractéristiques de séquence des V-DOMAINs sont deux cystéines conservées en positions 23 (1st-CYS) et 104 (2nd-CYS) (impliquées dans le pont disulfure caractéristique des V-DOMAINs), un tryptophane en position 41 (CONSERVED-TRP) et un acide aminé hydrophobe en position 89. Les caractéristiques de structure correspondent aux positions de début et de fin des brins (FR-IMGT) et des CDR (CDR-IMGT) ; la définition de leur longueur se base sur la longueur maximale observée au sein des séquences alignées. Chaque position est donc associée à une localisation structurale et parfois à une propriété physico-chimique. Ces relations sont visualisées sous forme de Colliers de Perles (Ruiz and Lefranc 2002) pour chaque séquence de V-DOMAIN (Fig. 1.6) ; cette représentation indique la topologie du domaine (1D) et son organisation sur deux feuillets (2D).

La numérotation unique des V-DOMAINs permet d'une part de décrire les mutations, le polymorphisme allélique et les hypermutations somatiques de façon standardisée, mais également de visualiser la topologie d'un domaine en l'absence de structure 3D.

La numérotation unique des C-DOMAINs a été établie par l'alignement de séquences de C-DOMAINs et de V-DOMAINs, et par l'identification de 72 positions structuralement équivalentes : 1-15 (A-STRAND), 16-26 (B-STRAND), 39-45 (C-STRAND), 77-84 (D-STRAND), 85-96 (E-STRAND), 97-104 (F-STRAND) et 118-128 (G-STRAND) (Lefranc et al. 2005b, Publication 1)

Les principales caractéristiques de séquence des C-DOMAINs sont deux cystéines conservées en positions 23 et 104 (impliquées dans le pont disulfure caractéristique des V-DOMAINs et C-DOMAINs), et deux acides aminés hydrophobes en positions 41 et 89. Les caractéristiques de structure des C-DOMAINs correspondent aux positions de début et de fin des brins (A, B, C, CD, D, E, F, G), des boucles (BC, FG), et des coudes (AB, DE, EF). Le brin transversal CD constitue la principale différence structurale entre les C-DOMAINs et les V-DOMAINs. Les zones d'insertions privilégiées au sein des C-DOMAINs comprennent la boucle FG, les coudes AB, DE, EF, et le brin transversal CD ; leurs positions sont décrites par des identifiants spécifiques, de telle sorte que la taille variable de ces zones n'affecte pas la numérotation des C-DOMAINs. Ces positions additionnelles sont : 1.1-1.9 à l'extrémité N-terminale du C-DOMAIN, 15.1-15.3 (AB-TURN), 45.1-45.9 (CD-STRAND), 84.1-84.7 et 85.7-85.1 (DE-TURN), 96.1-96.2 (EF-TURN), 111.1-111.6 et 112.6-112.1 (FG-LOOP).

La numérotation unique des C-DOMAINs permet de décrire les mutations et les sites d'interaction protéine-ligand de façon standardisée ; la représentation des séquences de C-DOMAINs sous forme de Colliers de Perles (Ruiz and Lefranc 2002) indique la topologie du domaine (1D) et son organisation sur deux feuillets (2D) (Fig. 1.7), en l'absence ou non de structure 3D.

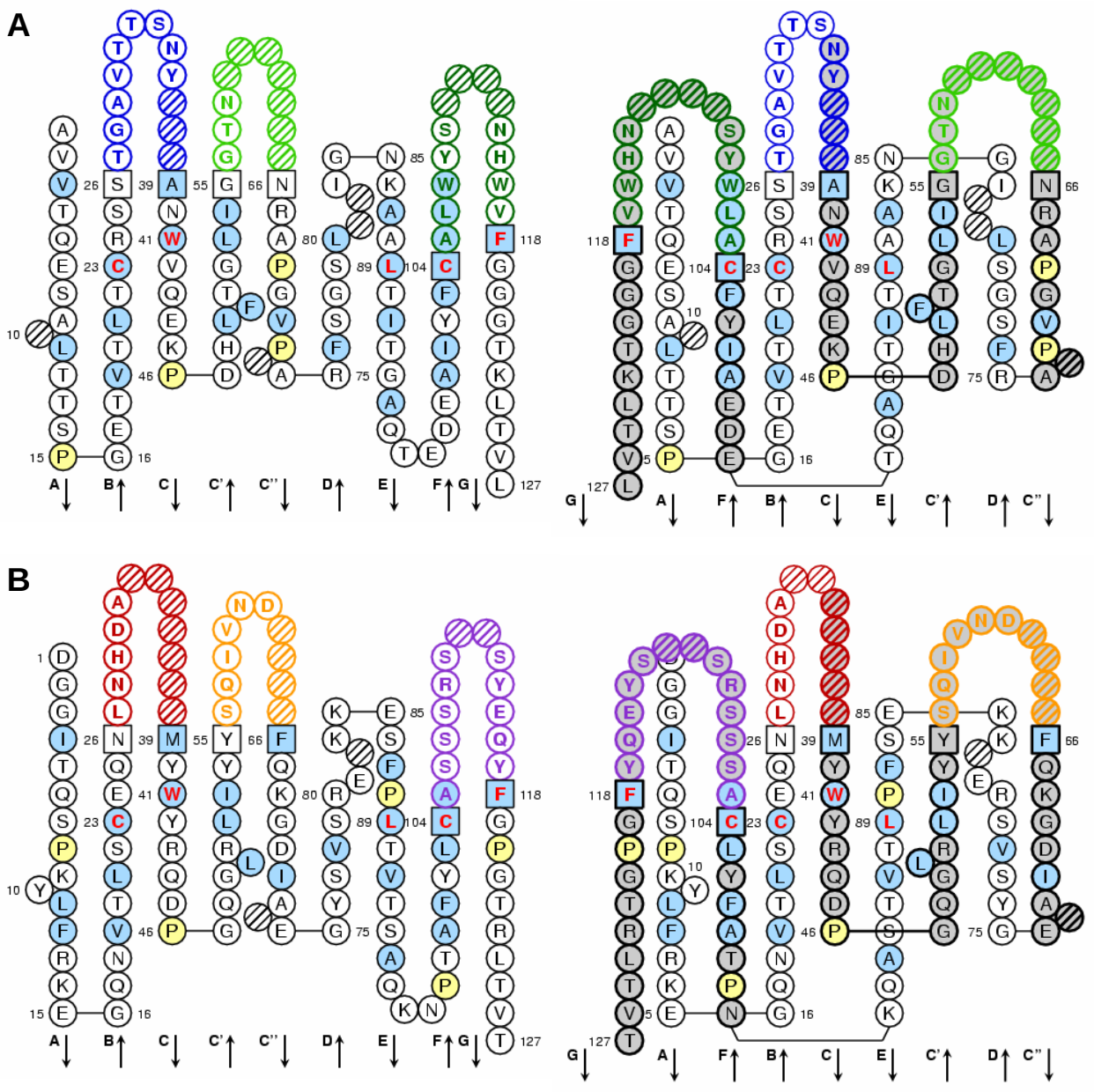


Figure 1.6. Colliers de Perles IMGT de V-DOMAINs sur un plan et deux plans. (A) Domaine V-LAMBDA d'une chaîne L-LAMBDA de *Mus musculus* (IMGT/3D-structure-DB 1a6u), représenté sur un et deux plans ; les CDR-IMGT sont représentés en bleu (CDR1), vert clair (CDR2) et vert foncé (CDR3). (B) Domaine V-BETA d'une chaîne TR-BETA-1 d'*Homo sapiens* (IMGT/3D-structure-DB 1oga), représenté sur un et deux plans ; les CDR-IMGT sont représentés en rouge (CDR1), orange (CDR2) et violet (CDR3). Les acides aminés hydrophobes (correspondant à une valeur positive de l'indice d'hydropathie, Kyte and Doolitle 1982) et les résidus Tryptophane présents dans plus de 50% des séquences alignées sont représentées en bleu. Les résidus Proline sont représentés en jaune. Les positions délimitant les CDR-IMGT sont représentées par des carrés ; ces positions appartiennent aux FR-IMGT. Les positions hachurées correspondent à des positions non occupées par rapport à la numérotation unique. Les flèches indiquent l'orientation des brins et leurs labels.

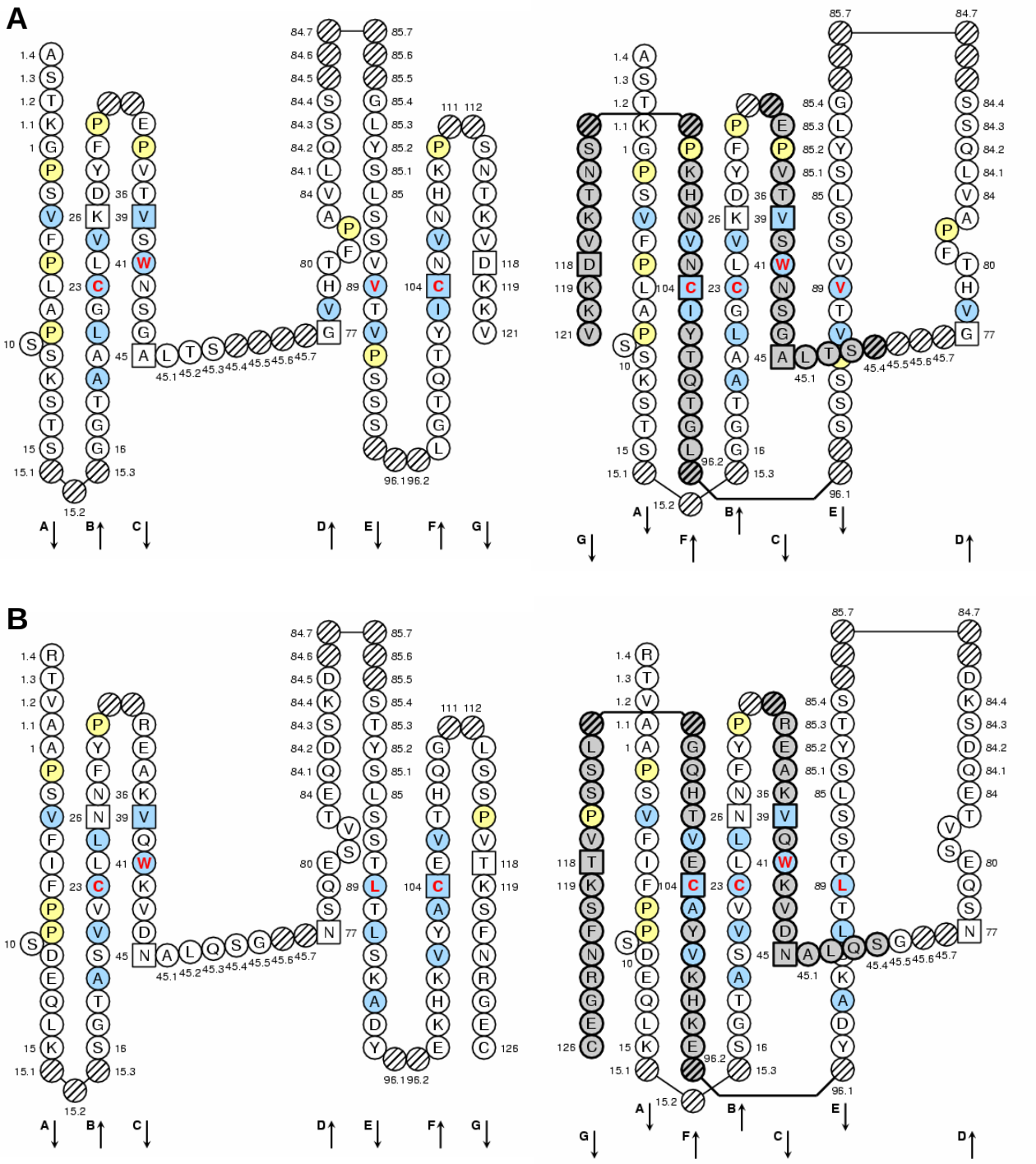


Figure 1.7. Colliers de Perles IMGT de C-DOMAINs sur un plan et deux plans (extrait de Lefranc et al. 2005b, Publication 1). (A) Domaine CH1 d'une chaîne H-GAMMA-1 d'*Homo sapiens* (IMGT/LIGM-DB J00228), représenté sur un et deux plans. (B) Domaine C-KAPPA d'une chaîne L-KAPPA d'*Homo sapiens* (IMGT/3D-structure-DB J00241), représenté sur un et deux plans. Les acides aminés hydrophobes (correspondant à une valeur positive de l'indice d'hydropathie, Kyte and Doolitle 1982) et les résidus Tryptophane présents dans plus de 50% des séquences alignées sont représentées en bleu. Les résidus Proline sont représentés en jaune. Les positions 26, 39, 104 et 118 sont représentées par des carrés par homologie avec les positions correspondantes dans les V-DOMAINs. Les positions hachurées correspondent à des positions non occupées par rapport à la numérotation unique. Les flèches indiquent l'orientation des brins et leurs labels.

1.3 MHC-I et MHC-II : DESCRIPTION et NUMEROTATION

Les protéines MHC-I et MHC-II qui présentent des peptides endogènes et exogènes aux cellules T appartiennent au MHC-I classique (MHC-Ia) ou au MHC-II classique (MHC-IIa). Le MHC-Ia comprend les sous-classes de protéines HLA-A, HLA-B et HLA-C chez l'homme, et les sous-classes H2-D, H2-K et H2-L chez la souris ; le MHC-IIa comprend les sous-classes HLA-DP, HLA-DQ et HLA-DR chez l'homme, et les sous-classes H2-A, H2-E et H2-P (non productive) chez la souris. Les voies d'assemblage des protéines MHC-Ia et MHC-IIa et des peptides (respectivement endogènes et exogènes) sont distinctes, tant par leur localisation cellulaire que par les protéines chaperones qu'elles impliquent (Duprat and Lefranc, IMGT Education MhcSF, http://imgt.cines.fr). Les protéines MHC-I non classiques (MHC-Ib) sont assemblées avec des peptides issus de la dégradation protéolytique des protéines MHC-Ia ; les protéines MHC-II non classiques (MHC-IIb) ne présentent pas de peptides aux cellules T. Le MHC-Ib comprend les sous-classes HLA-E, HLA-F et HLA-G chez l'homme, et les sous-classes H2-Q, H2-T et H2-M chez la souris ; le MHC-IIb comprend les sous-classes HLA-DM et HLA-DO chez l'homme, H2-DM et H2-DO chez la souris.

Il existe 11 sous-classes de MHC chez l'homme et chez la souris : 3 MHC-Ia, 3 MHC-IIa, 3 MHC-Ib et 2 MHC-IIb. Ces protéines sont codées par une famille multigénique localisée sur le bras court du chromosome 6 chez l'homme et 17 chez la souris ; cette région est la plus dense en gènes au sein du génome humain (The MHC sequencing consortium 1999, Horton et al. 2004). Les gènes MHC-Ia et MHC-IIa sont caractérisés par leur polymorphisme intraspécifique.

1.3.1 DESCRIPTION des récepteurs, chaînes et domaines protéiques

Les protéines MHC-I sont constituées de deux chaînes polypeptidiques liées de façon non covalente : une chaîne lourde transmembranaire (I-ALPHA) et la chaîne légère de la beta2-microglobuline (B2M) ; ce récepteur est nommé MHC-I-ALPHA_B2M. La liaison non covalente des chaînes I-ALPHA et B2M est nécessaire à l'assemblage peptide/MHC-I (Boyd et al. 1992), à la stabilisation de la conformation du complexe (Solheim et al. 1995) et à son expression à la surface cellulaire (Williams et al. 1989) ; l'échange de B2M intraspécifique et interspécifique à la surface cellulaire semble être un phénomène courant, mis en évidence *in vitro* (Bernabeu et al. 1984).

La chaîne lourde I-ALPHA est constituée de deux G-DOMAINs (G-ALPHA1 [D1] et G-ALPHA2 [D2]), d'un C-LIKE-DOMAIN [D3] et des régions : CONNECTING-REGION, TRANSMEMBRANE-REGION, INTRACYTOPLASMIC-REGION (Fig. 1.8 et 1.9,

Annexe 3) ; [D1], [D2] et [D3] désignent les domaines et leurs positions en partant de l'extrémité N-terminale de la chaîne polypeptidique. Chacun de ces domaines est lié de façon non covalente avec la chaîne B2M ; Collins et al. (1995) ont néanmoins mis en évidence que la déletion *in vitro* du C-LIKE-DOMAIN d'une chaîne I-ALPHA n'affecte pas sa capacité de liaison à la B2M ; la structure et la fonction de ce récepteur sont identiques à celles d'un récepteur MHC-I-ALPHA_B2M.

Les protéines MHC-II sont des hétérodimères constitués de deux chaînes transmembranaires II-ALPHA et II-BETA ; ce récepteur est nommé MHC-II-ALPHA_BETA. Chaque chaîne lourde de protéine MHC-II est constituée d'un G-DOMAIN (G-ALPHA [D1] pour la chaîne II-ALPHA, G-BETA [D1] pour la chaîne II-BETA) et d'un C-LIKE-DOMAIN [D2] (Fig. 1.8, Annexe 3) ; cette organisation modulaire empêche toute interaction des protéines MHC-II avec la B2M (Fig. 1.9).

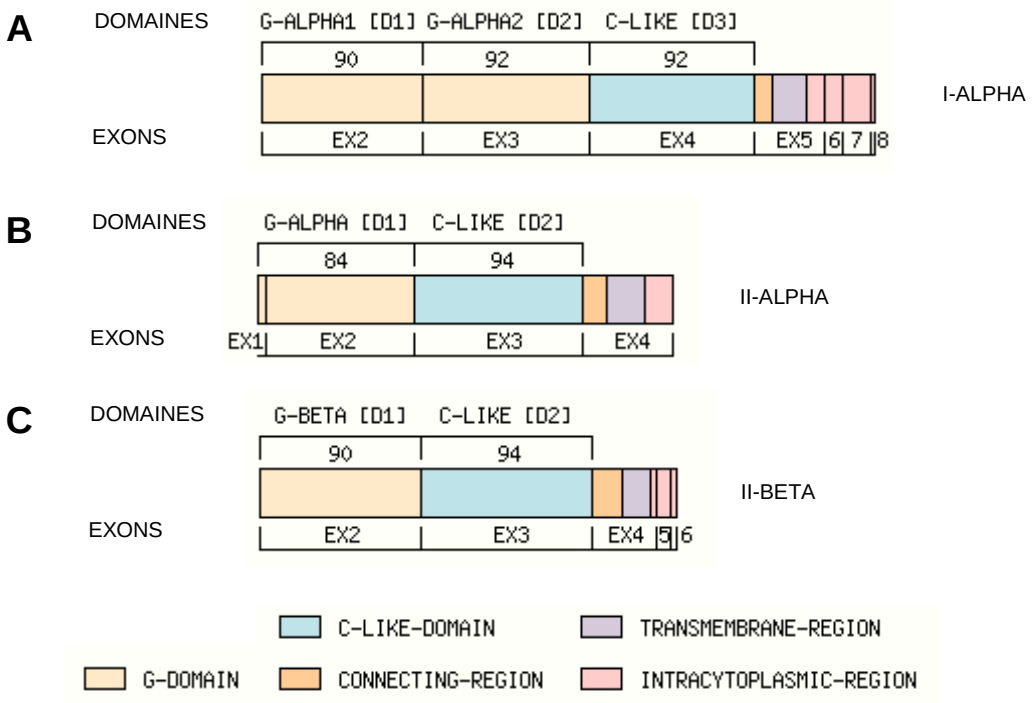


Figure 1.8. Correspondance entre les domaines et les exons pour MHC-I et MHC-II (extrait de Lefranc et al. 2005c, Publication 2). (A) Domaines de la chaîne I-ALPHA de la protéine MHC-I HLA-A d'*Homo sapiens* (EMBL/GenBanj/DDBJ K02883). (B) Domaines de la chaîne II-ALPHA de la protéine MHC-II HLA-DRA d'*Homo sapiens* (J00203, J00204). Le domaine G-ALPHA (84 acides aminés) est codé par l'extrémité 3' de EX1 (2 codons) et par EX2 (82 codons). (C) Domaines de la chaîne II-BETA de la protéine MHC-II HLA-DRB1 d'*Homo sapiens* (AL137064). La longueur des domaines est indiquée en nombre d'acides aminés. Hughes et Nei (1993) ont proposé un modèle de l'origine des protéines MHC-I et MHC-II, basé sur l'hypothèse de l'antériorité évolutive des chaînes II-ALPHA et II-BETA (Annexe 4).

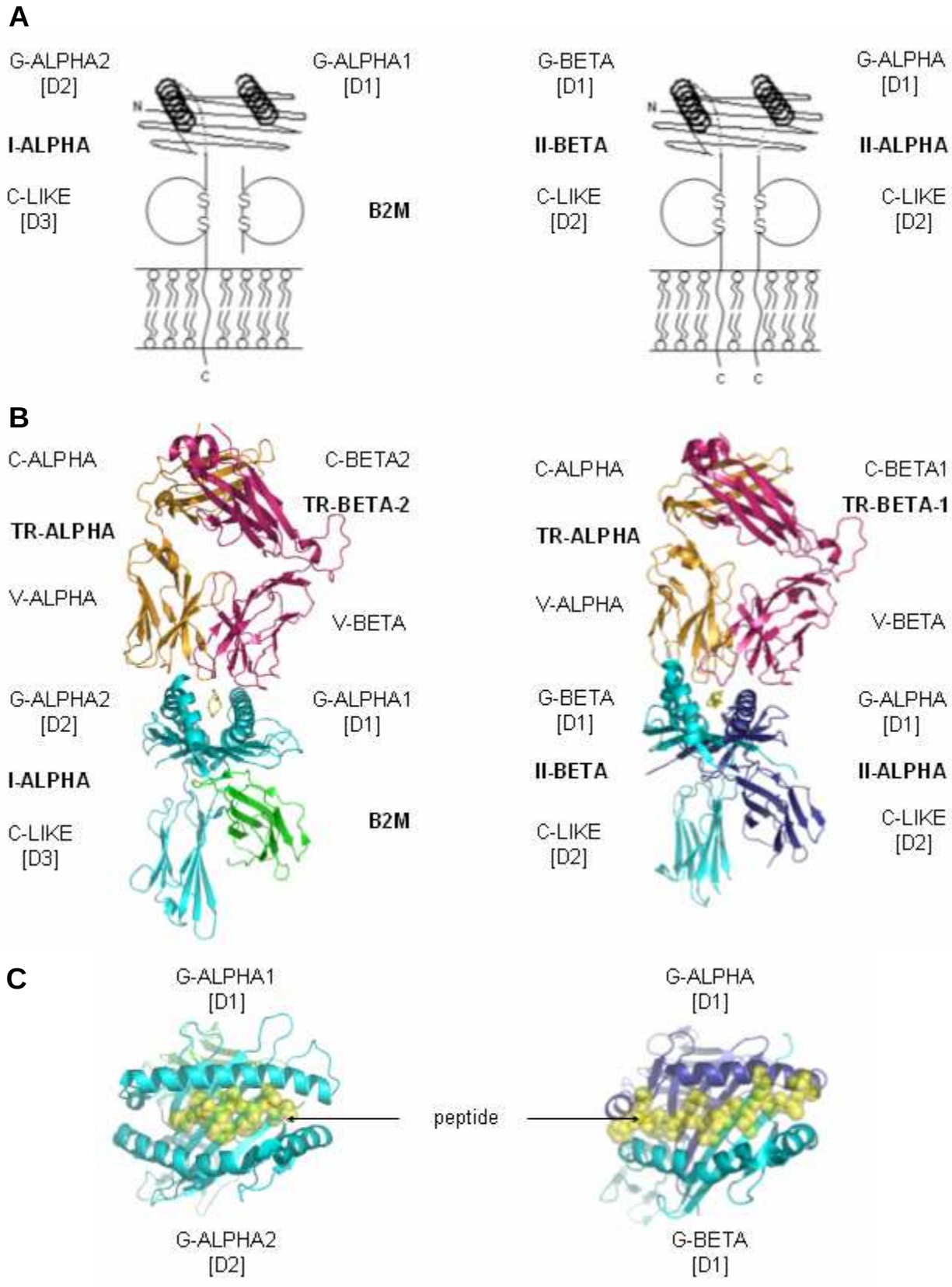


Figure 1.9. Représentation schématique et structure 3D des protéines MHC-I et MHC-II, en interaction ou non avec un TR. (A) Représentation schématique des protéines MHC-I et MHC-II à la surface d'une cellule cible. (B) Structure 3D de protéines MHC-I et MHC-II en interaction avec un TR. Les structures 3D correspondent aux récepteurs MHC-I-ALPHA_B2M et TR-ALPHA_BETA-2 en interaction (IMGT/3Dstructure-DB 1oga, à gauche), et MHC-II-ALPHA_BETA et TR-ALPHA_BETA-1 en interaction (1j8h, à droite). Le nom des chaînes est indiquée en gras. Les peptides localisés dans le sillon des protéines MHC-I et MHC-II sont représentés en jaune. (C) Vue supérieure du sillon des protéines MHC-I (à gauche) et MHC-II (à droite) présentant des peptides (IMGT/3Dstructure-DB 1oga et 1j8h). Les chaînes latérales des peptides sont représentées par des sphères.

Un G-DOMAIN est constitué d'un feuillet de quatre brins beta antiparallèles (A, B, C et D) et d'une longue hélice alpha. Les deux G-DOMAINs des récepteurs MHC-I-ALPHA_B2M et MHC-II-ALPHA_BETA forment un sillon, qui correspond au site de liaison des peptides (Fig. 1.9). Les hélices et les brins de ces G-DOMAINs constituent respectivement les « murs » et le « plancher » du sillon ; les 8 brins beta forment en effet un feuillet. Le sillon des protéines MHC-I et MHC-II diffère par la distance séparant les extrémités des deux hélices, qui conditionne la taille des peptides présentés (respectivement 8-10 et 10-15 acides aminés, Fig. 1.9). Les G-DOMAINs des protéines MHC-I et MHC-II interagissent avec les domaines V-ALPHA et V-BETA des TR lors de la présentation des peptides. Les domaines G-ALPHA2 et G-BETA sont caractérisés par un pont disulfure entre le brin A et l'hélice ; cette liaison est nécessaire au maintien de la structure du sillon (Saper et al. 1991).

1.3.2 NUMEROTATION des G-DOMAINs

La numérotation unique des G-DOMAINs a été établie par une combinaison d'alignements de séquences et de structures 3D des domaines G-ALPHA1 [D1], G-ALPHA2 [D2], G-ALPHA [D1] et G-BETA [D1] de chaînes I-ALPHA, II-ALPHA et II-BETA ; la faible similarité des séquences de ces domaines a nécessité la mise en place d'une procédure particulière, détaillée dans le chapitre 3. Cette numérotation est unique pour tous les G-DOMAINs, quels que soient le type de récepteur, le type de chaîne et l'espèce (Lefranc et al. 2005c, Publication 2).

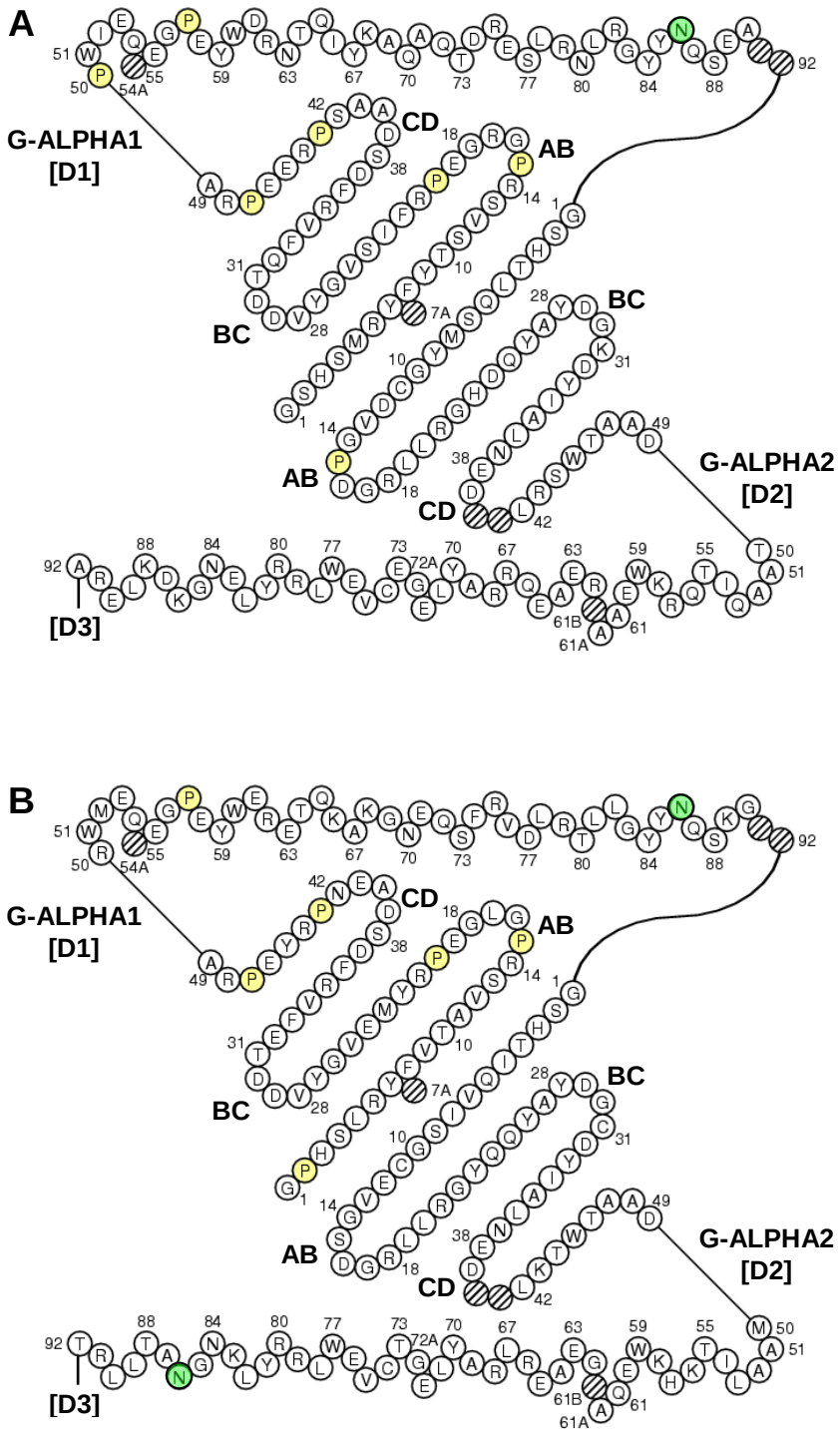
Les caractéristiques de structure des G-DOMAINs correspondent aux positions de début et de fin des brins (A-STRAND, B-STRAND, C-STRAND, D-STRAND) et de l'hélice (HELIX). Deux cyctéines sont conservées au sein des domaines G-ALPHA2 [D2] et G-BETA [D1], et sont impliquées dans un pont disulfure qui rapproche l'hélice du plancher du sillon : CYS-11 (A-STRAND) et CYS-74 (HELIX). Différents sites potentiels de glycosylation sont décrits : ASN-15 (A-STRAND) pour G-BETA [D1], et ASN-86 (HELIX) pour G-ALPHA1 [D1], G-ALPHA [D1], et G-ALPHA2 [D2] de *Mus musculus*.

Des positions additionnelles sont spécifiques de chaque type de G-DOMAIN et correspondent à des caractéristiques structurales locales :

- 7A (A-STRAND) décrit un bulge au niveau du brin A de domaines G-ALPHA [D1]
- 49.1-49.4 sont spécifiques de certains domaines G-BETA [D1], pour lesquels la région charnière entre le brin D et l'hélice est plus longue que pour les autres G-DOMAINs
- 61A et 72A (HELIX) décrivent deux cassures de l'hélice des domaines G-ALPHA2 [D2] et G-BETA [D1]

- 61B (HELIX) est spécifique des domaines G-BETA [D1] (à l'exception de celui de la chaîne II-BETA de la protéine H2-A), pour lesquels la cassure décrite en 61A s'étend sur deux acides aminés.
- 92A (HELIX) décrit le dernier acide aminé de certains domaines G-ALPHA [D1].

Les G-DOMAINs des protéines MHC-I et MHC-II sont représentés sous forme de Colliers de Perles sur un plan (Fig. 1.10).



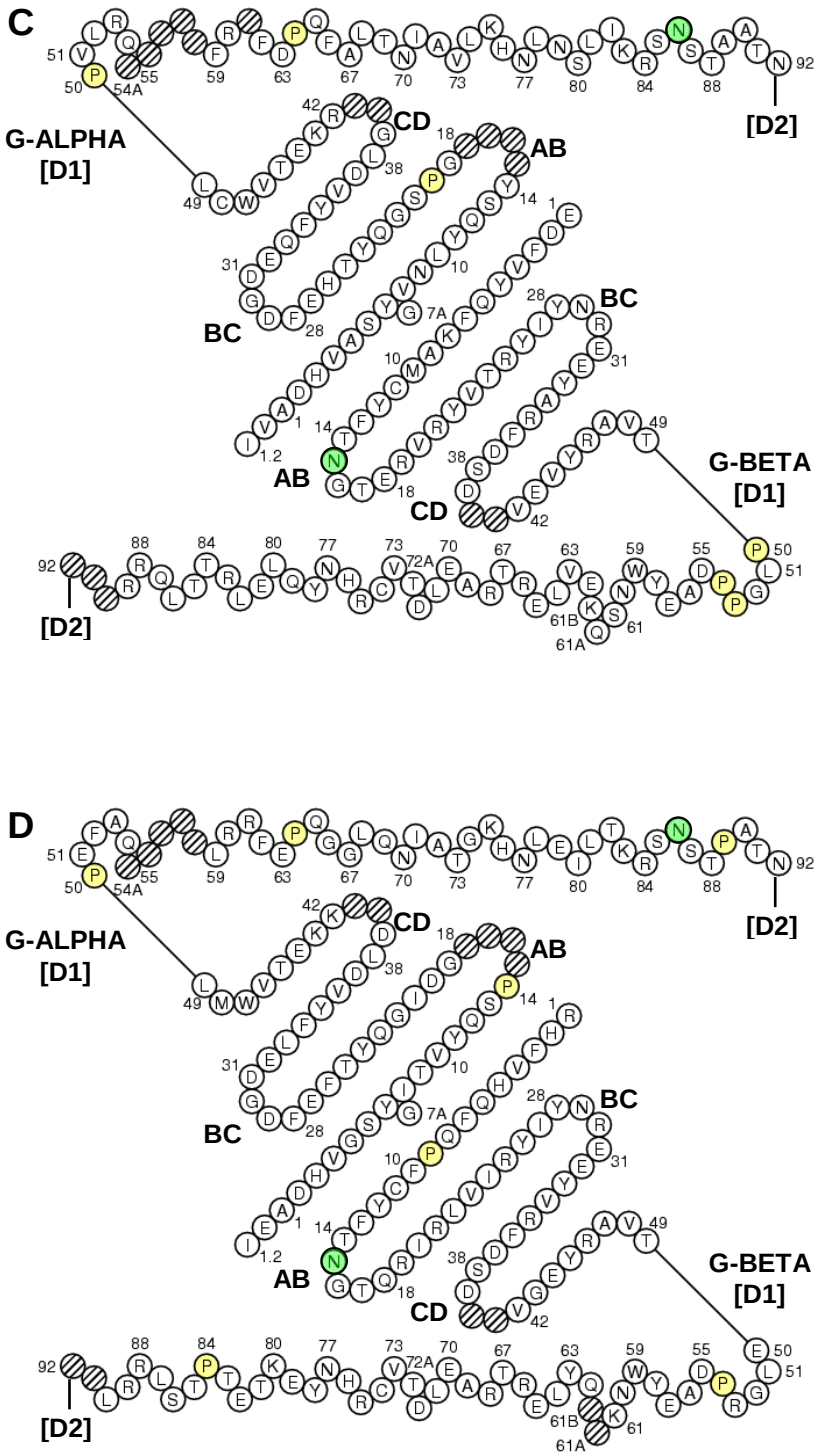


Figure 1.10. Colliers de Perles IMGT de G-DOMAINs (extrait de Lefranc et al. 2005c, Publication 2). Domaines G-ALPHA1 [D1] et G-ALPHA2 [D2] d'une chaîne I-ALPHA (A) d'*Homo sapiens* (HLA-B*0702) et (B) de *Mus musculus* (H2-K1*01). Domaines G-ALPHA [D1] et G-BETA [D1] de chaînes II-ALPHA et II-BETA (C) d'*Homo sapiens* (HLA-DQA1*0501/HLA-DQB1*0301) et (D) de *Mus musculus* (H2-AA*02/H2-AB*02). Les résidus Proline et les sites potentiels de N-glycosylation (N-X-S/T) sont représentés respectivement en jaune et vert. Les positions hachurées correspondent à des positions manquantes par rapport à la numérotation unique. L'extrémité N-terminale des peptides liés au niveau du sillon de ces domaines serait localisée à gauche sur cette représentation. [D3] indique la position du C-LIKE-DOMAIN dans la chaîne I-ALPHA ;.[D2] indique la position du C-LIKE-DOMAIN dans les chaînes II-ALPHA et II-BETA. La spécificité allélique des protéines MHC-I est localisée au niveau des acides aminés qui bordent le sillon (Matsumura et al. 1992) ; ces positions bordent également le sillon au niveau de cette représentation sous forme de Colliers de Perles.

CHAPITRE 2

Les superfamilles IgSF et MhcSF

Comme nous l'avons vu dans le chapitre précédent, les règles de standardisation issues d'IMGT-ONTOLOGY assurent la description de l'organisation modulaire des récepteurs et chaînes des IG, TR et MHC ; trois familles structurales de domaines protéiques sont définies : V-DOMAIN, C-DOMAIN et G-DOMAIN. Des domaines de structure similaire ont été identifiés au sein de protéines autres que les IG, TR et MHC, impliquées dans une grande variété de processus biologiques, appartenant ou non au système immunitaire ; ces domaines sont nommés V-LIKE-DOMAINs, C-LIKE-DOMAINs et G-LIKE-DOMAINs.

Nous définissons dans ce chapitre deux superfamilles de protéines : la superfamille des immunoglobulines (IgSF) et la superfamille du MHC (MhcSF) ; nous mettons en évidence la diversité fonctionnelle des protéines qui les composent, et nous démontrons que les règles de standardisation des IG, TR et MHC présentées dans le chapitre précédent sont appliquables à l'ensemble des protéines de ces superfamilles. Les caractéristiques de séquence, de structure et de fonction des protéines d'une même superfamille peuvent dorénavant être comparées.

2.1 La superfamille des immunoglobulines (IgSF)

Cette superfamille regroupe les IG et TR, et les protéines comprenant au moins un V-LIKE-DOMAIN ou C-LIKE-DOMAIN. Alors que les IG et TR sont spécifiques des vertébrés à mâchoire, les autres protéines de cette superfamille sont observées chez les vertébrés et invertébrés, chez les virus, les champignons et les plantes (Halaby and Mornon 1998, Du Pasquier 2001). Ces protéines sont essentiellement impliquées dans le système immunitaire (MHC-I, MHC-II, B2M, famille des Fc récepteurs, famille des protéines KIR) ou plus généralement dans des fonctions de structure et d'adhésion cellulaire (CEACAM1, MOG, MPZ, VCAM1) (Duprat and Lefranc, IMGT Education IgSF, http://imgt.cines.fr).

Nous avons montré à partir de quelques exemples que la numérotation unique des V-DOMAINs et des C-DOMAINs des IG et TR s'applique respectivement à tout V-LIKE-DOMAIN et à tout C-LIKE-DOMAIN (Duprat et al. 2004, Lefranc et al. 2005b, Publications 1 et 3). Les brins beta antiparallèles des V-LIKE-DOMAINs correspondent aux FR-IMGT décrits pour les V-DOMAINs. Des positions additionnelles 46A-C, 47A-B et 84A-C sont requises pour certains domaines, mais la numérotation unique est préservée. Les Colliers de Perles (Fig. 2.1 et 2.2) mettent en évidence les caractéristiques de séquence et de structure des

V-LIKE-DOMAINs et C-LIKE-DOMAINs. Cette standardisation nous a également permis de décrire le polymorphisme allélique du gène FCGR3B et les interactions protéine-ligand de ce Fc récepteur (Bertrand et al. 2004, Publication 4).

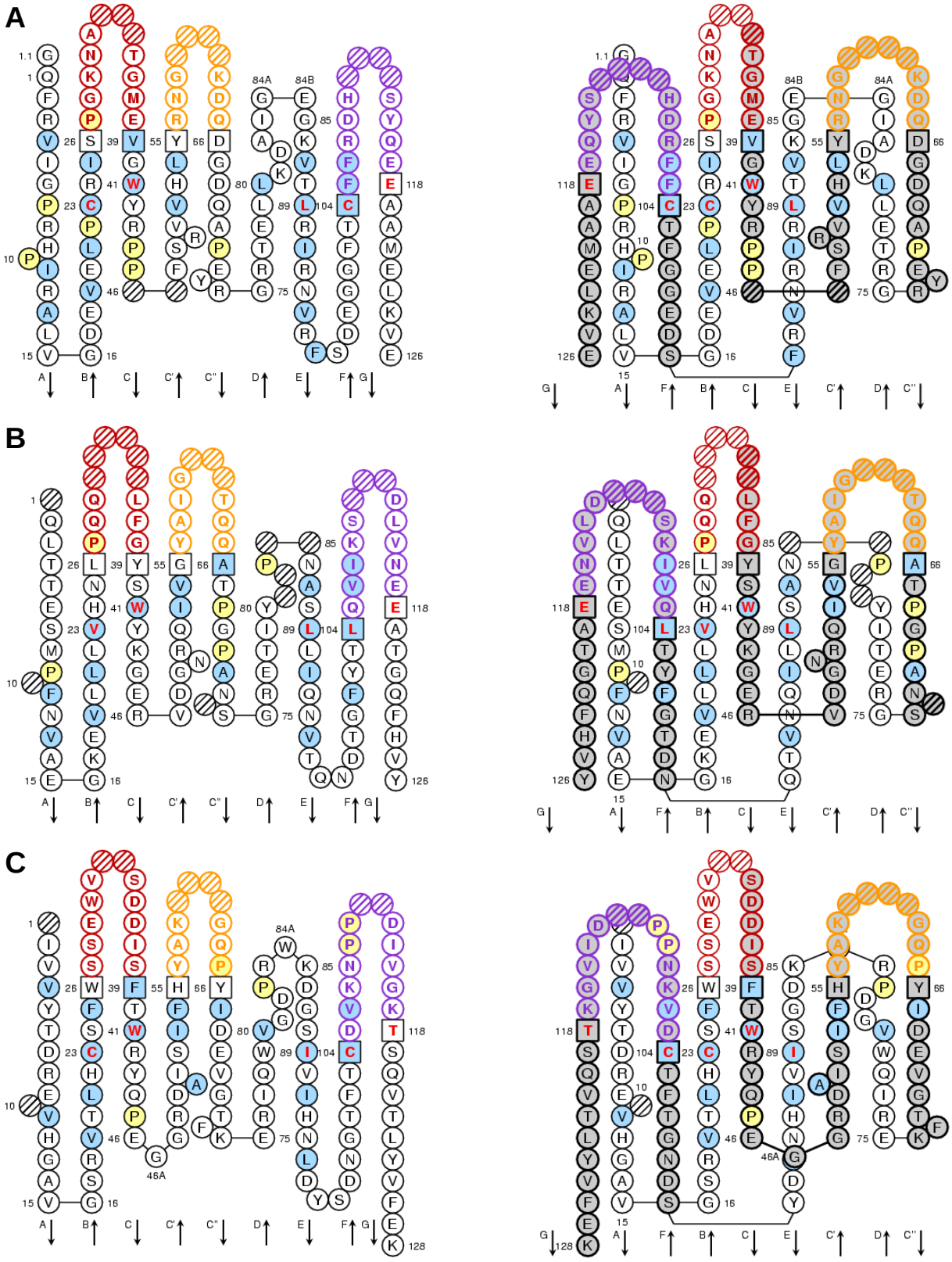


Figure 2.1. Colliers de Perles de V-LIKE-DOMAINs sur un plan et deux plans (extrait de Duprat et al. 2004, Publication 3). (A) *Homo sapiens* MOG [D]. (B) *Homo sapiens* CEACAM1 [D1]. (C) *Homo sapiens* MPZ [D] ; le V-LIKE-DOMAIN de MPZ présente deux insertions additionnelles (46A et 84A) localisées au niveau de l'apex de C'C''-TURN et DE-TURN, et qui ne perturbent donc pas l'architecture globale du domaine.

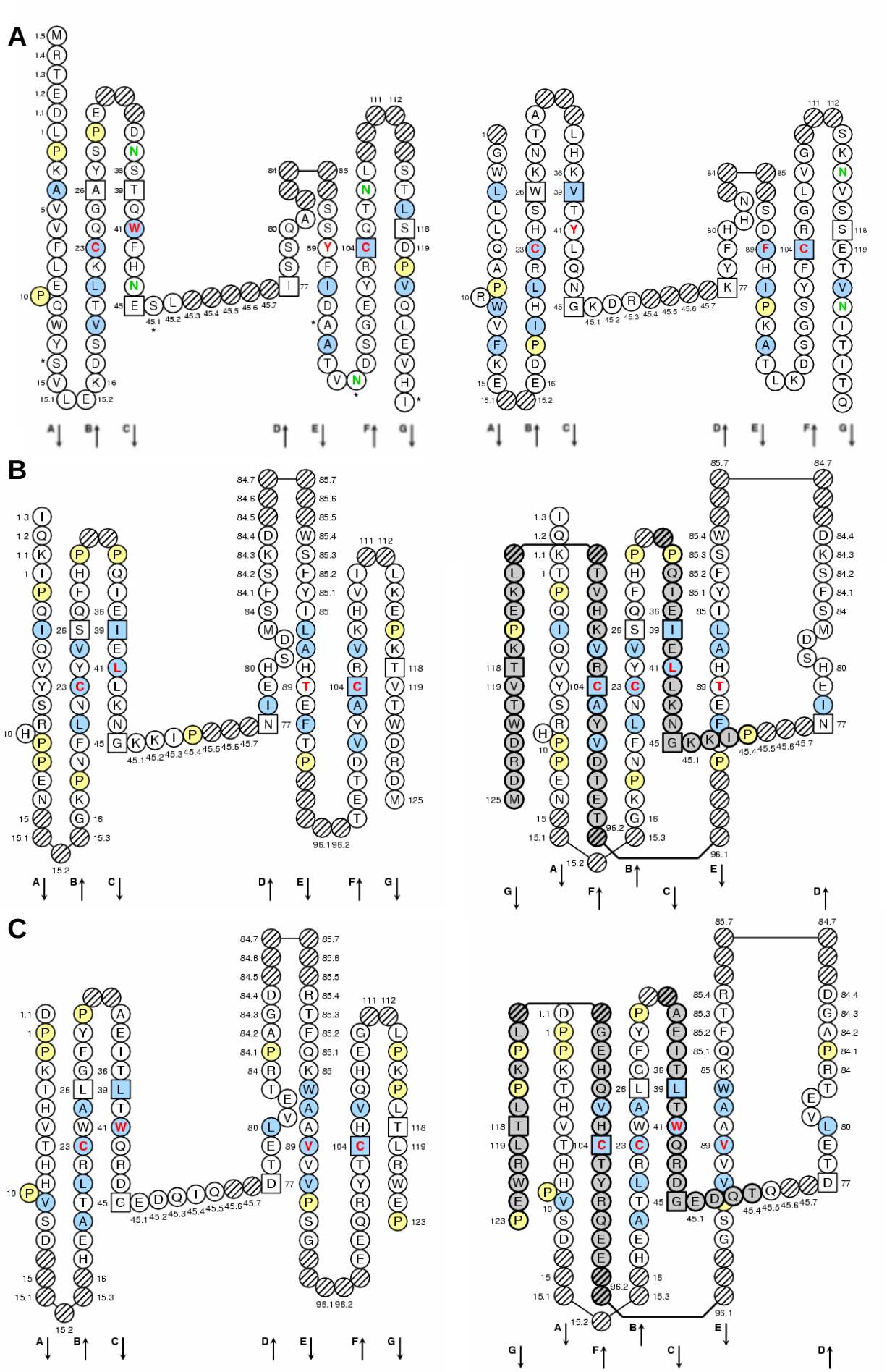


Figure 2.2. Colliers de Perles de C-LIKE-DOMAINs sur un plan et deux plans (extrait de Bertrand et al. 2004, Lefranc et al. 2005b, Publications 1 et 4). (A) *Homo sapiens* FCGR3B*02 [D1] et [D2]. (B) *Rattus norvegicus* B2M [D] sur un et deux plans. (C) *Homo sapiens* HLA-B [D3] (1a1m) sur un et deux plans.

2.2 La superfamille du MHC (MhcSF)

Cette superfamille regroupe les protéines MHC-I et MHC-II, et les protéines dont la chaîne lourde comprend deux G-LIKE-DOMAINs (MHC-I-like). Ces domaines sont nommés respectivement G-ALPHA1-LIKE [D1] et G-ALPHA2-LIKE [D2] ; un C-LIKE-DOMAIN [D3] est présent ou non. Cette organisation modulaire est similaire à celle des chaînes MHC-I-ALPHA ; la chaîne lourde des protéines MHC-I-like est par conséquent désignée par MHC-I-ALPHA-LIKE. De façon étonnante, aucune protéine de type MHC-II-like (c'est-à-dire constituée de deux chaînes comportant chacune un G-LIKE-DOMAIN) n'a été identifiée à ce jour (Maenaka and Jones 1999, Strong 2000). Les chaînes MHC-I-ALPHA-LIKE sont liées ou non à la chaîne de la beta2-microglobuline (B2M) ; les protéines MHC-I-like constituées uniquement d'une chaîne protéique sont correctement repliées et exprimées à la surface cellulaire. Nous avons classé ces protéines en 8 types de récepteurs selon l'organisation modulaire de leur chaîne lourde et leur fonction (Fig. 2.3, Annexes 3 et 5) :

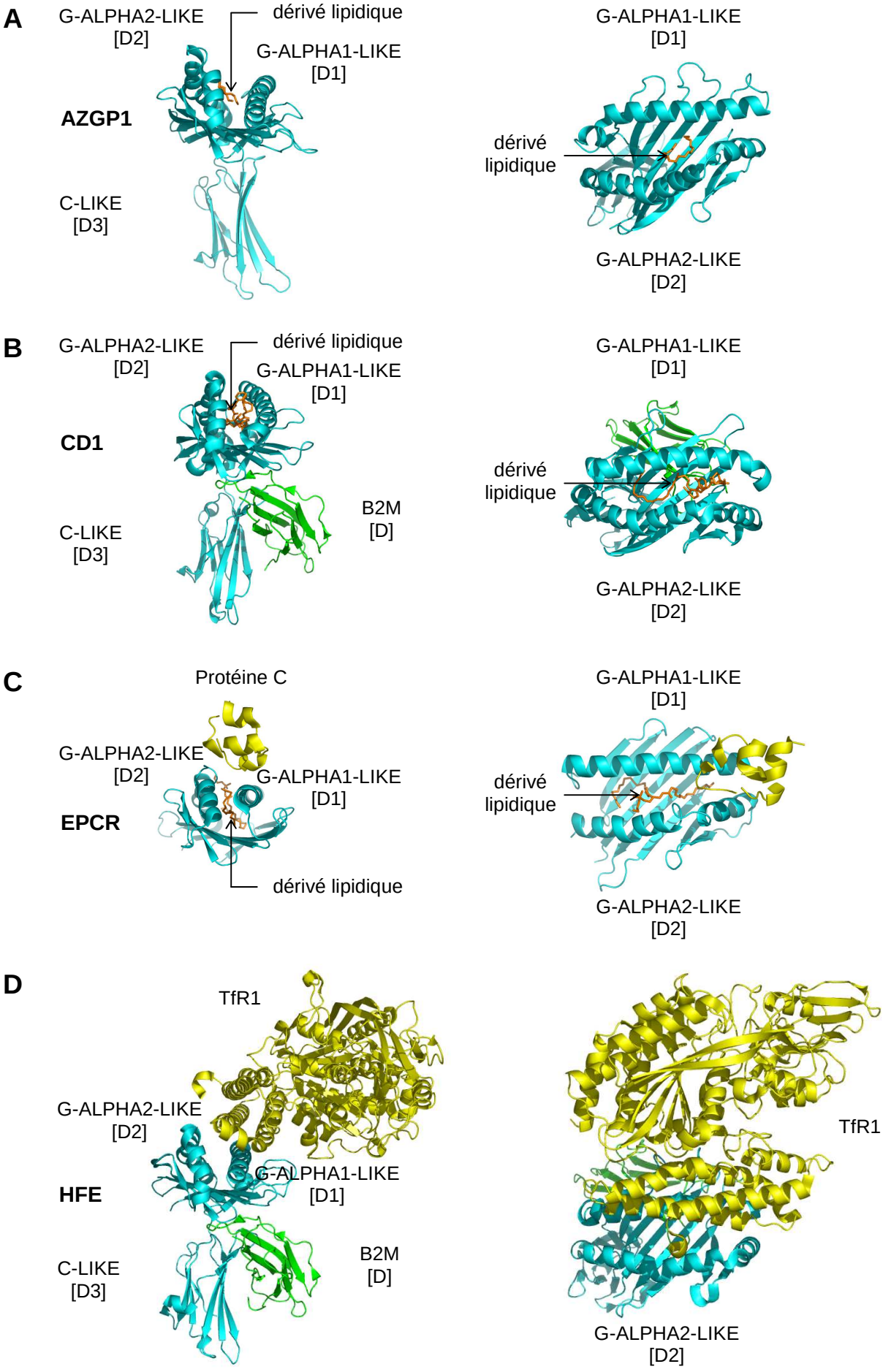
AZGP1. La chaîne lourde comprend un C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. Cette protéine (AZGP1 chez *Homo sapiens*, *Mus musculus* et *Rattus norvegicus*) se lie à des dérivés lipidiques et régule la dégradation des acides gras dans les adipocytes (Sanchez et al. 1999).

CD1. La chaîne lourde comprend un C-LIKE-DOMAIN [D3] et se lie à la B2M. Ces protéines (CD1A, CD1B, CD1C, CD1D, CD1E chez *Homo sapiens*, CD1D1 chez *Mus musculus* et *Rattus norvegicus*) présentent des phospholipides aux TR alpha-beta et gamma-delta et induisent une réponse immunitaire contre les pathogènes microbiens (Zeng et al. 1997).

EPCR. La chaîne lourde ne comporte pas de C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. Cette protéine (EPCR chez *Homo sapiens*, *Mus musculus* et *Bos taurus*) interagit avec des dérivés lipidiques et les protéines C activées ou non, et est impliquée dans la voie de coagulation du sang (Simmonds and Lane 1999).

FCGRT. La chaîne lourde est constituée d'un C-LIKE-DOMAIN [D3] et se lie à la B2M. Cette protéine (FCGRT chez *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* et *Bos taurus*) transporte les immunoglobulines à travers le placenta et gouverne l'immunité néonatale (West and Bjorkman 2000) ; elle reconnait spécifiquement les domaines CH2 et CH3 des IgG.

HFE. La chaîne lourde est constituée d'un C-LIKE-DOMAIN [D3] et se lie à la B2M. Cette protéine (HFE chez *Homo sapiens*, *Mus musculus* et *Rattus norvegicus*) interagit avec le


A
G-ALPHA2-LIKE
[D2]
dérivé lipidique
G-ALPHA1-LIKE
[D1]
AZGP1
C-LIKE
[D3]
G-ALPHA1-LIKE
[D1]
dérivé
lipidique
G-ALPHA2-LIKE
[D2]
B
G-ALPHA2-LIKE
[D2]
dérivé lipidique
G-ALPHA1-LIKE
[D1]
CD1
C-LIKE
[D3]
B2M
[D]
G-ALPHA1-LIKE
[D1]
dérivé
lipidique
G-ALPHA2-LIKE
[D2]
C
Protéine C
G-ALPHA2-LIKE
[D2]
G-ALPHA1-LIKE
[D1]
EPCR
dérivé lipidique
G-ALPHA1-LIKE
[D1]
dérivé
lipidique
G-ALPHA2-LIKE
[D2]
D
TfR1
G-ALPHA2-LIKE
[D2]
G-ALPHA1-LIKE
[D1]
HFE
C-LIKE
[D3]
B2M
[D]
TfR1
G-ALPHA2-LIKE
[D2]

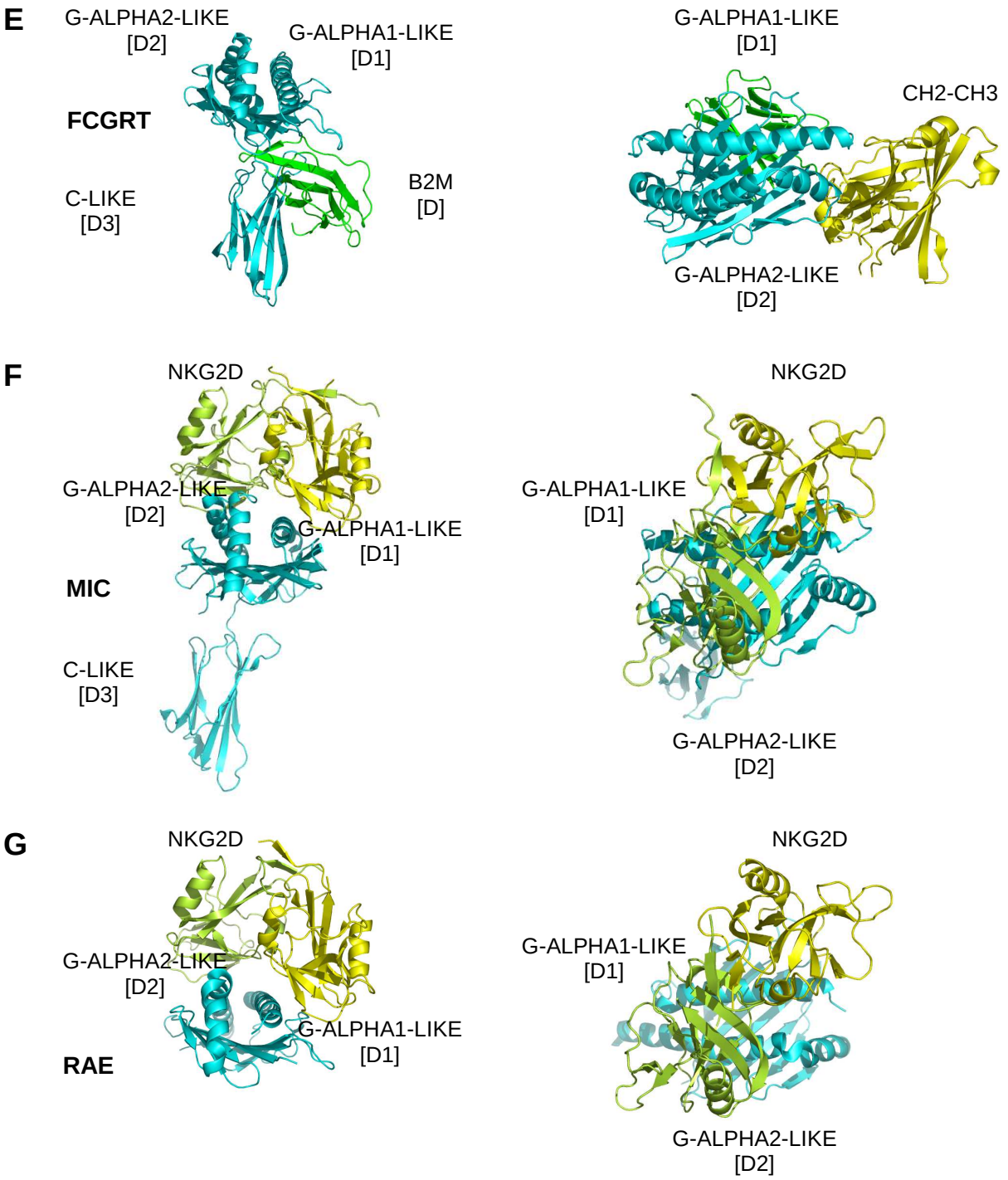


Figure 2.3. Diversité d'organisation modulaire et de ligand au sein de la MhcSF. La chaîne lourde des protéines de la MhcSF comprend au moins deux G-LIKE-DOMAINs (G-ALPHA1-LIKE [D1] et G-ALPHA2-LIKE [D2]). (A) *Homo sapiens* AZGP1 lié à un dérivé lipidique (IMGT/3Dstructure-DB 1t7v) ; la chaîne lourde des protéines AZGP1 comprend un C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. (B) *Homo sapiens* CD1A lié à un dérivé lipidique (1xz0) ; la chaîne lourde des protéines CD1 comprend un C-LIKE-DOMAIN [D3] et est liée de façon non covalente à la B2M. (C) *Homo sapiens* EPCR lié à un phospholipide (1lqv) ; la chaîne lourde des protéines EPCR ne comprend pas de C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. (D) *Homo sapiens* HFE lié au récepteur à la transferrine TfR1 (1de4) ; la chaîne lourde des protéines HFE comprend un C-LIKE-DOMAIN [D3] et est liée à la B2M. (E) *Rattus norvegicus* FCGRT lié aux domaines CH2 et CH3 d'une chaîne H-GAMMA (1frt) ; la chaîne lourde des protéines FCGRT comprend un C-LIKE-DOMAIN [D3] et est liée à la B2M. (F) *Homo sapiens* MICA lié au récepteur NKG2D inhibiteur des cellules NK (1hyr) ; la chaîne lourde des protéines MIC comprend un C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. (G) *Homo sapiens* RAET1N lié à NKG2D (1kcg) ; la chaîne lourde des protéines RAE ne comprend pas de de C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. Les ligands protéiques et lipidiques sont représentés respectivement en jaune et orange.

récepteur à la transferrine TfR1, et prend part à l'homéostase du fer en régulant son transport à travers les membranes cellulaires (Feder et al. 1998).

MIC. La chaîne lourde est constituée d'un C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. L'expression de ces protéines (MICA et MICB chez *Homo sapiens*) est induite par le stress lié aux tumeurs et infections virales ; leur interaction avec le récepteur NKG2D activateur des cellules NK assure la reconnaissance et la destruction des cellules infectées et tumorales par le système immunitaire (Holmes et al. 2002, Frigoul and Lefranc 2005).

MR1. La chaîne lourde est constituée d'un C-LIKE-DOMAIN [D3] et se lie à la B2M. La fonction de cette protéine (MR1 chez *Homo sapiens*, *Mus musculus* et *Rattus norvegicus*) est actuellement inconnue (Miley et al. 2003).

RAE. La chaîne lourde ne comporte pas de C-LIKE-DOMAIN [D3] et n'est pas liée à la B2M. L'expression de ces protéines (RAET1E, RAET1H, RAET1I, RAET1L, RAET1N chez *Homo sapiens*, RAE1B et RAE1G chez *Mus musculus* et *Rattus norvegicus*) est induite par l'acide rétinoïque ; ces protéines stimulent la production de cytokine/chemokine et l'activité cytotoxique des cellules NK, par leur interaction avec le récepteur NKG2D (Li et al. 2002).

Ces protéines sont caractérisées par une hétérogénéité de localisations cellulaires et de voies d'expression, de ligands et de zones d'interaction, et de fonctions appartenant ou non au système immunitaire (Duprat and Lefranc, IMGT Education MhcSF) ; les protéines MHC-I-like ne présentent pas de peptide aux TR. Ces protéines composent une même superfamille structurale mais diffèrent néanmoins par la distance qui sépare les hélices de leurs G-LIKE-DOMAINs (Fig. 2.3).

Pour les protéines qui se lient constitutivement à la B2M (CD1, FCGRT, HFE et MR1), cette interaction est nécessaire à la stabilité de leur structure et à leur expression à la surface cellulaire, comme dans le cas des protéines MHC-I ; il est par conséquent surprenant que les protéines non liées à la B2M soient fonctionnelles. L'absence constitutive de liaison à la B2M (AZGP1, EPCR, MIC et RAE) pourrait être imputable à une gêne stérique liée à l'orientation de chaînes latérales ou à la liaison d'une sucre au niveau d'un site N-glycosylé (Li et al. 1999), ou à des propriétés physico-chimiques incompatibles avec cette interaction au niveau des G-LIKE-DOMAINs (Li et al. 2002) ; des mécanismes de compensation assurant la stabilité structurale de la chaîne lourde en l'absence de B2M sont observés (Sanchez et al. 1999, Radaev et al. 2001), mais restent en partie à déterminer. La prédiction de l'interaction des protéines de la MhcSF avec la B2M, et la compréhension des caractéristiques de leurs

séquences qui favorisent ou défavorisent cette interaction constituent des enjeux biologiques majeurs.

Nous avons montré que la numérotation unique des G-DOMAINs s'applique aux G-LIKE-DOMAINs de ces protéines (Lefranc et al. 2005c, Publication 2) ; une unique position additionnelle (54A) a été introduite pour numéroter les G-LIKE-DOMAINs. La faible similarité des G-LIKE-DOMAINs de différents types de récepteurs a nécessité la mise en place d'une procédure particulière, détaillée dans le chapitre 3. Cette numérotation est unique pour tous les G-DOMAINs et G-LIKE-DOMAINs, quels que soient le type de récepteur, le type de chaîne et l'espèce.

Cette standardisation permet désormais de comparer les caractéristiques de séquence, de structure et de fonction de ces domaines, pour toutes les protéines de la MhcSF ; la représentation en Colliers de Perles constitue pour cela un outil précieux (Fig. 2.4). Comme nous le verrons par la suite, la comparaison des caractéristiques des protéines de la MhcSF est essentielle afin d'appréhender et de résoudre la problématique de classification des protéines de cette superfamille selon leur liaison ou non à la B2M.

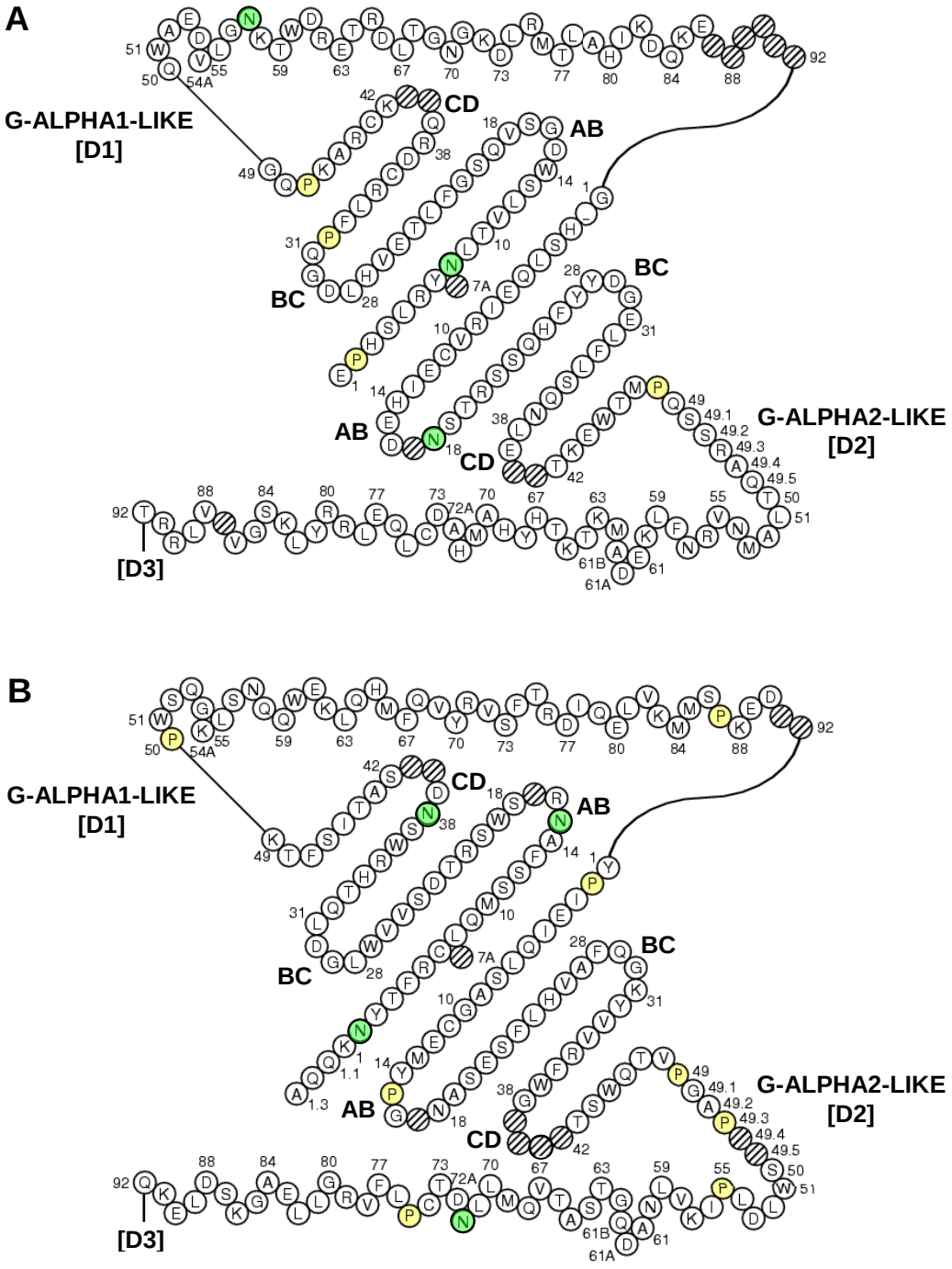


Figure 2.4. Colliers de Perles de G-LIKE-DOMAINs (extrait de Lefranc et al. 2005c, Publication 2). Domaines G-ALPHA1-LIKE [D1] et G-ALPHA2-LIKE [D2] d'une chaîne I-ALPHA-LIKE (A) d'*Homo sapiens* (MICA*01) et (B) de *Mus musculus* (CD1D1*01). Les résidus Proline et les sites potentiels de N-glycosylation (N-X-S/T) sont représentés respectivement en jaune et vert. Les positions hachurées correspondent à des positions manquantes par rapport à la numérotation unique. [D3] indique la position du C-LIKE-DOMAIN dans la chaîne I-ALPHA.

CHAPITRE 3

Standardisation et caractérisation des protéines de la MhcSF

Nous avons vu dans le chapitre précédent que les règles de DESCRIPTION des IG, TR et MHC, et de NUMEROTATION de leurs V-DOMAINs, C-DOMAINs et G-DOMAINs s'appliquent avec succès aux protéines des IgSF et MhcSF et à leurs V-LIKE-DOMAINs, C-LIKE-DOMAINs et G-LIKE-DOMAINs.

Ce chapitre présente tout d'abord la procédure que nous avons mise en place afin d'aligner les G-DOMAINs des protéines MHC-I et MHC-II et les G-LIKE-DOMAINs des protéines MHC-I-like ; cette procédure facilite l'identification et la standardisation des G-DOMAINs et G-LIKE-DOMAINs au sein de nouvelles protéines de la MhcSF. A partir de ces données standardisées, nous caractérisons ensuite les protéines de cette superfamille en terme de séquence, d'évolution, et d'interactions protéine-ligand ; ces analyses nous permettent également d'appréhender la problématique de classification des protéines de la MhcSF selon leur interaction ou non avec la B2M, en identifiant les contraintes inhérentes aux données.

Nous mettons en évidence que la diversité de fonction des protéines de la MhcSF, décrite dans le chapitre précédent, est liée à une grande diversité de séquences et d'interactions protéiques ; nous révélons également l'antériorité évolutive de la spécialisation des protéines de la MhcSF sur la spéciation, et la perte partielle de la capacité de liaison à la B2M au cours de l'évolution de cette superfamille. Enfin, nous déterminons la zone de contact potentiel entre les G-DOMAINs et G-LIKE-DOMAINs de la MhcSF et le C-LIKE-DOMAIN de la B2M, et concluons que les propriétés physico-chimiques requises pour l'interaction avec la B2M sont similaires quel que soit l'espèce et le type de récepteur de la MhcSF.

3.1 Procédure d'alignement des G-DOMAINs et G-LIKE-DOMAINs

Cette procédure consiste à combiner des alignements de séquences et de structures 3D des G-DOMAINs des protéines MHC-I et MHC-II, et G-LIKE-DOMAINs des protéines MHC-I-like (Figure 3.1). Chaque protéine MHC-I, MHC-II et MHC-I-like est représentée par une ou plusieurs séquences alléliques, nommées « allèles » dans ce contexte et dans la suite du texte ; ces séquences diffèrent par au moins un acide aminé, pour une protéine codée par un gène donné, dans une espèce donnée.

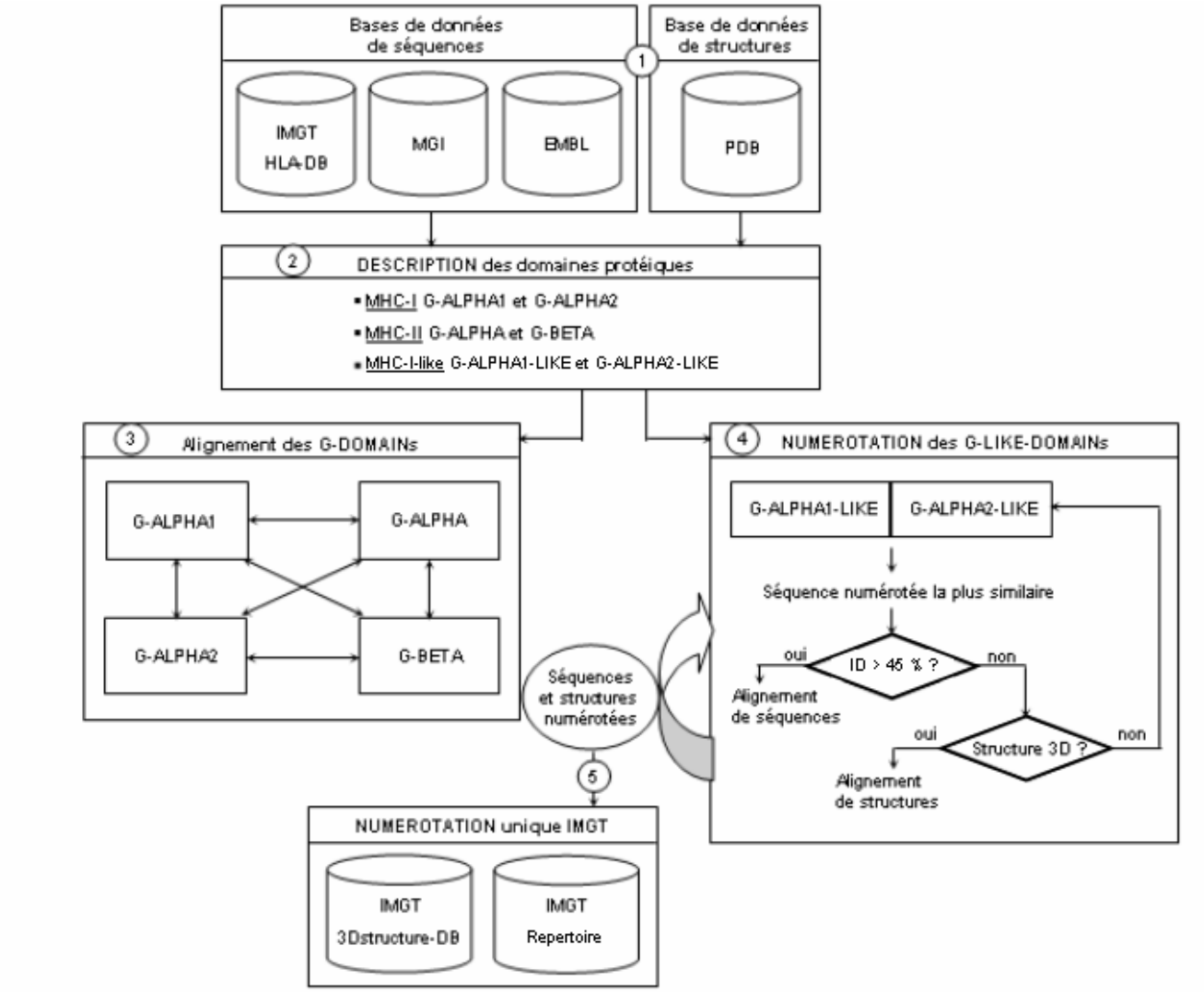


Figure 3.1. Procédure d'alignement des G-DOMAINs et G-LIKE-DOMAINs. (1) Récupération des données de séquence dans IMGT/HLA-DB (Marsh and Robinson 2001), MGI (Blake et al. 2003) et EMBL/DDBJ/Genbank (Kanz et al. 2005, Benson et al. 2005, Tateno et al. 2005), et des données de structure dans PDB (Berman et al. 2000). (2) Description des domaines protéiques. (3) Alignements multiples des séquences et structures des G-DOMAINs et numérotation. (4) Numérotation des G-LIKE-DOMAINs. Nous utilisons le programme Fasta2 d'alignement local de paires de séquences (Pearson and Lipman 1988) pour identifier la séquence numérotée la plus similaire (correspondant au score le plus élevé) pour chaque domaine considéré. (5) Obtention de la numérotation unique IMGT pour les G-DOMAINs et G-LIKE-DOMAINs ; les séquences et structures ainsi numérotées sont accessibles respectivement dans IMGT Repertoire (Lefranc et al. 2005a) et IMGT/3Dstructure-DB (Kaas et al. 2004). Les alignements multiples de séquences et de structures 3D sont effectués respectivement par MUSCLE (Edgar et al. 2004) et COMPARER (Sali and Blundell 1990).

3.1.1 Extraction des données

Les séquences nucléotidiques et protéiques des protéines MHC-I et MHC-II sont tout d'abord extraites de trois bases de données en fonction de leur espèce: IMGT/HLA-DB (Marsh and Robinson 2001), MGI (Blake et al. 2003) et EMBL/DDBJ/GenBank (Kanz et al. 2005, Benson et al. 2005, Tateno et al. 2005), respectivement pour *Homo sapiens*, *Mus musculus* et *Rattus norvegicus*. Nous obtenons ainsi 767 allèles de MHC-I et 511 allèles de MHC-II

(Annexe 5). Les structures 3D correspondant à ces séquences sont extraites de la base de données PDB (Berman et al. 2000) ; la correspondance entre les allèles et les structures 3D disponibles est présentée en Annexe 5.

Les protéines MHC-I-like ont été identifiées au sein du génome de nombreux vertébrés ; nous nous limitons néanmoins dans cette étude à quatre espèces de mammifères, pour lesquelles la littérature est la plus abondante : *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* et *Bos taurus*. Les séquences nucléotidiques et protéiques des protéines MHC-I-like sont extraites de MGI (*Mus musculus*) et EMBL/DDBJ/GenBank (*Homo sapiens*, *Rattus norvegicus*, *Bos taurus*). Nous obtenons ainsi 39 allèles de MHC-I-like, décrites en Annexe 5. Les structures 3D correspondant à ces séquences sont extraites de la base de données PDB ; à l'exception des protéines MR1, la structure 3D d'au moins une protéine de chaque type de récepteur MHC-I-like est connue.

3.1.2 Description des domaines protéiques

La délimitation des exons de chaque gène (au niveau des séquences) et la reconnaissance du repliement en domaines (au niveau des structures 3D) nous permettent d'identifier et de décrire les G-DOMAINs des protéines MHC-I et MHC-II (G-ALPHA1 [D1] et G-ALPHA2 [D2] pour les MHC-I, G-ALPHA [D1] et G-BETA [D1] pour les MHC-II) et les G-LIKE-DOMAINs des protéines MHC-I-like (G-ALPHA1-LIKE [D1] et G-ALPHA2-LIKE [D2]).

3.1.3 Alignement des G-DOMAINs et numérotation

L'objectif de cette étape est d'obtenir l'alignement multiple des G-DOMAINs des protéines MHC-I et MHC-II. Alors que les séquences d'un même type de G-DOMAIN (G-ALPHA1, G-ALPHA2, G-ALPHA ou G-BETA) sont très similaires et peuvent être alignées directement, la similarité des séquences de différents types de domaines est insuffisante (Table 3.1) ; le pourcentage d'identité de ces séquences est en effet inférieur à 45%, ce qui correspond à la « twilight zone » (Doolittle 1986, Rost 1999). L'unité structurale de ces domaines protéiques nous permet néanmoins de nous affranchir de cette hétérogénéité de séquence.

Nous sélectionnons tout d'abord les structures 3D représentatives de chaque type de G-DOMAIN : pour chaque sous-classe de protéine MHC-I et MHC-II, la structure 3D dont la résolution est la meilleure est retenue (Annexe 5). Nous alignons successivement les structures 3D représentatives de chaque type de G-DOMAIN avec celles de chacun des trois

	G-ALPHA1	G-ALPHA1-LIKE	G-ALPHA2	G-ALPHA2-LIKE	G-ALPHA	G-BETA
G-ALPHA1	73.1 (46.2-94.3)					
G-ALPHA1-LIKE	33.0 (18.9-53.8)	35.6 (18.9-98.1)				
G-ALPHA2	20.6 (17.0-25.5)	20.0 (14.2-29.3)	74.2 (55.7-90.6)			
G-ALPHA2-LIKE	19.6 (13.2-26.4)	18.5 (10.4-27.4)	33.9 (18.9-54.7)	35.5 (16.0-96.2)		
G-ALPHA	22.7 (15.1-28.3)	19.9 (13.2-27.4)	20.9 (16.0-26.4)	19.9 (11.3-28.3)	56.98 (30.2-84.0)	
G-BETA	20.6 (13.2-26.4)	19.8 (11.3-26.4)	21.9 (12.3-28.3)	21.1 (12.3-30.2)	18.2 (11.3-23.6)	57.1 (17.9-87.7)

Table 3.1. Pourcentage d'identité des séquences des différents types de G-DOMAINs. Les pourcentages moyen, minimum et maximum d'identité sont calculés entre les allèles de référence ou les allèles les plus fréquents de chaque protéine MHC-I, MHC-II (Annexe 5), à partir de l'alignement multiple IMGT. Les valeurs excédant 45% (soulignées) indiquent que les domaines correspondants peuvent être alignés en séquences. La séquence des domaines G-ALPHA et G-BETA de HLA-DM (MHC-IIb, HLA-DMA et HLA-DMB) correspondent aux valeurs minimales de pourcentage d'identité pour ces deux types de domaines (respectivement 30.2 et 17.9). Leur pourcentage d'identité avec les autres séquences de leur type respectif de domaine est toujours inférieur à 45% ; cette séquence correspond à une structure 3D représentative, et n'est donc pas alignée en séquence au cours de notre processus.

autres types ; nous obtenons ainsi un alignement multiple des séquences de G-DOMAINs correspondant à une structure 3D connue. Les séquences de chaque type de domaine sont alors alignées avec les séquences de leurs structures 3D représentatives, et sont ajoutées à l'alignement multiple des G-DOMAINs. Les positions de l'alignement multiple sont alors numérotées par des nombres croissants de l'extrémité N-terminale à C-terminale des domaines ; des positions additionnelles (décrites dans le chapitre 1) sont définies pour les zones d'insertion. Nous obtenons ainsi l'alignement multiple et la numérotation unique des G-DOMAINs ; cette numérotation est identique quels que soient le récepteur, la chaîne, le type de domaine et l'espèce. Les séquences et structures 3D numérotées sont ajoutées respectivement dans IMGT Repertoire (Lefranc et al. 2005a) et IMGT/3Dstructure-DB (Kaas et al. 2004).

3.1.4 Numérotation des G-LIKE-DOMAINs

L'objectif de cette étape est d'aligner les G-LIKE-DOMAINs des protéines MHC-I-like avec les G-DOMAINs des protéines MHC-I et MHC-II préalablement alignés, et d'obtenir ainsi un unique alignement multiple et la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs. Les domaines G-ALPHA1 et G-ALPHA2 des protéines MHC-I sont les plus similaires respectivement aux domaines G-ALPHA1-LIKE et G-ALPHA2-LIKE des protéines MHC-I-like (Table 3.1) ; cette similarité est cependant insuffisante pour obtenir un alignement de qualité (pourcentage d'identité inférieur à 45%) (Table 3.2). Pour un type de

domaine (G-ALPHA1-LIKE ou G-ALPHA2-LIKE), les séquences d'un même type de récepteur MHC-I-like (AZGP1, CD1, EPCR, FCGRT, HFE, MIC, MR1, RAE) sont très similaires et peuvent être alignées directement, alors que la similarité des séquences de différents types de récepteurs est insuffisante (Table 3.2). L'unité structurale de ces domaines protéiques nous permet néanmoins de nous affranchir de cette hétérogénéité de séquence.

[D1]	MHC-I	FCGRT	MR1	HFE	CD1	AZGP1	MIC	RAE	EPCR
FCGRT	35.8	81.6							
MR1	48.5	40.9	94.3						
HFE	43.4	37.6	47.1	90.6					
CD1	24.8	28.4	27.1	28.6	62.3				
AZGP1	41.7	40.4	38.1	43.9	29.7	83.2			
MIC	37.0	38.7	38.2	35.9	25.4	36.3	92.9		
RAE	28.7	28.5	31.5	31.2	23.9	30.1	34.3	51.0	
EPCR	24.0	23.6	24.5	22.6	33.0	26.2	22.6	24.6	78.0
[D2]	MHC-I	FCGRT	MR1	HFE	CD1	AZGP1	MIC	RAE	EPCR
FCGRT	33.5	81.3							
MR1	48.8	36.2	93.9						
HFE	41.2	38.3	47.5	88.3					
CD1	29.6	32.0	32.8	31.1	60.7				
AZGP1	45.8	36.7	51.4	45.0	30.4	89.3			
MIC	29.8	27.0	32.4	32.6	23.8	30.0	92.9		
RAE	28.7	28.9	31.5	30.3	24.0	30.6	25.1	48.2	
EPCR	28.5	30.8	29.5	28.3	32.2	26.1	18.6	22.4	80.9

Table 3.2. Pourcentage moyen d'identité des séquences des domaines G-ALPHA1 [D1] et G-ALPHA1-LIKE [D1], et G-ALPHA2 [D2] et G-ALPHA2-LIKE [D2] des différents types de récepteurs de la MhcSF. Les pourcentages d'identité sont calculés entre les allèles de référence ou les allèles les plus fréquents de chaque protéine MHC-I et MHC-I-like (Annexe 5), à partir de l'alignement multiple IMGT. Les valeurs excédant 45% (soulignées) indiquent que les domaines correspondants peuvent être alignés en séquences.

Cette étape correspond à un processus itératif, qui considère successivement chaque séquence de protéine MHC-I-like. Pour chaque G-LIKE-DOMAIN à aligner, nous identifions la séquence numérotée la plus similaire ; l'ensemble des séquences numérotées comprend uniquement des G-DOMAINs lors de la première itération (c'est-à-dire pour le premier G-LIKE-DOMAIN traité) et des G-DOMAINs et G-LIKE-DOMAINs dans la suite du processus. Deux cas sont envisagés, en fonction de la similarité de la séquence du G-LIKE-DOMAIN avec la séquence numérotée sélectionnée. (i) si leur similarité est suffisante (pourcentage d'identité supérieur à 45%), leur alignement permet de définir les positions homologues (qui seront donc décrites par le même numéro) ; ce cas correspond à un type de récepteur pour lequel au moins une séquence a été préalablement numérotée. (ii) si leur similarité ne permet pas d'obtenir un alignement fiable, le G-LIKE-DOMAIN considéré est aligné en structure avec les structures 3D représentatives numérotées ; ce cas correspond au traitement d'un nouveau type de récepteur, et nécessite qu'une structure 3D soit connue pour

chaque type de récepteur. Aucune structure 3D n'est connue pour le type de récepteur MR1 ; la protéine MR1 d'*Homo sapiens* peut cependant être alignée en séquence avec la protéine HLA-A (54% d'identité), et correspond donc au cas (i). Le processus itératif se termine lorsque tous les G-LIKE-DOMAINs des protéines MHC-I-like ont été ajoutés à l'alignement multiple des G-DOMAINs et G-LIKE-DOMAINs.

3.1.5 Mise à jour de l'alignement et numérotation de nouvelles séquences

Chaque nouvelle protéine de la MhcSF est dans un premier temps décrite en terme de domaines ; la numérotation unique est alors appliquée à chaque G-DOMAIN ou G-LIKE-DOMAIN identifié, par alignement de séquence (avec la séquence numérotée la plus similaire) ou de structure 3D (avec les structures 3D représentatives) selon les cas décrits précédemment. Des positions additionnelles peuvent être ajoutées à la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs ; les séquences et structures de G-LIKE-DOMAINs ainsi numérotées sont ajoutées respectivement à IMGT Repertoire et IMGT/3DstructureDB, et la numérotation unique est mise à jour.

3.2 Analyse position-dépendante des G-DOMAINs et G-LIKE-DOMAINs

La procédure de numérotation unique des G-DOMAINs et G-LIKE-DOMAINs présentée précédemment nous permet d'obtenir 806 séquences allèliques alignées de 47 protéines MHC-I et MHC-I-like (767 allèles de 13 protéines MHC-I d'*Homo sapiens*, *Mus musculus* et *Rattus norvegicus*, et 39 allèles de 34 protéines MHC-I-like d'*Homo sapiens*, *Mus musculus*, *Rattus norvegicus* et *Bos taurus*) ; le nombre d'allèles de chaque protéine MHC-I et MHC-I-like est indiqué en Annexe 5. Chaque séquence comporte deux G-DOMAINs (G-ALPHA1 [D1] et G-ALPHA2 [D2]) ou G-LIKE-DOMAINs (G-ALPHA1-LIKE [D1] et G-ALPHA2-LIKE [D2]).

La numérotation unique des G-DOMAINs et G-LIKE-DOMAINs nous permet de décrire la composition et la variabilité position-dépendante de leurs séquences. Pour chaque position de ces domaines, nous représentons d'une part la fréquence de 11 groupes d'acides aminés représentant leurs principales caractéristiques chimiques (Fig. 3.2), et d'autre part la variabilité associée à ces fréquences, évaluée par la mesure d'entropie de Shannon (1948, Annexe 8) (Fig. 3.3). Nous identifions ainsi pour chaque type de domaine, les positions fortement variables et les positions conservées ; alors que les G-DOMAINs des protéines MHC-I présentent une variabilité forte localisée sur certains sites spécifiques au niveau des

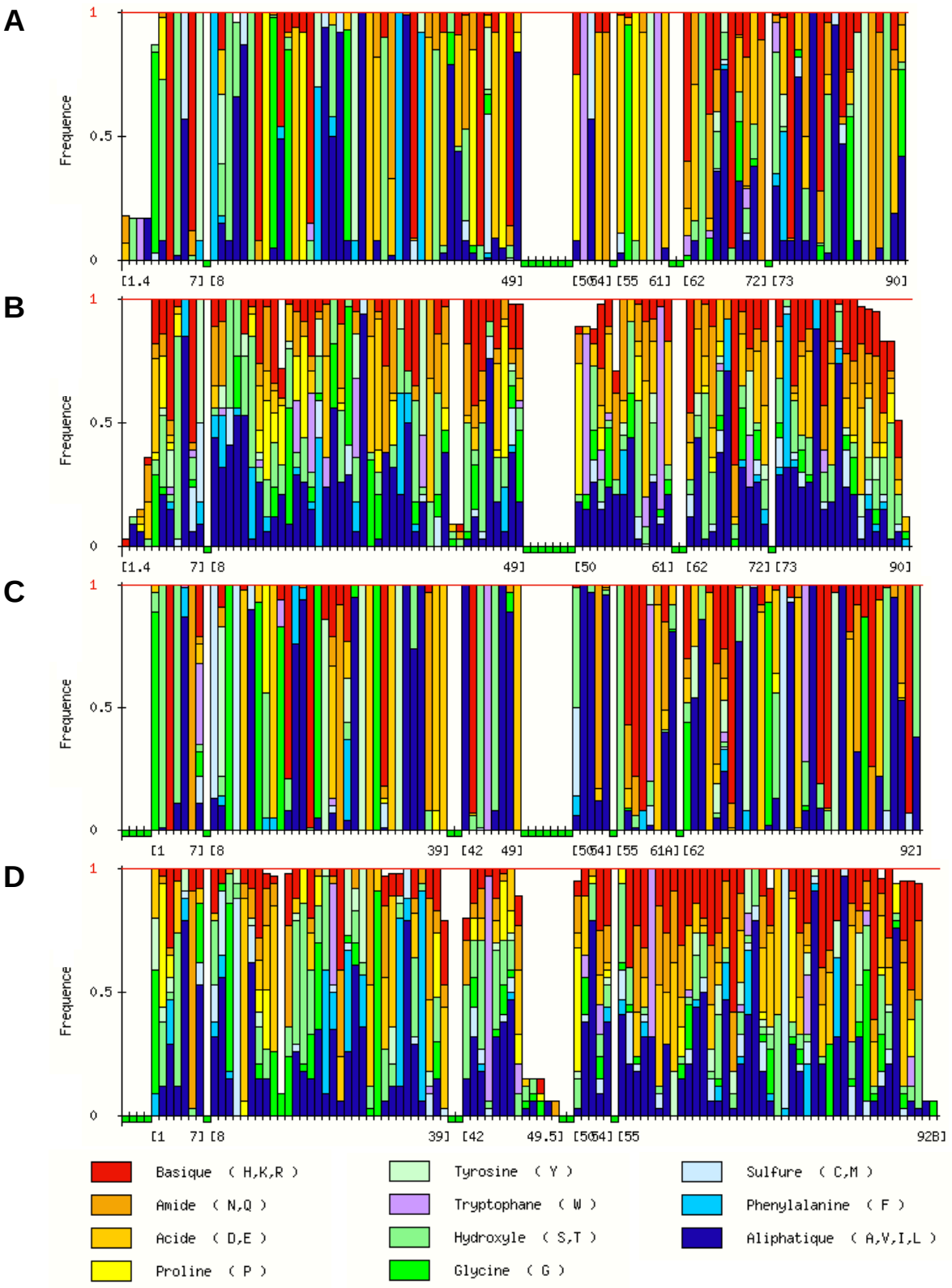


Figure 3.2. Fréquence position-dépendante de 11 caractéristiques chimiques des acides aminés au sein des G-DOMAINs et G-LIKE-DOMAINs. (A) Domaines G-ALPHA1 de 767 allèles de 13 protéines MHC-I. (B) Domaines G-ALPHA1-LIKE de 39 allèles de 34 protéines MHC-I-like. (C) Domaines G-ALPHA2 de 767 allèles de 13 protéines MHC-I. (D) Domaines G-ALPHA2-LIKE de 39 allèles de 34 protéines MHC-I-like. Les groupes d'acides aminés sont définis par IMGT (Pommié et al. 2004, Annexe 6). Les positions sont décrites d'après les règles de la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs. Le polymorphisme des protéines de la MhcSF est pris en compte par la pondération des allèles lors du calcul des fréquences : pour une protéine qui possède e allèles, un poids de $1/e$ est attribué à chaque séquence allélique.

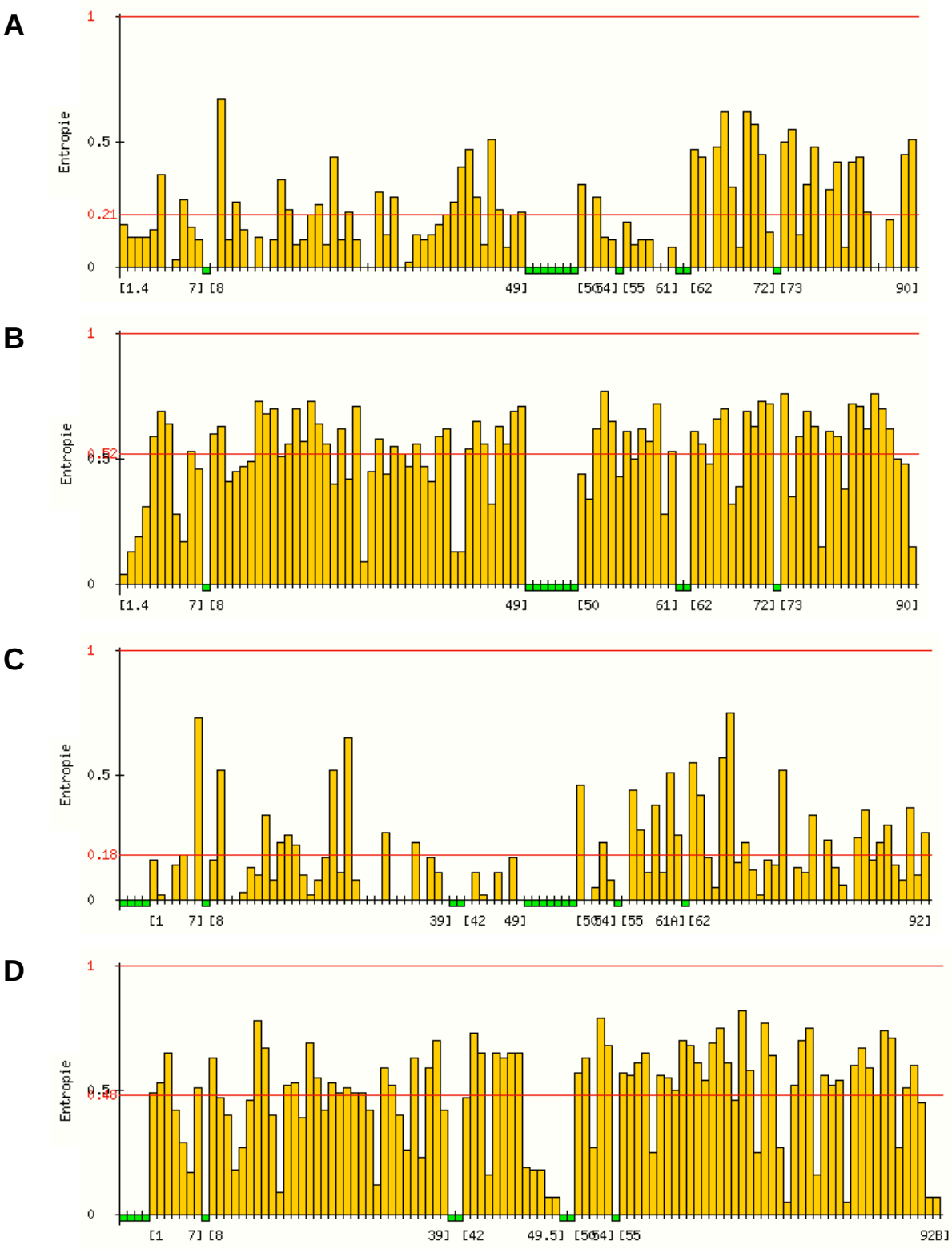


Figure 3.3. Entropie position-dépendante de 11 caractéristiques chimiques des acides aminés au sein des G-DOMAINs et G-LIKE-DOMAINs. (A) Domaines G-ALPHA1 de 767 allèles de 13 protéines MHC-I. (B) Domaines G-ALPHA1-LIKE de 39 allèles de 34 protéines MHC-I-like. (C) Domaines G-ALPHA2 de 767 allèles de 13 protéines MHC-I. (D) Domaines G-ALPHA2-LIKE de 39 allèles de 34 protéines MHC-I-like. Les groupes d'acides aminés sont définis par IMGT (Pommié et al. 2004, Annexe 6) ; les gaps sont considérés comme un groupe d'acides aminés additionnel pour le calcul de l'entropie, détaillé en Annexe 8. Les positions sont décrites d'après les règles de la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs. Les valeurs d'entropie sont calculées à partir des fréquences décrites en Fig. 3.2, et normalisées : elles sont comprises entre 0 (position conservée) et 1 (position totalement variable).

brins et de l'hélice ([D1] 9, 45, 66, 69, 70, 73, 74, 90, et G-ALPHA2 [D2] 7, 9, 24, 26, 61, 62, 66, 67, 73), la variabilité des G-LIKE-DOMAINs concerne l'ensemble de leurs positions. L'homogénéité des domaines des protéines MHC-I est reflétée par leur homogénéité de fonction, tandis que l'hétérogénéité des domaines des protéines MHC-I-like est liée à la diversité fonctionnelle des 8 types de récepteurs qui les composent. Différentes positions sont conservées pour ces groupes d'acides aminés entre les domaines G-ALPHA1 et G-ALPHA1-LIKE (4, 6, 26, 28, 30, 51, 60, 67, 68, 78), G-ALPHA2 et G-ALPHA2-LIKE (1, 4-6, 10-13, 16, 27, 30, 32, 34, 45, 52, 59, 68, 74, 78, 82, 86, 89), G-ALPHA1 et G-ALPHA2 (1-5, 21, 29, 34, 39, 52, 68, 78, 88). Les positions 4, 68 et 78 sont majoritairement constituées d'acides aminés respectivement hydroxyle, basique et aliphatique, quel que soit le type de domaine ; leur conservation indique qu'elles pourraient être impliquées dans le maintien de la structure des G-DOMAINs et G-LIKE-DOMAINs.

3.3 Caractéristiques évolutives

Nous avons construit la phylogénie des 47 protéines de la MhcSF afin de caractériser leurs relations évolutives. Chaque protéine est représentée par la séquence des deux G-DOMAINs ou G-LIKE-DOMAINs (issus de la procédure de numérotation unique présentée dans le paragraphe 3.1) de l'allèle le plus représenté dans les populations caucasiennes pour les protéines humaines ou de l'allèle de référence (Annexe 5). L'arbre est enraciné par la séquence des G-DOMAINs d'une protéine MHC-II (d'après l'hypothèse d'antériorité évolutive des protéines MHC-II sur les protéines MHC-I présentée en Annexe 4). L'arbre phylogénétique obtenu est représenté en Figure 3.4, et met en évidence deux caractéristiques évolutives de la MhcSF.

Chaque type de récepteur définit un clade regroupant les séquences des espèces étudiées (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus* et *Bos taurus*), ce qui laisse supposer que les différentes fonctions des protéines de la MhcSF sont apparues avant l'ancêtre commun de ces espèces de mammifères. Cette caractéristique est cohérente avec la faible similarité de séquence des G-DOMAINs et G-LIKE-DOMAINs pour différents types de récepteurs (détaillée dans le paragraphe 3.1), et est supportée par les valeurs élevées de bootstrap obtenues pour chaque clade. L'arbre phylogénétique de la MhcSF met donc en évidence l'antériorité évolutive de la spécialisation de ces protéines sur la spéciation. L'analyse des plus proches voisins dans l'arbre phylogénétique permettrait donc de classer les protéines de la MhcSF selon leur fonction (ou type de récepteur).

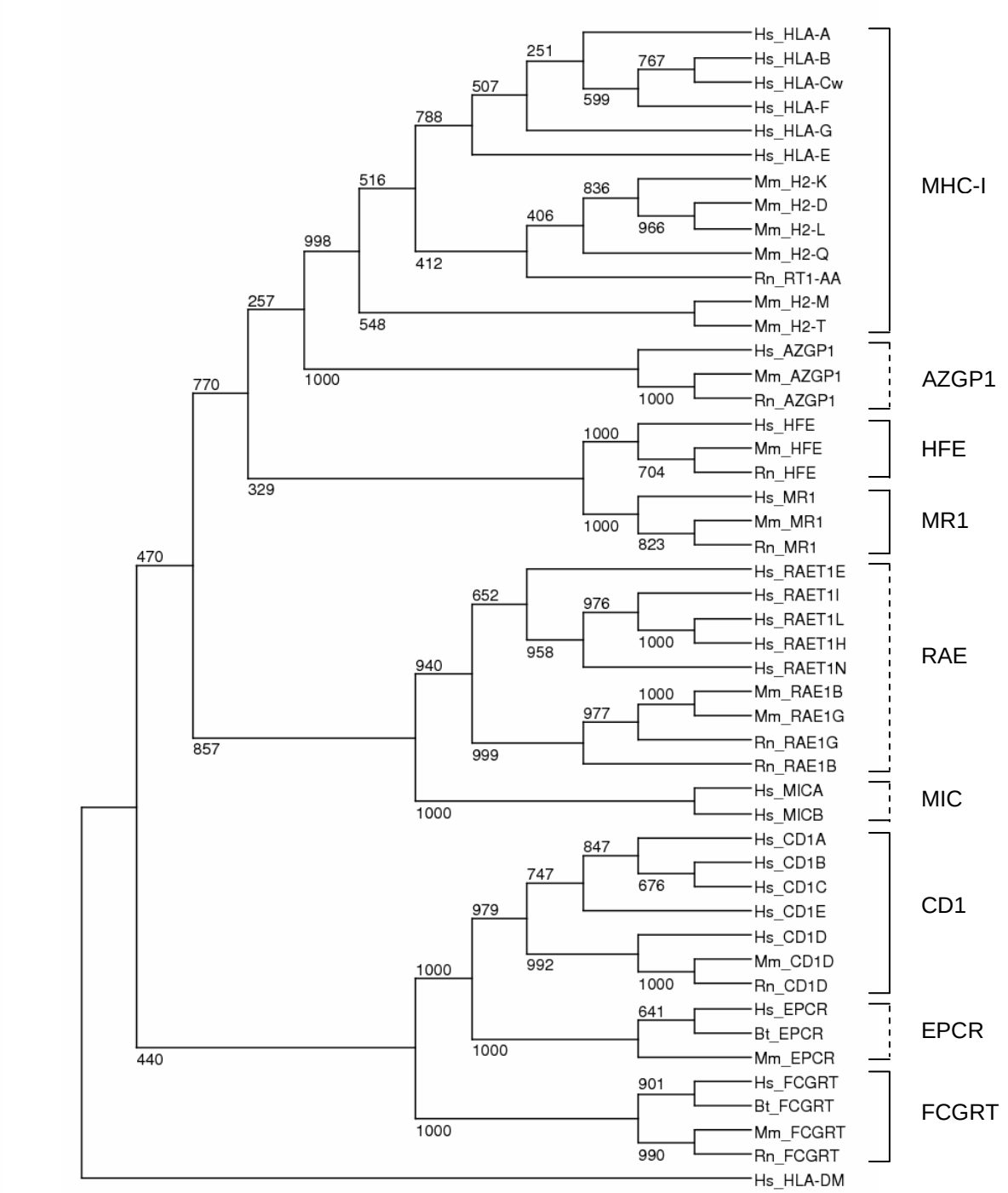


Figure 3.4. Arbre phylogénétique de 47 protéines de la MhcSF. Arbre phylogénétique inféré par le programme PHYML basé sur le maximum de vraisemblance (Guindon and Gascuel 2003), avec le modèle de substitution WAG, enraciné avec une séquence de protéine MHC-II (*Homo sapiens* HLA-DM) et représenté par NJplot (Perrière and Gouy 1996). Chaque protéine appartient à un type de récepteur parmi 9 (MHC-I, AZGP1, CD1, EPCR, FCGRT, HFE, MIC, MR1 and RAE) et à une espèce parmi 4 (Hs: *Homo sapiens*, Mm: *Mus musculus*, Rn: *Rattus norvegicus*, Bt: *Bos taurus*). Les 47 allèles pris en compte sont indiqués en Annexe 5. Les lignes verticales pleines et pointillées indiquent respectivement les types de récepteurs liés ou non à la B2M. Les valeurs de bootstrap ont été estimées pour 1000 réplicats.

Les deux classes de séquences correspondant aux protéines liées ou non à la B2M ne constituent pas deux clades monophylétiques, mais plusieurs clades non corrélés à la phylogénie. Comme nous l'avons indiqué précédemment, chaque type de récepteur définit un

clade ; les clades voisins peuvent correspondre à des protéines qui n'ont pas le même comportement par rapport à la B2M. Par exemple, le clade CD1 est le plus proche voisin du clade EPCR, alors que les protéines CD1 se lient à B2M tandis que les protéines EPCR ne s'y lient pas. Cette caractéristique peut être expliquée par la perte partielle de la capacité de liaison à la B2M au cours de l'évolution (à partir d'un ancêtre commun lié à la B2M), ou par le gain partiel de cette capacité (à partir d'un ancêtre commun non lié à la B2M) ; différents évènements indépendants d'acquisition de cette capacité semble cependant peu probables. L'analyse des plus proches voisins dans l'arbre phylogénétique serait donc inadaptée au classement des protéines d'un nouveau type de récepteur selon leur interaction ou non à la B2M. Lorsque le récepteur est déjà connu, le problème de classification semble plus simple car toutes les protéines d'un même récepteur ont le même comportement vis-à-vis de la B2M.

3.4 Les interactions protéine-ligand au sein de la MhcSF

A l'issue de la procédure d'alignement et de numérotation des G-DOMAINs et G-LIKE-DOMAINs, nous disposons de 186 structures 3D de la MhcSF standardisées dans IMGT/3Dstructure-DB (160 structures 3D correspondant à 23 allèles de 10 protéines MHC-I, et 26 structures 3D correspondant à 12 allèles de 12 protéines MHC-I-like) (Annexe 5).

A partir de l'analyse de ces structures 3D, nous décrivons par la suite les contacts impliquant les G-DOMAINs et G-LIKE-DOMAINs des protéines MHC-I et MHC-I-like avec leurs ligands respectifs, et en particulier avec le C-LIKE-DOMAIN de la B2M. Ces liaisons sont non covalentes, et sont considérées comme non polaires ou polaires selon les atomes impliqués ; une liaison polaire peut correspondre ou non à une liaison hydrogène. Nous considérons par la suite que deux résidus sont en contact si la distance entre les centres de deux atomes (un atome de chaque résidu considéré) est inférieure à la somme de leurs rayons de Van der Waals et du diamètre d'une molécule d'eau ; un contact polaire correspond à une liaison hydrogène si cette distance est inférieure à 3.5 Å et que les angles des atomes sont corrects (Kaas et al. 2004).

3.4.1 Diversité fonctionnelle

Les protéines de la MhcSF appartiennent à 9 types de récepteurs (MHC-I, AZGP1, CD1, EPCR, FCGRT, HFE, MIC, MR1 et RAE) ; comme nous l'avons indiqué dans le chapitre 1, chaque type de récepteur correspond à une fonction biologique et à des ligands protéiques,

peptidiques ou lipidiques distincts. Les positions des G-DOMAINs et G-LIKE-DOMAINs impliquées dans ces contacts sont représentées en Figure 3.5.

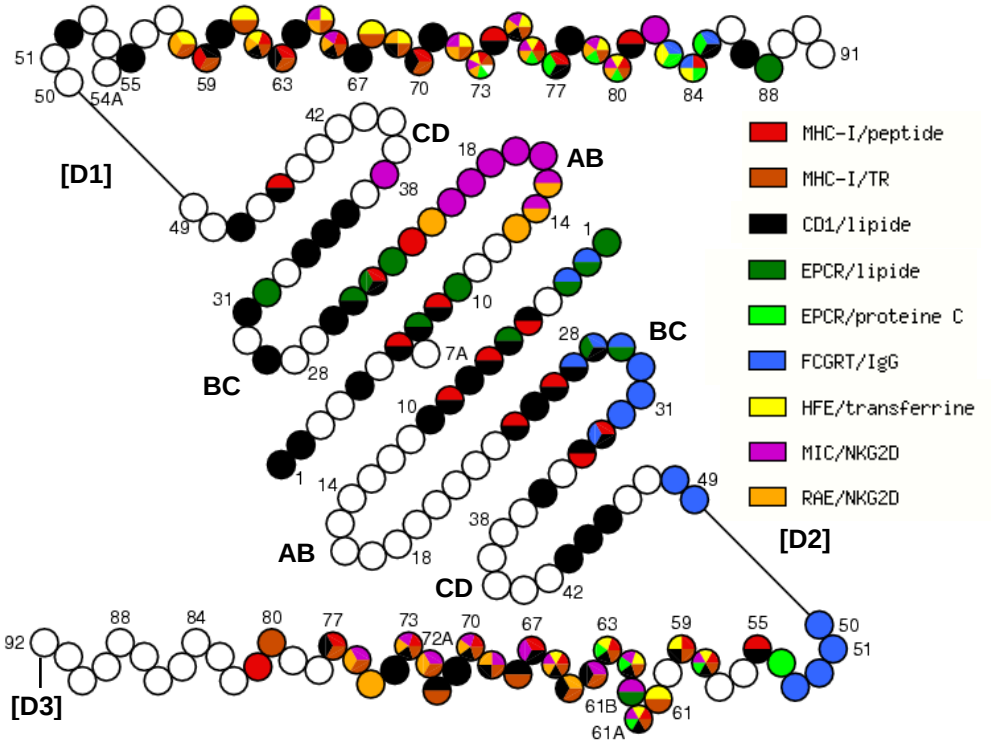


Figure 3.5. Sites des G-DOMAINs ou G-LIKE-DOMAINs des protéines de la MhcSF en contact avec leurs ligands. La chaîne I-ALPHA des protéines MHC-I et la chaîne I-ALPHA-LIKE des types de récepteurs AZGP1, CD1, FCGRT, HFE, MIC et MR1 comporte un C-LIKE-DOMAIN [D3]. Les ligands interagissent avec la face supérieure des G-DOMAINs et G-LIKE-DOMAINs. Les contacts MHC-I/peptide concernent des peptides de 8 acides aminés, et sont extraits de Kaas et al. 2005. Les contacts des protéines MR1 et AZGP1 ne sont pas représentés.

Le sillon formé par les G-DOMAINs des protéines MHC-I constitue le site de liaison de peptides endogènes de 8 à 10 acides aminés. Ces contacts sont décrits suite à l'analyse des 114 structures 3D de complexes peptide/MHC-I (pMHC-I) (Kaas et al. 2005, Publication 5). Nous avons défini onze sites de contact (C1 à C11), et leur correspondance avec les six poches (A à F) des protéines MHC-I décrites dans la littérature d'après leur conservation/polymorphisme en acides aminés (Matsumura et al. 1992, Madden 1995, Rammensee et al. 1995) ; les résidus des poches A et F sont majoritairement conservés et constituent les ancres des extrémités N et C-terminales des peptides, tandis que ceux des poches intermédiaires sont variables. Nous avons défini les sites C1 à C11 d'après des critères structuraux ; l'analyse de la variabilité de séquence des domaines G-ALPHA1 [D1] et G-ALPHA2 [D2] des protéines MHC-I (Fig. 2.2 et 2.3) révèle néanmoins une certaine concordance. Parmi les 17 positions des domaines G-ALPHA1 [D1] en contact avec les

peptides de 8 acides aminés, 14 présentent une variabilité supérieure à la moyenne des positions du domaine : 9 (C6), 22 (C6), 24 (C3), 45 (C3), 62 (C1), 63 (C1), 66 (C1), 70 (C6), 73 (C10), 74 (C6), 76 (C10), 77 (C10 et C11), 80 (C11) et 84 (C11) ; parmi les 17 positions des domaines G-ALPHA2 [D2] en contact avec les peptides de 8 acides aminés, 10 présentent une variabilité supérieure à la moyenne des positions du domaine : 7 (C6), 9 (C3, C4 et C6), 24 (C4 et C6), 26 (C6 et C11), 59 (C9 et C11), 61A (C9), 63 (C4 et C9), 67 (C4), 73 (C1) et 77 (C1).

Les G-DOMAINs et les peptides des complexes pMHC-I sont reconnus par les domaines V-ALPHA et V-BETA des TR ; ces contacts sont décrits suite à l'analyse des 9 structures 3D de complexe TR/pMHC-I (Kaas et al. 2005, Publication 5) (Annexe 5). Enfin, nous avons identifié les contacts impliquant les G-LIKE-DOMAINs des protéines CD1, EPCR, FCGRT, HFE, MIC et RAE, par l'analyse des 7 structures 3D des protéines MHC-I-like en complexe avec leurs ligands respectifs (Annexe 5). Chaque type de récepteur est caractérisé par des sites de contact distincts, couvrant la majorité des positions des G-LIKE-DOMAINs (Fig. 3.5) ; cette diversité rappelle celle décrite dans le paragraphe 3.2 au niveau des séquences de ces domaines.

3.4.2 B2M

Parmi les protéines de la MhcSF, seules celles appartenant aux types de récepteurs MHC-I, CD1, FCGRT et HFE sont liées à la B2M. Les positions des G-DOMAINs et G-LIKE-DOMAINs de ces protéines et celles du C-LIKE-DOMAIN de la B2M impliquées dans ces contacts sont représentées en Figures 3.6 et 3.7 et résumées en Table 3.3 ; les caractéristiques des liaisons atomiques impliquées dans ces contacts sont décrites en Table 3.4.

Les positions des G-DOMAINs et G-LIKE-DOMAINs impliquées dans l'interaction avec le C-LIKE-DOMAIN de la B2M diffèrent selon le type de récepteur ; ces sites d'interaction sont cependant chevauchant : 8 positions des domaines G-ALPHA1 et G-ALPHA1-LIKE [D1] (8-10, 23, 25, 27, 32 et 35), et 11 positions des domaines G-ALPHA2 et G-ALPHA2-LIKE [D2] (4, 6-8, 25-27, 29-32) sont impliquées dans l'interaction quel que soit le type de récepteur. Nous définissons la zone de contact potentiel MhcSF/B2M par l'ensemble des positions des G-DOMAINs et G-LIKE-DOMAINs de la MhcSF et du C-LIKE-DOMAIN de la B2M impliquées dans l'interaction dans au moins une des 171 structures 3D analysées (160 structures 3D de protéines MHC-I et 11 pour MHC-I-like, détaillées en Annexe 5). La localisation des sites de contact au niveau du C-LIKE-DOMAIN de la B2M est globalement

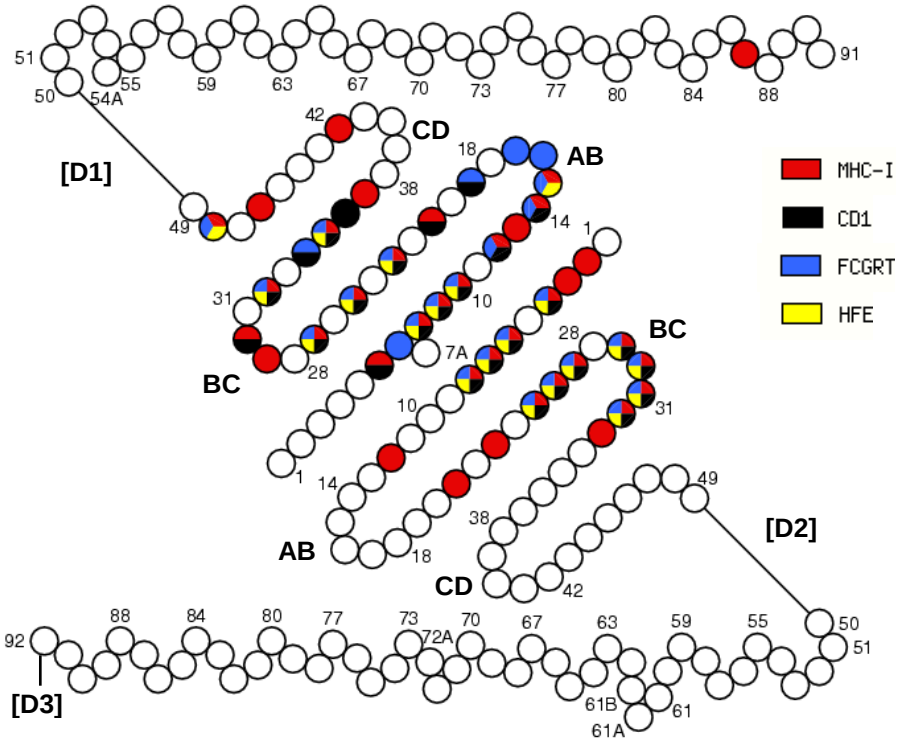


Figure 3.6. Sites des G-DOMAINs ou G-LIKE-DOMAINs des protéines de la MhcSF en contact avec le C-LIKE-DOMAIN de la B2M. Les sites de contact sont indiqués pour les types de récepteurs liés à la B2M : MHC-I, CD1, FCGRT et HFE ; ces positions sont en contact avec le C-LIKE-DOMAIN [D] de la B2M dans au moins une des 171 structures 3D de complexes MhcSF/B2M analysées (160 structures 3D de MHC-I et 11 structures 3D de MHC-I-like, détaillées en Annexe 5). La B2M interagit avec la face inférieure des G-DOMAINs et G-LIKE-DOMAINs.

conservée pour les différents types de récepteurs de la MhcSF, de même que la nature des contacts, essentiellement non polaires. Les résidus hydrophobes (IVLFCMAW) constituent 30-40% des résidus des deux G-DOMAINs ou G-LIKE-DOMAINs de la MhcSF, et 30-40% des résidus de ces domaines impliqués dans l'interaction avec la B2M ; 8-20% des résidus hydrophobes des deux G-DOMAINs ou G-LIKE-DOMAINs sont localisés dans cette zone de contact. Quel que soit le type de récepteur de la MhcSF, les résidus hydrophobes sont donc moins représentés dans la zone de contact qu'au sein des protéines complètes, et les interactions hydrophobes (impliquant une paire de résidus hydrophobes) contribuent peu à la liaison MhcSF/B2M. Ces observations nous laissent supposer que les propriétés physico-chimiques requises pour l'interaction MhcSF/B2M sont similaires quel que soit le type de récepteur.

La B2M est une protéine non polymorphe, dont la séquence diffère néanmoins entre les espèces. Comme nous l'avons indiqué dans le chapitre 1, le phénomène d'échange de B2M intraspécifique et interspécifique à la surface cellulaire a été décrit par Bernabeu et al. (1984) ;

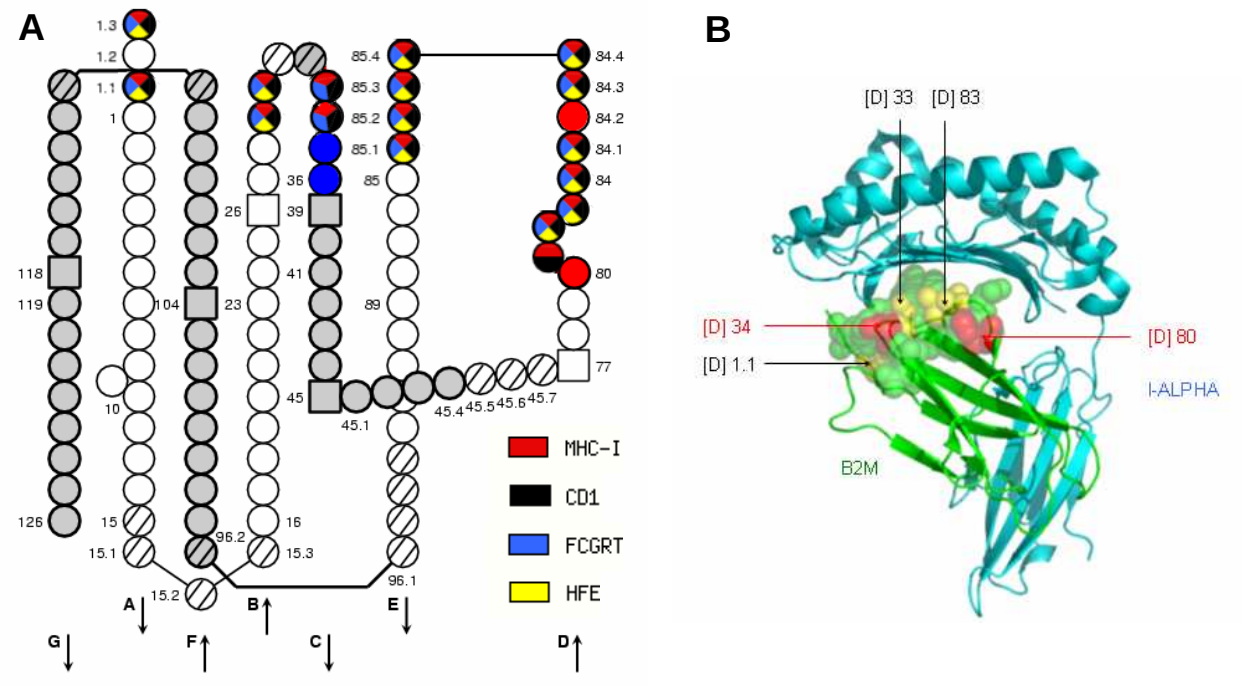


Figure 3.7. (A) Sites de la B2M en contact avec les G-DOMAINs ou G-LIKE-DOMAINs des protéines de la MhcSF et (B) variabilité interspécifique. (A) Collier de Perles sur 2 plans du C-LIKE-DOMAIN [D] de la B2M ; les positions hachurées correspondent à des positions manquantes par rapport à la numérotation unique des C-DOMAINs et C-LIKE-DOMAINs, pour les 4 espèces de mammifères étudiées (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*). Les sites de contact sont indiqués pour les types de récepteurs MhcSF liés à la B2M : MHC-I, CD1, FCGRT et HFE ; ces sites sont en contact avec les G-DOMAINs ou G-LIKE-DOMAINs dans au moins une des 171 structures 3D de complexes MhcSF/B2M analysées (160 structures 3D de MHC-I et 11 structures 3D de MHC-I-like, détaillées en Annexe 5). (B) Structure 3D d'une chaîne lourde I-ALPHA (MHC-I) liée à la B2M (*Homo sapiens*, IMGT/3Dstructure-DB 1oga). La chaîne I-ALPHA comprend les domaines G-ALPHA1 [D1] (devant), G-ALPHA2 [D2] (au fond) et C-LIKE [D3]. Les chaînes latérales des résidus localisés dans la zone de contact sont représentées par des sphères ; les sphères représentées en vert, jaune et rouge correspondent à des positions dont les acides aminés sont respectivement conservés, variables mais appartenant à un même groupe de propriétés physico-chimiques (parmi ceux définis en Annexe 6 d'après Wu et Brutlag 1995, et Pommié et al. 2004), et totalement variables.

Domaine	Label	Position
G-ALPHA1 [D1] et G-ALPHA1-LIKE [D1]	A-STRAND	6-10, 12-14
	AB-TURN	15-17
	B-STRAND	19, 21, 23, 25, 27
	BC-TURN	29, 30
	C-STRAND	32, 34-37
	D-STRAND	42, 46, 48
	HELIX	87
G-ALPHA2 [D2] et G-ALPHA2-LIKE [D2]	A-STRAND	2-4, 6-8, 12
	B-STRAND	21, 23, 25-27
	BC-TURN	29, 30
	C-STRAND	31-33
C-LIKE-DOMAIN [D]	A-STRAND	1.3, 1.1
	BC-LOOP	29, 30, 33-36
	D-STRAND	80-84
	DE-TURN	84.1-84.4, 85.4-85.1

Table 3.3. La zone de contact potentiel MhcSF/B2M. Les positions des G-DOMAINs ou G-LIKE-DOMAINs des protéines MHC-I, CD1, FCGRT et HFE (protéines de la MhcSF liées à la B2M) et du C-LIKE-DOMAIN de la B2M en contact sont décrites respectivement par la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs, et par la numérotation unique des C-DOMAINs et C-LIKE-DOMAINs. Ces positions sont en contact dans au moins une des 171 structures 3D de complexe MhcSF/B2M analysées (160 structures 3D de MHC-I et 11 structures 3D de MHC-I-like, détaillées en Annexe 5).

Espèce et nom de la protéine	H_{tot}	IMGT/ 3Dstructure -DB	Nombre de contacts atomiques MhcSF/B2M			H_{cont}	Nombre de contacts atomiques MhcSF/B2M impliquant une paire de résidus hydrophobes		
			Tot.	Pol.	Hydrog.		Tot.	Pol.	Hydrog.
Hs HLA-A	59 (31%)	1oga_A	410	34	7	7 (33%)	95	0	0
Hs HLA-B	59 (31%)	1k5n_A	384	31	5	6 (27%)	77	0	0
Hs HLA-Cw	56 (30%)	1qqd_A	412	38	6	8 (36%)	102	1	0
Hs HLA-E	58 (31%)	1mhe_A	566	29	7	8 (33%)	122	0	0
Mm H2-D1	58 (32%)	1wbx_A	407	40	5	7 (30%)	103	1	0
Mm H2-K1	58 (32%)	1lk2_A	423	38	6	9 (41%)	117	1	0
Mm H2-L	60 (33%)	1ld9_A	347	33	1	9 (38%)	107	1	0
Mm H2-Q7	58 (32%)	1k8d_A	431	41	6	9 (41%)	105	1	0
Mm H2-T3	60 (33%)	1nez_A	381	32	3	6 (30%)	55	0	0
Rn RT1-AA	58 (29%)	1kjv_A	426	52	5	8 (33%)	102	2	0
Hs CD1A	68 (38%)	1onq_A	403	43	5	8 (36%)	90	0	0
Hs CD1B	73 (41%)	1gzq_A	399	40	6	6 (26%)	72	1	0
Mm CD1D1	70 (37%)	1cd1_A	373	28	1	8 (38%)	100	0	0
Hs FCGRT	66 (38%)	1exu_A	432	37	7	9 (41%)	77	2	1
Rn FCGRT	68 (38%)	3fru_A	381	37	7	13 (50%)	82	2	1
Hs HFE	63 (35%)	1a6z_A	401	38	7	8 (40%)	86	0	0

Table 3.4. Les contacts MhcSF/B2M impliquant les G-DOMAINs ou G-LIKE-DOMAINs. Les contacts atomiques sont polaires (Pol.) ou non ; les liaisons hydrogènes (Hydrog.) correspondent à des contacts polaires. Les 4 types de récepteurs liés à la B2M sont représentés : MHC-I (*Homo sapiens* HLA-A, HLA-B, HLA-Cw, HLA-E, *Mus musculus* H2-D1, H2-K1, H2-L, H2-Q7, H2-T3 et *Rattus norvegicus* RT1-AA), CD1 (*Homo sapiens* CD1A et CD1B, *Mus musculus* CD1D1), FCGRT (*Homo sapiens* et *Rattus norvegicus*) et HFE (*Homo sapiens*). H_{tot} : nombre et pourcentage de résidus hydrophobes (IVLFCMAW) au sein des deux G-DOMAINs ou G-LIKE-DOMAINs de chaque protéine ; H_{cont} : nombre et pourcentage de résidus hydrophobes au sein de la zone de contact. *Hs* : *Homo sapiens*, *Mm* : *Mus musculus*, *Rn* : *Rattus norvegicus.*

cette observation indique que l'interaction des protéines MHC-I et de la B2M est régie par des propriétés physico-chimiques et structurales similaires, quelle que soit l'espèce considérée. L'échange de B2M n'a pas été décrit dans le cas des protéines MHC-I-like liées à la B2M ; l'analyse de la variabilité interspécifique du C-LIKE-DOMAIN de la B2M (Fig. 3.7 et 3.8) indique néanmoins une implication mineure des positions variables dans l'interaction avec les protéines de la MhcSF. Parmi les 21 positions de la zone de contact potentiel impliquée dans le contact avec les G-DOMAINs et G-LIKE-DOMAINs de la MhcSF, 5 ont une composition en acide aminé qui diffère selon l'espèce ([D] 1.1, 33, 34, 80 et 83). Seules les positions [D] 1.1 et 83 sont impliquées dans l'interaction quel que soit le type de récepteur ; les acides aminés observés pour ces positions appartiennent cependant à des groupes physico-chimiques identiques, respectivement hydrophile, large ou basique et hydrophobe ou large. Les positions [D] 34 et 80 comportent des acides aminés dont les propriétés physico-chimiques diffèrent selon l'espèce ; néanmoins, [D] 34 n'est pas impliqué dans l'interaction dans le cas des protéines HFE, et [D] 80 est impliquée uniquement dans le cas des protéines MHC-I.

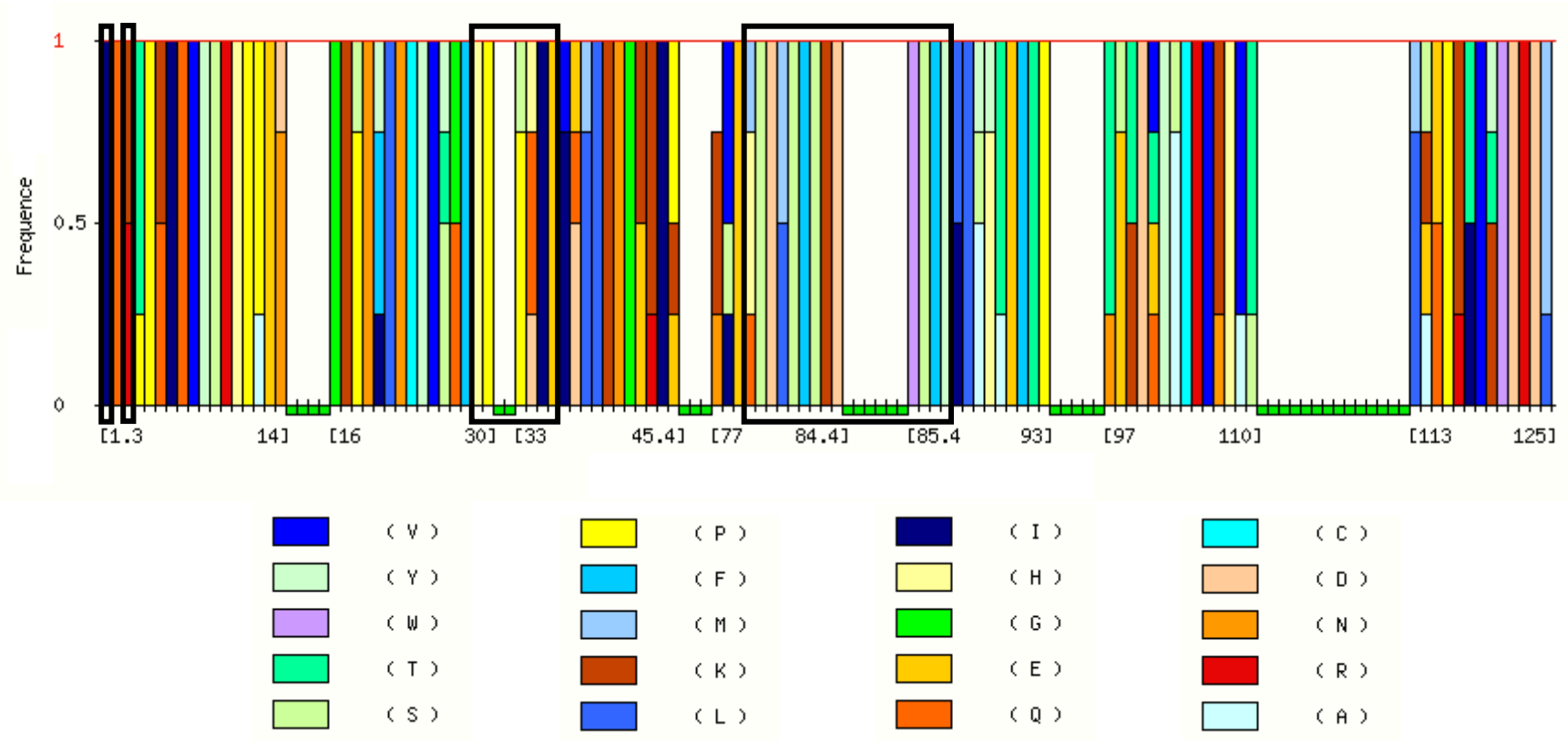


Figure 3.8. Variabilité interspécifique du C-LIKE-DOMAIN [D] de la B2M. La fréquence des 20 acides aminés est indiquée pour chaque position de l'alignement des C-LIKE-DOMAINs [D] de B2M d'*Homo sapiens*, *Mus musculus*, *Rattus norvegicus* et *Bos taurus* ; ces positions sont décrites par la numérotation unique des C-DOMAINs et C-LIKE-DOMAINs. Les positions pour lesquelles aucune fréquence n'est représentée correspondent à des positions manquantes (ou gaps) par rapport à la numérotation unique, pour les 4 espèces étudiées. La zone de contact potentiel (encadrée en noir) comprend les positions : 1.3 (I), 1.1 (acides aminés hydrophiles, larges et basiques ; *Hs* R, *Mm* K, *Rn* K, *Bt* R), 29 (H), 30 (P), 33 (neutre ; *Hs* S, *Mm* P, *Rn* P, *Bt* P), 34 (*Hs* D, *Mm* H, *Rn* Q, *Bt* Q), 35 (I), 36 (E), 80 (*Hs* H, *Mm* M, *Rn* H, *Bt* Q), 81 (S), 82 (D), 83 (hydrophobe et large ; *Hs* L, *Mm* M, *Rn* M, *Bt* L), 84 (S), 84.1-84.4 (FSKD), 85.4-85.1 (WSFY). Parmi ces sites, les positions 1.1, 33, 34, 80 et 83 présentent une spécificité d'espèce en terme d'acides aminés ; les acides aminés des positions 1.1, 33 et 83 correspondent à un même groupe d'acides aminés parmi ceux définis en Annexe 6 (Wu and Brutlag 1995, Pommié et al. 2004).

CHAPITRE 4

Classification fonctionnelle de familles protéiques

Nous avons décrit précédemment les caractéristiques de séquence et de structure des 47 protéines MHC-I et MHC-I-like, comportant deux G-DOMAINs ou G-LIKE-DOMAINs ; ces protéines correspondent à 4 espèces et à 9 types de récepteurs, et sont représentées par 806 allèles. Notre objectif est dorénavant de classer les protéines de la MhcSF selon qu'elles se lient ou non à la B2M et de déterminer les propriétés physico-chimiques qui, observées à une position donnée des G-DOMAINs et G-LIKE-DOMAINs, favorisent ou défavorisent cette interaction ; l'analyse des contacts MhcSF/B2M présentée dans le chapitre 3 nous laisse en effet supposer que les propriétés physico-chimiques requises pour cette interaction sont similaires quel que soit le type de récepteur et l'espèce considérée.

Les méthodes de classification automatique de données sont apparues avec l'expansion de l'informatique et la nécessité de gérer des volumes importants de données de structure complexe. La classification non supervisée de données n'est basée sur aucune connaissance à priori de leur structure, et consiste à identifier les classes sous-jacentes c'est-à-dire à déterminer le découpage optimal des données en classes. Au contraire, la classification supervisée consiste à apprendre une fonction de classification à partir d'un échantillon de données (nommé échantillon d'apprentissage) dont le découpage en classes est connu a priori ; cette fonction permet alors de classer automatiquement de nouvelles observations. La classification fonctionnelle de familles protéiques constitue une application de ces techniques ; chaque classe correspond à un ensemble de protéines partageant la même fonction et/ou les mêmes interactions protéine-ligand.

Nous présentons tout d'abord le principe des méthodes de classification non supervisée et quelques unes de leurs applications à la classification fonctionnelle de familles protéiques ; nous nous limitons aux approches dites hiérarchiques, et considérons l'apport des modèles d'évolution moléculaire. Nous exposons ensuite le principe et quelques applications des principales méthodes de classification supervisée, selon l'ordre chronologique de leur développement. Nous présentons enfin la méthodologie que nous avons mise en place afin de prédire l'interaction des protéines de la MhcSF avec la B2M à partir de leur séquence ; cette approche combine un classifieur Bayesien naïf et la numérotation unique IMGT des G-DOMAINs et G-LIKE-DOMAINs (Duprat et al. 2005a, 2005b, Publications 6 et 7).

4.1 Classification non supervisée

La méthode de classification hiérarchique UPGMA (Sneath and Snokal 1973) consiste à définir une hiérarchie de partitions à partir d'une matrice de distances ou de dissimilarités entre les observations, au cours d'un processus itératif. Chaque observation est tout d'abord considérée comme un groupe ; à chaque itération, les deux groupes les plus similaires sont agrégés et la matrice de distances est remise à jour. Le processus agglomératif se termine lorsque toutes les observations constituent un unique groupe. La classification hiérarchique peut être représentée par un diagramme en 2 dimensions nommé dendrogramme (Fig. 4.1) ; la fixation d'une valeur seuil de dissimilarité, en fonction du nombre de classes souhaité ou d'après l'optimisation de mesures objectives, aboutit à un découpage des données en classes.

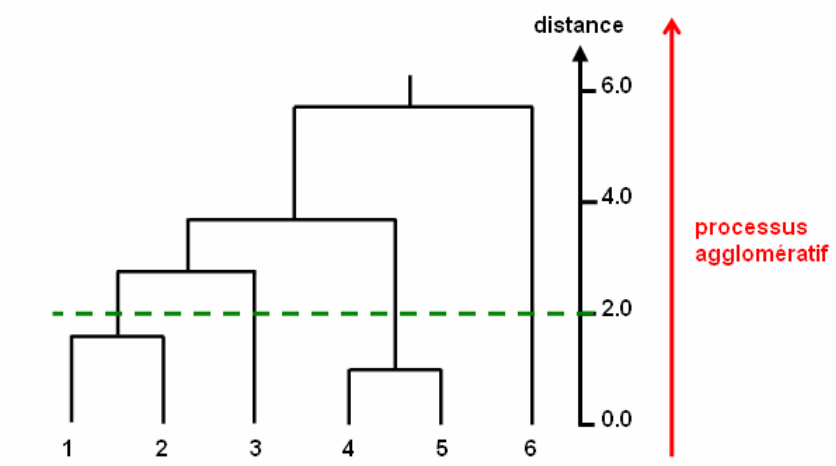


Figure 4.1. Dendrogramme et seuil de dissimilarité. Cette représentation illustre les différentes étapes du processus agglomératif. Si l'on considère un seuil de dissimilarité de 2.0, les 6 observations représentées sont découpées en quatre classes : C1 (comportant les observations 1 et 2), C2 (obs. 3), C3 (obs. 4 et 5) et C4 (obs. 6).

Dans le cas de familles de protéines, la matrice de distance initiale est en général dérivée d'un alignement multiple et représente la dissimilarité des paires de séquences alignées. La méthode UPGMA est la plus ancienne méthode de distance utilisée pour la reconstruction d'arbres phylogénétiques. Elle se base sur l'hypothèse d'horloge moléculaire (Zuckerland and Pauling 1962) et aboutit donc à une topologie inexacte pour des séquences caractérisées par des taux de mutation variables au cours du temps ; la méthode de distance Neighbor-Joining (Saitou and Nei 1987) corrige la méthode UPGMA en prenant en compte la divergence moyenne de chaque séquence avec les autres, et constitue une méthode de reconstruction plus fidèle de l'évolution moléculaire. La méthode BIONJ (Gascuel 1997) constitue une amélioration de l'algorithme Neighbor-Joining et est particulièrement adaptée lorsque les distances évolutives sont calculées à partir de séquences alignées.

Wicker et al. (2001) ont développé Secator, une méthode automatique de classification fonctionnelle de familles protéiques à partir d'un arbre phylogénétique non enraciné ; l'arbre est reconstruit par BIONJ à partir d'une matrice de distances issue de l'alignement multiple des séquences de la famille à classer, déterminée par défaut à partir du pourcentage d'identité des résidus. Le principe de cette approche est de collapser les branches courtes, voisines au sein de l'arbre (Fig. 4.2) ; les classes inférées sont séparées par des longues branches. Cette méthode est particulièrement adaptée à la classification des protéines de la MhcSF selon leurs fonctions, c'est-à-dire selon les 9 types de récepteurs (Fig. 4.3). A l'exception des protéines RAE qui sont séparés en deux classes fonctionnelles par Secator (RAE d'*Homo sapiens*, et RAE de *Mus musculus* et *Rattus norvegicus*), chaque classe fonctionnelle prédite correspond exactement à un type de récepteur.

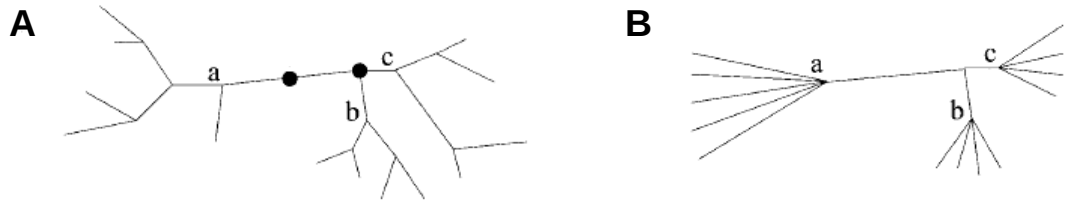


Figure 4.2. Principe de Secator (extrait de Wicker et al. 2001). (A) Représentation schématique d'un arbre reconstruit par BIONJ (Gascuel 1997). Secator identifie automatiquement les nœuds joignant différents sous-arbres (représentés par des points), et collapsent les branches depuis les feuilles jusqu'aux branches internes, jusqu'à ce qu'un tel nœud soit rencontré ; le processus de collapse des branches s'arrête donc ici aux nœuds a, b et c. (B) Représentation schématique du résultat de Secator, après collapse des branches ; 3 classes sont inférées.

La méthode de Trace Evolutive développée par Lichtarge et al. (1996) se base également sur un arbre phylogénétique pour inférer des classes fonctionnelles à partir d'une famille de protéines. Cette approche n'est pas automatique : le niveau de coupure de l'arbre est en effet déterminé en fonction d'informations a priori fournies par l'utilisateur, telles que le pourcentage d'identité de séquence minimum au sein des classes (Lichtarge et al. 1996) ou le nombre de classes attendues (Lichtarge and Sowa 2002). Outre l'inférence de classes fonctionnelles, cette méthode permet d'expliquer la partition des données en classes, par l'identification des résidus constituant la Trace Evolutive. Ces positions comprennent des acides aminés totalement conservés au sein de chaque classe, et distincts entre les classes ; de telles positions sont relativement rares voir absentes pour des familles protéiques hétérogènes, et les travaux de Mihalek et al. (2004) permettent dorénavant d'identifier les positions les plus discriminantes entre les classes, par l'évaluation de l'entropie relative position-dépendante au sein de l'alignement multiple.

Dans le cas de la MhcSF, la combinaison de Secator et de la méthode de Trace Evolutive nous permettrait de classer automatiquement les protéines MHC-I et MHC-I-like selon leur type de récepteur, et d'identifier les positions informatives qui confèrent à ces protéines leur fonction respective. Comme nous l'avons mis en évidence dans le chapitre 3, la MhcSF a subi une perte partielle de la capacité ancestrale de liaison à la B2M au cours du processus évolutif. La phylogénie d'une famille ne reflète donc pas toujours la spécialisation des protéines qui la constituent, et les approches de classification non supervisée présentées précédemment s'avèrent par conséquent inadaptées dans de tels cas ; les protéines de la MhcSF ne peuvent donc pas être classées selon leur interaction ou non avec la B2M par une approche non supervisée (Fig. 4.3). Cette problématique de classification des protéines de la MhcSF correspond néanmoins à des classes déterminées expérimentalement, et l'utilisation de méthodes de classification supervisée est alors recommandée.

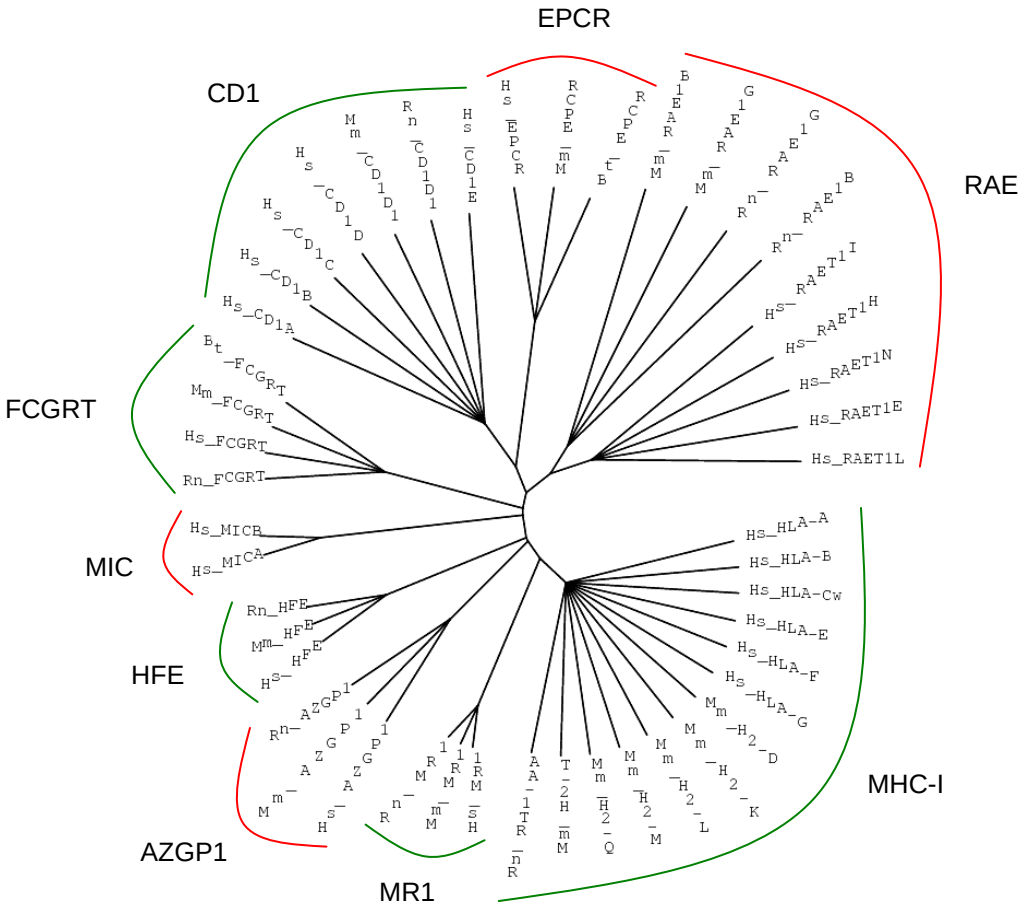


Figure 4.3. Résultat de Secator pour les 47 protéines de la MhcSF. Secator infère 10 classes de protéines. Parmi les 9 types de récepteurs de la MhcSF (MHC-I, AZGP1, CD1, EPCR, FCGRT, HFE, MIC, MR1 et RAE), 8 sont correctement prédits par Secator ; RAE est cependant séparé en deux classes, comprenant respectivement les protéines RAE d'*Homo sapiens*, et celles de *Mus musculus* et *Rattus norvegicus*. L'augmentation de la valeur du seuil de dissimilarité définissant les classes (déterminée automatiquement par défaut) permet à l'utilisateur de diminuer le nombre de classes inférées. Nous obtenons ainsi 8 classes : les types de récepteurs sont correctement classés, à l'exception des protéines CD1 et EPCR qui constituent une unique classe. Cette approche ne permet pas la classification des protéines de la MhcSF selon qu'elles se lient (en vert) ou non (en rouge) à la B2M.

4.2 Classification supervisée

Les méthodes de classification supervisée (Gascuel et al. 1998) se basent majoritairement sur la règle de Bayes. La règle de classement avec erreur minimale consiste en effet à assigner une nouvelle observation x à la classe C_l pour laquelle la probabilité conditionnelle $P(C_l \mid x)$ est maximale ; cette probabilité est exprimée par la règle de Bayes

$$P(C_l \mid x) = \frac{P(C_l)P(x \mid C_l)}{P(x)}. \qquad (1)$$

La règle de classement optimale correspond ainsi à la fonction discriminante $F_l(x) = P(C_l)P(x \mid C_l)$, où la classe C_l est assignée à x si $F_l(x) = \max_j F_j(x)$.

Les classifieurs paramétriques considèrent que l'expression de la distribution conditionnelle des observations est connue, mais que ses paramètres sont inconnus. Ces méthodes consistent donc à déterminer les paramètres de la fonction discriminante, qui peut-être linéaire ou non (Fig. 4.4) ; au contraire, les classifieurs non paramétriques estiment localement $P(x \mid C_l)$.

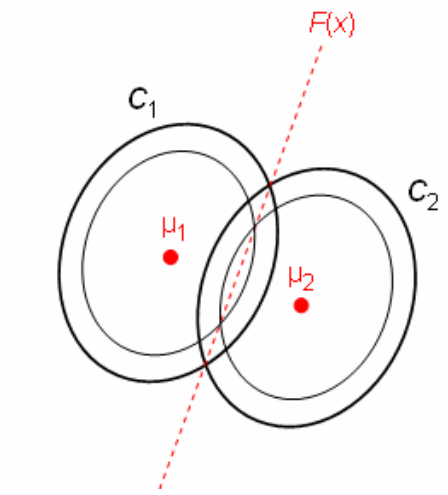


Figure 4.4. Principe des méthodes paramétriques binaires. Les paramètres de la distribution conditionnelle des observations correspondent à la moyenne μ_l et à la matrice de covariance de la population de chaque classe C_l. Ils sont estimés pour un échantillon d'apprentissage, et permettent d'établir la fonction discriminante optimale $F(x)$; cette fonction est linéaire si les matrices de covariance des deux classes sont identiques, et quadratique dans le cas contraire. Les stratégies destinées à étendre les procédures binaires aux cas multiclasses (k classes) consistent entre autres à construire (i) k classifieurs binaires, chacun discriminant une classe des autres, ou (ii) $k(k\text{-}1)$ classifieurs binaires, chacun discriminant deux classes.

La fonction discriminante est apprise par estimation de la distribution conditionnelle des observations (méthodes non paramétriques) ou de ses paramètres (méthodes paramétriques), pour un échantillon des observations nommé échantillon d'apprentissage ; cette fonction permet alors de classer automatiquement de nouvelles observations, et la performance du

classifieur est évaluée pour un échantillon de test indépendant de l'échantillon d'apprentissage (Annexe 7). Certaines méthodes de classification supervisée sont dites explicatives, car outre la prédiction de la classe de nouvelles observations, elles expliquent la partition des données en classes.

Chaque observation est représentée par un vecteur d'attributs qualitatifs ou quantitatifs. Comme nous allons le voir, la classification de familles protéiques se base sur différents types de recodage des données en vecteurs, en fonction des caractéristiques du problème biologique considéré ; les observations peuvent correspondre à des séquences alignées ou non, et les attributs peuvent représenter par exemple le nombre d'occurrence de motifs de *n* lettres parmi l'alphabet des acides aminés ou de leurs propriétés physico-chimiques (*n*-gram), ou l'hydrophobicité et la charge des positions en contact au sein des complexes protéine-ligand.

4.2.1 Analyse discriminante de Fisher

L'analyse linéaire discriminante de Fisher (1936) est la méthode de classification supervisée la plus ancienne. Chaque observation est représentée par un vecteur d'attributs quantitatifs. Cette approche suppose que la distribution des observations de chaque classe est gaussienne, avec comme paramètres la moyenne et la matrice de covariance de la population de la classe ; la phase d'apprentissage de cette méthode paramétrique consiste donc à estimer ces paramètres par la moyenne et la matrice de covariance des observations de chaque classe dans l'échantillon d'apprentissage. La fonction de discrimination est linéaire si les matrices de covariance des différentes classes sont identiques, et quadratique dans le cas contraire. Cette méthode n'est pas explicative.

4.2.2 *k* plus proches voisins (*k*-NN)

L'approche des k-NN est non paramétrique et basée sur une idée intuitive, qui consiste à assigner à une nouvelle observation x la classe la plus représentée parmi ses k plus proches voisins (Fix and Hodges 1951) ; chaque observation est représentée par un vecteur d'attributs quantitatifs. Cette méthode ne comprend pas de phase d'apprentissage, et l'ensemble des observations du jeu de données doit être considéré à nouveau pour chaque classement ; cette approche est par conséquent coûteuse en temps de calcul. Cette méthode n'est pas explicative.

La méthode des k-NN constitue l'approche la plus intuitive de classification supervisée des séquences d'une famille protéique : la distance entre la séquence à classer et chaque séquence du jeu de données est estimée par le score de similarité issu de leur alignement par Blast

(Altschul et al. 1990) ou Fasta (Pearson and Lipman 1988), et la classe prédite est celle de la séquence la plus similaire ($k=1$). Cette approche est néanmoins inefficace dans le cas où la spécificité de classe est conférée par des motifs particuliers (Bork and Koonin 1996) : deux séquences similaires peuvent en effet comporter des motifs fonctionnels différents, de même que des séquences fortement éloignées peuvent présenter un même motif. Nous avons mis en évidence dans le chapitre 3 qu'une séquence de la MhcSF appartenant à un nouveau type de récepteur n'a pas le même comportement vis-à-vis de la B2M que ses plus proches voisins.

4.2.3 Bayes naïf

Le classifieur Bayesien naïf (Good 1965) est dédié à la classification de données représentées par des vecteurs d'attributs qualitatifs. Une nouvelle observation est classée d'après la règle de classement avec erreur minimale exprimée par la règle de Bayes (1), en supposant l'indépendance conditionnelle des attributs ; la fonction de discrimination de ce classifieur est non linéaire. Cette hypothèse est simplificatrice mais s'est avérée efficace pour de très nombreux jeux de données réels, même avec des attributs fortement corrélés; cette propriété est expliquée par des arguments théoriques par Domingos et Pazzani (1996). Ce classifieur est sensible à un grand nombre d'attributs non discriminants, et nécessite par conséquent la sélection explicite d'attributs pour leur capacité à discriminer les observations des différentes classes ; les attributs sélectionnés (d'après des connaissances a priori des données ou par des méthodes objectives présentées en Annexe 8) sont nommées descripteurs. L'analyse des probabilités des descripteurs conditionnellement à la classe au sein de l'échantillon d'apprentissage permet de déterminer leur contribution relative dans la partition des données en classes. Outre la simplicité de sa mise en œuvre et son caractère explicatif, ce classifieur a l'avantage de s'accommoder d'un échantillon d'apprentissage de taille restreinte.

Bandyopadhyay et al. (2002) ont utilisé un classifieur Bayesien naïf pour prédire l'interaction de deux protéines à partir de leur séquence ; les auteurs s'intéressent plus particulièrement à la classification des couples de protéines SH3/PxxP selon leur interaction ou non. Les domaines protéiques SH3 jouent un rôle dans la transduction des signaux intracellulaires, en liant ou non des protéines dont la séquence comporte un motif palindromique PxxP (x représentant n'importe quel acide aminé) ; 3 classes de partenaires sont considérées, correspondant respectivement à une liaison de type I ou II selon l'orientation du motif par rapport au domaine SH3, ou à une absence de liaison. Chaque couple de protéines est représenté par un vecteur d'attributs qualitatifs décrivant l'hydrophobicité (+1 ou -1) et la charge (+1, 0 ou -1)

des acides aminés localisés au niveau des sites de contact potentiel entre deux partenaires de ces familles protéiques ; ces sites ont été identifiés au préalable par l'analyse des contacts au sein des structures 3D correspondant aux observations positives. Les auteurs évaluent la performance de leur classifieur à 90% de classement correct, par une procédure de validation croisée. Le caractère explicatif de ce classifieur permet aux auteurs d'identifier les sites et leurs valeurs de charge et d'hydrophobicité contribuant le plus à chaque type d'interaction.

4.2.4 Segmentation par arbre binaire (CART)

La méthode CART (Breiman 1984) est basée sur la construction d'un arbre de décision généralement binaire, par segmentations successives des observations de l'échantillon d'apprentissage. Ce classifieur non paramétrique est capable de classer des données représentées par des vecteurs d'attributs de différents types ; la fonction discriminante est non linéaire. L'apprentissage du classifieur CART est une procédure récursive, qui traite chaque nœud terminal de l'arbre en construction en 3 étapes : (i) énumération de l'ensemble des partitions binaires possibles pour chaque attribut des vecteurs de données, (ii) sélection de la partition qui maximise la réduction de l'impureté (un nœud pur comprend uniquement des observations de même classe), (iii) estimation des probabilités conditionnelles $P(t \mid C_l)$ pour chaque nouveau nœud terminal t. A la fin de la phase d'apprentissage, chaque nœud non terminal est associé à un attribut discriminant (nommé descripteur) et à une règle de décision, correspondant à une valeur seuil dans le cas d'un attribut quantitatif ; la capacité des descripteurs à discriminer les données des différentes classes est ainsi hiérarchisée au sein de l'arbre. Le classement d'une nouvelle observation consiste à lui faire parcourir l'arbre jusqu'à un nœud terminal, et à lui assigner la classe dont la probabilité conditionnellement à ce nœud (exprimée par la règle de Bayes) est maximale. Ce classifieur sélectionne explicitement les descripteurs et est donc insensible à la représentation des données par un grand nombre d'attributs non discriminants.

McLaughlin et Berman (2003) ont utilisé un classifieur CART pour prédire, à partir de la structure 3D d'une protéine comportant trois hélices successives, sa capacité de liaison à l'ADN ; il s'agit d'un problème de classification binaire d'une famille structurale de protéines. Les 152 structures 3D du jeu d'apprentissage comportent trois hélices successives notées respectivement H, $H-1$ et $H-2$; l'hélice H comporte le site de liaison à l'ADN pour les 76 observations positives. Les auteurs utilisent un classifieur explicatif afin de déterminer les caractéristiques structurales requises pour qu'un tel motif soit associé à la

capacité de liaison à l'ADN. Chaque structure 3D est représentée par un vecteur d'attributs quantitatifs décrivant différentes mesures géométriques des hélices et de leur environnement structural. Quatre descripteurs sont sélectionnés au cours de la phase d'apprentissage et pris en compte dans l'arbre final, selon la hiérarchie suivante : (i) aire de l'interaction hydrophobe entre les hélices *H* et *H*-2, supérieure ou non à 44 $Å^2$; (ii) accessibilité moyenne des résidus de *H* à la surface, supérieure ou non à 30% ; (iii) aire de l'interaction hydrophobe entre les hélices *H* et *H*-1, supérieure ou non à 37 $Å^2$; (iv) aire de l'interaction hydrophobe entre les hélices *H*-1 et *H*-2, supérieure ou non à 61 $Å^2$. Les auteurs évaluent la performance de leur classifieur à 91% de classement correct, par une procédure de validation croisée. Cette approche permet d'identifier ce motif de liaison à l'ADN au sein de protéines sans homologie de séquence avec les protéines de l'échantillon d'apprentissage ; l'apport de cette étude est également d'expliquer cette interaction, en indiquant quels critères structuraux sont requis.

4.2.5 Réseaux de neurones multicouches (ANN)

Ce classifieur est non paramétrique, et classe une nouvelle observation x par une fonction discriminante non linéaire construite par la combinaison de fonctions sigmoïdes ; chaque observation est représentée par un vecteur d'attributs quantitatifs. Les neurones d'un réseau multicouches sont organisés au minimum en trois couches (entrée, cachée et sortie). Le nombre de neurones des couches d'entrée et de sortie correspond respectivement au nombre d'attributs des vecteurs de données et au nombre de classes considérées ; la valeur d'un neurone l de la couche de sortie est une estimation de la probabilité conditionnelle $P(C_l \mid x)$. Chaque neurone du réseau est associé à une fonction sigmoïde et à un poids ; la phase d'apprentissage est une procédure itérative d'optimisation des poids, qui consiste à classer chaque observation de l'échantillon d'apprentissage, en évaluant puis rétro-propageant les erreurs de classement à travers le réseau. La règle d'arrêt de cette procédure est la convergence de l'erreur vers un minimum. L'algorithme de rétro-propagation qui a permis l'émergence des réseaux de neurones multicouches a été développé par Rumelhart et al. (1986). Ce classifieur tend à minimiser l'erreur de classement et est par conséquent très performant ; de plus, le classement de nouvelles observations est très rapide. Néanmoins, ce classifieur est performant pour des jeux de données de taille importante, n'est pas explicatif et nécessite une durée importante d'apprentissage (du fait de la convergence de l'erreur).

Wu et al. (1992) ont développé un système de classification de familles protéiques à partir de leurs séquences, basé sur des réseaux de neurones multicouches. Ils évaluent la capacité de ce

système ProCANS (Protein Classification Artificial Neural System) à classer les protéines de la base de données PIR d'après leur fonction. PIR comprend 6 catégories de protéines réparties en 620 classes fonctionnelles ; les auteurs entraînent un réseau (constitué de 3 couches) distinct pour chaque catégorie. Chaque séquence protéique est représentée par un vecteur d'attributs quantitatifs ; les auteurs évaluent la performance et la vitesse de convergence des classifieurs pour différents types d'attributs : (i) nombre d'occurrence d'un n-gram donné (motif de n lettres qui se suivent), pondéré par sa fréquence d'occurrence dans la nature, (ii) position moyenne d'un n-gram donné (les positions sont décrites par la numérotation des séquences alignées), (iii) combinaison de ces deux types d'attributs. Les auteurs mettent en évidence que l'encodage de type (iii) ralentit la convergence mais accroît la performance du classifieur. Deux types de classifieurs s'avèrent particulièrement performants, permettant de classer correctement 90% de 8309 protéines de la base de données PIR selon leur fonction ; les attributs combinent le nombre d'occurrence et la position moyenne des 2-gram pour l'alphabet des 20 acides aminés, ou des 3-gram pour un alphabet à 6 lettres correspondant à des groupes d'acides aminés similaires (d'après la matrice PAM).

4.2.6 Machines à vecteurs supports (SVM)

La méthode des SVM a été développée en 1995 par Vapnik. Chaque observation est représentée par un vecteur d'attributs quantitatifs. Dans la cas binaire, l'apprentissage de ce classifieur non paramétrique consiste à identifier l'hyperplan qui discrimine le mieux les observations des deux classes, en garantissant que la marge entre le vecteur le plus proche de chaque classe soit maximale ; la maximisation de la marge aboutit à la minimisation de l'erreur de classement (Vapnik and Chervonenkis 1971). Les vecteurs localisés au niveau des marges de l'hyperplan sont nommés les Vecteurs Supports ; ces vecteurs sont les plus informatifs et sont utilisés pour le classement de nouvelles observations. Les SVM constituent par conséquent une méthode rapide, qui minimise de plus l'erreur de classement ; cette approche n'est pas explicative, et nécessite des jeux de données de taille importante. Ce classifieur est capable d'effectuer des séparations non linéaires des données, en remplaçant les valeurs des attributs initiaux dans la fonction discriminante par une fonction de ces attributs ; cette fonction peut par exemple être polynomiale, et constitue le noyau du classifieur.

Huang et al. (2004) ont utilisé un classifieur SVM pour prédire si deux protéines de *Saccharomyces cerevisiae* interagissent ou non, à partir de leurs catégories fonctionnelles et structurales respectivement annotées par FUNCAT (Ruepp et al. 2004) et SCOP (Murzin et

al. 1995) et notées *F* et *S* par la suite ; il s'agit d'un problème de classification binaire d'interactions protéiques. Chaque couple de protéines est représenté par un vecteur d'attributs quantitatifs décrivant l'ensemble des combinaisons $F_1S_1F_2S_2$ possibles pour deux partenaires ; seul l'attribut associé à la combinaison observée a une valeur non nulle. Les 2594 couples de protéines qui constituent les observations positives sont extraits des bases de données d'interaction. Faute d'information concernant les couples de protéines qui n'interagissent pas, les auteurs génèrent 2571 observations négatives de façon pseudo-aléatoire ; les caractéristiques F_1S_1 et F_2S_2 des couples simulés sont sélectionnées séparément parmi celles représentées au sein des observations positives. La fonction gaussienne est utilisée comme noyau du classifieur. Les auteurs évaluent la performance de leur classifieur à 79% de classement correct, par une procédure de validation croisée.

4.3 Classifieur Bayesien naïf et numérotation unique IMGT

La classification des protéines de la MhcSF selon leur interaction ou non avec la B2M correspond à un problème de classification supervisée binaire, les deux classes correspondant aux protéines respectivement liées (30 protéines représentées par 784 allèles) et non liées (17 protéines, 22 allèles) à la B2M.

Parmi les méthodes de classification supervisée, le classifieur Bayesien naïf est le plus adapté à notre problématique car il s'adapte particulièrement bien aux petits jeux de données et est explicatif. La méthodologie que nous avons mise en place afin de prédire l'interaction des protéines de la MhcSF avec la B2M combine par conséquent un classifieur Bayesien naïf et la numérotation unique IMGT des G-DOMAINs et G-LIKE-DOMAINs (Duprat et al. 2005a, 2005b, Publications 6 et 7) ; nous souhaitons déterminer les propriétés physico-chimiques qui, observées à une position donnée des G-DOMAINs et G-LIKE-DOMAINs, favorisent ou défavorisent cette interaction, et représentons donc chaque descripteur par l'association d'une position dans l'alignement et d'un groupe d'acides aminés (parmi 38 groupes non redondants définis à partir des regroupements présentés en Annexe 6). Chaque descripteur est binaire, un groupe étant observé ou non à une position donnée d'une séquence.

Deux étapes sont nécessaires pour construire le classifieur : (i) la sélection d'un ensemble de descripteurs d'après leur capacité de discrimination entre les deux classes, évaluée par des critères statistiques (les mesures du χ^2 et de l'information mutuelle, décrites en Annexe 8, donnent des résultats identiques) ; (ii) l'apprentissage des descripteurs, c'est-à-dire l'estimation des fréquences des descripteurs conditionnellement à la classe. Nous évaluons la

performance de notre classifieur par 3 procédures de « leave-one-out », respectivement pour les prédictions concernant une nouvelle protéine, une espèce non référencée au sein des données, ou un nouveau type de récepteur ; cette stratégie de validation croisée est requise par la taille restreinte du jeu de données et leur similarité particulièrement importante pour un même type de récepteur.

Cette méthodologie permet de classer efficacement les protéines de la MhcSF selon leur liaison ou non à la B2M : 98%, 94% et 70% des données correspondant respectivement à une nouvelle protéine, une espèce non référencée au sein des données et un nouveau type de récepteur sont classées correctement ; ce dernier résultat est satisfaisant, car il correspond à une procédure au cours de laquelle le pourcentage d'identité des séquences de test avec les séquences d'apprentissage ne dépasse pas 45%. Ce classifieur est sélectionné pour sa performance optimale quelle que soit la procédure de leave-one-out ; il est constitué de 18 descripteurs, décrits en Table 4.1.

Domaine	Position	Groupe d'acides aminés	Type de descripteur
[D1]	8	CDPNT	3
	11	ILV	4
	12	MILKR	3
	21	W	3
	25	DNEQKR	3
	27	FYW	1
	32	EVQH	1
	35	EVQH	3
	51	W	2
	74	MILKR	4
	86	NQ	2
	88	CDPNT	4
[D2]	10	G	2
	27	AG	1
	32	DE	1
	39	EVQH	4
	83	DE	2
	85	G	2

Table 4.1. Les 18 descripteurs sélectionnés. [D1] et [D2] indiquent respectivement les domaines G-ALPHA1 [D1] et G-ALPHA1-LIKE [D1], et les domaines G-ALPHA2 [D2] et G-ALPHA2-LIKE [D2]. Les positions sont décrites selon la numérotation unique IMGT des G-DOMAINs et G-LIKE-DOMAINs. Les descripteurs de type 1 et 2 sont favorables à l'interaction des protéines de la MhcSF avec la B2M ; les descripteurs de type 3 et 4 sont défavorables à cette interaction. Les descripteurs de type 1 et 3 correspondent à une position localisée dans la zone de contact potentiel avec la B2M.

Les descripteurs peuvent être classés en quatre types, selon qu'ils sont favorables ou défavorables à la liaison à la B2M (c'est-à-dire que le groupe d'acide aminé discriminant à cette position est respectivement majoritaire ou conservé pour les protéines liées à la B2M et

minoritaire ou absent pour les protéines non liées, et inversement), et qu'ils sont localisés ou non dans la zone de contact potentiel à la B2M (définie dans le chapitre précédent). Afin de comprendre comment ces descripteurs favorisent ou défavorisent le contact des protéines de la MhcSF avec la B2M, nous avons analysé leur contexte structural pour deux protéines de cette superfamille, chacune étant représentative d'une classe (Fig. 4.5).

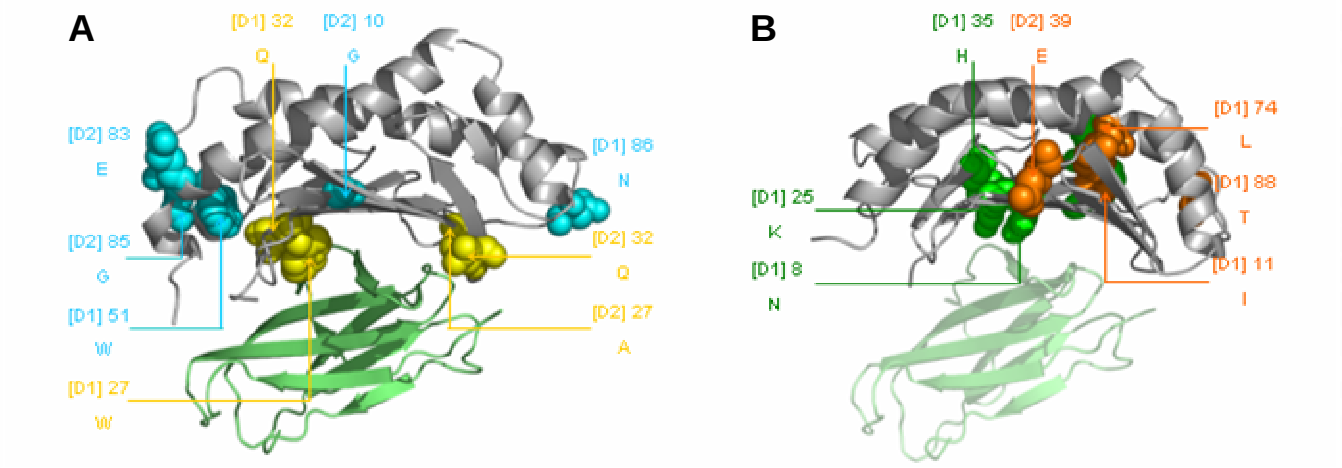


Figure 4.5. Contexte structural des quatre types de descripteurs pour les protéines (A) FCGRT de *Rattus norvegicus* et (B) RAE1B de *Mus musculus*. Chaque chaîne lourde MHC-I-like (en gris) comprend les domaines extracellulaires [D1] (au fond) et [D2] (devant). La B2M (en vert) est complexée à la protéine FCGRT, et placée virtuellement pour la protéine RAE1B (par superposition de la structure 3D de cette protéine avec celle de la protéine H2-D1 de *Mus musculus*, liée à la B2M ; le fichier de coordonnées de H2-D1 utilisé est 1wbx et la superposition est réalisée par le programme ProFit). Le domaine C-LIKE extracellulaire de la chaîne lourde de la protéine FCGRT n'est pas représenté. Le domaine, la position et l'acide aminé observé dans la structure 3D sont indiqués pour chaque descripteur ; les chaînes latérales de ces acides aminés sont mises en évidence par des sphères. (A) (Fichier de coordonnées 3fru) Les descripteurs localisés dans la zone de contact potentiel à la B2M sont représentés en jaune, et ceux localisés hors de cette zone sont représentés en cyan. (B) (Fichier de doordonnées 1jfm) Les descripteurs localisés dans la zone de contact potentiel à la B2M sont représentés en vert, et ceux localisés hors de cette zone sont représentés en orange.

Les 9 descripteurs favorables au contact sont analysés pour la protéine FCGRT de *Rattus norvegicus* ; cette protéine est en effet liée à la B2M et chacune des 9 positions définies par ces descripteurs comporte le groupe d'acides aminés caractéristique de cette classe. De même, les 9 descripteurs défavorables au contact sont analysés pour la protéine RAE1B de *Mus musculus*, non liée à la B2M. Globalement, les descripteurs favorables à l'interaction et localisés dans la zone de contact potentiel semblent correspondre à une orientation de chaîne latérale ou à une propriété physico-chimique favorable au contact direct avec la B2M, tel qu'un résidu large et aromatique F, W ou Y en position [D1] 27 (W pour FCGRT de *Rattus norvegicus*). Les descripteurs favorables localisés hors de cette zone semblent maintenir une structure adéquate au contact ; les résidus [D1] 51 et [D2] 85 pourraient en effet maintenir la fermeture du sillon (par le rapprochement des deux hélices) à une extrémité. Au contraire, les descripteurs défavorables situés dans cette zone semblent empêcher le contact direct par gêne

stérique, tels que les résidus K en positions [D1] 12 et 25 de RAE1B de *Mus musculus*. La déstabilisation de la structure propice à l'interaction par des résidus tels que E, V, Q ou H en [D2] 39 serait à analyser en détail, par des approches telles que la dynamique moléculaire.

Le caractère explicatif du classifieur Bayesien naïf et notre définition des descripteurs nous permettent donc d'identifier les propriétés physico-chimiques dont l'observation à une position semble être favorable ou non au contact direct avec la B2M (pour les positions localisées dans la zone potentielle de contact), ou stabiliser ou non la structure nécessaire au contact (pour les positions localisées hors de cette zone).

Nous avons enfin utilisé notre classifieur comme outil de prédiction de l'interaction ou non à la B2M pour 8 protéines MHC-I des vertébrés inférieurs *Salmon trutta*, *Ambystoma mexicanum*, *Oncorhynchus kisutch* et *Oncorhynchus mykiss*. De nombreux travaux concernent en effet le séquençage et l'étude de l'origine évolutive des gènes MHC d'amphibiens et de téléostéens (Sammut et al. 1999, Hansen et al. 1999), mais peu de données expérimentales en rapport avec leur expression à la surface cellulaire et leur interaction avec la B2M sont actuellement disponibles. Nous avons donc dans un premier temps aligné les G-DOMAINs de ces protéines avec l'alignement multiple IMGT et numéroté leurs positions (selon la procédure décrite dans le chapitre précédent), et nous avons ensuite utilisé notre classifieur pour prédire cette interaction. Le résultat de la prédiction est identique pour ces 8 protéines MHC-I, et indique qu'elles se lient très probablement à la B2M. Ce résultat suggère par conséquent qu'elles devraient être exprimées à la surface cellulaire par un processus similaire à celui des protéines MHC-I de mammifères.

La méthodologie que nous avons mise en place afin de classer les protéines de la MhcSF selon leur interaction ou non avec la B2M aboutit à des résultats biologiquement cohérents, et démontre son intérêt pour le traitement automatique des génomes de vertébrés nouvellement séquencés ; les informations fournies par cette étude devraient de plus être précieuses pour de futures expériences de mutagenèse dirigée. La résolution de cette problématique de classification supervisée, caractérisée par la faible similarité des protéines de la MhcSF, a mis en évidence que l'information de séquence peut-être suffisante pour prédire efficacement la fonction et l'interaction des protéines. Notre approche devrait s'appliquer avec succès à d'autres problématiques de classification des protéines des IgSF et MhcSF pour lesquelles on dispose de classes connues a priori, et plus généralement à la classification fonctionnelle de superfamilles protéiques à partir d'un alignement multiple.

DISCUSSION ET CONCLUSION

IMGT est le système d'information international en ImMunoGénéTique®, spécialisé dans la gestion des séquences et des structures 3D des IG, TR et MHC des vertébrés ; ces protéines assurent la reconnaissance antigénique et la spécificité du système immunitaire adaptatif. L'ontologie de référence en immunogénétique (IMGT-ONTOLOGY) fournit les règles de description de leurs récepteurs, chaînes et domaines : V-DOMAINs et C-DOMAINs des IG et TR, G-DOMAINs des protéines du MHC. La numérotation unique des V-DOMAINs comme celle des C-DOMAINs se base sur un alignement multiple de séquences dont la similarité est élevée et assure la description standardisée de leurs caractéristiques fonctionnelles et structurales. Des domaines de structure similaire ont été identifiés au sein de protéines impliquées dans une grande variété de processus biologiques, localisés dans des compartiments cellulaires différents et correspondant à différents sites d'interaction protéine-ligand ; ces domaines sont nommés V-LIKE-DOMAINs, C-LIKE-DOMAINs et G-LIKE-DOMAINs, et définissent les superfamilles des immunoglobulines (IgSF) et du MHC (MhcSF).

La première partie du travail réalisé pendant cette thèse a porté sur la standardisation des protéines des IgSF et MhcSF à partir des règles d'IMGT-ONTOLOGY existantes, ou en établissant de nouvelles règles. Nous avons d'une part démontré que les numérotations des V-DOMAINs et des C-DOMAINs étaient applicables respectivement aux V-LIKE-DOMAINs et aux C-LIKE-DOMAINs, et d'autre part mis en place la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs. L'alignement de ces domaines a posé le problème de la « twilight zone » et a nécessité le développement d'une stratégie adaptée à cette hétérogénéité de séquences, basée sur la combinaison d'alignements multiples de séquences et de structures 3D.

Cette approche permet de standardiser les nouvelles protéines des IgSF et MhcSF, et fournit dorénavant les informations nécessaires à la gestion de ces superfamilles par les différents outils et bases de données constituant IMGT.

Dans la seconde partie de cette thèse, nous avons développé une méthode de classification de ces protéines selon leur fonction et leur interaction, à partir de leur séquence. Nous nous sommes particulièrement intéressés à la classification des protéines de la MhcSF selon leur liaison ou non à la beta2-microglobuline (B2M), afin de prédire cette interaction pour de nouvelles protéines et d'identifier les propriétés physico-chimiques qui, observées à une

position donnée des G-DOMAINs ou G-LIKE-DOMAINs, la favorisent ou la défavorisent. La connaissance a priori des classes et l'analyse évolutive des protéines de la MhcSF ont mis en évidence la nécessité d'une approche de classification supervisée ; le classifieur Bayesien naïf s'est avéré le plus adapté. Nous avons représenté chaque séquence alignée, décrite par la numérotation unique des G-DOMAINs et G-LIKE-DOMAINs, par un ensemble de descripteurs binaires associant une position dans l'alignement et un groupe d'acides aminés ; l'analyse des contacts entre les G-DOMAINs ou G-LIKE-DOMAINs des protéines de la MhcSF et le C-LIKE-DOMAIN de la B2M ont mis en évidence la cohérence de cette représentation et la faisabilité d'une telle approche de classification. Cette méthodologie est performante quelle que soit la similarité de la séquence à classer avec les séquences du jeu de données, et présente un intérêt particulier dans le cas d'un nouveau type de récepteur ; ces protéines sont en effet faiblement similaires aux autres protéines du jeu de données, et leur comportement concernant la liaison ou non à la B2M est distinct de celui de leurs plus proches voisins.

Les perspectives de cette étude seraient la mise en œuvre d'expériences de mutagenèse dirigée couplées à une approche de dynamique moléculaire, afin de valider individuellement chaque descripteur et d'évaluer leur impact respectif sur l'interaction ; l'identification de corrélations entre ces descripteurs permettrait d'affiner le modèle. Cette approche permettrait par la suite la prédiction de l'interaction ou non à la B2M à grande échelle, pour toutes les protéines MHC-I de différents génomes de vertébrés inférieurs.

Notre méthode devrait s'appliquer avec succès à toute problématique de classification des protéines des IgSF et MhcSF pour lesquelles on dispose de classes fonctionnelles ou d'interaction connues a priori. La classification des protéines MHC-I selon leur fonction au sein du système immunitaire (classiques *vs.* non classiques) présente notamment un enjeu biologique majeur. Comme nous l'avons indiqué dans le premier chapitre, les protéines MHC-I classiques (MHC-Ia) présentent des peptides endogènes aux TR et permettent ainsi la reconnaissance des cellules saines et la destruction des cellules infectées. Les protéines MHC-I non classiques (MHC-Ib) sont assemblées avec des peptides issus de la dégradation protéolytique des protéines MHC-Ia, et témoignent du bon fonctionnement de la voie d'expression des protéines MHC-I ; leur activité permet de protéger les cellules saines de la lyse par les cellules NK. La prédiction successive de la liaison ou non à la B2M et de la fonction des protéines MHC-I (par la combinaison de deux classifieurs) nous permettrait d'annoter automatiquement l'ensemble des protéines MHC-I de vertébrés inférieurs connues à ce jour ; ces annotations constitueraient alors des informations précieuses pour appréhender

l’évolution de ce système performant de protection contre les agents pathogènes depuis l’apparition des vertébrés.

Enfin, cette thèse a démontré que l’information de séquence peut-être suffisante pour prédire efficacement la fonction et l’interaction des protéines, et a ainsi ouvert la voie pour la résolution de nombreuses problématiques de classification fonctionnelle de superfamilles protéiques à partir d’un alignement multiple.

BIBLIOGRAPHIE

Allison, T.J. and Garboczi, D.N., Structure of γδ T cell receptors and their recognition of non-peptide antigens, *Mol. Immunol.*, 2001, 38, 1051-1061.

Alt, F.W. and Baltimore, D., Joining of immunoglobulin heavy chain gene segments: implications for a chromosome with evidence of three D-JH fusions, *Proc. Natl. Acad. Sci. USA*, 1982, 79, 4118-4122.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 1990, 215, 403.

Bandyopadhyay, R., Tan, X.X., Matthews, K.S. and Subramanian, D, Predicting protein-ligand interactions from primary structure, Technical Report Rice University, 2002, TR02-398.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L., GenBank, *Nucl. Acids Res.*, 2005, 33, D34-38.

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

Bernabeu, C., van de Rijn, M., Lerch, P and Terhost, C., β2microglobulin from serum associates with class I antigens on the surface of cultured cells, *Nature*, 1984, 308, 642-645.

Bertrand, G., Duprat, E., Lefranc, M.-P., Marti, J. and Coste, J., Human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles, *Tissue Antigens*, 2004, 64, 119-131.

Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A. and Eppig, J.T., MGD: The Mouse Genome Database, *Nucl. Acids Res.*, 2003, 31, 193-195.

Bork, P., Holm, L. and Sander, C., The immunoglobulin fold. Structural classification, sequence patterns and common core, *J. Mol. Biol.*, 1994, 242, 309-320.

Bork, P. and Koonin, E.V., Protein sequence motifs, *Curr. Opin. Struct. Biol.*, 1996, 6, 366-376.

Boyd, L.F., Kozlowski, S. and Margulies, D.H., Solution binding of an antigenic peptide to a major histocompatibility complex class I molecule and the role of B2-microglobulin, *Proc. Natl. Acad. Sci.*, 1992, 89, 2242-2246.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., Classification and Regression Trees, Wadsworth International Group, Belmont, 1984.

Burton, D.R., Structure and function of antibodies, In: *Molecular genetics of immunoglobulin*, F. Calabi and M.S. Neuberger eds., Oxford, Elsevier, 1987, pp. 1-50.

Chothia, C., Gelfand, I. and Kister, A., Structural determinants in the sequences of immunoglobulin variable domain, *J. Mol. Biol.*, 1998, 278, 457-479.

Collins, E.J., Garboczi, D.N., Karpusas, M.N. and Wiley, D.C., The three-dimensional structure of a class I major histocompatibility complex molecule missing the α3 domain of the heavy chain, *Proc. Natl. Acad. Sci. USA*, 1995, 92, 1218-1221.

Domingos, P. and Pazzani, M., Beyond independence: conditions for the optimality of the simple Bayesian classifier, *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996, Bari, 105-112.

Doolittle, R.F., Of URFs and ORFs: a primer on how to analyze derived amino acid sequences, University Science Books, 1986, Mill Valley California.

Du Pasquier, L., The immune system of invertebrates and vertebrates, *Comp. Biochem. Physiol. B. Biochem. Mol. Biol.*, 2001, 129, 1-15.

Duprat, E., Kaas, Q., Garelle, V., Giudicelli, 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, In: *Recent Research Developments in Human Genetics*, Trivandium, India, Research Signpost, 2004, 2, 111-136.

Duprat, E., Lefranc, M.-P. and Gascuel, O., Prédire l'interaction des protéines de la superfamille du MHC avec la beta2-microglobuline en combinant classifieur Bayesien « naïf » et alignement multiple IMGT, *Actes des Journées Ouvertes Biologie Informatique Mathématiques*, 2005a.

Duprat, E., Lefranc, M.-P. and Gascuel, O., A simple method to predict protein binding from aligned sequences – application to MHC superfamily and beta2-microglobulin, *Bioinformatics*, 2005b (in press).

Edelman, G.M., Cunningham, B.A., Gall, W.E., Gottlieb, P.D., Rutishauser, U. and Waxdal, M.J., The covalent structure of an entire gammaG immunoglobulin molecule, *Proc. Natl. Acad. Sci. USA*, 1969, 63, 78-85.

Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucl. Acids Res.*, 2004, 32, 1792-1797.

Feder, J.N., Penny, D.M., Irrinki, A., Lee, V.K., Lebron, J.A., Watson, N., Tsuchihashi, Z., Sigal, E., Bjorkman, P.J. and Schatzman, R.C., The hemochromatosis gene product complexes with the transferrin receptor and lowers its affinity for ligand binding, *Proc. Natl. Acad. Sci. USA*, 1998, 95, 1472-1477.

Fisher, R.A., The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 1936, 7, 179-188.

Fix, E. and Hodges, J.L., Discriminatory analysis, non-parametric discrimination, *Technical report* 21-49-004, USAF School of Aviation Medicine, Randolf Field, 1951.

Frigoul, A. and Lefranc, M.-P., MICA: standardized IMGT allele nomenclature, polymorphisms and diseases, In: *Recent Research Developments In Human Genetics*, Trivandium, India, Research Signpost, 2005 (in press).

Gascuel, O., BIONJ, an improved version of the NJ algorithm based on a simple model of sequence data, *Mol. Biol. Evol.*, 1997, 14, 685-695.

Gascuel, O., Bouchon-Meunier, B., Caraux, G., Gallinari, P., Guénoche A., Guermeur, Y., Lechevallier, Y., Marsala, C., Miclet, L., Nicolas, J., Nock, R., Ramdani, M., Sebag, M., Tallur, B., Venturini, G. and Vitte, P., Twelve numerical, symbolic and hybrid supervised classification methods, *International Journal of Pattern Recognition and Artificial Intelligence*, 1998, 12, 517-571.

Gearhart, P.J., Johnson, N.D., Douglas, R. and Hood, L., IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts, *Nature*, 1981, 291, 29-34.

Giudicelli, V. and Lefranc, M.-P., Ontology for Immunogenetics: the IMGT-ONTOLOGY, *Bioinformatics*, 1999, 12, 1047-1054.

Good, I.J., The estimation of probabilities: an essay on modern Bayesian methods. *Research Monograph 30*, 1965, MIT Press, Cambridge, MA.

Gorer, P.A., The detection of antigenic differences in mouse erythrocytes by the employment of immune sera, *Br. J. Exp. Pathol.*, 1936, 17, 42-50.

Guindon, S. and Gascuel, O., A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst. Biol.*, 2003, 52, 696-704.

Halaby, D.M. and Mornon, J.P., The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity, *J. Mol. Evol.*, 1998, 46, 389-400.

Hansen, J.D., Strassburger, P., Thorgaard, G.H., Young, W.P. and Du Pasquier, L., Expression, linkage, and polymorphism of MHC-related genes in *Rainbow trout, Oncorhynchus mykiss*, *J. Immunol.*, 1999, 163, 774-786.

Holmes, M.A., Li, P., Petersdorf, E.W. and Strong, R.K., Structural studies of allelic diversity of the MHC class I homolog MIC-B, a stress-inducible ligand for the activating immunoreceptor NKG2D, *J. Immunol.*, 2002, 169, 1395-1400.

Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Wright, M.W., Wain, H.M., Trowsdale, J., Ziegler, A. and Beck, S., Gene map of the extended human MHC, *Nature Rev. Genet.*, 2004, 5, 889-899.

Huang, Y., Frishman, D. and Muchnik, I., Predicting protein-protein interactions by a supervised learning classifier, *Comput. Biol. Chem.*, 2004, 28, 291-301.

Hughes, A.L. and Nei, M., Evolutionary relationships of the classes of major histocompatibility complex genes, *Immunogenetics*, 1993, 37, 337-346.

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, *Nucl. Acids Res.*, 2004, 32, D208-D210.

Kaas, Q., Duprat, E., Tourneur, G. and Lefranc, M.-P., IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes, In: *Immunoinformatics*, V. Brusic and C. Schoenbach eds., Springer, The Netherlands, 2005 (in press).

Kanz, C., Aldebert, P., Althorpe, N., Baker., W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R., The EMBL Nucleotide Sequence Database, *Nucl. Acids Res.*, 2005, 33, D29-33.

Kyte, J. and Doolittle, R.F., A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 1982, 157, 105-132.

Landau, N.R., St John, T.P., Weissman, I.L., Wolf, S.C., Silverstone, A.E. and Baltimore, D., Cloning of terminal transferase cDNA by antibody screening, *Proc. Natl. Acad. Sci. USA*, 1984, 81, 5836-5840.
Lefranc, M.-P. and Lefranc, G., Molecular genetics of immunoglobulin allotype expression, In: *The human IgG subclasses*, F. Shakib eds., Oxford, Pergamon Press, 1990, pp. 43-78.

Lefranc, M.-P., The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains, *The Immunologist*, 1999, 7, 132-136.

Lefranc, M.-P. and Lefranc, G., *The Immunoglobulin FactsBook*, Academic Press, London, UK, 2001a.

Lefranc, M.-P. and Lefranc, G., *The T cell receptor FactsBook*, Academic Press, London, UK, 2001b.

Lefranc, M.-P., Pommié, 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.*, 2003, 27, 55-77.

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 (http://imgt.cines.fr), *In Silico Biol.*, 2004, 4, 17-29.

Lefranc, M.P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clement, O., Chaume, D. and Lefranc, G., IMGT, the international ImMunoGeneTics information system, *Nucl. Acids Res*., 2005a, 33, D593-597.

Lefranc, M.-P., Pommié, 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.*, 2005b, 29, 185-203.

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.*, 2005c, 29, 917-938.

Lesk, A.M. and Chothia, C., Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains, *J. Mol. Biol.*, 1982, 160, 325-342.

Li, P., Willie, S.T., Bauer, S., Morris, D.L., Spies, T. and Strong, R.K., Crystal structure of the MHC class I homolog MIC-A, a gammadelta T cell ligand, *Immunity*, 1999, 10, 577-584.

Li, P., McDermott, G. and Strong, R.K., Crystal structures of RAE-1β and its complex with the activating immunoreceptor NKG2D, *Immunity*, 2002, 16, 77-86.

Lieber, M., Antibody diversity: a link between switching and hypermutation, *Curr. Biol.*, 2000, 10, R798-800.

Lichtarge, O., Bourne, H.R. and Cohen, F.E., An evolutionary trace method defines binding surfaces common to protein families, *J. Mol. Biol.*, 1996, 257, 342-358.

Lichtarge, O. and Sowa, M.E., Evolutionary predictions of binding surfaces and interactions, *Curr. Opin. Struct. Biol.*, 2002, 12, 21-27.

Maenaka, K. and Jones, E.Y., MHC superfamily structure and the immune system, *Curr. Opin. Struct. Biol.*, 1999, 9, 745-753.

Madden, D.R., The three-dimensional structure of peptide-MHC complexes, *Annu. Rev. Immunol.*, 1995, 13, 587-622.

Marsh, S.G.E. and Robinson, J., The IMGT/HLA Sequence Database, *Rev. Immunogenet.*, 2001, 2, 518-531.

Matsumura, M., Fremont, D.H., Peterson, P.A. and Wilson, I.A., Emerging principles for the recognition of peptide antigens by MHC class I molecules, *Science*, 1992, 257, 927-934.

McLaughlin, W.A. and Berman, H.M., Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif, *J. Mol. Biol.*, 2003, 330, 33-55.

Michaëlsson, J., Achour, A., Rolle, A. and Karre, K., MHC class I recognition by NK receptors in the Ly49 family is strongly influenced by the beta 2-microglobulin subunit, *J. Immunol.*, 2001, 166, 7327-7334.

Mihalek, I., Res, I. and Lichtarge, O., A family of evolution-entropy hybrid methods for ranking protein residues by importance, *J. Mol. Biol.*, 2004, 336, 1265-1282.

Miley, M.J., Truscott, S.M., Yu, Y.Y.L., Gilfillan, S., Fremont, D.H., Hansen, T.H. and Lybarger, L., Biochemical features of the MHC-related protein 1 consistent with an immunological function, *J. Immunol.*, 2003, 170, 6090-6098.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 1995, 247, 536-540.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M., CATH--a hierarchic classification of protein domain structures, *Structure*, 1997, 5, 1093-1108.

Padlan, E.A., Anatomy of the antibody molecule, *Mol. Immunol.*, 1994, 31, 169-217.

Pearson, W.R. and Lipman, D.J., Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA*, 1988, 85, 2444-2448.

Perrière, G. and Gouy, M. WWW-Query: An on-line retrieval system for biological sequence banks, *Biochimie*, 1996, 78, 364-369.

Pommié, 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. Recogn.*, 2004, 17, 17-32.

Porcelli, S., Brenner, M.B. and Band, H., Biology of the human gamma delta T-cell receptor, *Immunol. Rev.*, 1991, 120, 137-183.

Porter, R.R., The hydrolysis of rabbit y-globulin and antibodies with crystalline papain, *Biochem J.*, 1959, 73, 119-126.

Radaev, S., Rostro, B., Brooks, A.G., Colonna, M. and Sun, P.D., Conformational plasticity revealed by the cocrystal structure of NKG2D and its class I MHC-like ligand ULBP3, *Immunity*, 2001, 15, 1039-1049.

Rammensee, H.G., Friede, T. and Stevanoviic, S., MHC ligands and peptide motifs: first listing, *Immunogenetics*, 1995, 41, 178-228.

Rose, F., Berissi, H., Weissenback, J., Manoteaux, L., Fellous, M. and Revel, M., The β2-microglobulin mRNA in human Daudi cells has a mutated initiation codon but is still inducuble by interferon, *EMBO J.*, 1983, 2, 239-243.

Rost, B., Twilight zone of protein sequence alignments, *Protein Eng.*, 1999, 12, 85-94.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. and Mewes, H.W., The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucl. Acids Res.*, 2004, 32, 5539-5545.

Ruiz, M. and Lefranc, M.-P., IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures, *Immunogenetics*, 2002, 53, 857-883.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., Learning representations by back-propagating errors, *Nature*, 1986, 323, 533-536.

Saitou, N. and Nei, M., The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 1987, 4, 406-425.

Sali, A. and Blundell, T.L., Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming, *J. Mol. Biol.*, 1990, 212, 403-428.

Sammut, B., Du Pasquier, L., Ducoroy, P., Laurens, V., Marcuz, A. and Tournefier, A., Axolotl MHC architecture and polymorphism, *Eur. J. Immunol.*, 1999, 29, 2897-2907.

Sanchez, L.M., Chirino, A.J. and Bjorkman, P., Crystal structure of human ZAG, a fat-depleting factor related to MHC molecules, *Science*, 1999, 283, 1914-1919.

Saper, M.A., Bjorkman, P.J. and Wiley, D.C., Refined structure of the human histocompatibility antigen HLA-A2 at 2.6Å resolution, *J. Mol. Biol.*, 1991, 219, 277-319.

Shannon, C.E., A mathematical theory of communication, *Bell Syst. Techn. J.*, 1948, 27, 379-423.

Simmonds, R.E. and Lane, D.A., Structural and functional implications of the intron/exon organization of the human endothelial cell protein C/activated protein C receptor (EPCR) gene: comparison with the structure of CD1/major histocompatibility complex alpha1 and alpha2 domains, *Blood*, 1999, 94, 632-641.

Sneath, P.H.A. and Snokal, R.R., *Numerical Taxonomy*, W. H. Freeman and Company, San Francisco, 1973.

Solheim, J.C., Cook, J.R. and Hansen, T.H., Conformational changes induced in the MHC class I molecule by peptide and beta 2-microglobulin, *Immunologic Research*, 1995, 14, 200-217.

Sonnhammer, E.L.L., Eddy, S.R. and Durbin, R., Pfam: a comprehensive database of protein domain families based on seed alignments, *Proteins*, 1997, 28, 405-420.

Strong, R.K., Class (I) will come to order—not, *Nature Struct. Biol.*, 2000, 7, 173-176.

Tateno, Y., Saitou, N., Okubo, K., Sugawara, H. and Gojobori, T., DDBJ in collaboration with mass-sequencing teams on annotation, *Nucl. Acids Res.*, 2005, 33, D25-28.

The MHC sequencing consortium, Complete sequence and gene map of a human major histocompatibility complex, *Nature*, 1999, 401, 921-923.

Tonegawa, S., Somatic generation of antibody diversity, *Nature*, 1983, 302, 575-581.

Vapnik, V.N. and Chervonenkis, A.Y., On the uniform convergence of relative frequencies of events to their probabilities, *Theory of probability and its applications*, 1971, 16, 264-280.

Vapnik, V.N., The Nature of Statistical Learning Theory, 1995, Springer, New York.

Wang, Z., Cao, Y., Albino, A.P., Zeff, R.A., Houghton, A. and Ferrone, S., Lack of HLA class I antigen expression by melanoma cells SK-MEL-33 caused by a reading frameshift in Beta2-microglobulin messenger RNA, *J. Clin. Invest.*, 1993, 91, 684-692.

Webb, E.C., *Enzyme Nomenclature 1992*. Recommendations of the Nomenclature Commitee of the International Union of Biochemistry and Molecular Biology, Academic Press, New York, 1992.

West, A.P. and Bjorkman, P.J., Crystal structure and immunoglobulin G binding properties of the human major histocompatibility complex-related Fc receptor, *Biochemistry*, 2000, 39, 9698-9708.

Wicker, N., Perrin, G.R., Thierry, J.C. and Poch, O., Secator: a program for inferring protein subfamilies from phylogenetic trees, *Mol. Biol. Evol.*, 2001, 18, 1435-1441.

Williams, A.F. and Barclay, A.N., The immunoglobulin superfamily—domains for cell surface recognition, *Annual Review of Immunology*, 1988, 6, 381-405.

Williams, D.B., Barber, B.H., Flavell, R.A. and Allen, H., Role of beta2-microglobulin in the intracellular transport and surface expression of murine class I histocompatibility molecules, *The Journal of Immunology*, 1989, 142, 2796-2806.

Wu, C., Whitson, G., McLarty, J., Ermongkonchai, A., Chang, T.C., Protein classification artificial neural system, *Protein Sci.*, 1992, 1, 667-677.

Wu, T.D. and Brutlag, D.L., Identification of protein motifs using conserved amino acids properties and partitioning techniques, *Proc. Thirteenth Int. Conf. Intell. Syst. Mol. Biol.*, 1995, 19, 402-410.

Zeng, Z., Castano, A.R., Segelke, B.W., Stura, E.A., Peterson, P.A. and Wilson, I.A., Crystal structure of mouse CD1: An MHC-like fold with a large hydrophobic binding groove, *Science*, 1997, 277, 339-345.

Zuckerland, E. and Pauling, L., Molecular disease, evolution, and genetic heterogeneity. In: *Horizons in Biochemistry*, Kasha and Pullman eds, 1962, pp. 189–225, Academic Press.

Annexes

Annexe 1. Les concepts d'IMGT-ONTOLOGY (extrait de Lefranc et al. 2005d).

IMGT-ONTOLOGY main concepts	IMGT Scientific chart rules (Lefranc et al. 1999)	Examples of IMGT expertised data concepts (Ruiz et al. 2000, Lefranc 2001)
IDENTIFICATION	Standardized keywords (Giudicelli et al. 1997)	Species, molecule type, receptor type, chain type, gene type, structure, functionality, specificity (Giudicelli et al. 1997, Lefranc et al. 1998)
DESCRIPTION	Standardized labels and annotations (Giudicelli et al. 1997)	Core (V-, D-, J-, C-REGION) Prototypes (Giudicelli et al. 1997) Labels for sequences Labels for 2D and 3D structures
CLASSIFICATION	Reference sequences Standardized IG and TR gene nomenclature (group, subgroup, gene, allele)	Nomenclature of the human IG and TR genes (entry in 1999 in GDB, HGNC (Wain et al. 2002) and LocusLink at NCBI) (Lefranc and Lefranc 2001a, 2001b) Alignment of alleles (Lefranc et al. 1998, Lefranc and Lefranc 2001a, 2001b) Nomenclature of the IG and TR genes of all vertebrate species (Lefranc et al. 1999)
NUMEROTATION	IMGT unique numbering (Lefranc et al. 1998, Lefranc 1997, Lefranc 1999) for: V- and V-LIKE-DOMAINs (Lefranc et al. 2003) C- and C-LIKE-DOMAINs (Lefranc et al. 2005b) G- and G-LIKE-DOMAINs (Lefranc et al. 2005c)	Protein displays (Lefranc et al. 1999) IMGT Colliers de Perles (Lefranc et al. 1998, Lefranc et al. 1999, Ruiz and Lefranc 2002) FR-IMGT and CDR-IMGT delimitations (Lefranc et al. 1999) Structural loops and beta strands delimitations (Lefranc et al. 2005b)
ORIENTATION	Orientation of genomic instances relative to each other	Chromosome orientation Locus orientation Gene orientation DNA strand orientation
OBTENTION	Standardized origin Standardized methodology (Lefranc et al. 2004)	

Giudicelli, V. et al., IMGT, the international ImMunoGeneTics database, *Nucl. Acids Res.*, 1997, 25, 206-211.

Lefranc, M.-P., Unique database numbering system for immunogenetic analysis, *Immunol. Today*, 1997, 18, 509.

Lefranc, M.-P. et al., IMGT, the international ImMunoGeneTics database, *Nucl. Acids Res.*, 1998, 26, 297-303.

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

Lefranc, M.-P., IMGT, the international ImMunoGeneTics database, *Nucl. Acids Res.*, 2001, 29, 207-209.

Lefranc, M.-P. et al., IMGT-Choreography for immunogenetics and immunoinformatics, *In Silico Biol.*, 2005d, 5, 45-60.

Ruiz, M. et al., IMGT, the international ImmunoGeneTics database, *Nucl. Acids Res.*, 2000, 28, 219-221.

Annexe 2. DESCRIPTION des récepteurs, chaînes et domaines des fragments protéolytiques d'IG.

Récepteur	Chaîne	Domaine
FAB-ALPHA_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-ALPHA-1_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-ALPHA-2_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-DELTA_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-EPSILON_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-GAMMA-1_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-GAMMA-2_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-GAMMA-2-A_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-GAMMA-2-B_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-GAMMA-2-C_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-GAMMA-3_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-GAMMA-4_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA
FAB-MU_LAMBDA	VH-CH1	VH, CH1
	L-LAMBDA	V-LAMBDA, C-LAMBDA

Table 1. Fragments Fab des IG comportant une chaîne L-LAMBDA.

Récepteur	Chaîne	Domaine
FAB-ALPHA_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-ALPHA-1_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-ALPHA-2_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-DELTA_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-EPSILON_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-GAMMA-1_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-GAMMA-2_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-GAMMA-2-A_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-GAMMA-2-B_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-GAMMA-2-C_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-GAMMA-3_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-GAMMA-4_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA
FAB-MU_KAPPA	VH-CH1	VH, CH1
	L-KAPPA	V-KAPPA, C-KAPPA

Table 2. Fragments Fab des IG comportant une chaîne L-KAPPA.

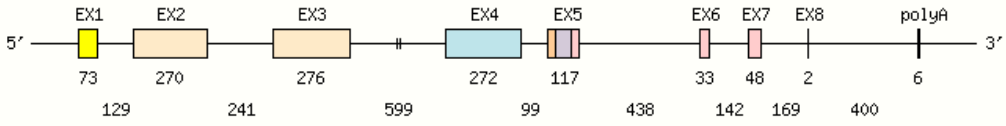
Récepteur	Chaîne	Domaine
FC-ALPHA	CH2-CH3	CH2, CH3
FC-ALPHA-1	CH2-CH3	CH2, CH3
FC-ALPHA-2	CH2-CH3	CH2, CH3
FC-DELTA	CH2-CH3	CH2, CH3
FC-EPSILON	CH2-CH3-CH4	CH2, CH3, CH4
FC-GAMMA-1	CH2-CH3	CH2, CH3
FC-GAMMA-2	CH2-CH3	CH2, CH3
FC-GAMMA-2-A	CH2-CH3	CH2, CH3
FC-GAMMA-2-B	CH2-CH3	CH2, CH3
FC-GAMMA-2-C	CH2-CH3	CH2, CH3
FC-GAMMA-3	CH2-CH3	CH2, CH3
FC-GAMMA-4	CH2-CH3	CH2, CH3
FC-MU	CH2-CH3-CH4	CH2, CH3, CH4

Table 3. Fragments Fc des IG.

Annexe 3. Organisation exon/intron des gènes MHC-I, MHC-II et MHC-I-like.

MHC-I

HLA-A (chaîne I-ALPHA, sous-classe HLA-A, MHC-Ia, *Homo sapiens*)

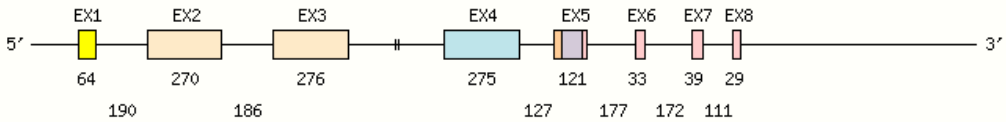


HLA-B (chaîne I-ALPHA, sous-classe HLA-B, MHC-Ia, *Homo sapiens*)

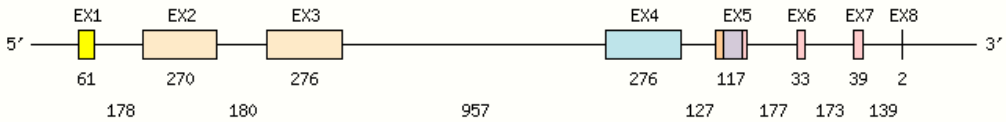
HLA-Cw (chaîne I-ALPHA, sous-classe HLA-C, MHC-Ia, *Homo sapiens*)

H2-D1 (chaîne I-ALPHA, sous-classe H2-D, MHC-Ia, *Mus musculus*)

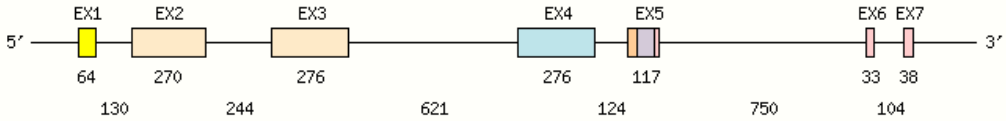
H2-K1 (chaîne I-ALPHA, sous-classe H2-K, MHC-Ia, *Mus musculus*)



H2-L (chaîne I-ALPHA, sous-classe H2-L, MHC-Ia, *Mus musculus*)



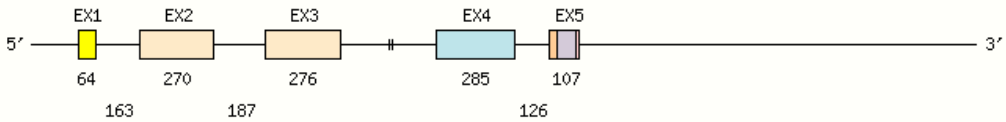
HLA-E (chaîne I-ALPHA, sous-classe HLA-E, MHC-Ib, *Homo sapiens*)



HLA-F (chaîne I-ALPHA, sous-classe HLA-F, MHC-Ib, *Homo sapiens*)

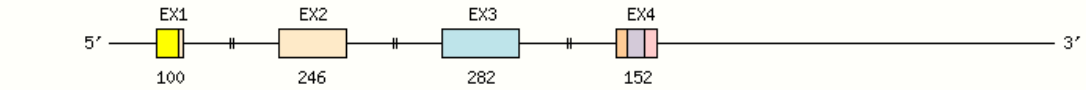
HLA-G (chaîne I-ALPHA, sous-classe HLA-G, MHC-Ib, *Homo sapiens*)

H2-Q7 (chaîne I-ALPHA, sous-classe H2-Q, MHC-Ib, *Mus musculus*)

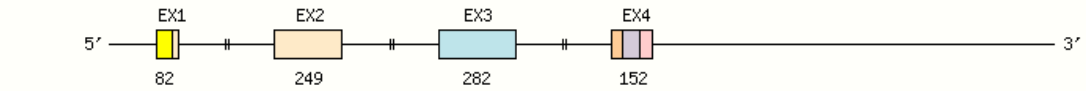


MHC-II

HLA-DPA1 (chaîne II-ALPHA, sous-classe HLA-DP, MHC-IIa, *Homo sapiens*)

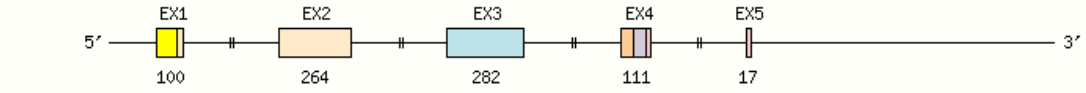


HLA-DQA1 (chaîne II-ALPHA, sous-classe HLA-DQ, MHC-IIa, *Homo sapiens*)

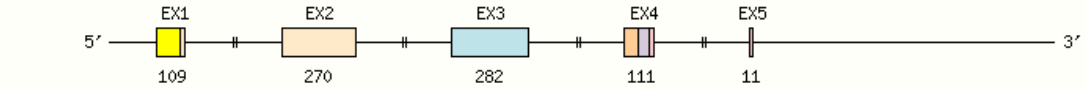


HLA-DRA (chaîne II-ALPHA, sous-classe HLA-DR, MHC-IIa, *Homo sapiens*)

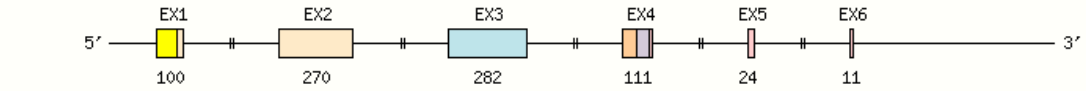
HLA-DPB1 (chaîne II-BETA, sous-classe HLA-DP, MHC-IIa, *Homo sapiens*)



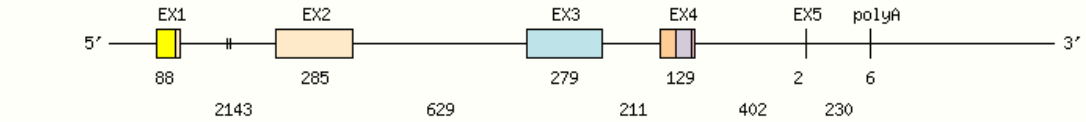
HLA-DQB1 (chaîne II-BETA, sous-classe HLA-DQ, MHC-IIa, *Homo sapiens*)



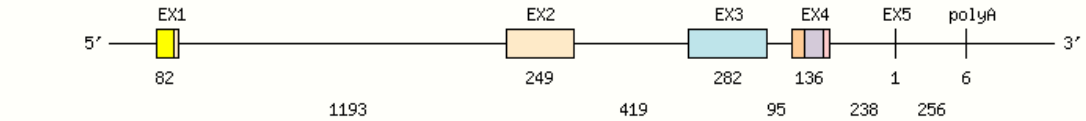
HLA-DRB1 (chaîne II-BETA, sous-classe HLA-DR, MHC-IIa, *Homo sapiens*)



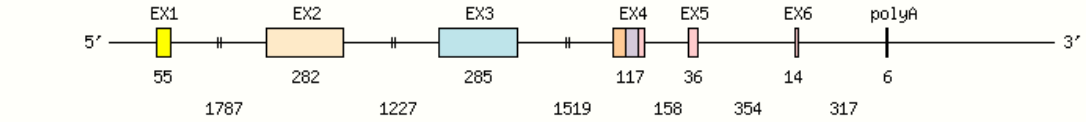
HLA-DMA (chaîne II-ALPHA, sous-classe HLA-DM, MHC-IIb, *Homo sapiens*)



HLA-DOA (chaîne II-ALPHA, sous-classe HLA-DO, MHC-IIb, *Homo sapiens*)



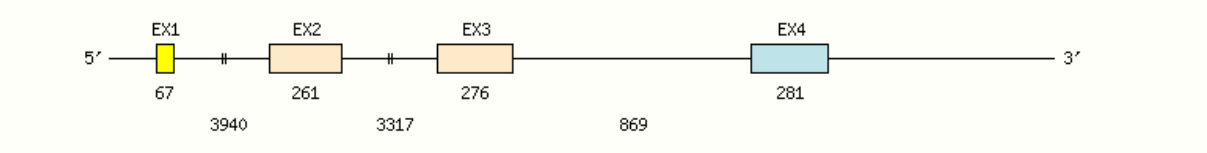
HLA-DMB (chaîne II-BETA, sous-classe HLA-DM, MHC-IIb, *Homo sapiens*)



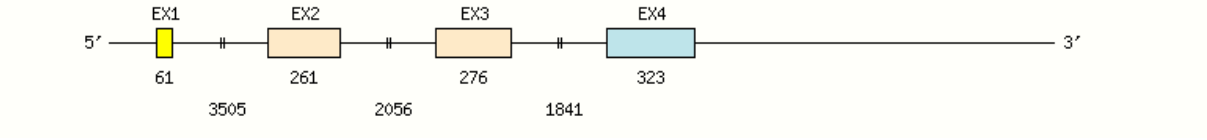
HLA-DOB (chaîne II-BETA, sous-classe HLA-DO, MHC-IIb, *Homo sapiens*)

MHC-I-like

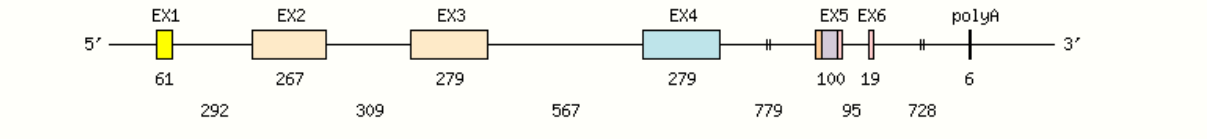
Homo sapiens AZGP1 (chaîne I-ALPHA-LIKE, type de récepteur AZGP1, *Homo sapiens*)



Mus musculus AZGP1 (chaîne I-ALPHA-LIKE, type de récepteur AZGP1, *Mus musculus*)

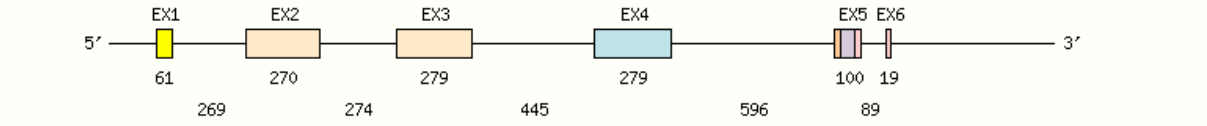


Homo sapiens CD1D (chaîne I-ALPHA-LIKE, type de récepteur CD1, *Homo sapiens*)

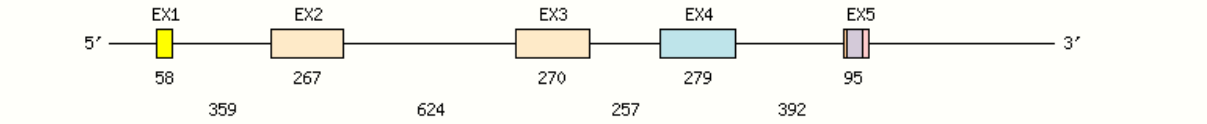


Mus musculus CD1D1 (chaîne I-ALPHA-LIKE, type de récepteur CD1, *Mus musculus*)

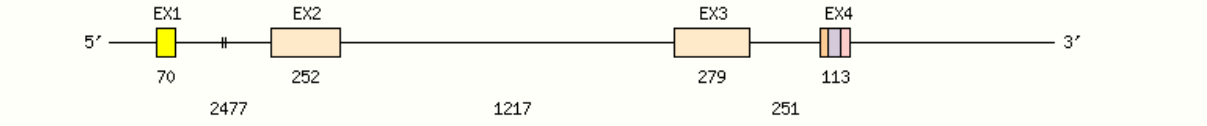
Rattus norvegicus CD1D1 (chaîne I-ALPHA-LIKE, type de récepteur CD1, *Rattus norvegicus*)



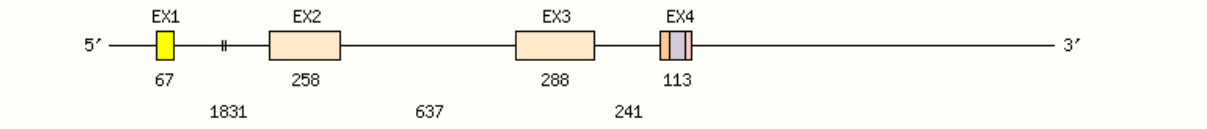
Homo sapiens CD1E (chaîne I-ALPHA-LIKE, type de récepteur CD1, *Homo sapiens*)



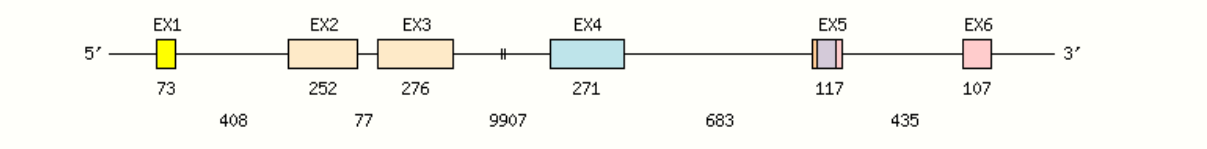
Homo sapiens EPCR (chaîne I-ALPHA-LIKE, type de récepteur EPCR, *Homo sapiens*)



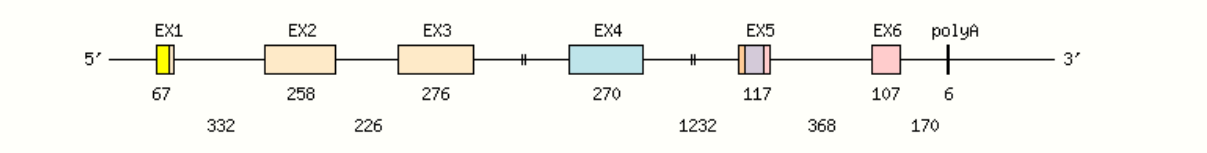
Mus musculus EPCR (chaîne I-ALPHA-LIKE, type de récepteur EPCR, *Mus musculus*)



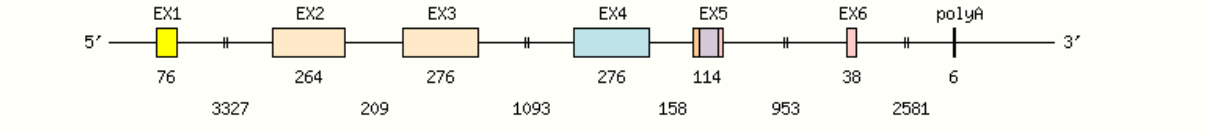
Homo sapiens FCGRT (chaîne I-ALPHA-LIKE, type de récepteur FCGRT, *Homo sapiens*)



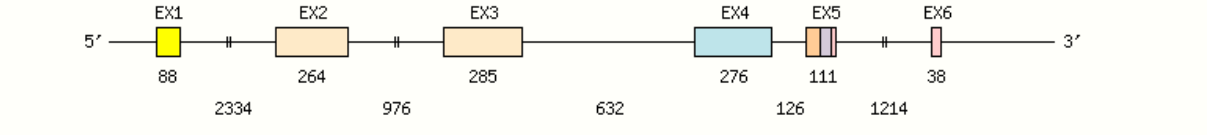
Mus musculus FCGRT (chaîne I-ALPHA-LIKE, type de récepteur FCGRT, *Mus musculus*)



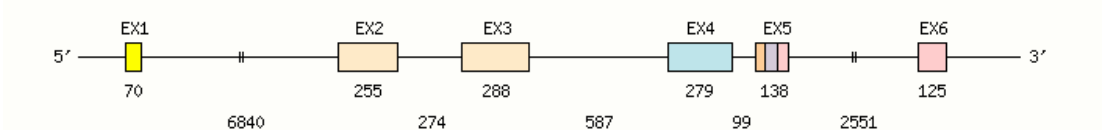
Homo sapiens HFE (chaîne I-ALPHA-LIKE, type de récepteur HFE, *Homo sapiens*)



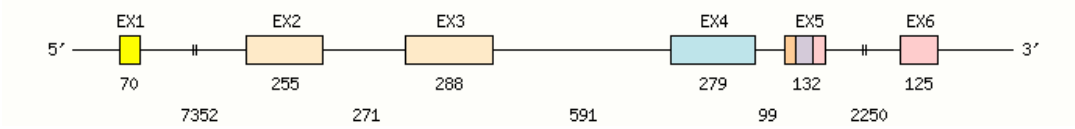
Mus musculus HFE (chaîne I-ALPHA-LIKE, type de récepteur HFE, *Mus musculus*)



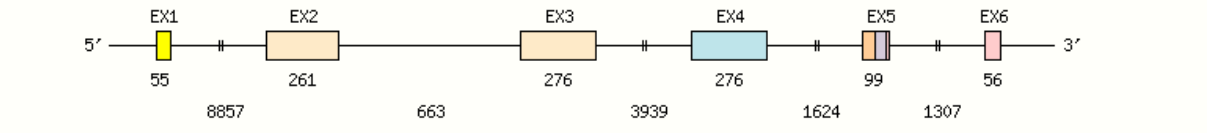
Homo sapiens MICA (chaîne I-ALPHA-LIKE, type de récepteur MIC, *Homo sapiens*)



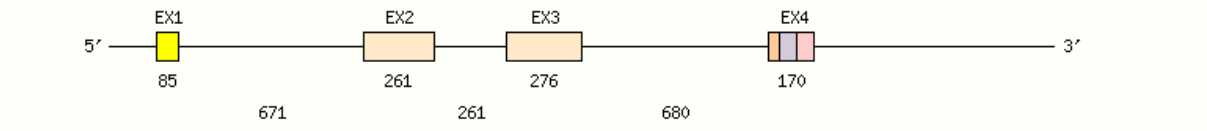
Homo sapiens MICB (chaîne I-ALPHA-LIKE, type de récepteur MIC, *Homo sapiens*)



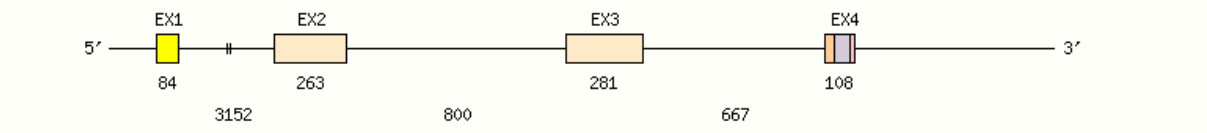
Mus musculus MR1 (chaîne I-ALPHA-LIKE, type de récepteur MR1, *Mus musculus*)



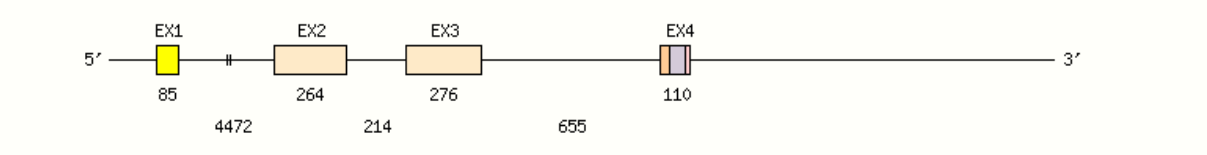
Homo sapiens RAET1E (chaîne I-ALPHA-LIKE, type de récepteur RAE, *Homo sapiens*)



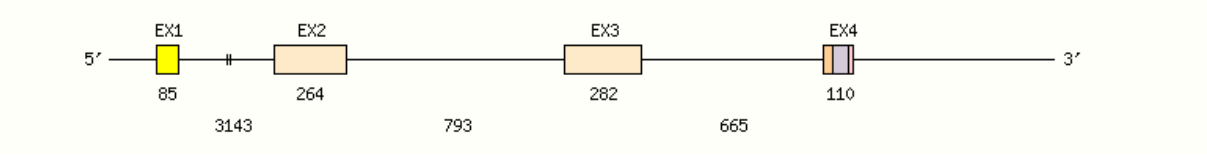
Homo sapiens RAET1H (chaîne I-ALPHA-LIKE, type de récepteur RAE, *Homo sapiens*)



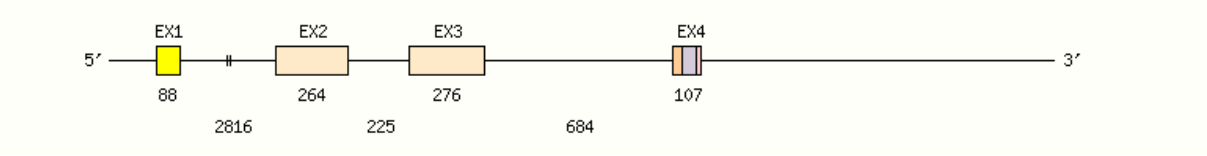
Homo sapiens RAET1I (chaîne I-ALPHA-LIKE, type de récepteur RAE, *Homo sapiens*)

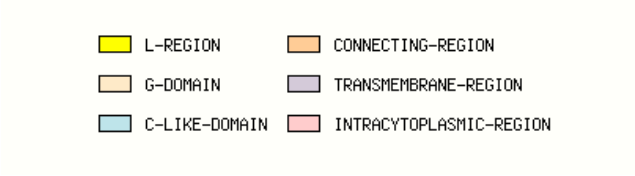


Homo sapiens RAET1L (chaîne I-ALPHA-LIKE, type de récepteur RAE, *Homo sapiens*)

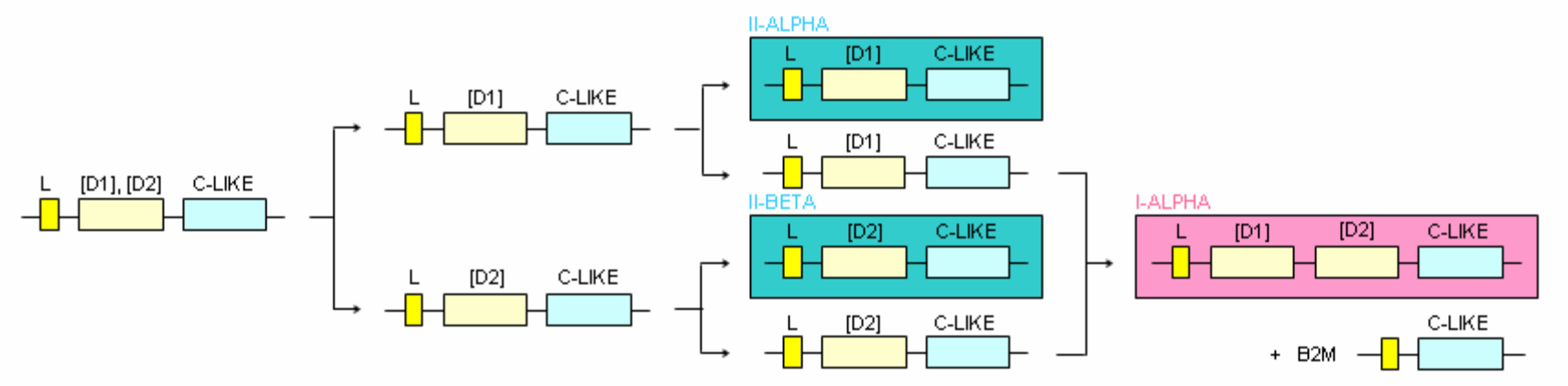


Homo sapiens RAET1N (chaîne I-ALPHA-LIKE, type de récepteur RAE, *Homo sapiens*)





Annexe 4. Modèle de l'origine évolutive des gènes MHC-I et MHC-II (d'après Hughes and Nei 1993).



Ce modèle se base sur l'hypothèse (supportée par l'analyse phylogénétique des séquences génomiques de MHC-I et MHC-II) de la précédence évolutive des gènes MHC-II par rapport aux gènes MHC-I ; les gènes du MHC-I et de la beta2-microglobuline (B2M) seraient issus de la recombinaison de deux gènes MHC-II. La représentation de chaque gène est ici simplifiée, et comporte uniquement les exons présents en 5'. L : exon leader. [D1] et [D2] indiquent les domaines G-ALPHA1 et G-ALPHA2 de la chaîne protéique I-ALPHA des protéines MHC-I ; les domaines G-ALPHA de la chaîne II-ALPHA et G-BETA de la chaîne II-BETA des protéines MHC-II sont supposés être leurs ancêtres respectifs, et sont notés ici [D1] et [D2] pour indiquer cette origine évolutive.

Annexe 5. Données de séquence et de structure 3D des protéines MHC-I, MHC-II et MHC-I-like.

Espèce	Nom des protéines	Nombre d'allèles	Numéros d'accès de séquences			Fichiers de coordonnées de structure 3D (IMGT/3Dstructure-DB)	
			Allèle	EMBL	SwissProt	Allèle	Code
Hs	HLA-A	212	HLA-A*0201	K02883	P01892	HLA-A*0201	1akj, 1ao7, 1aqd, 1b0g, 1b0r, 1bd2, 1duy, 1duz, 1eey, 1eez, 1hhg, 1hhh, 1hhi, 1hhj, 1hhk, 1hla, 1i1f, 1i1y, 1i4f, 1i7r, 1i7t, 1i7u, 1im3, 1jf1, 1jht, 1lp9, **1oga**, 1p7q, 1qew, 1qr1, 1qrn, 1qse, 1qsf, 1s9w, 1s9x, 1s9y, 1tvb, 1tvh, 2clr, 3hla
						HLA-A*1101	1q94, 1qvo
						HLA-A*6801	1hsb, 1tmc, 2hla
	HLA-B	428	HLA-B*0702	AJ292075	P01889	HLA-B*0801	1agb, 1agc, 1agd, 1age, 1agf , 1m05, 1mi5
						HLA-B*2705	1hsa, 1jge, 1ogt, 1rog, 1roh, 1roi, 1roj, 1rok, 1rol, 1uxs
						HLA-B*2709	**1k5n**, 1of2, 1uxw
						HLA-B*3501	1a1n, 1a9b, 1a9e, 1cg9, 1xh3
						HLA-B*4402	1m6o
						HLA-B*4403	1n2r, 1sys
						HLA-B*5101	1e27, 1e28
						HLA-B*5301	1a1m, 1a1o
	HLA-Cw	94	HLA-Cw*0701	Y18499, Y18533, Y18534	P10321	HLA-Cw*0401	1im9, **1qqd**
						HLA-Cw*0304	1efx
	HLA-E	3	HLA-E*0101	AF523277	P13747	HLA-E*0101	**1mhe**
						HLA-E*0103	1kpr, 1ktl
	HLA-F	1	HLA-F*0101	X17093	P30511		
	HLA-G	1	HLA-G*0101	J03027	P17693		
Mm	H2-D1	4	H2-D1*02	M18523	P01899	H2-D1*01	1bii, 1ddh, 1qo3
						H2-D1*02	1bz9, 1ce6, 1ffn, 1ffo, 1ffp, 1fg2, 1hoc, 1inq, 1jpf, 1jpg, 1juf, 1n3n, 1n5a, 1qlf, 1s7u, 1s7v, 1s7w, 1s7x, **1wbx**, 1wby
	H2-K1	7	H2-K1*01	V00746, V00747	P01901	H2-K1*02	1bqh, 1fo0, 1fzj, 1fzk, 1fzm, 1fzo, 1g6r, 1g7p, 1g7q, 1jtr, 1kbg, 1kj2, 1kj3, 1kpu, 1kpv, 1leg, 1lek, **1lk2**, 1mwa, 1n59, 1nam, 1nan, 1osz, 1p1z, 1p4l, 1rjy, 1rjz, 1rk0, 1rk1, 1s7q, 1s7r, 1s7s, 1s7t, 1t0m, 1t0n, 1vac, 1vad, 1wbz, 2ckb, 2mha, 2vaa, 2vab
						H2-K1*03	1l6q
	H2-L	2	H2-L*02	V00749-V00752	P01897	H2-L*02	**1ld9**, 1ldp
	H2-M5	1	H2-M5*01	L14279			
	H2-Q7	10	H2-Q7*02	X03210, X03441	P14429	H2-Q7*03	**1k8d**
	H2-T3	3	H2-T3*01	M13285	P14432	H2-T3*02	**1nez**, 1r3h
Rn	RT1-AA	1	RT1-AA*01	M31018	P16391	RT1-AA*01	1ed3, 1frt, 1i1a, 1kjm, **1kjv**, 3fru

MHC-I. *Hs* : *Homo sapiens*, *Mm* : *Mus musculus*, *Rn* : *Rattus norvegicus*. La colonne 4 indique les allèles les plus représentés dans les populations caucasiennes et les séquences de référence. Pour chaque protéine, la structure 3D dont la résolution est la meilleure est en gras : 1oga_A (HLA-A, 1.4 Å), 1k5n_A (HLA-B, 1.09 Å), 1qqd_A (HLA-Cw, 2.7 Å), 1mhe_A (HLA-E, 2.85 Å), 1wbx_A (H2-D1, 1.9 Å), 1lk2_A (H2-K1, 1.35 Å), 1ld9_A (H2-L, 2.4 Å), 1k8d_A (H2-Q7, 2.3 Å), 1nez_A (H2-T3, 2.1 Å), 1kjv_A (RT1-AA, 1.48 Å). Les codes des fichiers de coordonnées correspondant à un complexe TR/pMHC-I sont soulignés.

Espèce	Nom des protéines	Nombre d'allèles	Numéros d'accès de séquences			Fichiers de coordonnées de structure 3D (IMGT/3Dstructure-DB)	
			Allèle	EMBL	SwissProt	Allèle	Code
Hs	HLA-DPA1	12	HLA-DPA1*0103	X03100	O19686		
	HLA-DPB1	87	HLA-DPB1*0401	M23906-M23908	Q30161		
	HLA-DQA1	17	HLA-DQA1*0501	Z84489	P01909	HLA-DQA1*0102	**1uvq_A**
						HLA-DQA1*0302	1jk8_A
						HLA-DQA1*0501	1s9v_A
	HLA-DQB1	41	HLA-DQB1*0301	M25325 (c)	Q29965	HLA-DQB1*0201	1s9v_B
						HLA-DQB1*0302	1jk8_B
						HLA-DQB1*0602	**1uvq_B**
	HLA-DRA	2	HLA-DRA*0101	J00203, J00204	P01903	HLA-DRA*0101	1a6a_A, 1aqd_G, 1bx2_A, 1d5m_A, 1d5x_A, 1d5z_A, 1d6e_A, 1dlh_A, **1fv1_D**, 1fyt_A, 1h15_A, 1hqr_A, 1hxy_A, 1j8h_A, 1jws_A, 1jwm_A, 1jwu_A, 1kg0_A, 1klg_A, **1klu_A**, 1lo5_A, 1pyw_A, 1r5i_A, 1seb_A, 1sje_A, 1sjh_A, 1t5w_A, 1t5x_A, 1ymm_A, 2seb_A
	HLA-DRB1	268	HLA-DRB1*1402	AJ297583 (c)	Q9GIY2	HLA-DRB1*0101	1aqd_H, 1dlh_B, 1fyt_B, 1hxy_B, 1jwm_B, 1jws_B, 1jwu_B, 1kg0_B, 1kgl_B, **1klu_B**, 1lo5_B, 1pyw_B, 1r5i_B, 1seb_B, 1sje_B, 1sjh_B, 1t5w_B, 1t5x_B
						HLA-DRB1*0301	1a6a_B
						HLA-DRB1*0401	1d5m_B, 1d5x_B, 1d5z_B, 1d6e_B, 1j8h_B, 2seb_B
						HLA-DRB1*1501	1bx2_B, 1ymm_B
	HLA-DRB3	30	HLA-DRB3*0101	BC001023 (c)	P13761		
	HLA-DRB4	6	HLA-DRB4*0101	M16942 (c)	P14762		
	HLA-DRB5	12	HLA-DRB5*0101	M20429 (c)	Q30154	HLA-DRB5*0101	**1fv1_E**, 1h15_B, 1hqr_B
	HLA-DMA	2	HLA-DMA*01	X62744 (c) (X76775)	(Q31604)	HLA-DMA*01	**1hdm_A**
	HLA-DMB	3	HLA-DMB*01	X76776	P28068	HLA-DMB*01	**1hdm_B**
	HLA-DOA	1	HLA-DOA*01	X02882	P06340		
	HLA-DOB	4	HLA-DOB*01	X87344	Q8WLR4		
Mm	H2-AA	9	H2-AA*02	V00832 (c)	P01910	H2-AA*01	1lnu_A, 1muj_B
						H2-AA*02	1d9k_C, **1iak_A**, 1jl4_A
						H2-AA*03	1es0_A, 1f3j_A, 1iao_A, 2iad_A
						H2-AA*04	1k2d_A
	H2-AB	5	H2-AB*02	M13538 (c)	P06343	H2-AB*01	1lnu_B, 1muj_B
						H2-AB*02	1d9k_D, **1iak_B**, 1jl4_B
						H2-AB*03	1es0_B, 1f3j_B
						H2-AB*04	1iao_B, 2iad_B
						H2-AB*05	1k2d_B
	H2-EA	4	H2-EA*01	K00971	P01904	H2-EA*01	**1fne_C**, 1fng_C, 1i3r_E, 1iea_A, 1ieb_C, 1kt2_C, 1ktd_A, 1r5v_A, 1r5w_A
	H2-EB1	2	H2-EB1*01	AF050157	O78196	H2-EB1*01	**1fne_D**, 1fng_D, 1i3r_F, 1iea_B, 1ieb_D, 1kt2_D, 1ktd_B, 1r5v_B, 1r5w_B
	H2-DMA	1	H2-DMA*01	AF100956	Q31621	H2-DMA*01	**1k8i_A**
	H2-DMB1	2	H2-DMB1*02	U35323	Q31106		
	H2-DMB2	1				H2-DMB2*01	**1k8i_B**
	H2-DOA	1	H2-DOA*01	AK020594 (c)	Q9QWV1		
	H2-DOB	1	H2-DOB*01	M11800	Q31143		

MHC-II. *Hs* : *Homo sapiens*, *Mm* : *Mus musculus*. La colonne 4 indique les allèles les plus représentés dans les populations caucasiennes et les séquences de référence. Pour chaque protéine, la structure 3D dont la résolution est la meilleure est représentée en gras : 1uvq (1.8 Å), 1fv1 (1.9 Å), 1klu (1.93 Å), 1hdm (2.5 Å), 1iak (1.9 Å), 1fne (1.9 Å), 1k8i (3.1 Å).

Type de récepteur	Espèce	Nom des protéines	Nombre d'allèles	Numéros d'accès de séquences			Fichiers de coordonnées de structure 3D	
				Allèle	EMBL	SwissProt	Allèle	Code
AZGP1	*Hs*	AZGP1	1	AZGP1*01	D14034	P25311	AZGP1*01	**1t7v**, 1t7w, 1t7x, 1t7y, 1t7z, 1t80, 1zag
	Mm	AZGP1	1	AZGP1*01	AF281658	Q64726		
	Rn	AZGP1	1	AZGP1*01	X75309 (c)	Q63678		
CD1	*Hs*	CD1A	1	CD1A*01	M28825 (c)	P06126	CD1A*01	1xz0, **1onq**
		CD1B	1	CD1B*01	M28826 (c)	P29016	CD1B*01	1gzp, **1gzq**, 1uqs
		CD1C	1	CD1C*01	M28827 (c)	P29017		
		CD1D	1	CD1D*01	X14974	P15813		
		CD1E	1	CD1E*01	X14975	P15812		
	Mm	CD1D1	1	CD1D1*01	X13170	P11609	CD1D1*01	**1cd1**
	Rn	CD1D1	1	CD1D1*01	AB029486	Q9R1S6		
EPCR	*Hs*	EPCR	1	EPCR*01	AF106202	Q9UNN8	EPCR*01	1l8j, **1lqv**
	Mm	EPCR	1	EPCR*01	AF162695	Q64695		
	Bt	EPCR	1	EPCR*01	L39065 (c)	Q28105		
FCGRT	*Hs*	FCGRT	1	FCGRT*01	AF220542	P55899	FCGRT*01	**1exu**
	Mm	FCGRT	1	FCGRT*01	D37872, D37873	Q61559		
	Rn	FCGRT	1	FCGRT*01	X14323 (c)	P13599	FCGRT*01	1frt, **3fru**
	Bt	FCGRT	1	FCGRT*01	AF139106 (c)			
HFE	*Hs*	HFE	1	HFE*01	Z92910	Q30201	HFE*01	**1a6z**, 1de4
	Mm	HFE	1	HFE*01	AF007558	P70387		
	Rn	HFE	1	HFE*01	AJ001517 (c)	O35799		
MIC	*Hs*	MICA	1	MICA*01	L14848 (c)	Q29983	MICA*01	1b3j, **1hyr**
		MICB	1	MICB*01	X91625 (c)	Q29980	MICB*01	**1je6**
MR1	*Hs*	MR1	1	MR1*01	AL356267	Q9TQK3		
	Mm	MR1	1	MR1*01	AF035672	O19478		
	Rn	MR1	1	MR1*01	Y13972 (c)	O19477		
RAE	*Hs*	RAET1E	3	RAET1E*01	AL355312	Q8TD07		
		RAET1H	2	RAET1H*01	AL583835	Q9BZM5		
		RAET1I	2	RAET1I*01	AL355497	Q9BZM6		
		RAET1L	2	RAET1L*01	AL355497	Q5VY80		
		RAET1N	1	RAET1N*01	AL355497	Q9BZM4	RAET1N*01	**1kcg**
	Mm	RAE1B	1	RAE1B*01	D64161 (c)	O08603	RAE1B*01	**1jfm**, 1jsk
		RAE1G	1	RAE1G*01	D64162 (c)	O08604		
	Rn	RAE1B	1	RAE1B*01				
		RAE1G	1	RAE1G*01				

MHC-I-like. *Hs* : *Homo sapiens*, *Mm* : *Mus musculus*, *Rn* : *Rattus norvegicus*, *Bt* : *Bos taurus*. La colonne 5 indique les allèles correspondant aux séquences de référence dans IMGT. Pour chaque protéine, la structure 3D dont la résolution est la meilleure est représentée en gras : 1t7v_A (*Hs* AZGP1, 1.95 Å), 1onq_A (*Hs* CD1A, 2.15 Å), 1gzq_A (*Hs* CD1B, 2.26 Å), 1cd1_A (*Mm* CD1D1, 2.67 Å), 1lqv_A (*Hs* EPCR, 1.6 Å), 1exu_A (*Hs* FCGRT, 2.7 Å), 3fru_A (*Rn* FCGRT, 2.2 Å), 1a6z_A (*Hs* HFE, 2.6 Å), 1hyr_C (*Hs* MICA, 2.7 Å), 1je6_A (*Hs* MICB, 2.5 Å), 1kcg_C (Hs RAET1N, 2.6 Å), 1jfm_A (Mm RAE1B, 2.85 Å). Les gènes MICA et MICB sont hautement polymorphes ; les allèles de MICA sont décrites par Frigoul et Lefranc (2005). Les codes des fichiers de coordonnées correspondant à un complexe MHC-I-like/ligand sont soulignés.

Annexe 6. Regroupements des acides aminés selon leurs propriétés physico-chimiques.

Groupes d'acides aminés	Propriété physico-chimique	Référence bibliographique
DNEQKR	Hydrophile	IMGT amino acid hydropathy (Pommié et al. 2004)
IVLFCMAW	Hydrophobe	
GTSYPH	Neutre	
GAS	Très petit (60-90Å^3)	IMGT amino acid volume (Pommié et al. 2004)
CDPNT	Petit (108-117Å^3)	
EVQH	Moyen (138-154Å^3)	
MILKR	Large (162-174Å^3)	
FYW	Très large (189-228Å^3)	
DE	Acide	(Wu and Brutlag 1995)
NQ	Amide	
FYW	Aromatique	
RHK	Basique	
ST	Hydroxyle	
AGILPV	Neutre	
CM	Sulfure	
DE	Acide	(Wu and Brutlag 1995)
NQ	Amide	
FYW	Aromatique	
RHK	Basique	
ST	Hydroxyle	
CM	Sulfure	
ILV		
AG		
P	Proline	
DE	Acide	IMGT amino acid chemical characteristics (Pommié et al. 2004)
NQ	Amide	
F	Phenylalanine	
Y	Tyrosine	
W	Tryptophane	
RHK	Basique	
ST	Hydroxyle	
CM	Sulfure	
AVIL	Aliphatique	
G	Glycine	
P	Proline	

Annexe 7. Bootstrap et validation croisée.

Les classifieurs supervisés sont construits à partir d'un échantillon d'apprentissage, et leur performance est ensuite évaluée sur un échantillon de test ; ces deux types d'échantillons sont non redondants. Les jeux de données de taille restreinte ne peuvent pas être divisés en un échantillon d'apprentissage et un échantillon de test de taille suffisante, et nécessitent des stratégies alternatives :

- bootstrap (Efron 1979). Pour un jeu de n observations, le principe est de sélectionner aléatoirement L ensembles de n observations avec remise à chaque sélection ; chacun de ces ensembles est utilisé itérativement comme échantillon d'apprentissage. Chacun des L classifieurs est alors testé sur le jeu de données initial, et la performance globale est évaluée par la moyenne des L tests.

- validation croisée (Hand 1986). Pour un jeu de n observations, le principe est de réaliser la phase d'apprentissage sur $n-1$ observations, de tester le classifieur sur l'observation restante, et d'itérer le processus n fois. La performance globale du classifieur est évaluée par la moyenne des n tests.

La méthode de bootstrap est couramment utilisée pour évaluer la robustesse des arbres phylogénétiques ; les n observations du jeu de données correspondent alors aux n positions de l'alignement multiple.

Efron, B., Bootstrap methods: another look at the jackknife, *Annals Stat.*, 1979, 7, 1-26.

Hand, D.J., Recent advances in error rate estimation, *Pattern Recognition Letter*, 1986, 4, 335-346.

Annexe 8. Les mesures d'entropie relative, d'information mutuelle et du χ^2.

Ce sont les mesures les plus couramment utilisées pour identifier les sites discriminants entre différentes sous-familles de séquences protéiques alignées. Ces sites présentent des caractéristiques (telles que par exemple les acides aminés observés) majoritairement conservées au sein de chaque classe de séquences, et différentes entre les classes.

L'entropie de Shannon (1948) est une mesure standard de variabilité position-dépendante des séquences d'un alignement multiple :

$$S_p = -\sum_{i=1}^{20} f_p(i) \log f_p(i),$$

avec i représentant par exemple les 20 acides aminés, et $f_p(i)$ la fréquence de l'acide aminé i à la position p. Plus une position est variable, plus sa valeur d'entropie sera élevée ; le maximum d'entropie pour une position donnée dépend du nombre de séquences de l'alignement multiple. Shenkin et al. (1991) ont mis en évidence l'intérêt de cette mesure pour localiser les régions hypervariables (CDR) des récepteurs des cellules T, et son caractère généralisable à l'étude des protéines.

L'entropie relative est une mesure de distance entre deux distributions de probabilités ; cette mesure est issue de la théorie de l'information et est également connue sous le nom de distance de Kullback-Leibler (Cover and Thomas 1991). L'entropie relative est couramment utilisée pour comparer la distribution des fréquences d'acides aminés à une position donnée respectivement pour une classe de séquences et pour l'ensemble des séquences de l'alignement (Hannenhalli and Russel 2000) :

$$RE_p = \sum_{i=1}^{20} \sum_{j=1}^{L} f_p(i,j) \log \frac{f_p(i,j)}{f_p(i)},$$

avec $f_p(i,j)$ la fréquence de l'acide aminé i à la position p des séquences de la classe j (parmi L classes) ; la mesure d'entropie relative est non négative, et nulle si les deux distributions sont identiques.

L'information mutuelle (Cover and Thomas 1991) à la position p d'un alignement multiple reflète l'association statistique entre les variables i et j représentant par exemple les acides aminés et les classes de séquences, et est définie par :

$$MI_p = \sum_{i=1}^{20} \sum_{j=1}^{L} f_p(i,j) \log \frac{f_p(i,j)}{f_p(i)\, f(j)},$$

avec $f(j)$ la fraction de séquences de la famille appartenant à la classe j ; la mesure d'information mutuelle est non négative, et nulle si ces deux variables sont statistiquement indépendantes.

Des mesures de signification statistique permettent alors d'identifier les positions de l'alignement multiple caractérisées par des valeurs d'entropie relative ou d'information mutuelle significatives, et qui constituent ainsi des sites significativement discriminants entre les classes de séquences.

Si l'on considère la table de contingence suivante :

$$CT_p = \begin{array}{c|ccc} & C_1 & \dots & C_L \\ g_1 & n_{11} & & \\ \vdots & & n_{ij} & \\ g_M & & & n_{ML} \end{array},$$

avec n_{ij} le nombre d'occurrence de l'acide aminé (ou du groupe d'acides aminés) g_i dans les séquences de la sous-famille C_j pour la position p de l'alignement multiple de la famille d'intérêt, l'expression de l'information mutuelle de cette position est

$$MI_p = \sum_{i=1}^{M}\sum_{j=1}^{L} \frac{n_{ij}}{n_{\square\square}} \log\left(\frac{n_{ij}\times n_{\square\square}}{n_{i\square}\times n_{\square j}}\right).$$

La capacité de discrimination d'une position p peut également être estimée par la mesure du χ^2 à partir de la table de contingence CT_p :

$$\chi^2{}_p = \sum_{i=1}^{M}\sum_{j=1}^{L} \frac{\left(n_{ij} - x_{ij}\right)^2}{x_{ij}},$$

avec

$$x_{ij} = \frac{n_{i\square}\times n_{\square j}}{n_{\square\square}}.$$

La mesure du χ^2 est d'autant plus élevée que les variables g et C sont corrélées, c'est-à-dire que la distribution des acides aminés à la position p est corrélée à la partition des séquences de la famille en sous-familles ; une valeur de χ^2 élevée indique donc une position discriminante.

Cover, T.M. and Thomas, J.A., *Elements of information theory*, John Wiley and Sons, New York, 1991.

Hannenhalli, S.S. and Russel, R.B., Analysis and prediction of functional sub-types from protein sequence alignments, *J. Mol. Biol.*, 2000, 303, 61-76.

Shenkin, P.S., Erman, B. And Mastrandrea, L.D., Information-theorical entropy as a measure of sequence variability, *Proteins*, 1991, 11, 297-313.

Publications

Publication 1

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Developmental and Comparative Immunology 29 (2005) 185–203
Developmental & Comparative Immunology
www.elsevier.com/locate/devcompimm

IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains

Marie-Paule Lefranc*, Christelle Pommié, Quentin Kaas, Elodie Duprat, Nathalie Bosc, Delphine Guiraudou, Christelle Jean, Manuel Ruiz, Isabelle Da Piédade, Mathieu Rouard, Elodie Foulquier, Valérie Thouvenin, Gérard Lefranc

IMGT, the International ImMunoGeneTics Information System®, LIGM, Laboratoire d'ImmunoGénétique Moléculaire, Université Montpellier II, UPR CNRS 1142, IGH, 141 rue de la Cardonille, 34396 Montpellier cedex 5, France

Received 19 May 2004; accepted 16 July 2004
Available online 1 September 2004

Abstract

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) provides a common access to expertly annotated data on the genome, proteome, genetics and structure of immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC), and related proteins of the immune system (RPI) of human and other vertebrates. The NUMEROTATION concept of IMGT-ONTOLOGY has allowed to define a unique numbering for the variable domains (V-DOMAINs) and for the V-LIKE-DOMAINs. In this paper, this standardized characterization is extended to the constant domains (C-DOMAINs), and to the C-LIKE-DOMAINs, leading, for the first time, to their standardized description of mutations, allelic polymorphisms, two-dimensional (2D) representations and tridimensional (3D) structures. The IMGT unique numbering is, therefore, highly valuable for the comparative, structural or evolutionary studies of the immunoglobulin superfamily (IgSF) domains, V-DOMAINs and C-DOMAINs of IG and TR in vertebrates, and V-LIKE-DOMAINs and C-LIKE-DOMAINs of proteins other than IG and TR, in any species.

Keywords: IMGT; Immunoglobulin; T cell receptor; Constant domain; Immunoglobulin superfamily; V-set; C-set; Colliers de Perles

1. Introduction

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [1] is a high quality integrated knowledge resource specialized in immunoglobulins (IG), T cell receptors (TR), major

Abbreviations: 2D, two-dimensional; 3D, tridimensional; IG, immunoglobulin; IgSF, immunoglobulin superfamily; MHC, major histocompatibility complex; RPI, related proteins of the immune system; TR, T cell receptor.

* Corresponding author. Tel.: +33-4-9961-9965; fax: +33-4-9961-9901.
E-mail address: lefranc@ligm.igh.cnrs.fr (M.-P. Lefranc).

0145-305X/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dci.2004.07.003

histocompatibility complex (MHC), and related proteins of the immune system (RPI) of human and other vertebrates [1–14]. IMGT provides a common access to expertly annotated data on the genome, proteome, genetics and structure of the IG, TR, MHC, and RPI, based on the IMGT Scientific chart rules and on the IMGT-ONTOLOGY concepts [15]. More particularly, the IMGT unique numbering [16–18], based on the NUMEROTATION concept of IMGT-ONTOLOGY, has been set up to provide a standardized description of mutations, allelic polymorphisms, two-dimensional (2D) and tridimensional (3D) structure representations of the IG and TR variable domains (V-DOMAINs), whatever the antigen receptor, the chain type or the species [18]. The IMGT unique numbering for V-DOMAINs is used in all the IMGT components [3–8]: databases (IMGT/LIGM-DB [19], IMGT/PRIMER-DB [20], IMGT/GENE-DB [21], IMGT/3Dstructure-DB [22]), tools for sequence and structure analysis (IMGT/V-QUEST [23], IMGT/JunctionAnalysis [24], IMGT/Allele-Align, IMGT/PhyloGene [25], IMGT/StructuralQuery [22]), and Web resources ('IMGT Protein displays' [26,27], 'IMGT Colliers de Perles' 2D representations [28], and 'IMGT Alignments of Alleles' [29,30]; see IMGT Repertoire, http://imgt.cines.fr). Interestingly, the IMGT unique numbering for V-DOMAIN was fully applicable to the V-LIKE-DOMAINs of proteins other than IG and TR [18], although their genomic structure (V-LIKE-DOMAINs are often encoded by one exon) is different from that of the IG and TR V-DOMAINs (encoded by the rearranged V–(D–)J genes).

In this paper, the standardized IMGT unique numbering is extended to the IG and TR constant domains (C-DOMAINs) of the IG and TR of all jawed vertebrates, and to the C-LIKE-DOMAINs of proteins other than IG and TR of any species. The IMGT unique numbering represents therefore a major step forward for the comparative analysis and for the 2D and 3D structure and evolution studies of the immunoglobulin superfamily (IgSF) domains, V-DOMAINs and C-DOMAINs of IG and TR in vertebrates, and V-LIKE-DOMAINs and C-LIKE-DOMAINs of proteins other than IG and TR, in any species.

2. C-DOMAIN definition and relations with C-REGION

The C-DOMAIN of an IG or TR chain is a 3D structural unit comprising about 100 amino acids in seven antiparallel beta strands, on two sheets [31–33]. The seven strands of the C-DOMAIN, designated as A, B, C, D, E, F and G, have a topology and 3D structure similar to that of a V-DOMAIN without its C′ and C″ strands. Indeed, the C-DOMAIN beta sandwich fold is built up from the seven beta strands arranged so that four strands form one beta sheet, and three strands form a second sheet [31–33]. Depending from the topology of the D strand, which can be in one or the other sheet, the beta sandwich comprises ABE and GFCD, or ABED and GFC. ABE and GFC are closely packed against each other and joined together by a disulfide bridge from strand B in the first sheet with strand F in the second sheet [31], conserved in both the C-DOMAINs and V-DOMAINs. Amino acids with conserved physico-chemical characteristics form and stabilize the framework by packing the beta sheets through hydrophobic interactions giving a hydrophobic core [31,34].

Whereas the C-DOMAIN is a structural unit of an IG or TR chain, the C-REGION represents the part of an IG or TR chain encoded by the C-GENE [29,30]. Depending on the IG or TR chain type, the C-DOMAIN may correspond to a complete C-REGION, or to most of the C-REGION, or to only part of the C-REGION (if the C-REGION comprises several C-DOMAINs). As respective examples, (i) the domains C-KAPPA, C-LAMBDA and C-IOTA correspond to the complete C-REGION of an IG kappa, lambda, and iota chains, respectively (if the IG light chain type is not specified, the domain is designated as CL); (ii) the domains C-ALPHA, C-BETA, C-GAMMA and C-DELTA correspond to most (but not to the entirety) of the C-REGION of the TR alpha, beta, gamma and delta chains, respectively (indeed, the TR chains are transmembrane proteins whose C-REGION encoded by the C-GENE comprises, in addition to the C-DOMAIN, the CONNECTING-REGION, the TRANSMEMBRANE-REGION, and the INTRACYTOPLASMIC-REGION) [29,30]; (iii) the domains CH1, CH2, CH3 and,

if present, CH4, correspond to only part of the C-REGION of the IG heavy chains (e.g. the domains CH1, CH2 and CH3 of the human IGHG1 represent 98, 110 and 105 amino acids, respectively, on a total of 399 or 330 amino acids for the complete C-REGION of a membrane gamma 1 chain or that of a secreted gamma 1 chain, respectively [29]) (Fig. 1). It is worth to note that these relations between C-DOMAIN and C-REGION are quite different from those between V-DOMAIN and V-REGION, since the V-DOMAIN of an IG or TR chain results from the junction of two or three different regions: V and J (V–J-REGION of the IG light chains, and of the TR alpha and gamma chains), or V, D and J (V–D–J-REGION of the IG heavy chains, and of the TR beta and delta chains) [29,30] (Fig. 1).

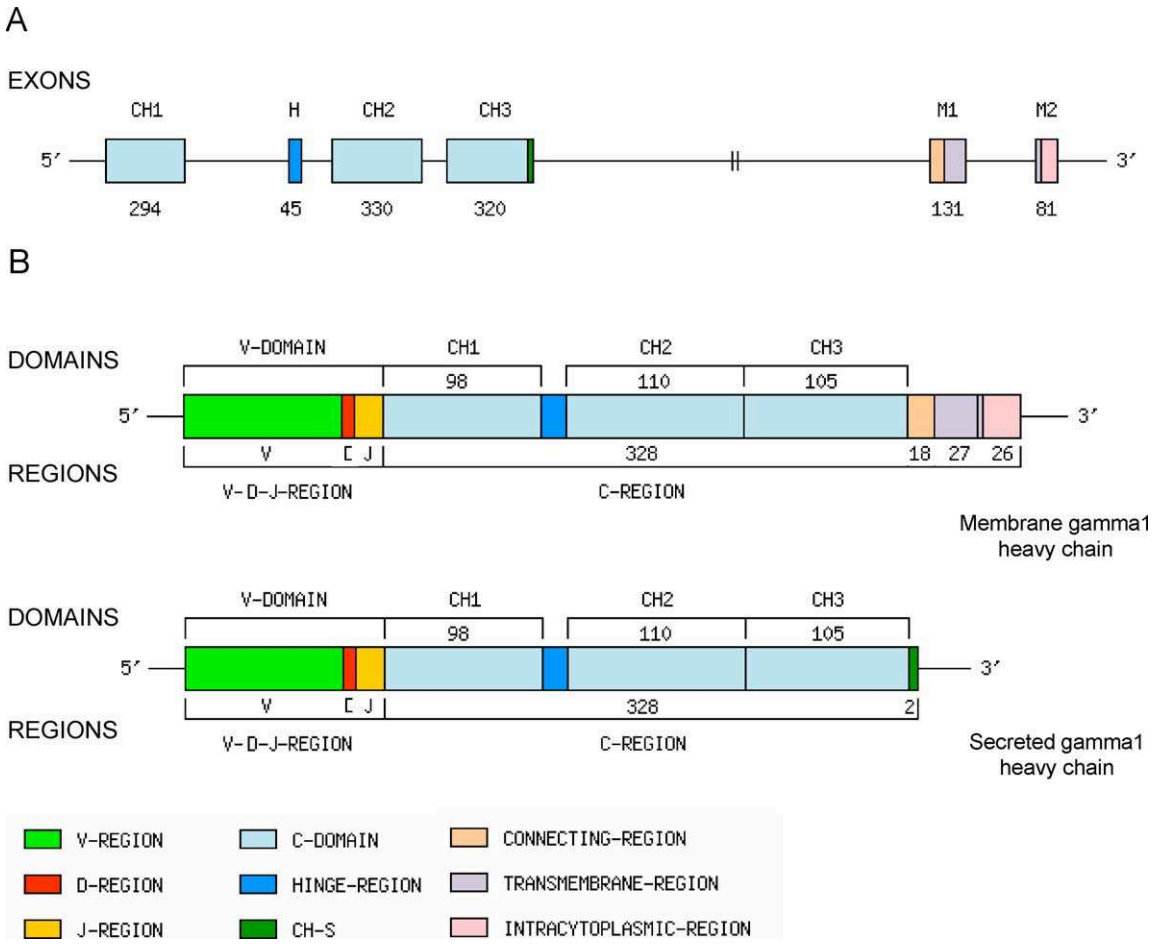


Fig. 1. Correspondence between exons, domains and regions. (A) Exons of the *Homo sapiens* IGHG1 gene shown as an example. Length of the exons are in base pairs. Introns are not at scale (see Ref. [29] for a representation at scale). (B) Domains and regions of a *Homo sapiens* membrane and secreted IG gamma 1 heavy chain shown as examples. Lengths of the domains and regions are in number of amino acids. The CH3 exon (320 nucleotides) encodes 107 amino acids (105 amino acids of the CH3 domain and 2 amino acids of CH-S, only present in the secreted gamma 1 chain). Exon M1 (131 nucleotides) encodes 44 amino acids (the 18 amino acids of the CONNECTING-REGION and 26 of the 27 amino acids of the TRANSMEMBRANE-REGION), exon M2 (81 nucleotides) encodes 27 amino acids (the last amino acid of the TRANSMEMBRANE-REGION and the 26 amino acids of the INTRACYTOPLASMIC-REGION). (Human IGHC 'Alignments of Alleles' and Protein displays in IMGT Repertoire, http://imgt.cines.fr and [29]).

3. IMGT unique numbering for C-DOMAIN

Owing to the high conservation of the structure of the immunoglobulin fold, the IMGT unique numbering for the C-DOMAINs of the IG and TR chains is derived from the IMGT unique numbering for V-DOMAIN [16–18], based on the NUMEROTATION concept of IMGT-ONTOLOGY. In the IMGT unique numbering, the conserved amino acids always have the same position, for instance Cysteine 23 (1st-CYS), Tryptophan 41 (CONSERVED-TRP), hydrophobic amino acid 89, Cysteine 104 (2nd-CYS). The hydrophobic amino acids of the antiparallel beta strands (framework regions) are also found in conserved positions [29,30].

In order to set up the IMGT unique numbering for C-DOMAIN, we first identified the amino acid positions, which correspond to equivalent positions in the V-DOMAIN. This correspondence was established by sequence alignment comparison of annotated IG and TR from the IMGT/LIGM-DB database [5,8,19] and by structural data analysis of IG and TR with known 3D structures from IMGT/3Dstructure-DB, http://imgt.cines.fr [22]. Seventy-two positions were identified as structurally equivalent between the C-DOMAIN and the V-DOMAIN, when the strands A to G were compared. They comprise: positions 1–15 (strand A), 16–26 (strand B), 39–45 (strand C), 77–84 (strand D), 85–96 (strand E), 97–104 (strand F), 118–128 (strand G) (Table 1). These positions are boxed in Fig. 2 and are indicated in the header upper line of the IMGT Protein display (Fig. 3) (Ruiz M., Martinez-Jean C. and Lefranc M.-P. IMGT Repertoire (for IG and TR) > Protein displays, on-line 27/02/2001, http://imgt.cines.fr).

We then identified the C-DOMAIN characteristic positions. These positions, shown in the header lower line of the IMGT Protein display (Fig. 3), comprise either additional or missing amino acid positions in the C-DOMAIN compared to the V-DOMAIN. Thirty-seven additional positions are characteristic of the C-DOMAIN numbering and are designated by a number followed by a dot and a number (Table 1): 1.1–1.9 at the N-terminal end of the C-DOMAIN, 15.1–15.3 at the AB turn, 45.1–45.9 which represent a characteristic transversal strand CD, 84.1–84.7, 85.7–85.1 at the DE loop (these positions correspond to longer antiparallel D and E strands in the C-DOMAIN), 96.1 and 96.2 at the EF turn. Interestingly, the additional positions 45.1–45.9 in the C-DOMAIN, compared to the V-DOMAIN, correspond to structural differences between the V- and C-DOMAINs. Indeed, these positions 45.1–45.9 represent a transversal strand between C and D in the C-DOMAIN whereas, in contrast, C′ and C″ in the V-DOMAIN are antiparallel strands. Thirty-three positions are missing in the C-DOMAIN compared to the V-DOMAIN. Thirty-one of these missing positions (46–76) correspond to the last amino acid of the C strand, to the two C′ and C″ strands, and to the C′C″ (or CDR2-IMGT) loop of the V-DOMAIN [18]. The last two missing positions (37 and 38) are in the BC loop. The C-DOMAIN BC loop has a maximum length of 10 amino acids (positions 27–36) compared to the maximum length of 12 amino acids for the equivalent V-DOMAIN CDR1-IMGT.

Table 1
Rules for gaps and additional positions in C-DOMAIN and C-LIKE-DOMAIN

A-Length of the A strand (positions 1.9-1.1, 1-15) and B strand (positions 16–26), and gaps at the AB turn								
Examples			A strand 1.9-1.1[a], 1–15			B strand 16–26		Number of gaps at the AB turn
Species	Gene	Domain	Number of additional positions at the N-terminus	Gap positions[b]	A strand length (16–24)[a]	Gap positions[b]	B strand length (8–11)	
Homo sapiens	IGHG1	CH1, CH3	4		19		11	0
Homo sapiens	IGHG1	CH2	6		21		11	0
Mus musculus	CD1D	C-LIKE [D3]	1	15	15	16	10	2
Homo sapiens	HLA-B	C-LIKE [D3]	1	14 15	14	16	10	3
Homo sapiens	TRAC	C-ALPHA	5	10 13 14 15	16	16	10	4
Homo sapiens	TRDC	C-DELTA	6	10 12 13 14 15	16	16 17 18	8	7

(*continued on next page*)

Table 1 (*continued*)

B-Length of the AB turn (positions 15.1–15.3)

Examples			AB turn 15.1–15.3	
Species	Gene	Domain	Additional positions[c]	AB turn length (0–3)[c]
Homo sapiens	IGHG1	CH1, CH3		0
Homo sapiens	TRBC2	C-BETA2	15.1	1
Homo sapiens	CD4	C-LIKE [D2]	15.1	1
Homo sapiens	IGHG1	CH2	15.1 15.2	2

C-Length of the BC loop (positions 27–36)

Examples			BC loop 27–36		
Species	Gene	Domain	Number of gaps	Gap positions[d]	BC loop length (6–10)[e]
Homo sapiens	IGHG1	CH2	0		10
Homo sapiens	IGHE	CH3	0		10
Mus musculus	IGHE	CH3	1	32	9
Homo sapiens	IGHG1	CH1, CH3	2	31 32	8
Homo sapiens	TRAC	C-ALPHA	3	31 32 33	7
Mus musculus	IGHD	CH1	4	30 31 32 33	6

D-Length of the C strand (positions 39–45) and CD transversal strand (positions 45.1–45.9)

Examples			C strand 39–45	CD transversal strand[g] 45.1–45.9	
Species	Gene	Domain	C strand length (7)[g]	Additional positions[f]	CD length (0–9)[f]
Homo sapiens	TRAC	C-ALPHA	7		0
Homo sapiens	CD4	C-LIKE [D4]	7		0
Homo sapiens	TRDC	C-DELTA	7	45.1	1
Homo sapiens	FCGR1A	C-LIKE [D1]	7	45.1,45.2	2
Homo sapiens	FCGR1A	C-LIKE [D2]	7	45.1–45.3	3
Homo sapiens	IGHG1	CH1	7	45.1–45.3	3
Homo sapiens	IGHG1	CH2, CH3	7	45.1–45.4	4
Homo sapiens	TRBC2	C-BETA2	7	45.1–45.5	5
Homo sapiens	TRGC1	C-GAMMA1	7	45.1–45.5	5
Homo sapiens	HLA-B	C-LIKE [D3]	7	45.1–45.5	5
Homo sapiens	IGHA1	CH3	7	45.1–45.6	6
Sus crofa	IGHE	CH3	7	45.1–45.7	7
Anarhichas minor	IGHC1S1	C-IOTA	7	45.1–45.9	9
Seriola quinqueradiata	IGIC1S22	C-IOTA	7	45.1–45.9	9
Siniperca chuatsi	IGIC1S3	C-IOTA	7	45.1–45.9	9
Siniperca chuatsi	IGIC1S5	C-IOTA	7	45.1–45.9	9

E-Length of the D strand (positions 77–84) and E strand (positions 85–96), and gaps at the DE

Examples			D strand 77–84[h]		E strand 85–96		Number of gaps at the DE turn
Species	Gene	Domain	Gap positions[i]	D strand length (5–8)[h]	Gap positions	E strand length (8–12)[j]	
Homo sapiens	IGHG1	CH2, CH3		8		12	0
Homo sapiens	FCGR1A	C-LIKE [D2]	83 84	6	85 86	10	4
Homo sapiens	FCGR1A	C-LIKE [Dl]	82 83 84	5	85 86	10	5

F-Length of the DE turn (positions 84.1–84.7, 85.7–85.1)

Examples			DE turn 84.1–84.7, 85.7–85.1	
Species	Gene	Domain	Additional positions[k]	DE turn length (6–14)[k]
Meleagris gallopavo	Telokin	C-LIKE [D]	84.1	1
Homo sapiens	CD4	C-LIKE [D4]	84.1	1
Homo sapiens	CD3E	C-LIKE [D]	84.1, 84.2, 85.1, 85.2	4

(*continued on next page*)

Table 1 (*continued*)

Examples			DE turn 84.1–84.7, 85.7–85.1	
Species	Gene	Domain	Additional positions[k]	DE turn length (6–14)[k]
Homo sapiens	ICAM1	C-LIKE [D1]	84.1–84.3, 85.1, 85.2	5
Mus musculus	IGHE	CH1	84.1–84.3, 85.1–85.3	6
Homo sapiens	TRDC	C-DELTA	84.1–84.4, 85.1–85.4	8
Homo sapiens	TRGC1	C-GAMMA1	84.1–84.4, 85.1–85.4	8
Homo sapiens	IGHG1	CH1, CH2, CH3	84.1–84.4, 85.1–85.4	8
Mus musculus	TRBC1	C-BETA1	84.1–84.5, 85.1–85.4	9
Canis familiaris	IGHA	CH3	84.1–84.5, 85.1–85.5	10
Homo sapiens	IGHA1	CH3	84.1–84.6, 85.1–85.5	11
Homo sapiens	IGHM	CH2	84.1–84.6, 85.1–85.6	12
Homo sapiens	TRBC2	C-BETA2	84.1–84.7, 85.1–85.6	13
Homo sapiens	TRAC	C-ALPHA	84.1–84.7, 85.1–85.7	14

G-Length of the E strand (positions 85–96) and F strand (positions 97–104), and gaps at the EF turn

Examples			E strand 85–96		F strand 97–104		Number of gaps at the EF turn
Species	Gene	Domain	Gap positions	E strand length (8–12)[l]	Gap positions	F strand length (4–8)[m]	
Homo sapiens	IGHG1	CH2, CH3		12		8	0
Homo sapiens	FCGR2A	C-LIKE [D1]	*85 86* 96	9		8	1
Homo sapiens	IGHG1	CH1	96	11		8	1
Bos taurus	IGHG1	CH1	96	11	97	7	2
Homo sapiens	TRGC1	C-GAMMA1	96	11	97	7	2
Homo sapiens	HLA-B	C-LIKE [D3]	95 96	10	97	7	3
Rattus norvegicus	IGHG2B	CH1	91 92[n]	10	97	7	3
Homo sapiens	TRDC	C-DELTA	95 96	10	97 98	6	4
Oryctolagus cuniculus	IGHG	CH1	95 96	10	97 98	6	4
Homo sapiens	TRAC	C-ALPHA	93 94 95 96	8	97 98 99 100	4	8

H-Length of the EF turn (positions 96.1–96.2)

Examples			EF turn 96.1–96.2	
Species	Gene	Domain	Additional positions[o]	EF turn length (0.2)[o]
Homo sapiens	IGHG1	CH1, CH2, CH3		0
Homo sapiens	TRBC2	C-BETA2	96.1	1
Homo sapiens	IGHM	CH1	96.1 96.2	2

I-Length of the FG loop (positions 105–117) (except for the C-BETA domains)

Examples			FG loop 105–117		
Species	Gene	Domain	Number of gaps	Gap positions	FG loop length (7–13)[p]
Homo sapiens	IGHE	CH1	0		13
Mus musculus	TRAC	C-ALPHA	1	111	12
Homo sapiens	IGHG1	CH3	1	111	12
Homo sapiens	IGHG1	CH1, CH2	2	111 112	11
Homo sapiens	FCGR2A	C-LIKE [D2]	3	110 111 112	10
Ovis aries	IGHA	CH2	4	110 111 112 113	9
Homo sapiens	FCGR2A	C-LIKE [D1]	6	109 110 111 112 113 114	7

J-Length of the FG loop of the C-BETA domains (positions 105–111.6, 112.6–117)

Examples			FG loop 105–111.6, 112.6–117		
Species	Gene	Domain	Number of additional positions	Additional positions[q]	FG loop length (25)[q]
Homo sapiens	TRBC2	C-BETA2	12	111.1–111.6, 112.1–112.6	25

(*continued on next page*)

K-Length of the G strand (positions 118–128)				
Examples			G strand positions 118–128	
Species	Gene	Domain	C-terminal positions	G strand length (4–11)[r]
Homo sapiens	IGHG1	CH1	121	4
Homo sapiens	HLA-A	C-LIKE [D3]	121	4
Mus musculus	IGHE	CH1	122	5
Homo sapiens	IGHG1	CH2, CH3	125	8
Mus musculus	IGHE	CH2, CH3, CH4	125	8
Mus musculus	TRBC1	C-BETA1	125	8
Homo sapiens	IGKC	C-KAPPA	126	9
Homo sapiens	FCGR2A	C-LIKE [D2]	126	9
Homo sapiens	IGLC1	C-LAMBDA1	127	10
Homo sapiens	FCGR1A	C-LIKE [D1]	127	10
Homo sapiens	FCGR2A	C-LIKE[D1]	127	10

Rules are described by comparison to the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [18]. Gaps and additional positions were confirmed for proteins with known 3D structures (PDB codes in IMGT Repertoire > Protein displays, http://imgt.cines.fr). Strand, turn and loop lengths are shown, in number of amino acids between parentheses, in the table headers.

[a] Up to nine additional positions (numbered 1.1–1.9, starting from position next to 1 towards the N-terminal end) may be found at the N-terminal end of the A strand. The maximal length of the A strand is 24 amino acids.

[b] Gap positions are based on 3D structures, and if 3D structures are not known, gaps are equally distributed on strands A and B with, for an odd number, one more gap on strand A.

[c] C-DOMAIN and C-LIKE-DOMAIN may have additional amino acids (potentially 3) at the AB turn, which define the AB turn length.

[d] Gap positions start with 32, then 31, 33, 30.

[e] The maximum length of the BC loop is 10 amino acids. If the number of amino acids is odd, there is one more amino acid position on the left. There are no positions 37 and 38 in C-DOMAINs and C-LIKE-DOMAINs.

[f] The maximum length of the C strand is seven amino acids. There is no position 46 in C-DOMAINs and C-LIKE-DOMAINs.

[g] The CD transversal strand is characteristic of the C-DOMAIN and C-LIKE-DOMAIN. The maximum length of the CD strand is 9 amino acids, found in Teleostei IGIC (available online in IMGT Repertoire http://imgt.cines.fr): IGIC1S1 (AF137397) gene of Spotted wolffish (*Anarhichas minor*), IGIC1S22 (AB062662) gene of Five-ray yellowtail (*Seriola quinqueradiata*), IGIC1S3 (AF454470) and IGIC1S5 (AY013294) genes of Chinese perch (*Siniperca chuatsi*). However, since this length is exceptional, usual IMGT Collier de Perles only display seven positions. Amino acid positions are added from left to right in sequence alignments.

[h] The maximal length of the D strand is eight amino acids. There are no positions 75 and 76 in C-DOMAINs and C-LIKE-DOMAINs.

[i] Gap positions are based on 3D structures, and if 3D structures are not known, gaps at the DE turn are equally distributed on strands D and E with, for an odd number, one more gap on strand D.

[j] The maximum length of the E strand is 12 amino acids. Note that the E strand length also depends from gaps found at the EF turn, which is reflected by E strand lengths of eight and nine amino acids, as described in Table 1G.

[k] Most of the C-DOMAINs and C-LIKE-DOMAINs have additional amino acids (potentially 14) at the DE turn which extend the D and E antiparallel beta strands. The number of additional positions defines the DE turn length. The numbering of the additional positions starts from positions next to 84 and 85, respectively, towards the top of the DE turn. If the number of additional amino acids is odd, there is one more position on the left.

[l] The maximal length of the E strand is 12 amino acids. Note that the E strand length also depends from gaps found at the DE turn (gap positions 85, 86 shown in italics, for the *Homo sapiens* FCGR2A) as described in Table 1E.

[m] The maximal length of the F strand is eight amino acids. Gap positions are based on 3D structures, but if 3D structures are not known, gaps are based on sequence alignments.

[n] Gaps were assigned by sequence alignment with the CH1 of the IGHG2A and IGHG2C.

[o] C-DOMAIN and C-LIKE-DOMAIN may have additional positions (potentially 2) at the EF turn, which define the EF length.

[p] Except for the C-BETA domains (TRBC sequences) described in Table 1J, the maximal length of the FG loop is 13 amino acids. For an odd number of gaps, there is one more gap on the left (starting with position 111).

[q] C-BETA domains (TRBC sequences) have an insertion of 12 positions between 111 and 112. The length of the FG loop in C-BETA domains is 25 amino acids. The numbering of the additional positions starts from positions next to 111 and 112, respectively, towards the top of the FG loop.

[r] If longer G strands are found, positions will be numbered consecutively.

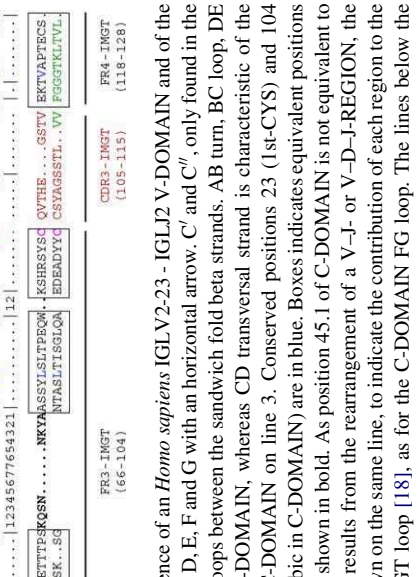


Fig. 2. Correspondence between the C-DOMAIN and V-DOMAIN IMGT unique numbering. Amino acid sequence of an *Homo sapiens* IGLV2-23 - IGLJ2 V-DOMAIN and of the *H. sapiens* IGLC1 C-DOMAIN are shown as examples. The upper line indicates the beta strands A, B, C, C′, C″, D, E, F and G with an horizontal arrow. C′ and C″, only found in the V-DOMAIN, are shown with dashed arrows. AB, BC, CD, C′C″, DE, EF and FG correspond to the turns and loops between the sandwich fold beta strands. AB turn, BC loop, DE and EF turns, and FG loop are found in both V-DOMAIN and C-DOMAIN. C′C″ loop is only found in V-DOMAIN, whereas CD transversal strand is characteristic of the C-DOMAIN. The IMGT unique numbering is shown on lines 2 and 3 with additional positions found in C-DOMAIN on line 3. Conserved positions 23 (1st-CYS) and 104 (2nd-CYS) are in magenta. Conserved positions 41 (CONSERVED-TRP), 89 (hydrophobic) and 121 (hydrophobic in C-DOMAIN) are in blue. Boxes indicates equivalent positions in both the V-DOMAIN and C-DOMAIN. Amino acids at additional positions in the C-DOMAIN sequence are shown in bold. As position 45.1 of C-DOMAIN is not equivalent to position 46 in V-DOMAIN, both positions 45.1 and 46 are shown in this figure. As a V-DOMAIN sequence results from the rearrangement of a V–J- or V–D–J-REGION, the sequence of the germline V-REGION and that of the germline J-REGION (in green) taken as examples are shown on the same line, to indicate the contribution of each region to the V-DOMAIN. Note that in a ‘true’ V–J rearrangement, the gaps would be placed at the top of the CDR3-IMGT loop [18], as for the C-DOMAIN FG loop. The lines below the sequences indicate the V-DOMAIN FR-IMGT and CDR-IMGT and their delimitations.

Rules for gaps and additional positions in C-DOMAIN and C-LIKE-DOMAIN are described in details in IMGT Scientific chart (http://imgt.cines.fr) and in Table 1. Correspondence between the C-DOMAIN and V-DOMAIN IMGT unique numbering is shown in Fig. 2 with an *Homo sapiens* IGLV2-23 - IGLJ2 V-DOMAIN and the *Homo sapiens* IGLC1 C-DOMAIN taken as examples.

It is worth to note that it is the analysis of structural data and sequence comparisons that we were carrying out to apply the IMGT unique numbering for the description of the IG and TR C-DOMAIN, which showed us that the standardized numbering of the FG loop of the C-DOMAIN could be applied to the CDR3-IMGT of rearranged IG and TR sequences (Fig. 3 in Ref. [18]). More precisely, the hydrogen bonds between second-CYS 104 and position 119 in the C-DOMAIN correspond structurally to the hydrogen bonds between second-CYS 104 and the Glycine which follows the J-TRP or J-PHE in the J-REGION. This Glycine was therefore numbered as 119, and as a consequence, J-TRP and J-PHE as 118.

4. IMGT unique numbering for C-DOMAIN and sequence data analysis

The IMGT unique numbering for C-DOMAIN allows for the first time a standardized comparison of the nucleotide substitutions and amino acid changes between different constant domains of a same gene or chain, or between constant domains of different IG and TR genes or chains, from either the same species, or from different species (Fig. 3). The IMGT unique numbering is also crucial for the standardization of the allele description. Indeed, in IMGT, the polymorphisms are described by comparison to the sequences from the IMGT reference directory. All the human IG and TR gene names from this IMGT reference directory [29,30], including the C-GENEs, were approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 [39], and entered in IMGT/GENE-DB [5,21], GDB [40], LocusLink at NCBI (USA) [41] and GeneCards [42]. This standardized nomenclature, based on the CLASSIFICATION concept of IMGT-ONTOLOGY [15], represented

a major step in the setting up of the 'Tables of alleles' and 'Alignments of alleles' of the IG and TR genes (http://imgt.cines.fr). V-GENE polymorphisms in IMGT [29,30] have been described from the start according to the IMGT unique numbering for V-REGION set up in 1997 [16–18]. In contrast, C-GENE polymorphisms were initially described according to the exon numbering [29,30] and sequence comparison between exons of different lengths was not easy. The implementation of the IMGT unique numbering for the C-DOMAIN represents therefore a new major step in the setting up of standardized 'Alignments of alleles' whatever the receptor, the chain, or the species (IMGT Repertoire, http://imgt.cines.fr) (Fig. 3). Owing to that standardization, the sequence polymorphisms of any C-DOMAIN of any IG or TR can very easily be compared and analysed.

5. IMGT unique numbering for C-DOMAIN and structural data comparison

Beyond sequence data comparison, the IMGT unique numbering for C-DOMAIN provides information on the strand and loop lengths (Table 2) and allows standardized IMGT Protein displays (Fig. 3) and IMGT Colliers de Perles (Fig. 4) for the IG and TR C-DOMAINs of any chain type from any species. Practically, structural data comparison of strand or loop of the same length can be done directly using the IMGT unique numbering. For example, all codons (or amino acids) at position 28 can be compared between domains with a BC loop of a given length. This standardization allows the structural characterization of a position inside a domain, and the statistical analysis of amino acid properties, position per position, between domains, as this has been demonstrated for the V-DOMAIN [34]. Fig. 4 shows the IMGT Colliers de Perles for the *Homo sapiens* IGHG1 CH1, C-KAPPA, C-LAMBDA1 and *Mus musculus* C-BETA1 domains, as examples.

As soon as the first IMGT Collier de Perles was set up on the Web site in December 1997, the enormous potential of the IMGT unique numbering as a means to control data coherence was obvious. For new sequences, for which no 3D structures are available, the IMGT Colliers de Perles allow to precisely delimit the strands and loops and give information on the topological organization of the domain. The IMGT unique numbering is also used in more sophisticated representations of the IMGT Colliers de Perles on two layers (Fig. 4) which allow, when 3D structures are available, the visualisation of the hydrogen bonds between amino acids belonging to beta strands from the same sheet or from different sheets (IMGT/3Dstructure-DB, http://imgt.cines.fr) [22].

6. IMGT unique numbering for C-LIKE-DOMAIN

A C-LIKE-DOMAIN is a domain of similar structure to a C-DOMAIN, found in chains other than IG and TR [43–52]. The IMGT unique numbering for the C-LIKE-DOMAIN follows exactly the same rules as those of the C-DOMAIN (Table 1). Strand and loop lengths of 40 examples of C-LIKE-DOMAINs are given in Table 2. The IMGT Protein display of the corresponding C-LIKE-DOMAIN sequences are shown in Fig. 3. The IMGT Colliers de Perles of four representative C-LIKE-DOMAINs (*Homo sapiens* HLA-B [D3], B2M [D], FCGR2A [D1], and *Meleagris gallopavo* telokin [D]) are represented in Fig. 5. Detailed IMGT Alignment of alleles of *Homo sapiens* FCGR3B and IMGT Colliers de Perles of [D1] and [D2] of the FCGR3B*02 allele [53] further highlight the importance of the IMGT unique numbering standardization for the polymorphism and structure analysis and comparison of the C-LIKE-DOMAINs.

The IMGT unique numbering provides, for the first time, a standardized approach to analyse the sequences and structures of any domain belonging to the C-set of the IgSF [32] (the C-set comprises the IMGT C-DOMAINs and C-LIKE-DOMAINs). Three features are worth noting: (i) In IMGT, any C-DOMAIN or C-LIKE-DOMAIN is characterized by its strand and loop lengths (Table 2). Examples are shown in Figs. 3–5. This first feature of the IMGT standardization based on the IMGT unique numbering shows that the distinction between the C1, C2, I1 and I2 types found in the literature and in the databases to describe the IgSF C-set domains [32,33,54–56] is unapplicable when dealing with sequences for which no structural data are known. Indeed, the four domain

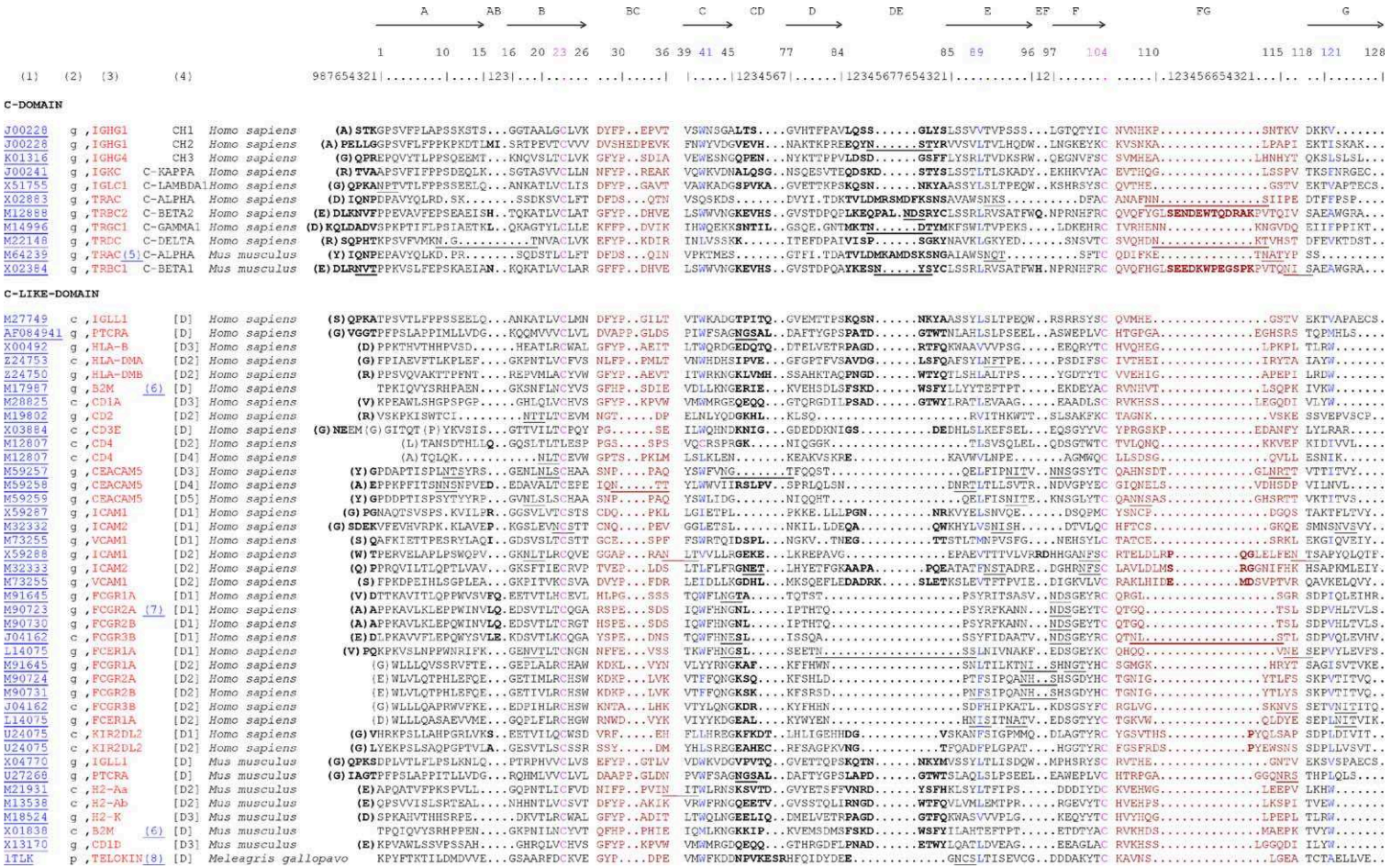


Fig. 3. IMGT Protein display of examples of C-DOMAINs (IG and TR) and C-LIKE-DOMAINs (proteins other than IG and TR). The protein display is according to the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN, based on the NUMEROTATION concept of IMGT-ONTOLOGY [15]. Sandwich fold beta strands are shown by horizontal arrows. Dots indicate missing amino acids according to the IMGT unique numbering. Amino acids resulting from a splicing with a preceding exon are shown between parentheses (for *Homo sapiens* FCGR3B [D2], the information is from M90745, for *H. sapiens* HLA-DMA [D2] from NT_007592, for *H. sapiens* CD3E [D] from NT_033899, for *H. sapiens* CD4 [D2] and [D4] from NT_009759, for *H. sapiens* CEACAM5 [D3], [D4] and [D5] from NT_011109, and for *Mus musculus* H2-Aa [D2] and H2-Ab [D2] from NT_039649). Putative N-glycosylation sites (N-X-S/T) are underlined. (1) Accession numbers are from IMGT/LIGM-DB (http://imgt.cines.fr) [5,19] for IG and TR and from EMBL/GenBank/DDBJ [35–37] for proteins other than IG and TR; Telokin identifier is from PDB [38] and IMGT/3Dstructure-DB [22]. (2) Molecule type. c: cDNA; g: genomic DNA; p: protein. (3) Gene names (symbols) for IG and TR are according to the IMGT Nomenclature committee (IMGT-NC) [29,30] and the HUGO Nomenclature Committee (HGNC) [39]. Full gene designations are the following: IGHG1: Immunoglobulin heavy constant gamma 1; IGHG4: Immunoglobulin heavy constant gamma 4; TRAC: T cell

receptor alpha constant; TRBC2: T cell receptor beta constant 2; TRGC1: T cell receptor gamma constant 1; TRDC: T cell receptor delta constant; IGLL1: Immunoglobulin lambda-like polypeptide 1; PTCRA: pre T-cell antigen receptor alpha; HLA-B: Major histocompatibility complex, class I, B; HLA-DMA: MHC class II, DM alpha; HLA-DMB: MHC class II, DM beta; B2M: Beta-2-microglobulin; CD1A: CD1A antigen, a polypeptide; CD2: CD2 antigen (p50), sheep red blood cell receptor; CD3E: CD3E antigen, epsilon polypeptide (TiT3 complex); CD4: CD4 antigen (p55); CEACAM5: Carcinoembryonic antigen-related cell adhesion molecule 5; ICAM1: intercellular adhesion molecule 1 (CD54), human rhinovirus receptor; ICAM2: intercellular adhesion molecule 2; VCAM1: vascular cell adhesion molecule 1; FCGR1A: Fc fragment of IgG, high affinity Ia, receptor for (CD64); FCGR2A: Fc fragment of IgG, low affinity IIa, receptor for (CD32); FCGR2B: Fc fragment of IgG, low affinity IIb, receptor for (CD32); FCGR3B: Fc fragment of IgG, low affinity IIIb, receptor for (CD16); FCER1A: Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide; H2-Aa: histocompatibility 2, class II antigen A, alpha; H2-Ab: histocompatibility 2, class II antigen A, beta; H2-K: histocompatibility 2, K region; CD1D: CD1D antigen, d polypeptide. (4) Domain name. The C-DOMAINs are designated with the IMGT labels (IMGT Scientific chart, http://imgt.cines.fr) The C-LIKE-DOMAINs are designated by the letter D between brackets with a number, corresponding to the position of the domain from the N-terminal end of the protein, and relative to the other domains. Membrane proteins quoted in this figure are of type I, that is with the N-terminal end being extracellular. There is no number if there is a unique C-LIKE domain in the chain. (5) Q at position 6 is according to M64239 (and replace A in PDB and IMGT/3Dstructure-DB entries 1tcr_A). (6) B2M [D] is encoded by EX2 and is preceded at the N-terminal end by SGLEGIQR or TGLYAIQK encoded by the EX1 end in *Homo sapiens* or *Mus musculus*, respectively (EX1 encodes 23 amino acids, including the L-REGION). The splicing site is between the last EX1 codon (encoding R in *Mus musculus*) and the first EX2 codon (encoding T in both species). The proteolytic cleavage site of the L-REGION is not known. (7) [D1] is encoded by EX3 and is preceded at the N-terminal end by seven amino acids (A)SADSQA encoded by EX2. The proteolytic cleavage site of the L-REGION is not known. (8) The N-terminal and C-terminal ends of Telokin [D] need to be confirmed by genomic sequence. Amino acid one-letter abbreviation: A (Ala), alanine; C (Cys), cysteine; D (Asp), aspartic acid; E (Glu), glutamic acid; F (Phe), phenylalanine; G (Gly), glycine; H (His), histidine; I (Ileu), isoleucine; K (Lys), lysine; L (Leu), leucine; M (Met), methionine; N (Asn), asparagine; P (Pro), proline; Q (Gln), glutamine; R (Arg), arginine; S (Ser), serine; T (Thr), threonine; V (Val), valine; W (Trp), tryptophan; Y (Tyr), tyrosine.

types C1, C2 (also known as H [57]), I1 and I2, are based on structural differences [55,58]. C1 and C2 differ by the location of the fourth strand D, which is in the ABE sheet for C1, and in the GFC sheet for C2 (the D strand is then designated as C′ by some authors) (Fig. 6). Thus C1 contains ABED and GFC sheets, whereas C2 is characterized by ABE and GFCD (or GFCC′) sheets (Table 3). The C1 type initially defined by the topology observed in the IG constant domain was described as being shared by the TR constant domain, and the C-like domain of MHC class I, MHC class II and B2M, and thought to be a characteristic of the immunoreceptor of the adaptative immune response. In the literature, the assignment of C-set domains to C1 or C2 is often done by extrapolation, without available 3D structures. Moreover, the 3D structure of the 2C T cell receptor C-ALPHA domain (PDB and IMGT/3Dstructure-DB: 1tcr_A) displays a quite different conformation than that of the C1 type with a beta sheet containing the GABED strands [59]. Therefore, the distinction between C1 and C2 is not so straightforward and as more 3D structures become available, more intermediate or variant structures will be described.

The structural analysis of telokin, the C-terminal domain of myosin light chain kinase [60] revealed a new domain type designated as 'I' [55], which was defined as an intermediate between the V and C1 types because, as frequently found in the V domain type, the A strand, in its second part, is located on the GFC sheet (and therefore designated as A′ strand), and as C1, it lacks the V-DOMAIN C′ and C″ strands. This I type was later divided in I1 and I2 [58] which differ by the location of the fourth strand, which is in the ABE sheet for I1, and in the A′GFC sheet for I2. Thus I1 contains ABED and A′GFC sheets, whereas I2 is characterized by ABE and A′GFCD (or A′GFCC″) sheets (I1 and I2 are described as I and E types in [61]). In the absence of 3D structures, a classification of the constant domains based on such structural differences cannot be taken into account, and the IMGT description based on strand and loop lengths and IMGT Colliers de Perles is more pertinent. (ii) A second feature of the IMGT standardization is the comparison of cDNA and/or amino acid sequences with genomic sequences, and the identification of the splicing sites, to delimit precisely the limits of the C-LIKE-DOMAIN,

Table 2
Strand and loop lengths of examples of C-DOMAINs and C-LIKE-DOMAINS

Species	Gene name	PDB code	Domain	A 1.9–1.1 1–15 (6–23)		AB 15.1–15.3 (0–2)	B 16–26 (7–11)	BC 27–36 (4–10)	C 39–45 (7)	CD 45.1–45.7 (0–7)	D 77–84 (4–8)	DE 84.1–84.7 85.7–86.1 (0–14)	E 85–96 (8–12)	EF 96.1–96.2 (0–2)	F 97–104 (4–8)	FG 105–117 111.1–111.6, 112–1–112.6 (7–25)	G 118–128 (4–10)	Total
C-DOMAINS																		
Homo sapiens	IGHG1	1hzh_H	CH1	+4	19		11	8	7	3	8	8	11		8	11	4	98
	IGHG1	1hzh_H	CH2	+6	21	2	11	10	7	4	8	8	12		8	11	8	110
	IGHG4	1adq_A	CH3	+4	19		11	8	7	4	8	8	12		8	12	8	105
	IGKC	1dfb_L	C–KAPPA	+4	19		11	8	7	5	8	9	12		8	11	9	107
	IGLC1	1a8j_L	C–LAMBDA1	+5	20		11	8	7	5	8	8	12		8	9	10	106
	TRAC	1qrn_D	C–ALPHA	+5,− 4	16		10	7	7		7	14	8		4	11	7	91
	TRBC2	1qm_E	C–BETA2	+7	22	1	11	8	7	5	8	13	12	1	8	25	8	129
	TRGC1	1hxm_B	C–GAMMA1	+8	23	1	11	8	7	5	7	8	11		7	13	9	110
	TRDC	1hxm_A	C–DELTA	+6,− 5	16		8	8	7	1	8	8	10		6	12	9	93
Mus musculus	TRAC	1tcr_A	C–ALPHA	+5,− 4	16		10	7	7		7	14	8		4	12	2	87
	TRBC1	1tcr_B	C–BETA1	+7	22	1	11	8	7	5	8	9	12	1	8	25	8	125
C-LIKE-DOMAINs																		
Homo sapiens	IGLL1		[D]	+5	20		11	8	7	5	8	8	12		8	9	10	106
	PTCRA		[D]	+5	20		11	9	7	5	8	8	12		8	12	7	107
	HLA–B	1a1m–A	[D3]	+1,− 2	14		10	8	7	5	8	8	10		7	11	4	92
	HLA–DMA	1hdm_A	[D2]	+1,− 1	15		11	8	7	4	8	8	10		7	11	4	93
	HLA–DMB	1hdm_B	[D2]	+1,− 1	15		11	8	7	5	8	8	10		7	11	4	94
	B2M	1lds_A	[D]	− 1	14		11	8	7	4	8	8	10		7	11	4	92
	CD1A	1onq_A	[D3]	+1,− 1	15		10	8	7	4	8	8	10		7	12	4	93
	CD2	1hnf	[D2]	+1,− 4	12		9	5	7	4	4		9		8	9	9	76
	CD3E		[D]	+3	18		11	4	7	4	8	4	11		8	13	6	94
	CD4	1wio_A	[D2]	− 5	10	1	11	6	7	2	6		9		8	11	7	78
	CD4	1wio_A	[D4]	− 5,− 4	6		7	8	7		8	1	9		6	10	5	67
Homo sapiens	CEACAM5		[D3]	+2	17		11	6	7		6		10		8	13	7	85
	CEACAM5		[D4]	+2	17	1	11	5	7	5	8		12		8	13	6	93
	CEACAM5		[D5]	+2	17		11	6	7		6		10		8	13	7	85
	ICAM1	1d3l_A	[D1]	+3,− 1	17	1	11	6	7		7	5	10		6	9	9	88
	ICAM2	1zxq	[D1]	+5,− 1	19	1	11	6	7		7	4	10		6	9	9	89
	VCAMI	1vsc_A	[D1]	+2	17	1	11	6	7	4	6	4	11		7	9	9	92
	ICAM1	1d3l_A	[D2]	+2	17		11	7	7	4	8		12	2	8	16	10	102
	ICAM2	1zxq	[D2]	+2	17		11	7	7	4	8	7	12		8	16	10	107
	VCAM1	1vsc_A	[D2]	+1	16		11	7	7	4	8	9	12		8	16	9	107
	FCGR1A		[D1]	+2	17	2	11	7	7	2	5		10		8	7	10	86
	FCGR2A	1fcg_A	[D1]	+2	17	2	11	7	7	2	6		9		8	7	10	86
	FCGR2B	2fcb_A	[D1]	+2	17	2	11	7	7	2	6		9		8	7	10	86
	FCGR3B	1e4k_C	[D1]	+2	17	2	11	7	7	2	5		10		8	7	10	86
	FCER1A	1f6a_A	[D1]	+3	18		11	7	7	2	5		10		8	7	10	85
	FCGR1A		[D2]	− 1	14		11	7	7	3	6		10		8	9	10	85
	FCGR2A	1fcg_A	[D2]	− 1	14		11	7	7	3	6		10		8	10	9	85
	FCGR2B	2fcb_A	[D2]	− 1	14		11	7	7	3	6		10		8	10	9	85
	FCGR3B	le4k_C	[D2]	− 1	14		11	7	7	3	6		10		8	10	10	86
	FCER1A	1f6a_A	[D2]	− 1	14		11	7	7	3	6		10		8	10	10	86
	KIR2DL2	1efx_D	[D1]	+2	17	1	11	5	7	5	8	3	12		8	14	9	100
	KIR2DL2	1efx_D	[D2]	+2	17	1	11	5	7	5	8	3	11		7	14	9	98

Musmuscu-lus	IGLL1		[D]	+5	20	11	8	7	5	8	8	12	8	9	10	106
	PTCRA		[D]	+5	20	11	9	7	5	8	8	12	8	12	7	107
	H2–Aa	1d9k_C	[D2]	+1, −1	15	11	8	7	5	8	8	10	7	11	4	94
	H2–Ab	1d9k_D	[D2]	+1, −1	15	11	8	7	5	8	8	10	7	11	4	94
	H2–K	2vaa_A	[D3]	+1, −2	14	10	8	7	5	8	8	10	7	11	4	92
	B2M	2vaa_B	[D]	−1	14	11	8	7	4	8	8	10	7	11	4	92
	CD1D	1cd1_A	[D3]	+1, −1	15	10	8	7	4	8	8	10	7	12	4	93
Meleagris gallopavo	Telokin	1fhg_A	[D]		15[a]	11	6	7	7	8	1	12	8	9	9[a]	93

The range of lengths observed in the selected examples of C-DOMAINs and C-LIKE-DOMAINs are shown between parentheses in the header. The total number of amino acids of each domain is indicated in the column 'Total'.
[a] The N-terminal and C-terminal ends of telokin [D] need to be confirmed by genomic sequence.

a C-LIKE-DOMAIN being frequently encoded by a unique exon, as this is the case for the C-DOMAIN (Fig. 1). This IMGT standardization for the domain delimitations explains the discrepancies observed with the generalist Swiss-Prot database which does not take into account this criteria. (iii) At last, a third feature is the C-LIKE-DOMAIN IMGT Collier de Perles, which, in the absence of available 3D structures, is particularly useful to compare domains of very diverse families, and to characterize them by their strand and loop lengths.

7. Conclusion

The IMGT unique numbering allows, for the first time, to compare any C-DOMAIN of IG and TR and C-LIKE-DOMAIN of proteins other than IG or TR, between them, and to any V-DOMAIN and V-LIKE-DOMAIN [18], that is to compare any domain belonging to the IgSF C-set or V-set. Sequences and 3D structures can be analysed whatever the domain (C-DOMAIN, C-LIKE-DOMAIN, V-DOMAIN, V-LIKE-DOMAIN), the receptor (IG, TR, or more generally IgSF), the chain type (heavy or light for IG; alpha, beta, gamma or delta for TR; or more generally IgSF chain), or the species. The IMGT unique numbering has many advantages. The strand and loop lengths (the number of codons or amino acids, that is the number of occupied positions) become crucial information, which characterizes the domains (V-DOMAINs, V-LIKE-DOMAINs, C-DOMAINs and C-LIKE-DOMAINs). The IMGT unique numbering allows standardized representations of nucleotide and amino acid sequences in IMGT Repertoire (http://imgt.cines.fr): Tables of strand and loop lengths, Tables of alleles, Alignments of alleles, Protein displays, IMGT Colliers de Perles, 3D structures. The IMGT unique numbering is applied through all the components (databases, tools and Web resources) of the IMGT information system® (http://imgt.cines.fr) [5], for the standardized label annotations [62] and database queries [19–22]. The IMGT unique numbering represents, therefore, a major step forward in the analysis and comparison of the sequence evolution and structure of the IgSF domains.

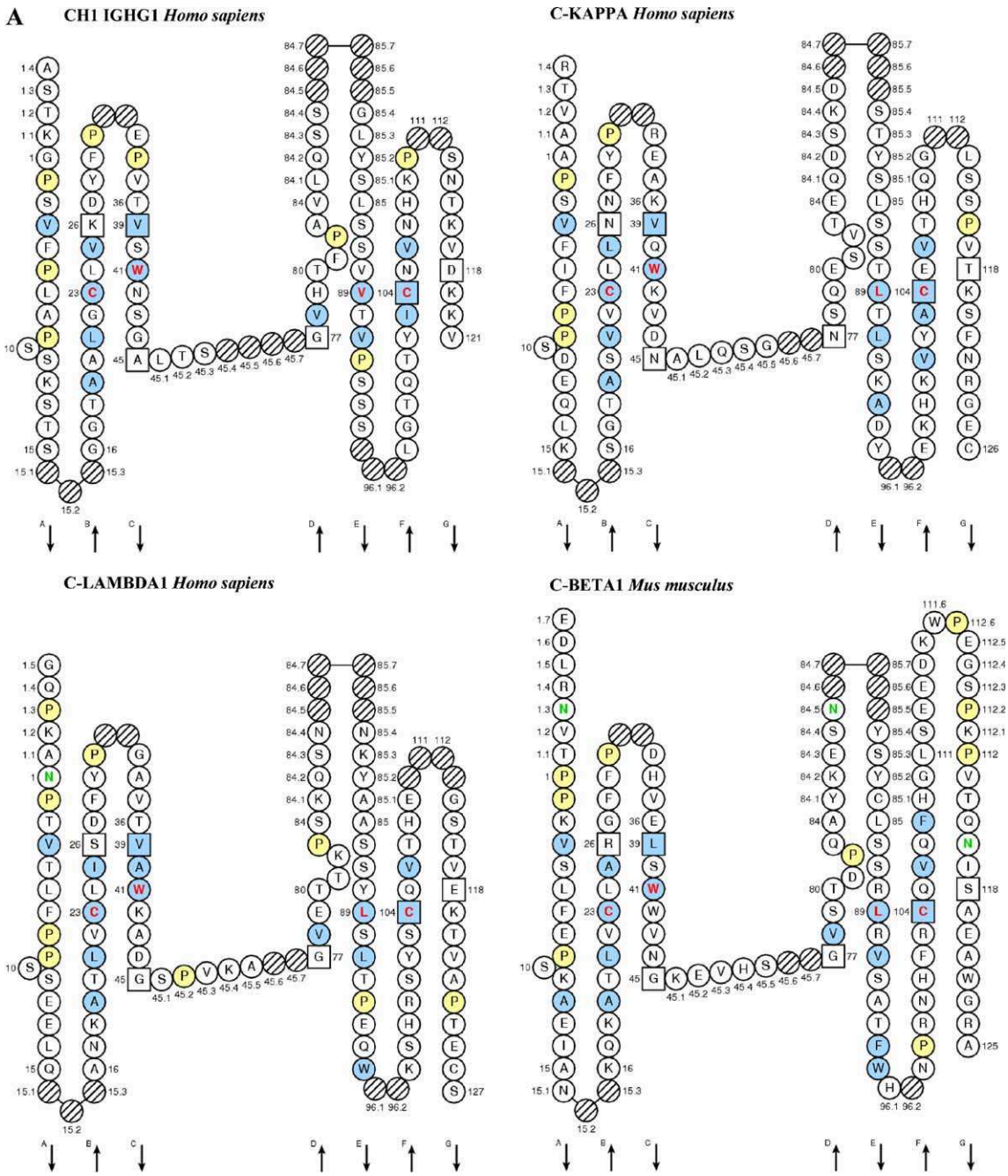


Fig. 4. IMGT Collier de Perles of C-DOMAINs. (A) on one layer (B) on two layers. CH1 of *Homo sapiens* IGHG1 (IMGT/LIGM-DB: J00228); C-KAPPA of *H. sapiens* IGKC (IMGT/LIGM-DB: J00241); C-LAMBDA1 of *H. sapiens* IGLC1 (IMGT/LIGM-DB: X51755); C-BETA1 of *Mus musculus* TRBC1 (IMGT/LIGM-DB: X02384). The first amino acids (A1.4, R1.4, G1.5 and E1.7) are encoded by a codon which results from the splicing with an IGHJ, IGKJ, IGLJ and TRBJ, respectively [29,30]. Amino acids are shown in the one-letter abbreviation. Positions at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and Tryptophan (W) are found in more than 50% of analysed IG and TR sequences are shown in blue. All Proline (P) are shown in yellow. The positions 26, 39, 104 and 118 are shown in squares by

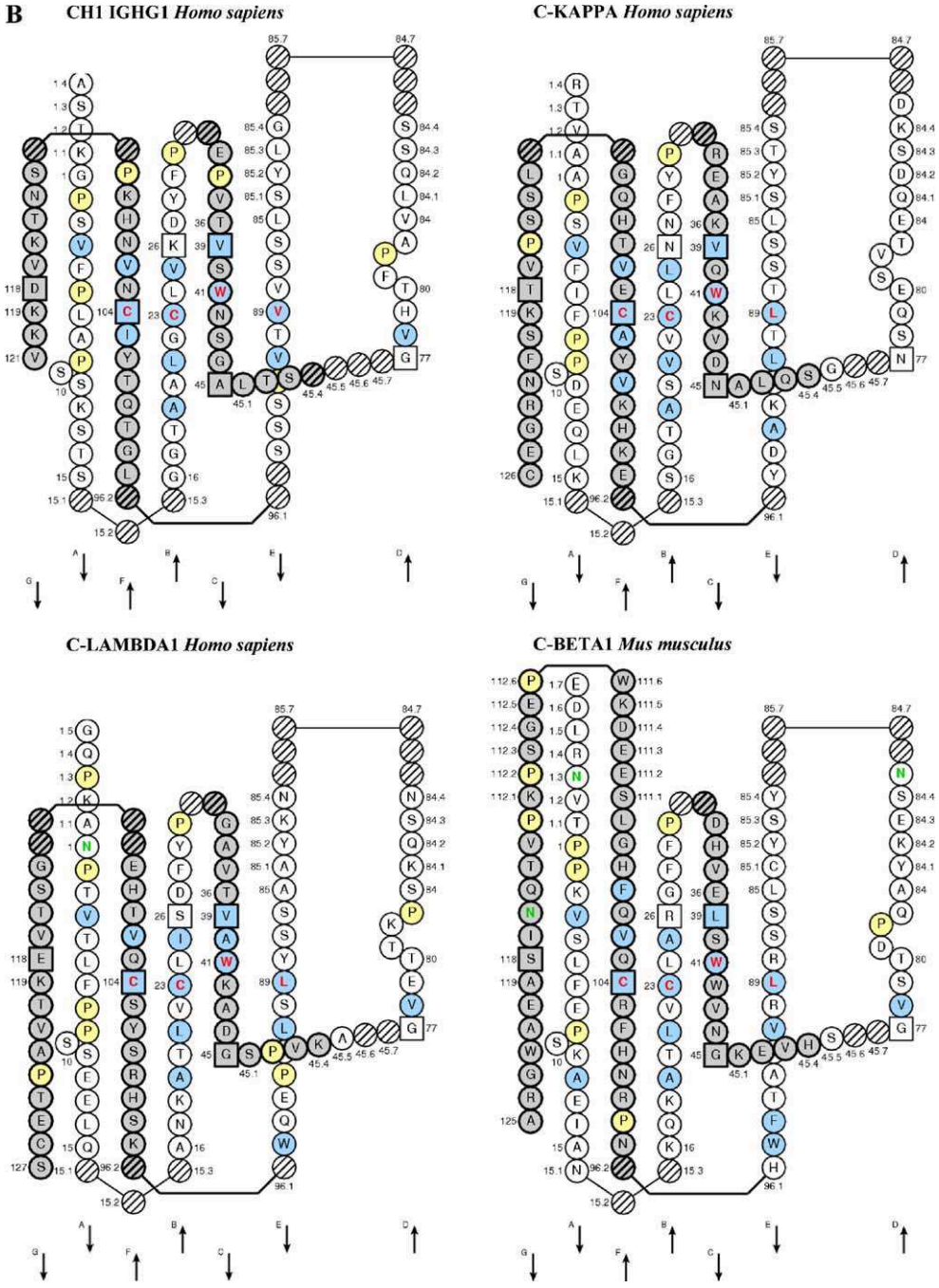


Fig. 4 (*continued*)

homology with the corresponding positions in the V-DOMAINs. Positions 45 and 77 which delimit the characteristic transversal 'CD' strand of the C-DOMAINs are also shown in squares. Hatched circles correspond to missing positions according to the IMGT unique numbering for C-DOMAINs. Arrows indicate the direction of the beta strands (IMGT Repertoire, http://imgt.cines.fr).

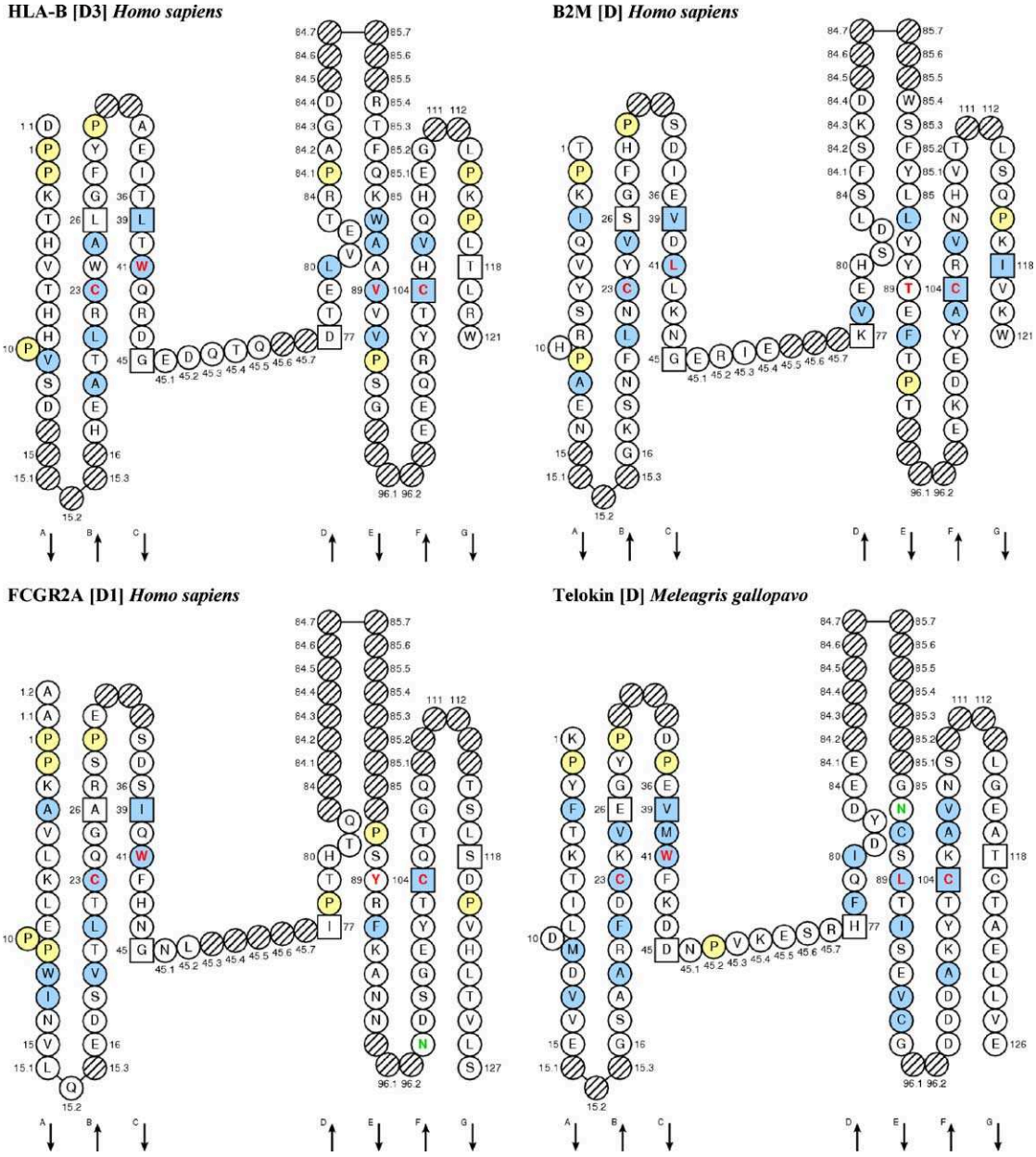


Fig. 5. IMGT Colliers de Perles of C-LIKE-DOMAINs. [D3] of *Homo sapiens* HLA-B (EMBL/GenBank/DDBJ: AB088084; PDB and IMGT/3Dstructure-DB: 1alm_A); [D] of *H. sapiens* B2M (EMBL/GenBank/DDBJ: M17987; PDB and IMGT/3Dstructure-DB: 1alm_B); [D1] of *H. sapiens* FCGR2A (EMBL/GenBank/DDBJ: M90723; PDB and IMGT/3Dstructure-DB: 1fcg_A); [D] of *Meleagris gallopavo* (turkey) telokin (PDB and IMGT/3Dstructure-DB: 1fhg_A). Amino acids are shown in the one-letter abbreviation. Positions at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and Tryptophan (W) are found in more than 50% of analysed IG and TR sequences are shown in blue. All Proline (P) are shown in yellow. The positions 26, 39, 104 and 118 are shown in squares by homology with the corresponding positions in the V-DOMAINs. Positions 45 and 77 which delimit the characteristic transversal 'CD' strand of the C-LIKE-DOMAINs are also shown in squares. Hatched circles correspond to missing positions according to the IMGT unique numbering for C-DOMAINs and C-LIKE-DOMAINs. Arrows indicate the direction of the beta strands (IMGT Repertoire, http://imgt.cines.fr).

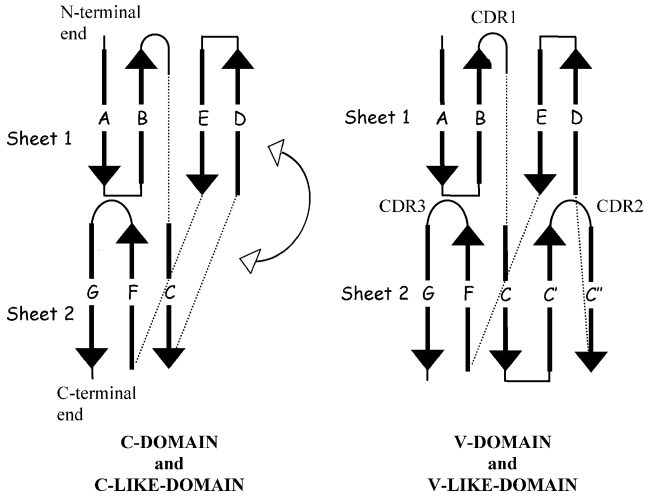


Fig. 6. Schematic representations of the C-DOMAIN and C-LIKE-DOMAIN, and of the V-DOMAIN and V-LIKE-DOMAIN. A double-headed arrow shows that the D strand can be localized in sheet 1 (on the back) or sheet 2 (on the front) depending from the length of the CD transversal strand. The second part of the A strand can be located in sheet 2 and is then designated as A′. This feature, described as 'strand A switching' in the literature, is not shown in IMGT Colliers de Perles, as this can be determined or verified only if 3D structures are available.

Table 3
Correspondence between the IMGT classification of the IgSF domains (C-DOMAIN, C-LIKE-DOMAIN, V-DOMAIN and V-LIKE-DOMAIN) and the diverse designations found in the literature

IMGT IgSF domains	C-DOMAIN (for IG and TR)	C-LIKE-DOMAIN (for proteins other than IG and TR)				V-DOMAIN (for IG and TR)	V-LIKE-DOMAIN (for proteins other than IG and TR)
Literature	C1[a] [32]	C1[a] [32]	C2 [32] or H [57]	I1 [58] or I [55,61]	I2 [58] or E [61]	V [32]	V [32]
Sheet 1	ABED	ABED	ABE	ABED	ABE	ABED	ABED
Sheet 2	GFC	GFC	GFCD (or GFCC′)	A′GFC	A′GFCD (or A′GFCC′)	A′GFCC′C″	A′GFCC′C″

[a] In the literature C1 is used for the IG and TR C-DOMAINs and for the MHC C-like domains. The diverse designations found in the literature for the C-set domain (C1, C2, Il, 12) are based on strand localisation in sheets 1 and 2. These designations are not used for the assignment of IMGT C-DOMAIN and C-LIKE-DOMAIN sequences, for the following reasons: (i) they are frequently extrapolated to sequences for which no 3D structures are available, therefore leading to misinterpretation, (ii) the designation 'strand C′' used in the literature for the C2 (or H) and I2 (or E) domain types is not the equivalent of strand C′ of the V-DOMAINs and V-LIKE-DOMAINs, (iii) the distinction made in the literature between C1 (as found in proteins of the immune adaptative response) and C2 (as found in other proteins) may suffer exceptions, (iv) as more 3D structures become available, and given the heterogeneity of the loop and strand lengths for C-DOMAINs and C-LIKE-DOMAINs, more intermediate or variant structures may be described (as shown by the 2C T cell receptor 3D structure IMGT/3Dstructure-DB: 1tcr_A).

Acknowledgements

We are grateful to Géraldine Folch, Chantal Ginestoux, Véronique Giudicelli, Joumana Jabado-Michaloud, Céline Protat, Dominique Scaviner and Denys Chaume for their helpful discussions. We thank Laurent Douchy and Bertrand Monnier for contribution to the tables, and Nora Bonnet-Saidali and Anita Gomez for editorial work. IMGT is funded by the European Union's 5th PCRDT (QLG2-2000-01287) program, the Centre National de la Recherche Scientifique (CNRS), and the Ministère de l'Education Nationale et de de la Recherche.

References

[1] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2003;31:307–10.

[2] Warr GW, Clem LW, Soderhall K. The international ImMunoGeneTics database IMGT. Dev Comp Immunol 2003;27:1.

[3] Lefranc M-P. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev Comp Immunol 2002; 26:697–705.

[4] Lefranc M-P. IMGT-ONTOLOGY databases, tools and web resources for immunogenetics and immunoinformatics. Mol Immunol 2004;40:647–59.

[5] 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, Lefranc G, IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. In Silico Biology 2003;4,0004. http://www.bioinfo.de/isb/2003/04/0004/. *In Silico* Biology 2004;4: 17–29.

[6] Lefranc M-P. IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis. Leukemia 2003;17:260–6 http://imgt.cines.fr.

[7] Lefranc M-P. IMGT, the international ImMunoGeneTics information system®. http://imgt.cines.fr. In: Bock G, Goode J, editors. Immunoinformatics: bioinformatics strategies for better understanding of immune function. Novartis Foundation Symposium 254. Chichester, UK: Wiley; 2003. p.126–36 [discussion p. 136–42, 216–22, 250–52].

[8] Lefranc M-P. IMGT, the international ImMunoGeneTics information system®. http://imgt.cines.fr. In: B.K.C. Lo Antibody engineering: methods and protocols. Methods in Molecular 248 Biology, 2nd ed. Totowa, NJ: Humana press; 2003;248:27–49 [chapter 3].

[9] Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, Bontrop R, Lemaitre M, Malik A, Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 1997;25:206–11.

[10] Lefranc M-P, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Chaume D. IMGT, the International ImMunoGeneTics database. Nucl Acids Res 1998;26:297–303.

[11] Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Barbié V, Chaume D. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 1999;27:209–12.

[12] Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, Marsh SG, Bontrop R, Lemaitre M, Lefranc G, Chaume D, Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2000;28:219–21.

[13] Lefranc M-P. IMGT ImMunoGeneTics Database. International BIOforum 2000;4:98–100.

[14] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2001;29:207–9.

[15] Giudicelli V, Lefranc M-P. Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 1999;15:1047–54.

[16] Lefranc M-P. Unique database numbering system for immunogenetic analysis. Immunol Today 1997;18:509.

[17] Lefranc M-P. The IGMT unique numbering for immunoglobulins. T cell receptors and Ig-like domains. The Immunologist 1999;7:132–6.

[18] Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 2003;27:55–77.

[19] Chaume D, Giudicelli V, Lefranc M-P. IMGT/LIGM-DB. In: The molecular biology database collection. Nucl Acids Res 2004;32 http://www3.oup.co.uk/nar/database/summary/504.

[20] Folch G, Bertrand J, Lemaitre M, Lefranc M-P. IMGT/PRIMER-DB. In: The molecular biology database collection. Nucl Acids Res 2004;32 http://www3.oup.co.uk/nar/database/summary/505.

[21] Giudicelli V, Lefranc M-P. IMGT/GENE-DB. In: The molecular biology database collection. Nucl Acids Res 2004;32. http://www3.oup.co.uk/nar/database/summary/503.

[22] Kaas Q, Ruiz M, Lefranc M-P. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucl Acids Res 2004;32:D208–D210.

[23] Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V–J and V–D–J rearrangement analysis. Nucl Acids Res 2004; 32:W435–W440.

[24] Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP. 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.

[25] Elemento O, Lefranc M-P. IMGT/PhyloGene: an online software package for phylogenetic analysis of immunoglobulin and T cell receptor genes. Dev Comp Immunol 2003;27: 763–79.

[26] Scaviner D, Barbié V, Ruiz M, Lefranc M-P. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp Clin Immunogenet 1999;16: 234–40.

[27] Folch G, Scaviner D, Contet V, Lefranc M-P. Protein displays of the human T cell receptor alpha, beta, gamma and delta variable and joining regions. Exp Clin Immunogenet 2000;17: 205–15.

[28] Ruiz M, Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 2002;53:857–83.

[29] Lefranc M-P, Lefranc G. The immunoglobulin FactsBook. London, UK: Academic Press; 2001, pp. 458.

[30] Lefranc M-P, Lefranc G. The T cell receptor FactsBook. London, UK: Academic Press; 2001, pp. 398.

[31] Lesk AM, Chothia C. Evolution of proteins formed by beta-sheets II. The core of the immunoglobulin domains. J Mol Biol 1982;160:325–42.

[32] Williams AF, Barclay AM. The immunoglobulin family: domains for cell surface recognition. Annu Rev Immunol 1988;6:381–405.
[33] Bork P, Holm L, Sander C. The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol 1994;242:309–20.
[34] Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc M-P. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recogn 2004;17:17–32.
[35] Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, Faruque N, Garcia-Pastor M, Harte N, Kanz C, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Stoehr P, Stoesser G, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R, The EMBL. Nucleotide sequence database. Nucl Acids Res 2004;32:D27–D30.
[36] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. Nucl Acids Res 2004;32: D23–D26.
[37] Miyazaki S, Sugawara H, Ikeo K, Gojobori T, Tateno Y. DDBJ in the stream of various biological data. Nucl Acids Res 2004;32:D31–D34.
[38] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000;28:235–42.
[39] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. Genomics 2002;79:464–70.
[40] Letovsky SI, Cottingham RW, Porter CJ, Li PW. GDB: the human genome database. Nucl Acids Res 1998;26:94–9.
[41] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucl Acids Res 2001;29:137–40.
[42] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet 1997;13:163.
[43] Garrett TPJ, Wang J, Yan Y, Liu J, Harrison SC. Refinement and analysis of the structure of the first two domains of human CD4. J Mol Biol 1993;234:763–78.
[44] Bajorath J, Peach RJ, Linsley PS. Immunoglobulin fold characteristics of B7-1 (CD80) and B7-2 (CD86). Protein Sci 1994;3:2148–50.
[45] Huang Z, Li S, Korngold R. Immunoglobulin superfamily proteins: structure, mechanisms, and drug discovery. Biopolymers 1997;43:367–82.
[46] Samaridis J, Colonna M. Cloning of novel immunoglobulin superfamily receptors expressed on human myeloid and lymphoid cells: structural evidence for new stimulatory and inhibitory pathways. Eur J Immunol 1997;27:660–5.
[47] Chretien I, Marcuz A, Courtet M, Katevuo K, Vainio O, Heath JK, White SJ, Du Pasquier L. CTX, a Xenopus thymocyte receptor, defines a molecular family conserved throughout vertebrates. Eur J Immunol 1998;28:4094–104.
[48] Halaby DM, Mornon JP. The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity. J Mol Evol 1998;46:389–400.
[49] Ioerger TR, Du C, Linthicum DS. Conservation of cys–cys trp structural triads and their geometry in the protein domains of immunoglobulin superfamily members. Mol Immunol 1999; 36:373–86.
[50] Davis RS, Dennis Jr G, Odom MR, Gibson AW, Kimberly RP, Burrows PD, Cooper MD. Fc receptor homologs: newest members of a remarkably diverse Fc receptor gene family. Immunol Rev 2002;190:123–36.
[51] Guethlein LA, Flodin LR, Adams EJ, Parham P. NK cell receptors of the orangutan (*Pongo pygmaeus*): a pivotal species for tracking the coevolution of killer cell Ig-like receptors with MHC-C. J Immunol 2002;169:220–9.
[52] Guselnikov SV, Ershova SA, Mechetina LV, Najakshin AM, Volkova OY, Alabyev BY, Taranin AV. A family of highly diverse human and mouse genes structurally links leukocyte FcR, gp42 and PECAM-1. Immunogenetics 2002;54:87–95.
[53] Bertrand G, Duprat E, Lefranc M-P, Marti J, Coste J. Human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. Tissue Antigens 2004;64: 119–31.
[54] Jones EY. The immunoglobulin superfamily. Curr Opin Struct Biol 1993;3:846–52.
[55] Harpaz Y, Chothia C. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J Mol Biol 1994;238: 528–39.
[56] Smith DK, Xue H. Sequence profiles of immunoglobulin and immunoglobulin-like domains. J Mol Biol 1997;274:530–45.
[57] Hunkapiller T, Hood L. Diversity of the immunoglobulin gene superfamily. Adv Immunol 1989;44:1–63.
[58] Casasnovas JM, Stehle T, Liu JH, Wang JH, Springer TA. A dimeric crystal structure for the N-terminal two domains of intercellular adhesion molecule-1. Proc Natl Acad Sci USA 1998;95:4134–9.
[59] Garcia KC, Degano M, Stanfield RL, Brunmark A, Jackson MR, Peterson PA, Teyton L, Wilson IA. An alphabeta T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. Science 1996;274:209–19.
[60] Holden HM, Ito M, Hartshorne DJ, Rayment I. X-ray structure determination of telokin, the C-terminal domain of myosin light chain kinase, at 2.8 Å resolution. J Mol Biol 1992;227: 840–51.
[61] Sun P, Boyington J. Overview of protein folds in the immune system In: Curr protocols immunology. New York, NY: Wiley; 2000, A.1N.1–A.1N.45.
[62] Giudicelli V, Protat C, Lefranc M-P. The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: Proceeding of the European Conference on Computational Biology (ECCB'2003), Paris, France; 2003 INRIA (DISC, Spid) DKB-31,103–104. http://www.inra.fr/eccb2003/posters/pdf/Annot_Giudicelli_20030528_160703.pdf.

Publication 2

Available online at www.sciencedirect.com
SCIENCE DIRECT
Developmental and Comparative Immunology 29 (2005) 917–938
Developmental & Comparative Immunology
www.elsevier.com/locate/devcompimm

IMGT Locus in focus

IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN

Marie-Paule Lefranc*, Elodie Duprat, Quentin Kaas, Madeleine Tranne, Aude Thiriot, Gérard Lefranc

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

Received 8 February 2005; accepted 10 March 2005
Available online 27 April 2005

Abstract

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) provides a common access to expertly annotated data on the genome, proteome, genetics and structure of immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC), and related proteins of the immune system (RPI) of human and other vertebrates. The NUMEROTATION concept of IMGT-ONTOLOGY has allowed to define a unique numbering for the variable domains (V-DOMAINs) and constant domains (C-DOMAINs) of the IG and TR, which has been extended to the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily (IgSF) proteins other than the IG and TR (*Dev Comp Immunol* 27:55–77, 2003; 29:185–203, 2005). In this paper, we describe the IMGT unique numbering for the groove domains (G-DOMAINs) of the MHC and for the G-LIKE-DOMAINs of the MHC superfamily (MhcSF) proteins other than MHC. This IMGT unique numbering leads, for the first time, to the standardized description of the mutations, allelic polymorphisms, two-dimensional (2D) representations and three-dimensional (3D) structures of the G-DOMAINs and G-LIKE-DOMAINs in any species, and therefore, is highly valuable for their comparative, structural, functional and evolutionary studies.

Keywords: IMGT; Colliers de Perles; Groove domain; MHC; Major histocompatibility complex; MhcSF; Superfamily; Immune system

1. Introduction

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [1] is a high quality integrated knowledge resource specialized in the immunoglobulins (IG), T cell receptors (TR) and major histocompatibility complex (MHC) of human and other vertebrates, the immunoglobulin superfamily (IgSF) and MHC superfamily (MhcSF), and the related proteins of the immune system (RPI)

Abbreviations 2D, two-dimensional; 3D, three-dimensional; MHC, major histocompatibility complex; MhcSF, MHC superfamily; IG, immunoglobulin; TR, T cell receptor; IgSF, immunoglobulin superfamily; IMGT, the international ImMunoGeneTics information system®; RPI, related proteins of the immune system.

* Corresponding author. Tel.: +33 4 9961 9965; fax: +33 4 9961 9901.
E-mail address: lefranc@ligm.igh.cnrs.fr (M.-P. Lefranc).
URL: http://imgt.cines.fr.

0145-305X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dci.2005.03.003

[1–16]. IMGT provides a common access to expertly annotated data on the genome, proteome, genetics and structure of the IG, TR, MHC, IgSF, MhcSF and RPI, according to the IMGT Scientific chart rules and to the IMGT-ONTOLOGY concepts [17]. More particularly, the IMGT unique numbering [18–21], based on the NUMEROTATION concept of IMGT-ONTOLOGY, has been set up to provide a standardized description of mutations, allelic polymorphisms, two-dimensional (2D) and three-dimensional (3D) structure representations of the IG and TR variable domains (V-DOMAINs), and constant domains (C-DOMAINs) whatever the antigen receptor, the chain type or the species [20,21]. The IMGT unique numbering for V-DOMAINs and C-DOMAINs is used in all the IMGT components [1,6], and more particularly in the databases (IMGT/LIGM-DB [22], IMGT/PRIMER-DB [23], IMGT/GENE-DB [24], IMGT/3Dstructure-DB [25]), in the tools for sequence and structure analysis (IMGT/V-QUEST [26], IMGT/JunctionAnalysis [27], IMGT/Allele-Align, IMGT/PhyloGene [28], IMGT/StructuralQuery [25]), and in the IMGT Repertoire Web resources ('IMGT Protein displays' [29,30], 'IMGT Colliers de Perles' 2D representations [31], and 'IMGT Alignments of Alleles' [32,33]; see http://imgt.cines.fr). Interestingly, the IMGT unique numbering for V-DOMAIN and for C-DOMAIN has been fully extended to the V-LIKE-DOMAINs and C-LIKE-DOMAINs of IgSF proteins other than the IG and TR [20,21,34,35]. This is particularly remarkable for the V-DOMAINs and V-LIKE-DOMAINs, the genomic structures of which are strikingly different. Indeed, the IG and TR V-DOMAINs are encoded by rearranged V-(D-)J genes [32,33], whereas the V-LIKE-DOMAINs are often encoded by a single exon [20,35].

In this paper, we define a standardized IMGT unique numbering for the MHC groove domains (G-DOMAINs) of the MHC of all jawed vertebrates. We show that this IMGT unique numbering for G-DOMAINs can be extended to the G-LIKE-DOMAINs of the MHC-like proteins (MhcSF proteins other than MHC), of any species. The IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN represents, therefore, a major step forward for the comparative analysis of the sequences and structures of these domains, and for the study of their evolution.

2. MHC chain and G-DOMAIN definition

The MHC proteins that present peptides to the T cells belong to the 'classical MHC class I' (MHC-Ia) or to the 'classical MHC class II' (MHC-IIa) [36]. The MHC-Ia comprises, in human, the HLA-A, HLA-B and HLA-C subclasses, and in mouse, the H2-D, H2-K and H2-L subclasses. The MHC-IIa comprises, in human, the HLA-DP, HLA-DQ and HLA-DR subclasses, and in mouse, the H2-A, H2-E and H2-P subclasses (H2-P being unproductive). The MHC proteins with more specific functions or which do not present peptides to the T cells belong to the 'nonclassical MHC-I' (MHC-Ib) or to the 'nonclassical MHC-II' (MHC-IIb). The MHC-Ib comprises, in human, the HLA-E, HLA-F and HLA-G subclasses (each one represented by a unique isotype), and, in mouse, the H2-Q, H2-M and H2-T subclasses (each one comprising several isotypes). The MHC-IIb comprises, in human, the HLA-DM and HLA-DO subclasses and, in mouse, the H2-DM and H2-DO subclasses (IMGT Repertoire MHC, http://imgt.cines.fr). Thus there are 11 MHC subclasses in human and in mouse (3 MHC-Ia, 3 MHC-IIa, 3 MHC-Ib, and 2 MHC-IIb).

The MHC-I proteins, expressed on the cell surface of most cells, are formed by the association of a transmembrane heavy chain (I-ALPHA chain) and a noncovalently linked light chain beta-2-microglobulin (B2M) (Fig. 1). The MHC-II proteins, expressed on the cell surface of professional antigen presenting cells (APC), are heterodimers formed by the association of two transmembrane chains, an alpha chain (II-ALPHA chain) and a beta chain (II-BETA chain) (Fig. 1).

The I-ALPHA chain of the MHC-I, and the II-ALPHA and II-BETA chains of the MHC-II proteins, comprise an extracellular region made of three domains for the MHC-I chain and of two domains for each MHC-II chain, a connecting region, a transmembrane region and an intracytoplamic region (Fig. 2). The I-ALPHA chain comprises two groove domains (G-DOMAINs), the G-ALPHA1 [D1] and G-ALPHA2 [D2] domains, and one C-LIKE domain [D3] [21,36]. The II-ALPHA chain and the II-BETA chain each comprises two domains, the G-ALPHA [D1] and C-LIKE [D2] domains, and the G-BETA [D1] and C-LIKE [D2] domains, respectively (Fig. 2). The four

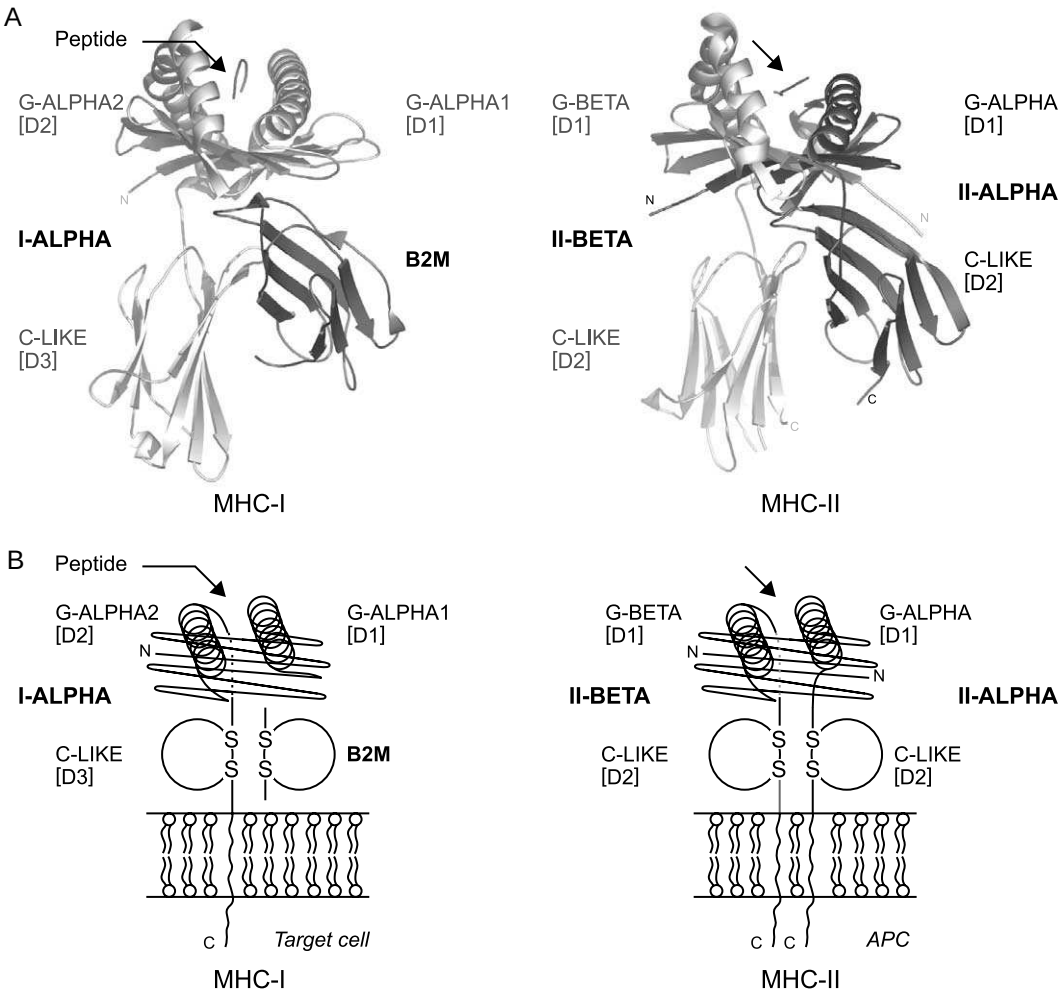


Fig. 1. 3D structures and schematic representations of the MHC class I (MHC-I) and MHC class II (MHC-II) proteins. (A) 3D structures of MHC-I and MHC-II (1oga and 1j8 h annotated coordinate files from IMGT/3Dstructure-DB [25], http://imgt.cines.fr, that include crystallographic data from the Protein DataBank PDB [37]). The MHC-I comprises the I-ALPHA and the beta-2-microglobulin (B2M) chains. The I-ALPHA chain is shown with its extracellular domains (G-ALPHA1, G-ALPHA2 and C-LIKE) [36]. The MHC-II comprises the II-ALPHA and II-BETA chains that are shown with their extracellular domains (G-ALPHA and C-LIKE for the II-ALPHA chain, G-BETA and C-LIKE for the II-BETA chain). (B) Schematic representations of the MHC-I and MHC-II proteins. The MHC-I and MHC-II are shown as transmembrane proteins, at the surface of a target cell and of an antigen presenting cell (APC), respectively. Complete MHC-I and MHC-II chains comprise the extracellular domains (shown in A) and the connecting, transmembrane and cytoplamic regions (not present in 3D structures [36], for details see Fig. 2). [D1], [D2] and [D3] indicate the position of the domains from the N-terminal end of the chains. Arrows indicate the peptide localization in the MHC groove (the N-terminal end of the peptide is in the back).

G-DOMAINs, G-ALPHA1 and G-ALPHA2 of the MHC-I proteins, and G-ALPHA and G-BETA of the MHC-II proteins have a similar groove 3D structure that consists of one sheet of four antiparallel beta strands (‘floor’ of the groove or platform) and one long helical region (‘wall’ of the groove). This groove is part of the cleft that is the peptide binding site of the classical MHC-Ia and MHC-IIa proteins [36] (Fig. 1).

Owing to the conserved structure between classical and nonclassical MHC, and according to the DESCRIPTION concept of IMGT-ONTOLOGY [5,6,17], the same labels are used, for MHC-Ia

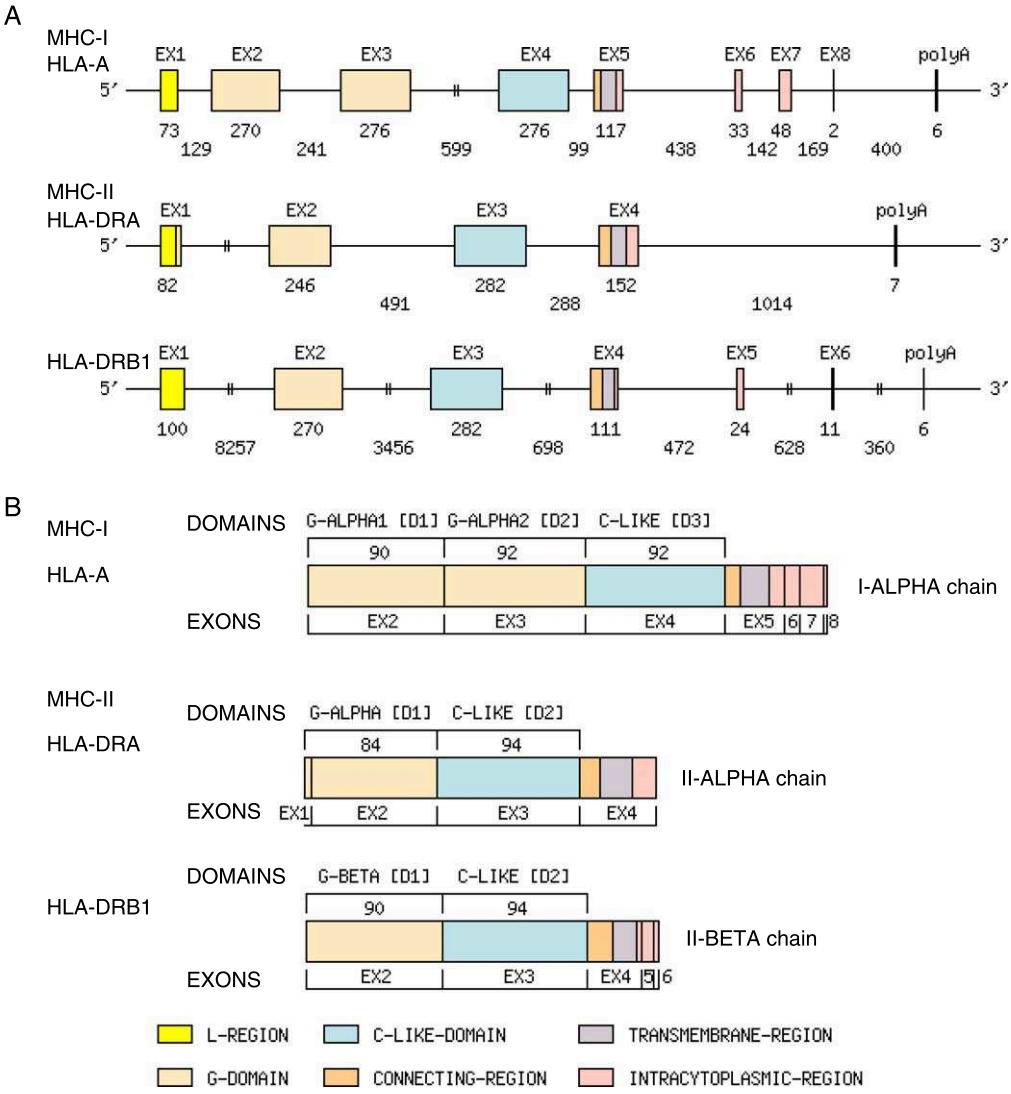


Fig. 2. Correspondence between exons and domains for the MHC-I and MHC-II. (A) Exons of the *Homo sapiens* MHC-I HLA-A gene and MHC-II HLA-DRA and HLA-DRB1 genes, shown as examples. Lengths of the exons, introns and polyA signal are in base pairs. Introns indicated with ‖ are not at scale. (B) Domains of the *Homo sapiens* MHC-I HLA-A (I-ALPHA) chain and of the MHC-II HLA-DRA (II-ALPHA) and HLA-DRB1 (II-BETA) chains, shown as examples. Lengths of the domains are in number of amino acids. The G-ALPHA domain of HLA-DRA (84 amino acids) is encoded by EX2 (82 codons) and the 3′ end of EX1 (2 codons). EMBL/GenBank/DDBJ accession numbers: HLA-A (K02883), HLA-DRA (J00203 and J00204) and HLA-DRB1 (AL137064) (Table 1).

and MHC-Ib, in the description of their heavy chain and domains: I-ALPHA chain, and G-ALPHA1 and G-ALPHA2 domains. Similarly, the same labels are used, for MHC-IIa and MHC-IIb, in the description of their respective alpha and beta chains and domains: II-ALPHA and II-BETA chains, and G-ALPHA and G-BETA domains.

3. IMGT unique numbering for G-DOMAIN

Correspondence between the four G-DOMAINs was established by extensive sequence alignment comparison of annotated MHC chains from the IMGT Repertoire [1,6] and by structural data analysis and alignment of MHC proteins with known 3D structures

from IMGT/3Dstructure-DB, http://imgt.cines.fr [25]. As each G-DOMAIN is usually encoded by a single exon, the delimitation of the domains in IMGT takes into account the limits of the exons in the genomic structure of the MHC genes (Fig. 2). In Table 1 are indicated the EMBL/GenBank/DDBJ [38–40] accession numbers of the sequences whose domains are reported in the IMGT Protein display (Fig. 3), the IMGT/3Dstructure-DB [25] entries of the representative alleles, as well as the accession numbers of genomic sequences from other alleles that were necessary to identify the splicing sites.

The IMGT Protein display (Fig. 3), based on the IMGT unique numbering for G-DOMAIN, shows, for the first time, a standardized amino acid alignment of G-DOMAINs that belong to the same or to different chain types, from the classical and nonclassical MHC-I and MHC-II, and from different species. Indeed, this IMGT Protein display includes (i) the G-ALPHA1 [D1] domains of MHC-Ia (human HLA-A*0201, HLA-B*0702, HLA-Cw*0701 alleles found in a frequent haplotype in caucasian populations, and mouse H2-D1*02, H2-K1*01, H2-L*02), (ii) the G-ALPHA1 [D1] domains of MHC-Ib (human HLA-E*01, HLA-F*01, HLA-G*01, and mouse H2-M5*02, H2-Q7*02, H2-T3*01), (iii) the G-ALPHA [D1] domains of MHC-IIa (human HLA-DPA1*0103, HLA-DQA1*0501, HLA-DRA*0101 alleles found in a frequent haplotype in caucasian populations, and mouse H2-AA*02, H2-EA*02) (iv) the G-ALPHA [D1] domains of MHC-IIb (HLA-DMA*01, HLA-DOA*01, and mouse H2-DMA*01, H2-DOA*01) (Fig. 3A), (v) the G-ALPHA2 [D2] domains of MHC-Ia and MHC-Ib (same chains as described above for the G-ALPHA1 [D1] domains), (vi) the G-BETA [D1] domains of MHC-IIa (human HLA-DPB1*0401, HLA-DQB1*0301, HLA-DRB1*1402, and mouse H2-AB*02, H2-EB1*01), and (vii) the G-BETA [D1] domains of MHC-IIb (human HLA-DMB*01, HLA-DOB*01), and mouse H2-DMB1*02, H2-DOB*01) (Fig. 3B).

For each G-DOMAIN, the positions that contribute to the groove floor comprise positions 1–49, with the A strand from positions 1–14, the AB turn positions 15–17, the B strand positions 18–28, the BC turn positions 29 and 30, the C strand positions 31–38, the CD turn positions 39–41 and the D strand positions 42–49 (Fig. 3 and Table 2). The additional position 7A represents a bulge in 3D structures and is present in some G-ALPHA domains, for instance those of the HLA-DQA1, H2-AA, HLA-DOA and H2-DOA chains. This position 7A is added if G-ALPHA sequences are introduced in G-DOMAIN alignments (Fig. 3). The gaps of the floor are localized in the turns. The AB turn (positions 15–17) comprises three amino acids in the G-ALPHA1, G-ALPHA2 and G-BETA domains but these positions are unoccupied in the G-ALPHA domains (as well as position 18 of strand B). The BC turn (positions 29 and 30) comprises two positions that are occupied in all G-DOMAINs. The CD turn (positions 39–41) is occupied by three amino acids in the G-ALPHA1 domains and only one in the other domains (G-ALPHA, G-ALPHA2 and G-BETA) (Fig. 3). Additional positions at the N-terminus of strand A or at the C-terminus of strand D can be added if necessary. Thus, two additional positions (1.2 and 1.1) are added at the N-terminus of the A strand of the G-ALPHA domains as the presence of these two amino acids was demonstrated by protein sequencing of the HLA-DRA [42] (Fig. 3A, Table 2A). In each [D1] G-DOMAIN, except the G-ALPHA domain of HLA-DMA and H2-DMA, the amino acid at position 1 (shown within parentheses in Fig. 3) is encoded by the codon that results from the splicing between the first exon (EX1) that encodes the L-REGION, and the second exon (EX2) that encodes [D1]. The two amino acids, isoleucine (I) and lysine (K) at positions 1.2 and 1.1 of HLA-DRA G-ALPHA [D1], are therefore encoded by EX1. By extrapolation, two amino acids have been added at positions 1.2 and 1.1 for the other G-ALPHA domains, but in those cases, the proteolytic cleavage site of the leader peptide (L-REGION) needs to be confirmed experimentally. It is also necessary to confirm if the amino acids at positions 1.10–1.1 of HLA-DMA and H2-DMA belong, or not, to the mature protein (Fig. 3A, Table 2A). Four additional positions (49.1–49.4) are observed at the C-terminus of strand D of the G-BETA domain of HLA-DMB and H2-DMB1 (Fig. 3B, Table 2B).

The numbering of the alpha helix starts at position 50 and ends at position 92, with five additional positions at 54A, 61A, 61B, 72A and 92A. Three of them (61A, 61B, 72A) characterize the G-ALPHA2 and/or G-BETA domains (Fig. 3B, Table 2B). Indeed, positions 61A and 72A are occupied in

Table 1
Representative MHC and MHC-I-like genes and chains

A	Species	IMGT gene and allele names[a]	gDNA[b]	Mouse strain	Mouse H2 haplotype[c]	cDNA[b]	IMGT/3D structure-DB[d]	gDNA EX1 (same allele)[b,e]	gDNA (other alleles)[b,e]
MHC-Ia	*Homo sapiens*	HLA-A*0201	K02883				1oga_A		
		HLA-B*0702	AJ292075						
		HLA-Cw*0701	Y18533					Y18499	
			Y18534						
	Mus musculus	H2-D1*02	M18523	C57BL/10	b		1juf_A		
		H2-K1*01	V00746	C57BL/10	b		1lk2_A		
		H2-L*02	L00127	BALB/c	d		(1ld9_A)		
MHC-Ib	*Homo sapiens*	HLA-E*01	AF523277				1mhe_A		
		HLA-F*01	X17093						
		HLA-G*01	J03027						
	Mus musculus	H2-M5*02	L14279	BALB/c	d				
		H2-Q7*02	X03210						
		H2-T3*01	M11742	C57BL/6	b				
MHC-I-like	*Homo sapiens*	MICA*01				U56940	1hyr_C		L29411
		MR1*01	AL356267						
		RAET1N*01	AL355497				1kcg_C		
	Mus musculus	AZGP1*01	AF281658	129/Svj	129				
		CD1D1*01	X13170				1cd1_A		
		FCGRT*01	D37872	BALB/c	d				

B		IMGT gene and allele names[a]	gDNA[b]	Mouse strain	Mouse H2 haplotype[c]	cDNA[b]	Mouse strain	Mouse H2 haplotype[c]	IMGT/3D structure-DB[d]	gDNA EX1 (same allele)[b,e]	gDNA (other alleles)[b,e]
MHC-IIa	*Hs*	HLA-DPA1*0103	X03100								
		HLA-DPB1*0401	M23907							M23906	
		HLA-DQA1*0501	Z84489						1s9v_A		
		HLA-DQB1*0301				M25325					U92032
		HLA-DRA*0101	J00204						1fv1_D	J00203	Z84814
		HLA-DRB1*1402				AJ297583					AL137064
	Mm	H2-AA*02	AY740451	B10.MOL1	(w12)	V00832	B10A	k/d2?	1iak_A		AF027865
		H2-AB*02	AY740477	B10.SNA70	(w8)	M13538	B10A	k/d2?	1iak_B		AF027865
		H2-EA*02	K00971	BALB/c	d				1fng_C		
			AY303782	BALB.K	k						
		H2-EB1*01	AF050157	129	bc						
MHC-IIb	*Hs*	HLA-DMA*01				X62744			(1 hdm_A)		X76775
		HLA-DMB*01	X76776						(1 hdm_B)		
		HLA-DOA*01	X02882								

	HLA-DOB*01	X87344				1k8i_A	
Mm	H2-DMA*01	AF100956	129/Svj	129			
	H2-DMB1*02	U35323	BALB/c	d			
	H2-DOA*01		C57BL/6J	b	AK020594		AF100956
	H2-DOB*01		C57BL/6J	b	AK053233		M11800

(A) Classical MHC-I (MHC-Ia), nonclassical MHC-I (MHC-Ib), and MHC-I-like. (B) Classical MHC-II (MHC-IIa) and nonclassical MHC-II (MHC-IIb).

[a] Owing to a lesser degree of polymorphism of the nonclassical *Homo sapiens* (*Hs*) MHC-Ib and MHC-IIb genes, compared to the classical MHC-Ia and MHC-IIa genes, a 2-digit is used for their allele description (IMGT Scientific chart, http://imgt.cines.fr): *01 refers to *0101 found in the literature. The allele nomenclature of the highly polymorphic *H. sapiens* classical MHC-Ia and MHC-IIa is according to HLA-DB [13]. *Mus musculus* (*Mm*) H2 alleles are numbered starting from allele *01 (sequence from strain C57BL/6). *H. sapiens* and *M. musculus* MHC-I-like alleles are numbered starting from allele *01 that corresponds to the IMGT reference sequence (IMGT Repertoire, http://imgt.cines.fr). The H2-PA gene (H2-PA*01 accession number D64112, strain C57BL/6) that is a pseudogene, and the H2-PB gene that has not yet been identified, are not included in the table.

[b] EMBL/GenBank/DDBJ nucleotide sequence accession numbers. gDNA: genomic DNA; cDNA: complementary DNA.

[c] From 'Mouse H2 haplotypes and polymorphisms' (IMGT Index > Strain, and IMGT Repertoire for MHC, http://imgt.cines.fr [1,6].

[d] Chain ID from IMGT/3Dstructure-DB, http://imgt.cines.fr [25]. The chain ID is shown within parentheses for 3D structures in which the chain sequence differs from the translation of the corresponding representative allele in column 3.

[e] Accession numbers that were necessary for splicing site identification: 'same allele' refers to the allele in column 3, 'other alleles' refers to alleles other than the ones in column 3.

the G-ALPHA2 domain, whereas positions 61A, 61B and 72A are occupied in G-BETA domains, at the exception of H2-AB (Table 2B). The position 92A is only occupied in the HLA-DMA and H2-DMA G-ALPHA domains (Fig. 3A, Table 2A). It is worthwhile to note that position 54A is the only additional position needed to extend the IMGT numbering for G-DOMAINs to the G-LIKE-DOMAINs of the MHC-I-like proteins (described in next paragraph and shown in Fig. 3A).

The helix (positions 50–92) seats on the beta sheet and its axis forms an angle of about 40° with the beta strands. The helix is split into two parts separated by a kink, positions 58 of G-ALPHA1, 61 of G-ALPHA2, 63 of G-ALPHA, and 62 of G-BETA being the 'highest' points on the groove floor [36].

Two cysteine, CYS-11 (in strand A) and CYS-74 (in the helix) are well conserved in the G-ALPHA2 and G-BETA domains where they participate to a disulfide bridge that fastens the helix on the groove floor (Fig. 3B). The G-ALPHA1 and G-ALPHA domains have a conserved *N*-glycosylation site at position 86 (N–X–S/T, where N is asparagine, X any amino acid except proline, S is serine and T is threonine) (Fig. 3A). A *N*-glycosylation site is also found at that position in the G-ALPHA2 domain of the mouse MHC-Ia chains (H2-D1, H2-K1 and H2-L) (Fig. 3B). The G-BETA domains (except for the HLA-DMB and H2-DMB1 chains) have a conserved potential *N*-glycosylation site at position 15 (AB turn) (Fig. 3B). Interestingly, the G-ALPHA domains of the HLA-DMA and H2-DMA chains have specific features compared to the other G-ALPHA domains and share common characteristics with the G-ALPHA2 and G-BETA domains: there is a conserved CYS-11_CYS-74 disulfide bridge, positions 61A and 61B are occupied (as in the G-BETA domains) and there is no *N*-glycosylation site at position 86 (Fig. 3).

4. IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN and sequence data analysis

The IMGT unique numbering for G-DOMAIN allows, for the first time, a standardized comparison of the amino acid (and corresponding codons) changes between the different groove domains

that belong to a same chain (G-ALPHA1 and G-ALPHA2), or to different MHC chains, and this whatever the species. Practically, the IMGT unique numbering for positions 1–39 and 73–92 of the G-ALPHA2 domains can be obtained very easily by substracting 90 from the mature protein numbering (91–129 and 163–182) (Table 3). Between these positions, the two gaps (at positions 40 and 41) and the two insertions (at positions 61A and 72A) are necessary, in the IMGT unique

```
A
                                 A                B            C           D                        helix
                           1        10  14     18       28    31     38    42 45  49     50      60         70         80        90
                        321|......A..|...| ... |.........| .. |......| ... |..|...|12345 |....A.....|.AB........|..A.......|.........|..A
G-ALPHA1 [D1]
MHC-Ia
HLA-A*0201    Hs     (G)SHSMRY.FFTSVSR PGR GEPRFIAVGYV DD TQFVRFDS DAA SQRMEPRA..... PWIEQ.EGPEYWD..GETRKVKAHSQ.THRVDLGTLRGYYNQSEA...
HLA-B*0702    Hs     (G)SHSMRY.FYTSVSR PGR GEPRFISVGYV DD TQFVRFDS DAA SPREEPRA..... PWIEQ.EGPEYWD..RNTQIYKAQAQ.TDRESLRNLRGYYNQSEA...
HLA-Cw*0701   Hs     (C)SHSMRY.FDTAVSR PGR GEPRFISVGYV DD TQFVRFDS DAA SPRGEPRA..... PWVEQ.EGPEYWD..RETQNYKRQAQ.ADRVSLRNLRGYYNQSED...

H2-D1*02      Mm     (G)PHSMRY.FETAVSR PGL EEPRYISVGYV DN KEFVRFDS DAE NPRYEPRA..... PWMEQ.EGPEYWE..RETQKAKGQEQ.WFRVSLRNLLGYYNQSAG...
H2-K1*01      Mm     (G)PHSLRY.FVTAVSR PGL GEPRYMEVGYV DD TEFVRFDS DAE NPRYEPRA..... RWMEQ.EGPEYWE..RETQKAKGNEQ.SFRVDLRTLLGYYNQSKG...
H2-L*02       Mm     (G)PHSMRY.FETAVSR PGL GEPRYISVGYV DN KEFVRFDS DAE NPRYEPQA..... PWMEQ.EGPEYWE..RITQIAKGQEQ.WFRVNLRTLLGYYNQSAG...
MHC-Ib
HLA-E*01      Hs     (G)SHSLKY.FHTSVSR PGR GEPRFISVGYV DD TQFVRFDN DAA SPRMVPRA..... PWMEQ.EGSEYWD..RETRSARDTAQ.IFRVNLRTLRGYYNQSEA...
HLA-F*01      Hs     (G)SHSLRY.FSTAVSR PGR GEPRYIAVEYV DD TQFLRFDS DAA IPRMEPRE..... PWVEQ.EGPQYWE..WTTGYAKANAQ.TDRVALRNLLRRYNQSEA...
HLA-G*01      Hs     (G)SHSMRY.FSAAVSR PGR GEPRFIAMGYV DD TQFVRFDS DSA CPRMEPRA..... PWVEQ.EGPEYWE..EETRNTKAHAQ.TDRMNLQTLRGYYNQSEA...

H2-M5*02      Mm     (G)IHSLQF.FATTMTQ PGL REHSFIFVVFV DA TQFLCYNN KGK NQRMEPRP..... LWVKQ.MGPEYWE..QQTRTVKVIEK.IALVNLQEAMDIYNHSKD...
H2-Q7*02      Mm     (G)QHSLQY.FHTAVSR PGL GEPWFISVGYV DD TQFVRFDS DAE NPRMEPRA..... RWMEQ.EGPEYWE..RETQIAKGHEQ.SFRGSLRTAQSYYNQSKG...
H2-T3*01      Mm     (G)SHSLRY.FYTALSR PAI SEPWYIAVGYL DD TQFVRFNS SGE TATYKLSA..... PWVEQ.EGPEYWA..RETEIVTSNAQ.FFRENLQTMLDYYNLSQN...
G-ALPHA1-LIKE [D1]
MHC-I-like
MICA*01       Hs     (E)PHSLRY.NLTVLSW DGS VQSGFLTEVHL DG QPFLRCDR Q.. KCRAKPQG..... QWAEDVLGNKTWD..RETRDLTGNGK.DLRMTLAHIKDQKE.......
MR1*01        Hs     (R)THSLRY.FRLGVSD PIH GVPEFISVGYV DS HPITTYDS V.. TRQKEPRA..... PWMAENLAPDHWE..RYTQLLRGWQQ.MFKVELKRLQRHYNHS.....
RAET1N*01     Hs     (D)AHSLWY.NFTIIHL PRH GQQWCEVQSQV DQ KNFLSYDC G.. SDKVLSMG..... HLEEQLYATDAWG..KQLEMLREVGQ.RLRLELADTELEDFTPS....

AZGP1*01      Mm     (G)SYYLTF.LYTGLSR PSK GFPRFQATAFL ND QAFFHYNS N.. SGKAEPVG..... PWSQV.EGMEDWE..KESQLQRAREE.IFLVTLKDIMDYYKDTT....
CD1D1*01      Mm (A)QQKNYTFRC.LQMSSFA NR. SWSRTDSVVWL GD LQTHRWSN D.. SATISFTK..... PWSQGKLSNQQWE..KLQHMFQVYRV.SFTRDIQELVKMMSPKED...
FCGRT*01      Mm    (E)TRPPLMY.HLTAVSN PST GLPSFWATGWL GP QQYLTYNS L.. RQEADPCG..... AWMWENQVSWYWE..KETTDLKSKEQ.LFLEALKTLEKILN.......
G-ALPHA [D1]
MHC-IIa
HLA-DPA1*0103 Hs   IK(A)DHVSTY.AAFVQTH ... .RPTGEFMFEF DE DEMFYVDL D.. KKETVWHL..... EEFGQ.....AFS..FEAQGGLANIA.ILNNNLNTLIQRSNHTQATN.
HLA-DQA1*0501 Hs   IV(A)DHVASYGVNLYQSY ... .GPSGQYTHEF DG DEQFYVDL G.. RKETVWCL..... PVLRQ.....FR...FDPQFALTNIA.VLKHNLNSLIKRSNSTAATN.
HLA-DRA*0101  Hs   IK(E)EHVIIQ.AEFYLNP ... .DQSGEFMFDF DG DEIFHVDM A.. KKETVWRL..... EEFGR.....FAS..FEAQGALANIA.VDKANLEIMTKRSNYTPITN.

H2-AA*02      Mm   IE(A)DHVGSYGITVYQSP ... .GDIGQYTFEF DG DELFYVDL D.. KKETVWML..... PEFAQ.....LRR..FEPQGGLQNIA.TGKHNLEILTKRSNSTPATN.
H2-EA*02      Mm   IK(E)EHTIIQ.AEFYLLP ... .DKRGEFMFDF DG DEIFHVDI E.. KSETIWRL..... EEFAK.....FAS..FEAQGALANIA.VDKANLDVMKERSNNTPDAN.
MHC-IIb
HLA-DMA*01   (A)PTPMWPDDLQNHTFLH.TVYCQDG ... .SPSVGLSEAY DE DQLFFFDF S.. QNTRVPRL..... PEFAD.....WAQEQGDAPAILFDKE.FCEWMIQQIGPKLDGKIPVSR
HLA-DOA*01         Hs TK(A)DHMGSYGPAFYQSY ... .GASGQFTHEF DE EQLFSVDL K.. KSEAVWRL..... PEFGD.....FAR..FDPQGGLAGIA.AIKAHLDILVERSNRSRAIN.

H2-DMA*01    (A)STPVFWDDPQNHTFRH.TLFCQDG ... .IPNIGLSETY DE DELFSFDF S.. QNTRVPRL..... PDFAE.....WAQGQGDASAIAFDKS.FCEMLMREVSPKLEGQIPVSR
H2-DOA*01          Mm IK(A)DHMGSYGPAFYQSY ... .DASGQFTHEF DG EQIFSVDL K.. NEEVVWRL..... PEFGD.....FAH..SDFQSGLMSIS.MIKAHLDILVERSNRTRAVS.
```

Fig. 3. IMGT Protein display of G-DOMAINs and G-LIKE-DOMAINs of representative MHC and MHC-I-like chains. (A) G-ALPHA1 [D1], G-ALPHA1-LIKE [D1] and G-ALPHA [D1] domains from classical (MHC-Ia) and nonclassical (MHC-Ib) MHC-I, from MHC-I-like, and from classical (MHC-IIa) and nonclassical (MHC-IIb) MHC-II, respectively. (B) G-ALPHA2 [D2], G-ALPHA2-LIKE [D2] and G-BETA [D1] domains from classical (MHC-Ia) and nonclassical (MHC-Ib) MHC-I, from MHC-I-like, and from classical (MHC-IIa) and nonclassical (MHC-IIb) MHC-II, respectively. [D1], [D2] and [D3] indicate the position of the domains from the N-terminal end of the chains. Membrane proteins quoted in this figure are of type I, that is with the chain N-terminal end being extracellular. Sequences are from *Homo sapiens* (*Hs*) and from *Mus musculus* (*Mm*). The IMGT Protein display is according to the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN, based on the NUMEROTATION concept of IMGT-ONTOLOGY [17]. The G-DOMAINs and G-LIKE-DOMAINs are designated with the IMGT labels (IMGT Scientific chart, http://imgt.cines.fr). Beta strands are shown by horizontal arrows. Dots indicate missing amino acids according to the IMGT unique numbering. Amino acids resulting from a splicing with a preceding exon are shown within parentheses. EMBL/GenBank/DDBJ accession numbers are reported in Table 1. Potential *N*-glycosylation sites (N–X–S/T) are underlined. Note that the C-LIKE-DOMAIN [D2] of *M. musculus* H2-AA (NT_039649, H2-AA*01), *H. sapiens* HLA-DMA (NT_007592, HLA-DMA*01) and *M. musculus* H2-AB (NT_039649, H2-AB*01) were reported in Fig. 3 of Ref. [21]. Gene names (symbols) are according to the IMGT Nomenclature committee (IMGT-NC) [1] and to the HUGO Nomenclature Committee (HGNC) [41]. Full gene designations for the MHC genes are based on the examples shown within parentheses: human MHC-I (HLA-A: Major histocompatibility complex, class I, A), mouse MHC-I (H2-K1: histocompatibility 2, class I, K1), human MHC-II (HLA-DPA1: MHC class II, DP alpha1; HLA-DPB1: MHC class II, DP beta1), and mouse MHC-II (H2-AA: histocompatibility 2, class II, A alpha; H2-AB: histocompatibility 2, class II, A beta). Full gene designations for the MHC-I-like genes are the following: MICA, MHC class I polypeptide-related sequence A; MR1: major histocompatibility complex, class I-related; RAET1N: retinoic acid early transcript 1 N, UL16 binding protein 3; CD1D1: CD1D antigen, polypeptide 1; FCGRT: Fc fragment of IgG, receptor, transporter, alpha; AZGP1: alpha-2-glycoprotein 1, zinc. Amino acid one-letter abbreviation: A (Ala), alanine; C (Cys), cysteine; D (Asp), aspartic acid; E (Glu), glutamic acid; F (Phe), phenylalanine; G (Gly), glycine; H (His), histidine; I (Ileu), isoleucine; K (Lys), lysine; L (Leu), leucine; M (Met), methionine; N (Asn), asparagine; P (Pro), proline; Q (Gln), glutamine; R (Arg), arginine; S (Ser), serine; T (Thr), threonine; V (Val), valine; W (Trp), tryptophan; Y (Tyr), tyrosine.

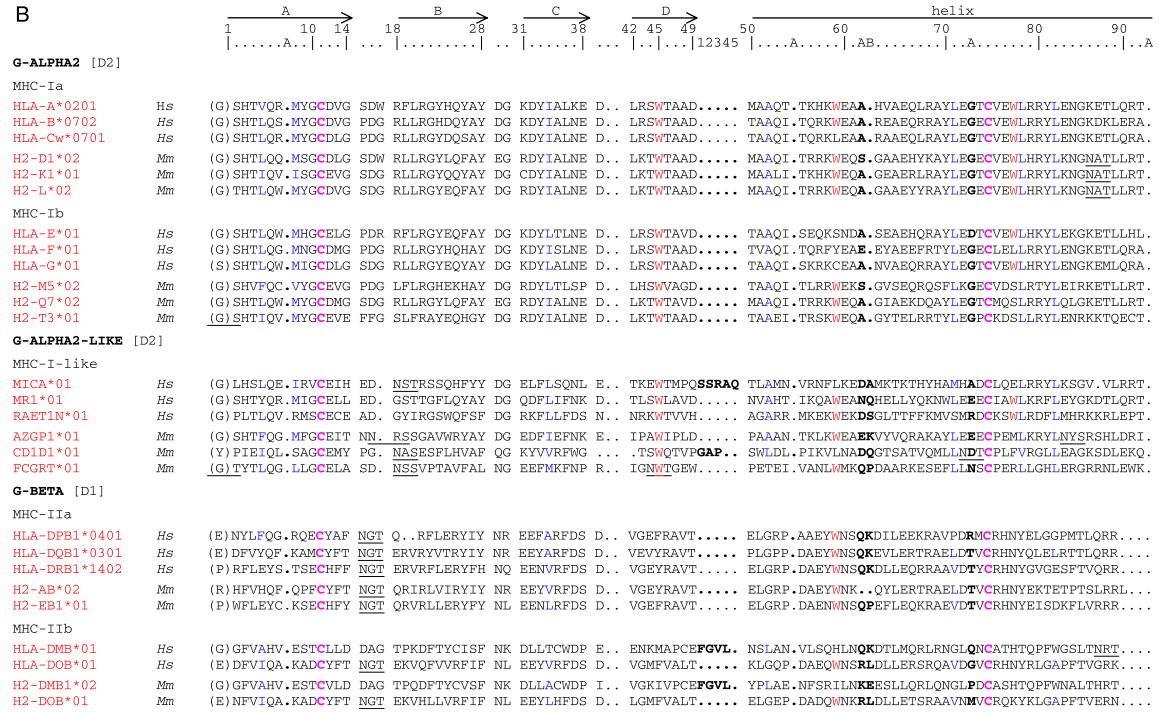


Fig. 3 (*continued*)

numbering, to allow meaningful sequence and structure alignment and comparison between the G-ALPHA1 and G-ALPHA2 sequences. Correspondence between the IMGT unique numbering for G-DOMAIN and different MHC chain numberings is shown in Table 3. Examples include the G-ALPHA1 [D1] and G-ALPHA2 [D2] domains of the classical MHC-I (MHC-Ia) HLA-A, the G-ALPHA [D1] and G-BETA [D1] domains of the classical MHC-II (MHC-IIa) HLA-DRA and HLA-DRB1, and the G-ALPHA [D1] and G-BETA [D1] domains of the nonclassical MHC-II (MHC-IIb) HLA-DMA and HLA-DMB.

The IMGT unique numbering for G-DOMAIN has been extended to the G-LIKE-DOMAINs of MhcSF proteins other than MHC. So far, only MHC-I-like proteins have been identified in the MhcSF [43–46]. The examples of MHC-I-like chains shown in Fig. 3 include the human MICA*01, MR1*01 and RAET1N*01, and the mouse AZGP1*01, CD1D1*01 and FCGRT*01. The G-ALPHA1-LIKE [D1] and G-ALPHA2-LIKE [D2] domains of these proteins show a striking structural homology with the MHC G-ALPHA1 and G-ALPHA2 domains and this, despite a high sequence divergence [46]. The implementation of the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN represents therefore a major step in the standardization of the MhcSF amino acid sequence alignments (Fig. 3). As the nucleotide positions are derived from the codon numbering, the IMGT unique numbering allows a standardized allele description and the setting up of 'Tables of alleles' and 'Alignments of alleles' for MhcSF proteins, whatever the receptor, the chain or the species (IMGT Repertoire for MHC, IMGT Repertoire for RPI, http://imgt.cines.fr). Owing to that standardization, the sequence polymorphisms of any G-DOMAIN (of any MHC gene or chain, from any vertebrate species) and the sequence polymorphisms of any G-LIKE-DOMAIN (of any MHC-I-like gene or chain, from any species) can easily be described and analysed.

Table 2
Lengths of the G-DOMAIN and G-LIKE-DOMAIN labels, according to the IMGT unique numbering

A			G-ALPHA1 [D1]				G-ALPHA1-LIKE [D1]					
			MHC-Ia		MHC-Ib		MHC-I-like					
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-A HLA-B HLA-C	H2-D1 H2-K1 H2-L	HLA-E HLA-F HLA-G	H2-M5 H2-Q7 H2-T3	MICA	MR1	RAET1N	AZGP1	CD1D1	FCGRT
			Hs	*Mm*	*Hs*	*Mm*	*Hs*	*Hs*	*Hs*	*Mm*	*Mm*	*Mm*
A-STRAND	1.3-1.1[a]	+3									+3	+1
	1-14	14	14	14	14	14	14	14	14	14	14	14
	7A	+1										
AB-TURN	15–17	3	3	3	3	3	3	3	3	3	2	3
B-STRAND	18–28	11	11	11	11	11	11	11	11	11	11	11
BC-TURN	29–30	2	2	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	3	3	3	3	1	1	1	1	1	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8	8	8
	49.1–49.5	+5										
HELIX	50–92	43	41	41	41	41	37	39	40	40	41	37
	54A	+1					+1	+1	+1		+1	+1
	61A	+1										
	61B	+1										
	72A	+1										
	92A	+1										
Total length		92 (+14)	90	90	90	90	86 (+1)	88 (+1)	89 (+1)	89	90 (+4)	86 (+2)

A			G-ALPHA [D1]							
			MHC-IIa				MHC-IIb			
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-DPA1 HLA-DRA	HLA-DQA1	H2-AA	H2-EA	HLA-DMA	HLA-DOA	H2-DMA	H2-DOA
			Hs	*Hs*	*Mm*	*Mm*	*Hs*	*Hs*	*Mm*	*Mm*
A-STRAND	1.10–1.1[a]	+10	+2	+2	+2	+2	+10	+2	+10	+2
	1–14	14	14	14	14	14	14	14	14	14
	7A	+1		+1	+1			+1		+1
AB-TURN	15–17	3	0	0	0	0	0	0	0	0
B-STRAND	18–28	11	10	10	10	10	10	10	10	10
BC-TURN	29–30	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	1	1	1	1	1	1	1	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8
	49.1–49.5	+5								

HELIX	50–92	43	39	38	39	39	39	39	39	39
	54A	+1								
	61A	+1					+1		+1	
	61B	+1					+1		+1	
	72A	+1								
	92A	+1					+1		+1	
Total length		92 (+21)	82 (+2)	81 (+3)	82 (+3)	82 (+2)	82 (+13)	82 (+3)	82 (+13)	82 (+3)

B			G-ALPHA [D2]				G-ALPHA2-LIKE [D2]					
			MHC-Ia		MHC-Ib		MHC-I-like					
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-A HLA-B HLA-C	H2-D1 H2-K1 H2-L	HLA-E HLA-F HLA-G	H2-M5 H2-Q7 H2-T3	MICA	MR1	RAET1N	AZGP1	CD1D1	FCGRT
			Hs	*Mm*	*Hs*	*Mm*	*Hs*	*Hs*	*Hs*	*Mm*	*Mm*	*Mm*
A-STRAND	1–14	14	14	14	14	14	14	14	14	14	14	14
	7A	+1										
AB-TURN	15–17	3	3	3	3	3	2	2	2	2	2	2
B-STRAND	18–28	11	11	11	11	11	11	11	11	11	11	11
BC-TURN	29–30	2	2	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	1	1	1	1	1	1	1	1	0	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8	7	8
	49.1–49.5	+5					+5				+3	
HELIX	50–92	43	43	43	43	43	42	43	43	43	43	43
	54A	+1										
	61A	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
	61B	+1					+1	+1	+1	+1	+1	+1
	72A	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
	92A	+1										
Total length		92 (+11)	90 (+2)	90 (+2)	90 (+2)	90 (+2)	88 (+8)	89 (+3)	89 (+3)	89 (+3)	87 (+6)	89 (+3)

B			G-BETA [D1]							
			MHC-IIa				MHC-IIb			
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-DPB1	HLA-DQB1 HLA-DRB1	H2-AB	H2-EB1	HLA-DMB	HLA-DOB	H2-DMB1	H2-DOB
			Hs	*Hs*	*Mm*	*Mm*	*Hs*	*Hs*	*Mm*	*Mm*
A-STRAND	1–14	14	14	14	14	14	14	14	14	14
	7A	+1								
AB-TURN	15–17	3	3	3	3	3	3	3	3	3
B-STRAND	18–28	11	9	11	11	11	11	11	11	11

(*continued on next page*)

Table 2 (*continued*)

B			G-BETA [D1]							
			MHC-IIa				MHC-IIb			
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-DPB1	HLA-DQB1 HLA-DRB1	H2-AB	H2-EB1	HLA-DMB	HLA-DOB	H2-DMB1	H2-DOB
			Hs	*Hs*	*Mm*	*Mm*	*Hs*	*Hs*	*Mm*	*Mm*
BC-TURN	29–30	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	1	1	1	1	1	1	1	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8
	49.1–49.5	+5					+4		+4	
HELIX	50–92	43	40	40	41	40	40	40	40	40
	54A	+1								
	61A	+1	+1	+1		+1	+1	+1	+1	+1
	61B	+1	+1	+1		+1	+1	+1	+1	+1
	72A	+1	+1	+1	+1	+1	+1	+1	+1	+1
	92A	+1								
Total length		92 (+11)	85 (+3)	87 (+3)	88 (+1)	87 (+3)	87 (+7)	87 (+3)	87 (+7)	87 (+3)

(A) G-ALPHA1 [D1] domains of MHC-Ia and MHC-Ib chains, G-ALPHA1-LIKE [D1] domains of MHC-I-like chains, and G-ALPHA [D1] domains of MHC-IIa and MHC-IIb chains. (B) G-ALPHA2 [D2] domains of MHC-Ia and MHC-Ib chains, G-ALPHA2-LIKE [D2] domains of MHC-I-like chains, and G-BETA [D1] domains of MHC-IIa and MHC-IIb chains. A plus (+) sign indicates additional positions (see text).

[a] The amino acids at positions 1.2 and 1.1 for HLA-DRA were identified by amino acid sequencing [42] and extrapolated for the G-ALPHA [D1] domains of the other MHC-IIa chains and for the MHC-IIb HLA-DQA and H2-DQA chains. These additional amino acids, and those at positions 1.10–1.1 of the G-ALPHA [D1] domain of the HLA-DMA and H2-DMA chains, need to be confirmed experimentally in the mature chain. This also concerns positions 1.3–1.1 of the G-ALPHA1-LIKE [D1] domain of the *M. musculus* CD1D1 and positions 1.1 of *M. musculus* FCGRT.

Table 3
Correspondence between the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN, and MHC-Ia, MHC-IIa and MHC-IIb chain numberings

IMGT G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN[a]	MHC-Ia		MHC-IIa		MHC-IIb	
		I-ALPHA chain		II-ALPHA chain	II-BETA chain	II-ALPHA chain	II-BETA chain
		HLA-A*0201		HLA-DRA*0101	HLA-DRB1*1402	HLA-DMA*01	HLA-DMB*01
		G-ALPHA1 [D1] domain	G-ALPHA2 [D2] domain	G-ALPHA [D1] domain	G-BETA [D1] domain	G-ALPHA [D1] domain	G-BETA [D1] domain
A-STRAND	1.10					1 gct ALA (A)	
	1.9					2 cct PRO P	
	1.8					3 act THR T	
	1.7					4 cca PRO P	
	1.6					5 atg MET M	
	1.5					6 tgg TRP W	
	1.4					7 cca PRO P	
	1.3					8 gat ASP D	
	1.2			1 atc ILE I		9 gac ASP D	
	1.1			2 aaa LYS K		10 ctg LEU L	
	1	1 ggc GLY (G)	91 ggt GLY (G)	3 gaa GLU (E)	1 cca PRO (P)	11 caa GLN Q	1 ggt GLY (G)
	2	2 tct SER S	92 tct SER S	4 gaa GLU E	2 cgt ARG R	12 aac ASN N	2 ggc GLY G
	3	3 cac HIS H	93 cac HIS H	5 cat HIS H	3 ttc PHE F	13 cac HIS H	3 ttc PHE F
	4	4 tcc SER S	94 acc THR T	6 gtg VAL V	4 ttg LEU L	14 aca THR T	4 gtg VAL V
	5	5 atg MET M	95 gtc VAL V	7 atc ILE I	5 gag GLU E	15 ttc PHE F	5 gcc ALA A
	6	6 agg ARG R	96 cag GLN Q	8 atc ILE I	6 tac TYR Y	16 ctg LEU L	6 cat HIS H
	7	7 tat TYR Y	97 agg ARG R	9 cag GLN Q	7 tct SER S	17 cac HIS H	7 gtg VAL V
	7A	–	–	–	–	–	–
	8	8 ttc PHE F	98 atg MET M	10 gcc ALA A	8 acg THR T	18 aca THR T	8 gaa GLU E
	9	9 ttc PHE F	99 tat TYR Y	11 gag GLU E	9 tct SER S	19 gtg VAL V	9 agc SER S
	10	10 aca THR T	100 ggc GLY G	12 ttc PHE F	10 gag GLU E	20 tac TYR Y	10 acc THR T
	11	11 tcc SER S	101 tgc CYS C	13 tat TYR Y	11 tgt CYS C	21 tgc CYS C	11 tgt CYS C
	12	12 gtg VAL V	102 gac ASP D	14 ctg LEU L	12 cat HIS H	22 cag GLN Q	12 ctg LEU L
	13	13 tcc SER S	103 gtg VAL V	15 aat ASN N	13 ttc PHE F	23 gat ASP D	13 ttg LEU L
	14	14 cgg ARG R	104 ggg GLY G	16 cct PRO P	14 ttc PHE F	24 ggg GLY G	14 gat ASP D
AB-TURN	15	15 ccc PRO P	105 tcg SER S	–	15 aat ASN N	–	15 gat ASP D
	16	16 ggc GLY G	106 gac ASP D	–	16 ggg GLY G	–	16 gct ALA A
	17	17 cgc ARG R	107 tgg TRP W	–	17 acg THR T	–	17 ggg GLY G
B-STRAND	18	18 ggg GLY G	108 cgc ARG R	–	18 gag GLU E	–	18 act THR T
	19	19 gag GLU E	109 ttc PHE F	17 gac ASP D	19 cgg ARG R	25 agt SER S	19 cca PRO P
	20	20 ccc PRO P	110 ctc LEU L	18 caa GLN Q	20 gtg VAL V	26 ccc PRO P	20 aag LYS K
	21	21 cgc ARG R	111 cgc ARG R	19 tca SER S	21 cgg ARG R	27 agt SER S	21 gat ASP D
	22	22 ttc PHE F	112 ggg GLY G	20 ggc GLY G	22 ttc PHE F	28 gtg VAL V	22 ttc PHE F
	23	23 atc ILE I	113 tac TYR Y	21 gag GLU E	23 ctg LEU L	29 gga GLY G	23 aca THR T

(*continued on next page*)

Table 3 (*continued*)

IMGT G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN[a]	MHC-Ia		MHC-IIa		MHC-IIb	
		I-ALPHA chain		II-ALPHA chain	II-BETA chain	II-ALPHA chain	II-BETA chain
		HLA-A*0201		HLA-DRA*0101	HLA-DRB1*1402	HLA-DMA*01	HLA-DMB*01
		G-ALPHA1 [D1] domain	G-ALPHA2 [D2] domain	G-ALPHA [D1] domain	G-BETA [D1] domain	G-ALPHA [D1] domain	G-BETA [D1] domain
	24	24 gca ALA A	114 cac HIS H	22 ttt PHE F	24 gag GLU E	30 ctc LEU L	24 tac TYR Y
	25	25 gtg VAL V	115 cag GLN Q	23 atg MET M	25 aga ARG R	31 tct SER S	25 tgc CYS C
	26	26 ggc GLY G	116 tac TYR Y	24 ttt PHE F	26 tac TYR Y	32 gag GLU E	26 atc ILE I
	27	27 tac TYR Y	117 gcc ALA A	25 gac ASP D	27 ttc PHE F	33 gcc ALA A	27 tcc SER S
	28	28 gtg VAL V	118 tac TYR Y	26 ttt PHE F	28 cat HIS H	34 tac TYR Y	28 ttc PHE F
BC-TURN	29	29 gac ASP D	119 gac ASP D	27 gat ASP D	29 aac ASN N	35 gac ASP D	29 aac ASN N
	30	30 gac ASP D	120 ggc GLY G	28 ggt GLY G	30 cag GLN Q	36 gag GLU E	30 aag LYS K
C-STRAND	31	31 acg THR T	121 aag LYS K	29 gat ASP D	31 gag GLU E	37 gac ASP D	31 gat ASP D
	32	32 cag GLN Q	122 gat ASP D	30 gag GLU E	32 gag GLU E	38 cag GLN Q	32 ctg LEU L
	33	33 ttc PHE F	123 tac TYR Y	31 att ILE I	33 aac ASN N	39 ctt LEU L	33 ctg LEU L
	34	34 gtg VAL V	124 atc ILE I	32 ttc PHE F	34 gtg VAL V	40 ttc PHE F	34 acc THR T
	35	35 cgg ARG R	125 gcc ALA A	33 cat HIS H	35 cgc ARG R	41 ttc PHE F	35 tgc CYS C
	36	36 ttc PHE F	126 ctg LEU L	34 gtg VAL V	36 ttc PHE F	42 ttc PHE F	36 tgg TRP W
	37	37 gac ASP D	127 aaa LYS K	35 gat ASP D	37 gac ASP D	43 gac ASP D	37 gat ASP D
	38	38 agc SER S	128 gag GLU E	36 atg MET M	38 agc SER S	44 ttt PHE F	38 cca PRO P
CD-TURN	39	39 gac ASP D	129 gac ASP D	37 gca ALA A	39 gac ASP D	45 tcc SER S	39 gag GLU E
	40	40 gcc ALA A	–	–	–	–	–
	41	41 gcg ALA A	–	–	–	–	–
D-STRAND	42	42 agc SER S	130 ctg LEU L	38 aag LYS K	40 gtg VAL V	46 cag GLN Q	40 gag GLU E
	43	43 cag GLN Q	131 cgc ARG R	39 aag LYS K	41 ggg GLY G	47 aac ASN N	41 aat ASN N
	44	44 agg ARG R	132 tct SER S	40 gag GLU E	42 gag GLU E	48 act THR T	42 aag LYS K
	45	45 atg MET M	133 tgg TRP W	41 acg THR T	43 tac TYR Y	49 cgg ARG R	43 atg MET M
	46	46 gag GLU E	134 acc THR T	42 gtc VAL V	44 cgg ARG R	50 gtg VAL V	44 gcc ALA A
	47	47 ccg PRO P	135 gcg ALA A	43 tgg TRP W	45 gcg ALA A	51 cct PRO P	45 cct PRO P
	48	48 cgg ARG R	136 gcg ALA A	44 cgg ARG R	46 gtg VAL V	52 cgc ARG R	46 tgc CYS C
	49	49 gcg ALA A	137 gac ASP D	45 ctt LEU L	47 acg THR T	53 ctg LEU L	47 gaa GLU E
	49.1	–	–	–	–	–	48 ttt PHE F
	49.2	–	–	–	–	–	49 ggg GLY G
	49.3	–	–	–	–	–	50 gtg VAL V
	49.4	–	–	–	–	–	51 ctg LEU L
	49.5	–	–	–	–	–	–
HELIX	50	50 ccg PRO P	138 atg MET M	46 gaa GLU E	48 gag GLU E	54 ccc PRO P	52 aat ASN N
	51	51 tgg TRP W	139 gca ALA A	47 gaa GLU E	49 ctg LEU L	55 gaa GLU E	53 agc SER S
	52	52 ata ILE I	140 gct ALA A	48 ttt PHE F	50 ggg GLY G	56 ttt PHE F	54 ttg LEU L

53	53 gag GLU E	141 cag GLN Q	49 gga GLY G	51 cgg ARG R	57 gct ALA A	55 gcg ALA A
54	54 cag GLN Q	142 acc THR T	50 cga ARG R	52 cct PRO P	58 gac ASP D	56 aat ASN N
54A	–	–	–	–	–	-
55	55 gag GLU E	143 acc THR T	–	53 gat ASP D	–	57 gtc VAL V
56	56 ggt GLY G	144 aag LYS K	–	54 gcc ALA A	–	58 ctc LEU L
57	57 ccg PRO P	145 cac HIS H	–	55 gag GLU E	–	59 tca SER S
58	58 gag GLU E	146 aag LYS K	–	56 tac TYR Y	–	60 cag GLN Q
59	59 tat TYR Y	147 tgg TRP W	51 ttt PHE F	57 tgg TRP W	59 tgg TRP W	61 cac HIS H
60	60 tgg TRP W	148 gag GLU E	52 gcc ALA A	58 aac ASN N	60 gct ALA A	62 ctc LEU L
61	61 gac ASP D	149 gcg ALA A	53 agc SER S	59 agc SER S	61 cag GLN Q	63 aac ASN N
61A	–	150 gcc ALA A	–	60 cag GLN Q	62 gaa GLU E	64 caa GLN Q
61B	–	–	–	61 aag LYS K	63 cag GLN Q	65 aaa LYS K
62	62 ggg GLY G	151 cat HIS H	54 ttt PHE F	62 gac ASP D	64 gga GLY G	66 gac ASP D
63	63 gag GLU E	152 gtg VAL V	55 gag GLU E	63 ctc LEU L	65 gat ASP D	67 acc THR T
64	64 aca THR T	153 gcg ALA A	56 gct ALA A	64 ctg LEU L	66 gct ALA A	68 ctg LEU L
65	65 cgg ARG R	154 gag GLU E	57 caa GLN Q	65 gag GLU E	67 cct PRO P	69 atg MET M
66	66 aaa LYS K	155 cag GLN Q	58 ggt GLY G	66 cag GLN Q	68 gcc ALA A	70 cag GLN Q
67	67 gtg VAL V	156 ttg LEU L	59 gca ALA A	67 agg ARG R	69 att ILE I	71 cgc ARG R
68	68 aag LYS K	157 aga ARG R	60 ttg LEU L	68 cgg ARG R	70 tta LEU L	72 ttg LEU L
69	69 gcc ALA A	158 gcc ALA A	61 gcc ALA A	69 gcc ALA A	71 ttt PHE F	73 cgc ARG R
70	70 cac HIS H	159 tac TYR Y	62 aac ASN N	70 gcg ALA A	72 gac ASP D	74 aat ASN N
71	71 tca SER S	160 ctg LEU L	63 ata ILE I	71 gtg VAL V	73 aaa LYS K	75 ggg GLY G
72	72 cag GLN Q	161 gag GLU E	64 gct ALA A	72 gac ASP D	74 gag GLU E	76 ctt LEU L
72A	–	162 ggc GLY G	–	73 acc THR T	–	77 cag GLN Q
73	73 act THR T	163 acg THR T	65 gtg VAL V	74 tac TYR Y	75 ttc PHE F	78 aat ASN N
74	74 cac HIS H	164 tgc CYS C	66 gac ASP D	75 tgc CYS C	76 tgc CYS C	79 tgt CYS C
75	75 cga ARG R	165 gtg VAL V	67 aaa LYS K	76 aga ARG R	77 gag GLU E	80 gcc ALA A
76	76 gtg VAL V	166 gag GLU E	68 gcc ALA A	77 cac HIS H	78 tgg TRP W	81 aca THR T
77	77 gac ASP D	167 tgg TRP W	69 aac ASN N	78 aac ASN N	79 atg MET M	82 cac HIS H
78	78 ctg LEU L	168 ctc LEU L	70 ctg LEU L	79 tac TYR Y	80 atc ILE I	83 acc THR T
79	79 ggg GLY G	169 cgc ARG R	71 gaa GLU E	80 ggg GLY G	81 cag GLN Q	84 cag GLN Q
80	80 acc THR T	170 aga ARG R	72 atc ILE I	81 gtt VAL V	82 caa GLN Q	85 ccc PRO P
81	81 ctg LEU L	171 tac TYR Y	73 atg MET M	82 ggt GLY G	83 ata ILE I	86 ttc PHE F
82	82 cgc ARG R	172 ctg LEU L	74 aca THR T	83 gag GLU E	84 ggg GLY G	87 tgg TRP W
83	83 ggc GLY G	173 gag GLU E	75 aag LYS K	84 agc SER S	85 cca PRO P	88 gga GLY G
84	84 tac TYR Y	174 aac ASN N	76 cgc ARG R	85 ttc PHE F	86 aaa LYS K	89 tca SER S
85	85 tac TYR Y	175 ggg GLY G	77 tcc SER S	86 aca THR T	87 ctt LEU L	90 ctg LEU L
86	86 aac ASN N	176 aag LYS K	78 aac ASN N	87 gtg VAL V	88 gat ASP D	91 acc THR T
87	87 cag GLN Q	177 gag GLU E	79 tat TYR Y	88 cag GLN Q	89 ggg GLY G	92 aac ASN N
88	88 agc SER S	178 acg THR T	80 act THR T	89 cgg ARG R	90 aaa LYS K	93 agg ARG R
89	89 gag GLU E	179 ctg LEU L	81 ccg PRO P	90 cga ARG R	91 atc ILE I	94 aca THR T
90	90 gcc ALA A	180 cag GLN Q	82 atc ILE I	–	92 ccg PRO P	–

(*continued on next page*)

Table 3 (*continued*)

IMGT G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN[a]	MHC-Ia		MHC-IIa		MHC-IIb	
		I-ALPHA chain		II-ALPHA chain	II-BETA chain	II-ALPHA chain	II-BETA chain
		HLA-A*0201		HLA-DRA*0101	HLA-DRB1*1402	HLA-DMA*01	HLA-DMB*01
		G-ALPHA1 [D1] domain	G-ALPHA2 [D2] domain	G-ALPHA [D1] domain	G-BETA [D1] domain	G-ALPHA [D1] domain	G-BETA [D1] domain
	91	–	181 cgc ARG R	83 acc THR T	–	93 gtg VAL V	–
	92	–	182 acg THR T	84 aat ASN N	–	94 tcc SER S	–
	92A	–	–	–	–	95 aga ARG R	–

Unoccupied positions according to the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN are shown with dashes. The codon encoding the amino acid at position 1 results from the splicing between EX1 and EX2, except for the HLA-DMA G-ALPHA domain, where it is the codon at position 1.10. The presence of I (1.2) and K (1.1) in the HLA-DRA G-ALPHA [D1] domain has been demonstrated experimentally by amino acid sequencing [42]. EMBL/GenBank/DDBJ accession numbers of (HLA-A*020101): K02883, HLA-DRA*0101: J00204, HLA-DRB1*1402: AJ297583, HLA-DMA*01: X62744, HLA-DMB*01: X76776. The allele nomenclature of the *H. sapiens* classical MHC-Ia and MHC-IIa is according to HLA-DB [13]. Owing to a lesser degree of polymorphism of the nonclassical *Homo sapiens* MHC-Ib and MHC-IIb genes, compared to the classical MHC-Ia and MHC-IIa genes, a 2-digit is used for their allele description (IMGT Scientific chart, http://imgt.cines.fr): *01 refers to *0101 found in the literature.

[a] IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN, first defined in 2002 by Marie-Paule Lefranc, Université Montpellier II, CNRS (IMGT http://imgt.cines.fr, online 15/05/2002), and this paper.

5. IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN and structural data comparison

Beyond sequence data comparison, the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN provides information on the strand, turn and helix lengths (Table 2) and allows standardized 2D graphical representations or IMGT Colliers de Perles for G-DOMAIN and G-LIKE-DOMAIN. Fig. 4 shows, as examples, the IMGT Colliers de Perles for the G-ALPHA1 and G-ALPHA2 domains of the MHC-Ia *Homo sapiens* HLA-B*0702 and *Mus musculus* H2-K1*01 (Fig. 4A), for the G-ALPHA and G-BETA domains of the MHC-IIa *Homo sapiens* HLA-DQA1*0501/HLA-DQB1*0301 and *Mus musculus* H2-AA*02/H2-AB*02 (Fig. 4B), and for the G-ALPHA1-LIKE and G-ALPHA2-LIKE domains of the MHC-I-like *Homo sapiens* MICA*01 and *Mus musculus* CD1D1*01 (Fig.4C).

Structural data comparison is straightforward using the IMGT unique numbering. Indeed, intra- and inter-domain contact analysis can be analysed and compared for any position between any G-DOMAIN or G-LIKE-DOMAIN [25,36]. This standardization not only allows the structural characterization of a position inside a domain, but also the statistical analysis of amino acid properties, position per position, between domains, as this has been demonstrated for the IG V-DOMAINs [47]. Eleven IMGT amino acid classes have been defined, based on the hydropathy, the volume and the chemical characteristics of the 20 common amino acids [47]. These classes, initially defined for a standardized comparison of the properties of the IG, TR and IgSF chains and domains will be used, with the IMGT unique numbering, for the comparison of the MHC and MhcSF chains and domains.

6. Conclusion

The IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN allows, for the first time, to compare any G-DOMAIN of MHC and G-LIKE-DOMAIN of MHC-I-like proteins, or in other words, to compare any G-set domain of MhcSF proteins. This is the third major breakthrough for domain analysis

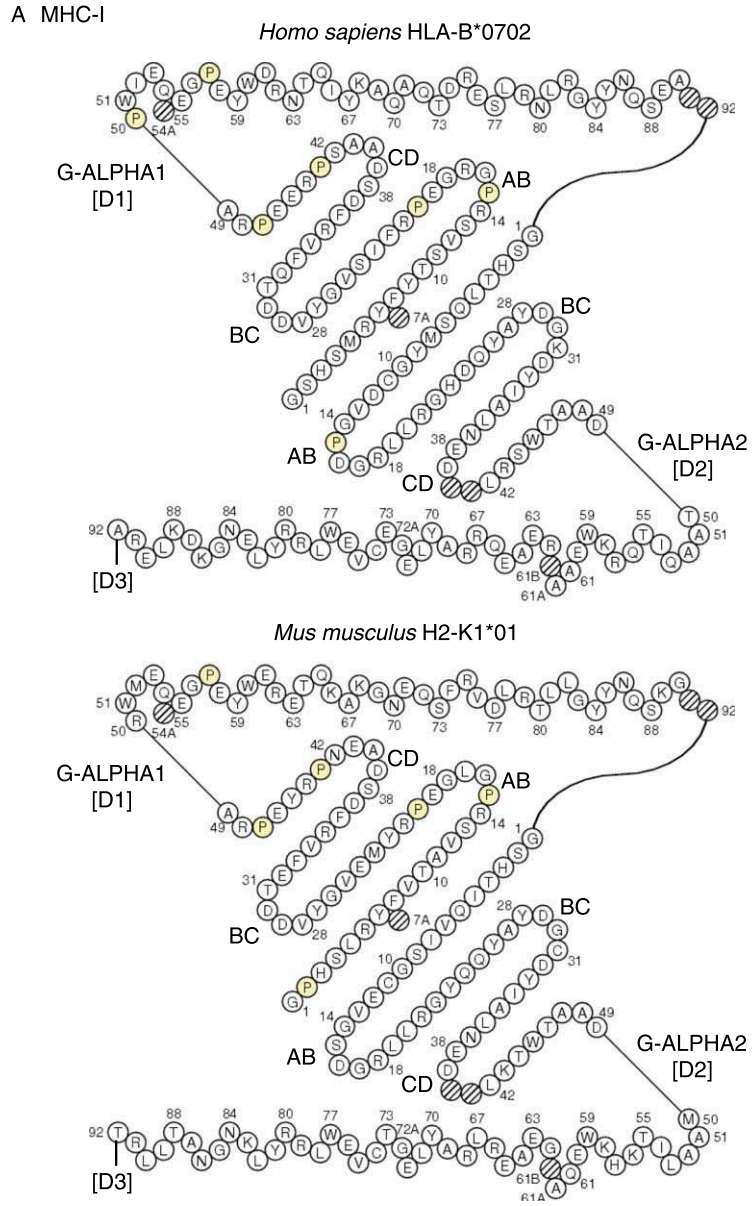


Fig. 4. IMGT Collier de Perles of G-DOMAINs and G-LIKE-DOMAINs. (A) MHC-I G-ALPHA1 [D1] and G-ALPHA2 [D2] domains of the *Homo sapiens* HLA-B*0702 and *Mus musculus* H2-K1*01 chains. (B) MHC-II G-ALPHA [D1] and G-BETA D1] domains of the *H. sapiens* HLA-DQA1*0501/HLA-DQB1*0301 and *M. musculus* H2-AA*02/H2-AA*02 chains. (C) MHC-I-like G-ALPHA1-LIKE [D1] and G-ALPHA2-LIKE [D2] domains of the *H. sapiens* MICA*01 and *M. musculus* CD1D1*01 chains. The amino acid at position 1 is encoded by a codon which results from the splicing with the preceding exon, except for the CD1D1 G-ALPHA1-LIKE domain, for which it is position 1.3. Amino acids are shown in the one-letter abbreviation. Hatched circles correspond to missing positions according to the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN. In IMGT Colliers de Perles, position 7A is only displayed in the G-ALPHA and G-ALPHA1-LIKE domains, and positions 61A and 61B in the G-BETA and G-ALPHA2-LIKE domains. As position 54A is only occupied in G-ALPHA1-LIKE of MHC-I-like proteins, this position can be omitted in IMGT Colliers de Perles, if only MHC chains are compared [36]. Position 92A is only added for MHC-DMA and H2-DMA IMGT Colliers de Perles. Note that the N-terminal end of a peptide in the cleft would be on the left hand side.

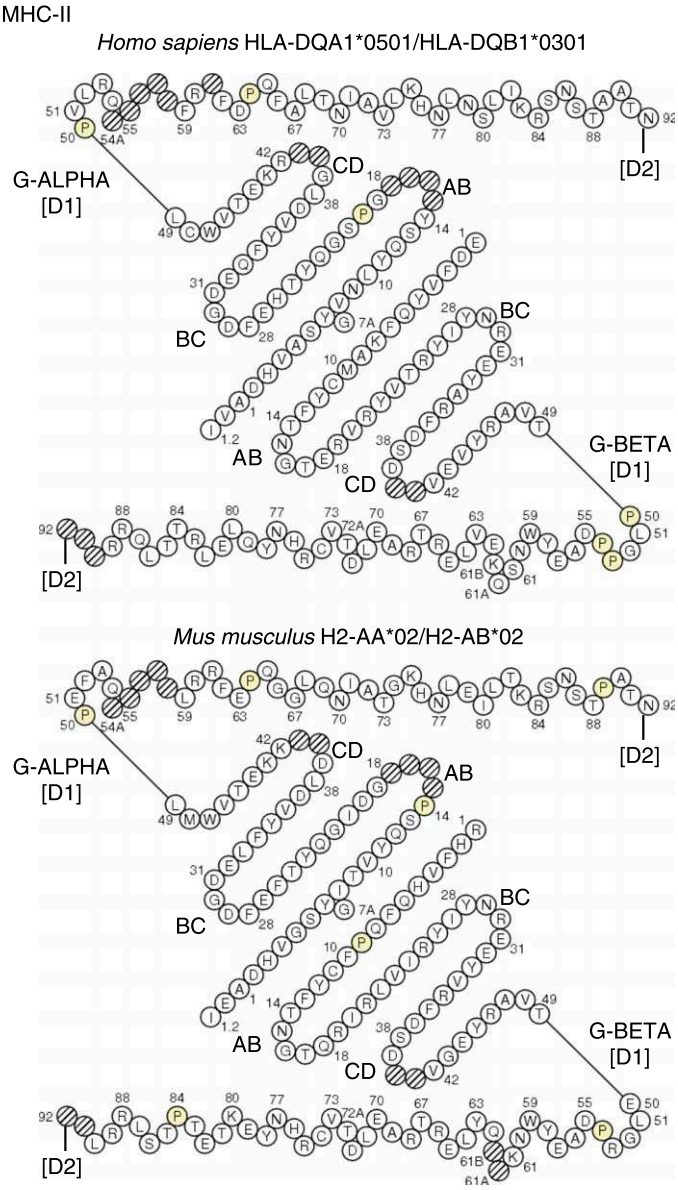


Fig. 4 (*continued*)

using the IMGT unique numbering, based on the NUMEROTATION concept of IMGT-ONTOLOGY [17]. Indeed, this completes the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN that allows to compare any V-set domain of IgSF proteins (V-DOMAIN of IG and TR, and V-LIKE-DOMAIN of IgSF proteins other than the IG or TR) [20,35], and the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN that allows to compare any C-set domain of IgSF proteins (C-DOMAIN of IG and TR, and C-LIKE-DOMAIN of IgSF other than the IG or TR) [21,35].

The IMGT unique numbering has many advantages. Three features are worth noting: (i) In IMGT,

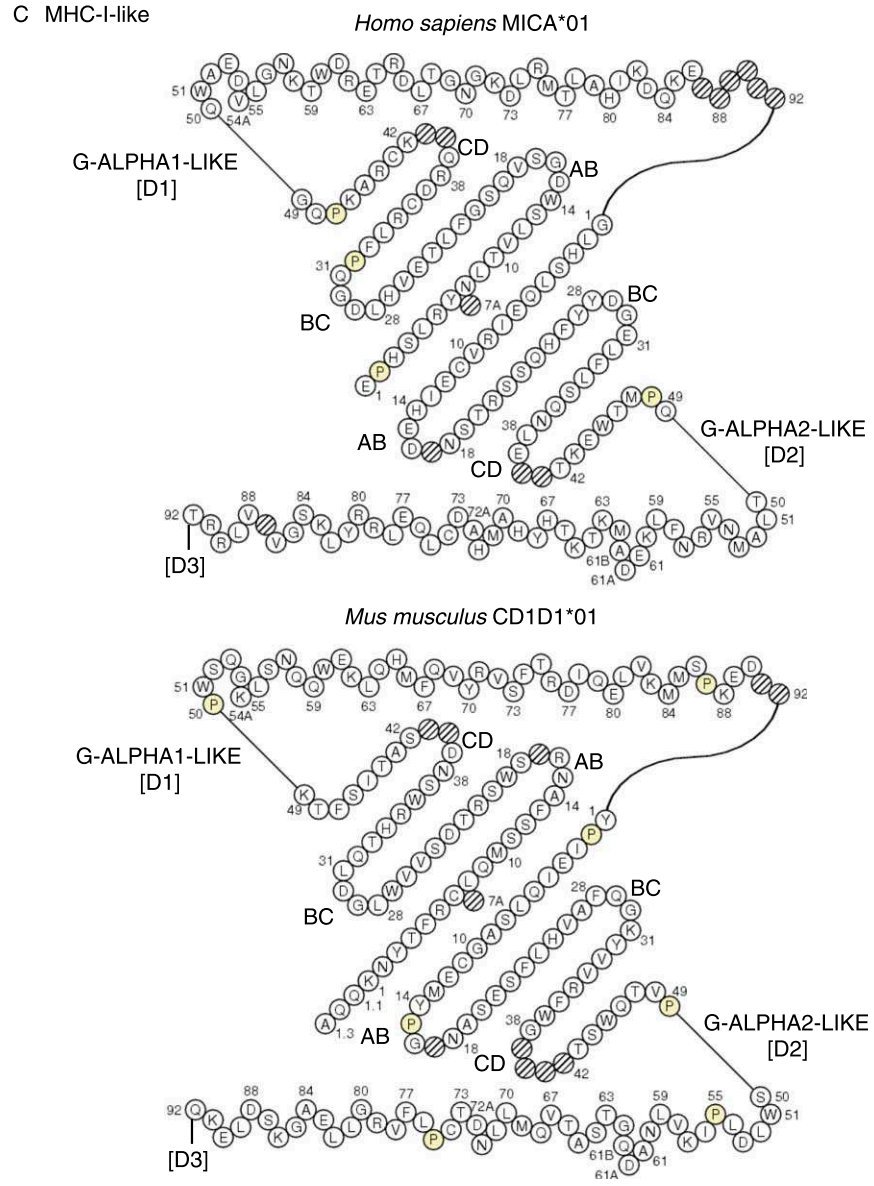


Fig. 4 (*continued*)

any domain is characterized by the length of its strands, loops and turns and, for the G-set, by the length of its helix. The strand, loop, turn or helix lengths (the number of codons or amino acids, that is the number of occupied positions) become crucial information which characterizes the domains. This first feature of the IMGT standardization based on the IMGT unique numbering allowed, for instance, to show that the distinction between the C1, C2, I1 and I2 types found in the literature and in the databases to describe the IgSF C-set domains is unapplicable when dealing with sequences for which no structural data are known (discussed in [21]). (ii) A second feature of the IMGT standardization is the comparison of cDNA

and/or amino acid sequences with genomic sequences, and the identification of the splicing sites, to delimit precisely the domains: a G-DOMAIN or a G-LIKE-DOMAIN is frequently encoded by a unique exon, as this is the case for the V-LIKE-DOMAINs [20,35], C-DOMAINs and C-LIKE-DOMAINs [21,35]. This IMGT standardization for the domain delimitation explains the discrepancies observed with the generalist Swiss-Prot database which does not take into account this criteria. (iii) At last, a third feature is the IMGT Collier de Perles which, in the absence of available 3D structures, is particularly useful to compare domains of very diverse families.

The IMGT unique numbering allows standardized representations of nucleotide and amino acid sequences in the IMGT Web resources, and more particularly in the IMGT Repertoire (http://imgt.cines.fr) (Tables of alleles, Alignments of alleles, IMGT Protein displays, IMGT Colliers de Perles, 3D structures). The IMGT unique numbering is extensively used in the IMGT databases [1,22–25] and sequence and structure analysis tools [25–28]. It is a key element of the IMGT strategy for the automatic annotation of nucleotide sequences and standardized label assignment [48]. The IMGT unique numbering represents, therefore, a major step forward in the analysis and comparison of the MhcSF domains as this was already demonstrated for the IgSF domains [20,21,35]. By providing a unique frame for the structural analysis of the V-set, C-set and G-set domains, the IMGT standardized approach opens new insight on the evolution of these domains and of their functional interactions.

Acknowledgements

We are grateful to Chantal Ginestoux, Véronique Giudicelli, Joumana Jabado–Michaloud, Géraldine Folch, Vincent Negre, Oliver Clément and Denys Chaume for helpful discussion. E.D. is holder of a doctoral grant from the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR). K.Q. was the recipient of a doctoral grant from the MENESR and is currently supported by a grant from the Association pour la Recherche sur le Cancer (ARC). IMGT is a registered mark of the Centre National de la Recherche Scientifique (CNRS). IMGT is a RIO platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287) programmes of the European Union and received subventions from ARC and from the Génopole-Montpellier-Languedoc-Roussillon. IMGT is currently supported by the CNRS, the MENESR (Université Montpellier II Plan Pluri-Formation, BIOSTIC-LR2004 Région Languedoc-Roussillon and ACI-IMPBIO IMP82-2004). Part of this work was carried out in the frame of the European Science Foundation Scientific Network Myelin Structure and its role in autoimmunity (MARIE).

References

[1] Lefranc M-P, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, Scaviner D, et al. IMGT, the international ImMunoGeneTics information system®. Nucleic Acids Res 2005;33:D593–D7.

[2] Warr GW, Clem LW, Soderhall K. The international ImMunoGeneTics database IMGT. Dev Comp Immunol 2003;27:1.

[3] Lefranc M-P. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev Comp Immunol 2002; 26:697–705.

[4] Lefranc M-P. IMGT-ONTOLOGY and MGT databases, tools and web resources for immunogenetics and immunoinformatics. Mol Immunol 2004;40:647–59.

[5] Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, et al. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics (http://imgt.cines.fr). *In Silico* Biology 2003;4,0004. http://www.bioinfo.de/isb/2003/04/2004/. *In Silico* Biology 2004;4,17–29.

[6] Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, et al. IMGT-Choreography for Immunogenetics and Immunoinformatics (http://imgt.cines.fr). *In Silico* Biology. http://www.bioinfo.de/isb/2004/05/0006/.

[7] Lefranc M-P. IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis Leukemia. 2003;17:260–6.

[8] Lefranc M-P. IMGT, the international ImMunoGeneTics information system®, http://imgt.cines.fr<http://imgt.cines.fr/>. In: Bock G, Goode J, editors. Immunoinformatics: bioinformatics strategies for better understanding of immune function, Novartis foundation symposium 254. Chichester, UK: Wiley; 2003. p. 126–36 (discussion pp.136–142, 216–222, 250–252).

[9] Lefranc M-P. IMGT, the international ImMunoGeneTics information system® http://imgt.cines.fr <http://imgt.cines.fr/>. In: Lo BKC, editor. Antibody engineering: methods and protocols. 2nd ed. Methods in molecular biology, 2nd ed, vol. 248. Totowa, NJ: Humana Press; 2003. p. 27–49 chap 3.
[10] Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, et al. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 1997;25:206–11.
[11] Lefranc M-P, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, et al. IMGT, the International ImMunoGeneTics database. Nucleic Acids Res 1998;26:297–303.
[12] Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, et al. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 1999;27:209–12.
[13] Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, et al. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 2000;28:219–21.
[14] Lefranc M-P. IMGT ImMunoGeneTics Database. Int BIOforum 2000;4:98–100.
[15] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 2001;29:207–9.
[16] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 2003;31:307–10.
[17] Giudicelli V, Lefranc M-P. Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 1999;15:1047–54.
[18] Lefranc M-P. Unique database numbering system for immunogenetic analysis. Immunol Today 1997;18:509.
[19] Lefranc M-P. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 1999;7:132–6.
[20] Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 2003;27:55–77.
[21] Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, et al. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol 2005;29:185–203.
[22] Chaume D, Giudicelli V, Lefranc M-P. IMGT/LIGM-DB. In: The molecular biology database collection. Nucleic Acids Res 2004;32. http://www3.oup.co.uk/nar/database/summary/504.
[23] Folch G, Bertrand J, Lemaitre M, Lefranc M-P. IMGT/PRIMER-DB. In: The molecular biology database Collection. Nucleic Acids Res 2004;32. http://www3.oup.co.uk/nar/database/summary/505.
[24] 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–D61.
[25] Kaas Q, Ruiz M, 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 2004;32:D208–D10.
[26] Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res 2004;32:W435–W40.
[27] Yousfi Monod M, Giudicelli V, Chaume D, 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.
[28] Elemento O, Lefranc M-P. IMGT/PhyloGene: an online software package for phylogenetic analysis of immunoglobulin and T cell receptor genes. Dev Comp Immunol 2003;27:763–79.
[29] Scaviner D, Barbié V, Ruiz M, Lefranc M-P. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp Clin Immunogenet 1999;16:234–40.
[30] Folch G, Scaviner D, Contet V, Lefranc M-P. Protein displays of the human T cell receptor alpha, beta, gamma and delta variable and joining regions. Exp Clin Immunogenet 2000;17:205–15.
[31] Ruiz M, Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 2002;53:857–83.
[32] Lefranc M-P, Lefranc G. The immunoglobulin FactsBook. London, UK: Academic Press; 2001, 458 pages.
[33] Lefranc M-P, Lefranc G. The T cell receptor FactsBook. London, UK: Academic Press; 2001, 398 pages.
[34] Bertrand G, Duprat E, Lefranc M-P, Marti J, Coste J. Human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. Tissue Antigens 2004;64:119–31.
[35] Duprat E, Kaas Q, Garelle V, Giudicelli V, Lefranc G, Lefranc M-P. IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In: Recent research developments in human genetics. Trivandrum, India: Research Signpost; 2004;2: p. 111–136.
[36] Kaas Q, Duprat E, Tourneur G, Lefranc M-P. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: Immunoinformatics (Brusic V, Schoenbach C eds). Springer, The Netherlands; in press.
[37] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[38] Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL nucleotide sequence database. Nucleic Acids Res 2005;33:D29–D33.
[39] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2005;33:D34–D8.
[40] Tateno Y, Saitou N, Okubo K, Sugawara H, Gojobori T. DDBJ in collaboration with mass-sequencing teams on annotation. Nucleic Acids Res 2005;33:D25–D8.
[41] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. Genomics 2002;79:464–70.
[42] Larhammar D, Gustafsson K, Claesson L, Bill P, Wiman K, Schenning L, et al. Alpha chain of HLA-DR transplantation antigens is a member of the same protein superfamily as the immunoglobulins. Cell 1982;30:153–61.

[43] Maenaka K, Jones EY. MHC superfamily structure and the immune system. Curr Opin Struct Biol 1999;9:745–53.
[44] Wilson IA, Bjorkman PJ. Unusual MHC-like molecules: CD1, Fc receptor, the hemochromatosis gene product, and viral homologs. Curr Opin Immunol 1998;10:67–73.
[45] Braud VM, Allan DS, McMichael AJ. Functions of non-classical MHC and non-MHC- encoded class I molecules. Curr Opin Immunol 1999;11:100–8.
[46] Delker SL, West Jr AP, McDermott L, Kennedy MW, Bjorkman PJ. Crystallographic studies of ligand binding by Zn-alpha2-glycoprotein. J Struct Biol 2004;148:205–13.
[47] Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc M-P. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit 2004;17:17–32.
[48] Giudicelli V, Protat C, Lefranc M-P. The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: Proceeding of the European Conference on Computational Biology (ECCB'2003), Paris, France, 2003 INRIA (DISC, Spid) DKB-31, p. 103–4. http://www.inra.fr/eccb2003/posters/pdf/Annot_Giudicelli_20030528_160703.pdf.

Publication 3

Research Signpost
37/661 (2), Fort P.O., Trivandrum-695 023, Kerala, India

Recent Res. Devel. Human Genet., 2(2004): 111-136 ISBN: 81-7736-212-7

6

IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily

Elodie Duprat, Quentin Kaas, Valérie Garelle, Véronique Giudicelli Gérard Lefranc and Marie-Paule Lefranc[1]
IMGT, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Université Montpellier II, UPR CNRS 1142, Institut de Génétique Humaine, IGH Montpellier, France

Abstract

The immunoglobulin superfamily (IgSF) comprises the immunoglobulins (IG) and T cell receptors (TR) involved in antigen recognition, and also a great number of proteins other than IG and TR that are involved in many different functions (in ligand-receptor interactions in development, differentiation, activation, adhesion, regulation, etc.). The IgSF proteins

[1]Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France
Correspondence/Reprint request: Dr. Marie-Paule Lefranc, IMGT, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire, LIGM UPR CNRS 1142, IGH, 141 rue de la Cardonille 34396 Montpellier Cedex 5, France. E-mail: lefranc@ligm.igh.cnrs.fr

are defined by having at least one immunoglobulin-like (Ig-like) domain. Despite a large divergence in the amino acid sequences, the Ig-like domains share the IG structural fold which typically consists of about one hundred amino acids in antiparallel beta strands, linked by beta turns or loops, and located on two layers maintained by a disulfide bridge. The number of antiparallel beta strands defines two sets: 9 strands for the V-set (which comprises the V-DOMAINs of the IG and TR, and the V-LIKE-DOMAINs of the IgSF proteins other than the IG or TR) and 7 strands for the C-set (which comprises the C-DOMAINs of the IG and TR, and the C-LIKE-DOMAINs of the IgSF proteins other than the IG or TR). IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr), has set up a unique numbering system which takes into account the structural features of the Ig-like domains. In this paper, we describe the IMGT Scientific chart rules for the description of the IgSF V-set and C-set domains, that are applicable for the sequence and structure analysis, whatever the species, the IgSF protein or the chain type. We present examples of 2D graphical representations (IMGT Colliers de Perles) based on the IMGT unique numbering, that are particularly useful for the visualization and comparison of the positions of mutations and polymorphisms in the Ig-like domains.

Introduction

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [1] is a high quality integrated knowledge resource specializing in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC), and related proteins of the immune system (RPI) of human and other vertebrates [2-12]. IMGT provides a common access to expertly annotated data on the genome, proteome, genetics and structure of the IG, TR, MHC and RPI, based on the IMGT Scientific chart and IMGT-ONTOLOGY [13]. The IMGT standardized description of mutations, allelic polymorphisms, 2D and 3D structure representations, is based on the IMGT unique numbering [14-17]. The IMGT unique numbering is used for the IG and TR variable (V-DOMAIN) and constant domain (C-DOMAIN) of all jawed vertebrates whatever the species, the antigen receptor, or the chain type [18-35]. The IMGT unique numbering led to the first complete description of the human IG and TR genes and alleles [18,19], the standardized 2D graphical representations or IMGT Colliers de Perles [16,17,34,36], the standardized definition of the framework (FR-IMGT) and complementarity-determining-regions (CDR-IMGT) of the V-DOMAINs [16] and the description of the strands and loops of the C-DOMAINs [17]. The many advantages of the IMGT unique numbering naturally led us to extend it to members of the immunoglobulin superfamily (IgSF) other than IG or TR [37]. Indeed, the IgSF not only comprises the IG and TR involved in antigen recognition, but

also a great number of proteins that are involved in many different functions (in ligand-receptor interactions in development, differentiation, activation, adhesion, regulation, etc.) [38]. The common feature of the IgSF proteins is to have at least one immunoglobulin-like (Ig-like) domain [37-41]. Despite a large divergence in the amino acid sequences, the Ig-like domains share the IG structural fold which typically consists of about one hundred amino acids in antiparallel beta strands, linked by beta turns or loops, and located on two layers maintained by a disulfide bridge. The number of antiparallel beta strands defines two sets: 9 strands for the V-set (which comprises the V-DOMAINs of the IG and TR, and the V-LIKE-DOMAINs of the IgSF proteins other than the IG or TR) [16], and 7 strands for the C-set (which comprises the C-DOMAINs of the IG and TR, and the C-LIKE-DOMAINs of the IgSF proteins other than the IG or TR) [17]. By taking into account the structural features of the Ig-like domains, the IMGT unique numbering [14-17] and its graphical representations, the IMGT Colliers de Perles [34,36,37], allow to fill in the gap between linear amino acid sequences and three-dimensional (3D) structures.

In this paper, we describe the IMGT Scientific chart rules for the description of the IgSF V-set and C-set domains, which are applicable for the sequence and structure analysis, whatever the species, the IgSF protein or the chain type. We present examples of IMGT Colliers de Perles based on the IMGT unique numbering. This standardization is particularly useful in the absence of 3D structural data, for the visualization and comparison of mutation and polymorphism positions in the Ig-like domains.

1. IMGT Colliers de Perles for V-LIKE-DOMAIN

The IMGT Colliers de Perles for V-LIKE-DOMAIN is based on the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [16]. Indeed, the 3D structure of a V-LIKE-DOMAIN is very similar to that of an IG and TR V-DOMAIN (Fig. 1). Both domain types are made of 9 antiparallel beta strands (A, B, C, C', C", D, E, F and G) linked by beta turns (AB, CC', C"D, DE and EF) or loops (BC, C'C" and FG) forming a sandwich of two sheets (Figure 1). The sheets are closely packed against each other through hydrophobic interactions giving a hydrophobic core, and joined together by a disulfide bridge between strand B in the first sheet and strand F in the second sheet. In the IMGT unique numbering, the conserved amino acids always have the same position, for instance cysteine 23 (1st-CYS), tryptophan 41 (CONSERVED-TRP), conserved hydrophobic (leucine) 89, cysteine 104 (2nd-CYS). The hydrophobic amino acids of the framework regions are also found in conserved positions [14-17]. It is remarkable that the IG fold 3D structure has been conserved through evolution, despite the particularities of the IG and TR synthesis compared to the other proteins [18,19] and the sequence divergence of the IgSF domains. Indeed, the V-LIKE-DOMAIN is usually encoded by a

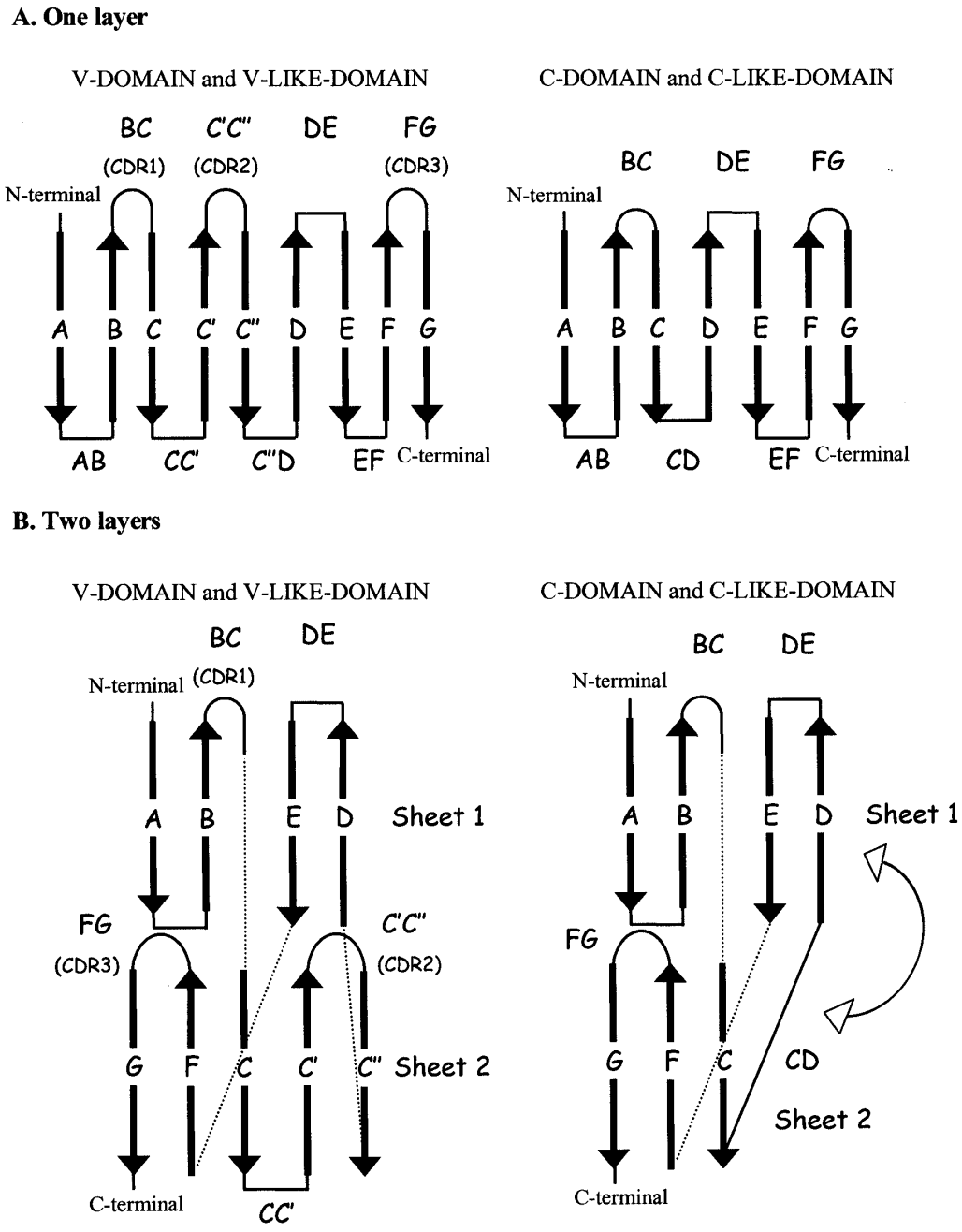
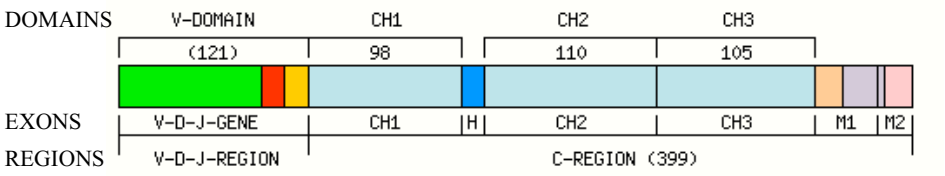


Figure 1. Schematic representation of the V-DOMAIN and V-LIKE-DOMAIN (V-set) and C-DOMAIN and C-LIKE-DOMAIN (C-set). (A) Representation on one layer. (B) Representation on two layers. A double-headed arrow shows that the D strand of the C-DOMAIN and C-LIKE-DOMAIN can be localized in sheet 1 (on the back) or in sheet 2 (on the front) depending from the length of the CD transversal strand.

unique exon, whereas the IG and TR V-DOMAIN results from the rearrangement of two (V, J) or three (V, D, J) genes [18,19] (Figure 2). The V-LIKE-DOMAIN is usually, as the IG and TR V-DOMAIN, the most N-terminal (and extracellular)

A. IG and TR

Homo sapiens membrane IG gamma 1



B. IgSF other than IG or TR

Homo sapiens CD4

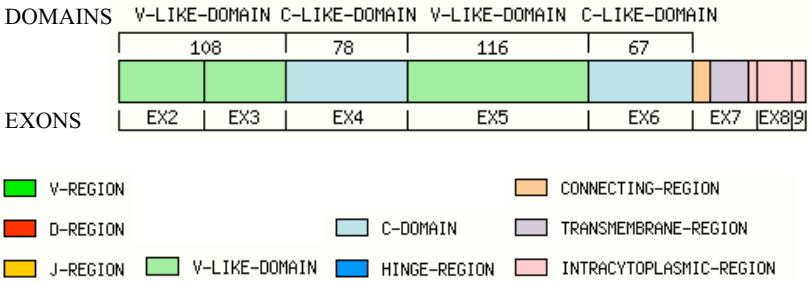


Figure 2. Correspondence between domains and exons. (A) IG and TR, (B) IgSF other than IG and TR. Lengths of the domains and exons are in number of amino acids or codons, respectively. IMGT standardized labels are in capital letters and are described in the IMGT Scientific chart (http://imgt.cines.fr). (A) *Homo sapiens* membrane IG gamma 1 heavy chain (IMGT/LIGM-DB M98324) as example of IG and TR. An IG or TR chain comprises two types of structural units: one V-DOMAIN and one (for the IG light chains and TR chains) or several (for the IG heavy chains) C-DOMAINs (CH1, CH2 and CH3). The unique V-DOMAIN (encoded by a rearranged V-J or V-D-J gene) of a IG or TR chain corresponds to the V-J-REGION or V-D-J-REGION, and is associated to a C-REGION encoded by the C-GENE [18, 19]. (B) *Homo sapiens* CD4 (EMBL/GenBank/DDBJ NT_009759) as example of an IgSF protein other than IG or TR. The general organization of the IgSF other than IG and TR is more diverse and follows the modular shuffling between domains ranging from a unique V-LIKE-DOMAIN or a unique C-LIKE-DOMAIN or to any combination of those domains [38].

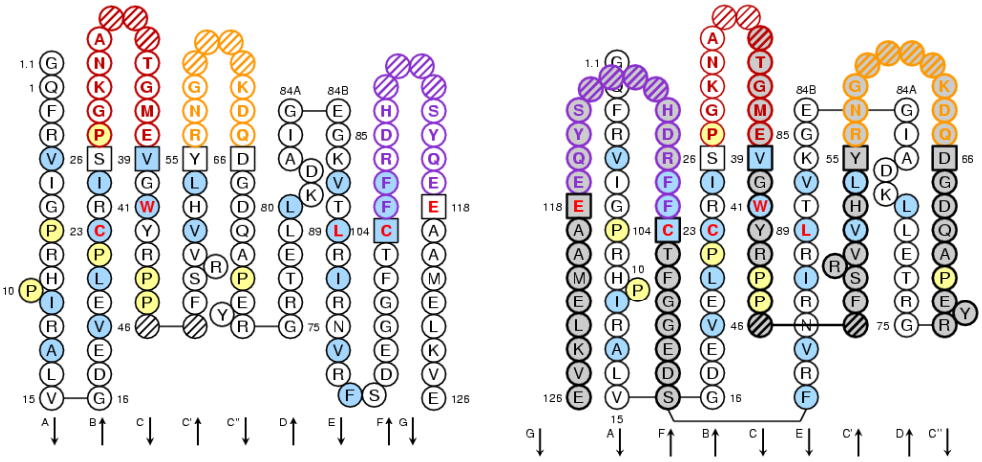
domain of the chain. However, in contrast to the IG and TR V-DOMAIN which is always unique, the V-LIKE-DOMAIN may be present in several copies in the same chain and interspersed with C-LIKE-DOMAINs (Figure 2) or with domains of other superfamilies [39].

1.1 Strands and loops of the V-LIKE-DOMAINs

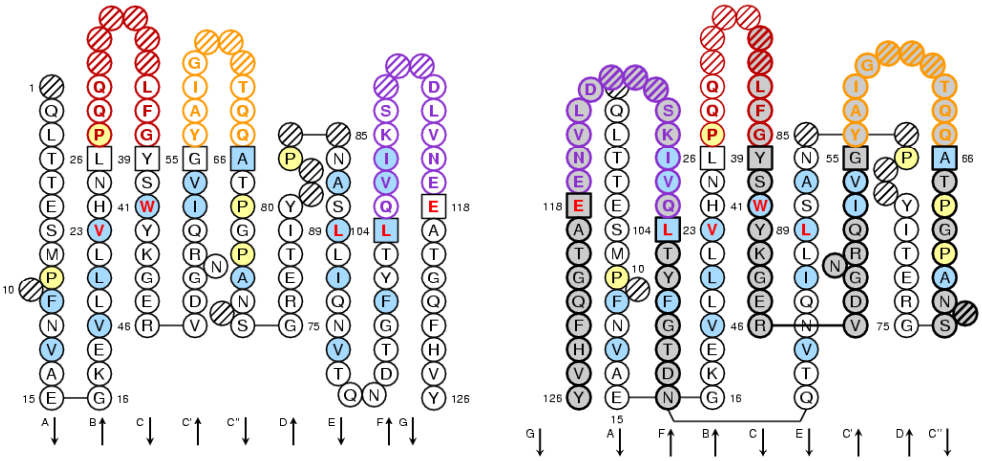
Three examples of V-LIKE-DOMAINs are shown in Figure 3: the myelin oligodendrocyte glycoprotein (MOG) [D], the carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1) [D1] and the myelin protein zero (MPZ) [D].

Figure 3

A. *Homo sapiens* MOG [D] V-LIKE-DOMAIN



B. *Homo sapiens* CEACAM1 [D1] V-LIKE-DOMAIN



C. *Homo sapiens* MPZ [D] V-LIKE-DOMAIN

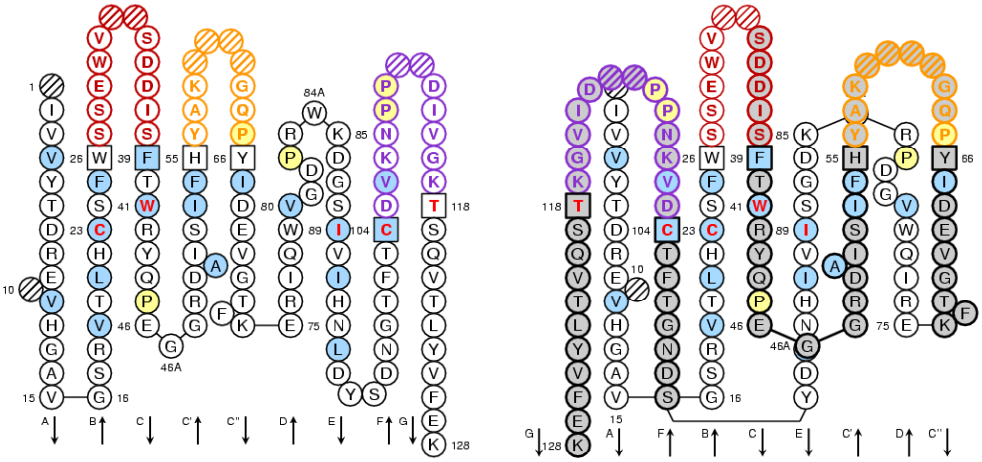


Figure 3. IMGT Colliers de Perles of V-LIKE-DOMAINs on one and two layers. (A) *Homo sapiens* MOG [D], (B) *Homo sapiens* CEACAM1 [D1], (C) *Homo sapiens* MPZ [D]. Amino acids are shown in the one-letter abbreviation. Position at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and tryptophan (W) are found in more than 50% of analysed sequences are shown in blue. All proline (P) are shown in yellow. The loops BC, C'C" and FG (corresponding to the CDR-IMGT) are limited by amino acids shown in squares (anchor positions), which belong to the neighbouring strands (FR-IMGT in V-DOMAINs). Arrows indicate the direction of the beta strands and their different designations in 3D structures (from IMGT Repertoire, http://imgt.cines.fr). BC loops are represented in red, C'C" loops in orange and FG loops in purple. The IMGT Colliers de Perles on two layers (on the right hand) show, on the forefront, the GFCC'C" strands and, on the back, the ABED strands. Hatched circles or squares correspond to missing positions according to the IMGT unique numbering. MPZ has two additional positions (46A and 84A) that interestingly are located at the apex of beta turns and do not disturb the general architecture of the domain.
Swiss-Prot accession numbers: Q16653 for the *Homo sapiens* MOG protein, P13688 for the *Homo sapiens* CEACAM1 protein, and P25189 for the *Homo sapiens* MPZ protein. IMGT Colliers de Perles were checked with the following PDB [42] entries: 1py9_A (*Mus musculus* MOG [D1]), 1pkq_E (*Rattus norvegicus* MOG [D1]), 1l6z_A (*Mus musculus* CEACAM1 [D1]), 1neu (*Rattus norvegicus* MPZ [D1]), as the human MOG [D], CEACAM1 [D1] and MPZ [D] have not yet been crystallized.

1.1.1 Strands

The antiparallel beta strands of the V-LIKE-DOMAIN correspond to the conserved regions or framework (FR-IMGT) described in the V-DOMAIN [16]. The A strand (A-STRAND, positions 1 to 15) and the B strand (B-STRAND, positions 16 to 26) with the conserved cysteine (1st-CYS) at position 23 correspond to the FR1-IMGT (Table 1). The C strand (C-STRAND, positions 39 to 46) with the tryptophan (CONSERVED-TRP) at position 41 and the C' strand (C'-STRAND, positions 47 to 55) correspond to FR2-IMGT. The C" strand (C"-STRAND, positions 66 to 74), the D strand (D-STRAND, positions 75 to 84), the E strand (E-STRAND, positions 85 to 96) with a conserved hydrophobic amino acid at position 89 and the F strand (F-STRAND, positions 97 to 104) with 2nd-CYS at position 104 correspond to the FR3-IMGT. The G strand (G-STRAND, positions 118 to 128) corresponds to FR4-IMGT (in the IG and TR V-DOMAINs, the G strand is the C-terminal part of the J-REGION, with J-PHE or J-TRP 118 and the canonical motif F/W-G-X-G at positions 118-121). Hatched circles or squares in Figure 3 correspond to missing positions according to the IMGT unique numbering. For example, unoccupied positions 46 and 47 in MOG [D], 10 or 73 in CEACAM1 [D1] and 10 in MPZ [D], are shown as hatched circles. In the IMGT Protein display (Figure 4), unoccupied positions according to the IMGT unique numbering are shown by dots.

1.1.2 Loops

The BC, C'C" and FG loops of the V-LIKE-DOMAIN correspond to the complementarity-determining regions (CDR-IMGT) described in the IG and TR V-DOMAINs [16]. The BC loop (BC-LOOP) comprises positions 27 to 38; the longest BC loops have 12 amino acids. For BC loops shorter than 12 amino acids, gaps are created at the apex (missing positions, hatched in IMGT Collier de Perles (Fig. 3), or not shown in structural data representations). The gaps are placed at the apex of the loop with an equal number of codons (or amino acids) on both sides if the loop length is an even number, or with one more codon (or amino acid) in the left part if it is an odd number. As an example, in a BC loop with 9 amino acids (MOG in Figure 3), positions 27 to 31 and 35 to 38 are present, whereas positions 32 to 34 are missing. The C'C" loop (C'C"-LOOP) comprises positions 56 to 65. The longest C'C" loops have 10 amino

Table 1. Gaps and additional positions for FG loop

A - Gaps for FG loops less than 13 amino acids

FG loop (CDR3-IMGT) lengths													
13	105	106	107	108	109	110	111	112	113	114	115	116	117
12	105	106	107	108	109	110	-	112	113	114	115	116	117
11	105	106	107	108	109	110	-	-	113	114	115	116	117
10	105	106	107	108	109	-	-	-	113	114	115	116	117
9	105	106	107	108	109	-	-	-	-	114	115	116	117
8	105	106	107	108	-	-	-	-	-	114	115	116	117
7	105	106	107	108	-	-	-	-	-	-	115	116	117
6	105	106	107	-	-	-	-	-	-	-	115	116	117
5	105	106	107	-	-	-	-	-	-	-	-	116	117

B - Additional positions for FG loops more than 13 amino acids

FG loop (CDR3-IMGT) lengths										

21	**111**	111.1	111.2	111.3	111.4	112.4	112.3	112.2	112.1	**112**
20	**111**	111.1	111.2	111.3	-	112.4	112.3	112.2	112.1	**112**
19	**111**	111.1	111.2	111.3	-	-	112.3	112.2	112.1	**112**
18	**111**	111.1	111.2	-	-	-	112.3	112.2	112.1	**112**
17	**111**	111.1	111.2	-	-	-	-	112.2	112.1	**112**
16	**111**	111.1	-	-	-	-	-	112.2	112.1	**112**
15	**111**	111.1	-	-	-	-	-	-	112.1	**112**
14	**111**	-	-	-	-	-	-	-	112.1	**112**

For FG loops (CDR3-IMGT) more than 13 amino acids, additional positions are created between positions 111 and 112 (in bold). In a given sequence set with FG loops more than 13 amino acids, gaps are created based on the FG loops in the set. As an example, gaps are shown by comparison to a 21 amino acid long FG loop.

Figure 4

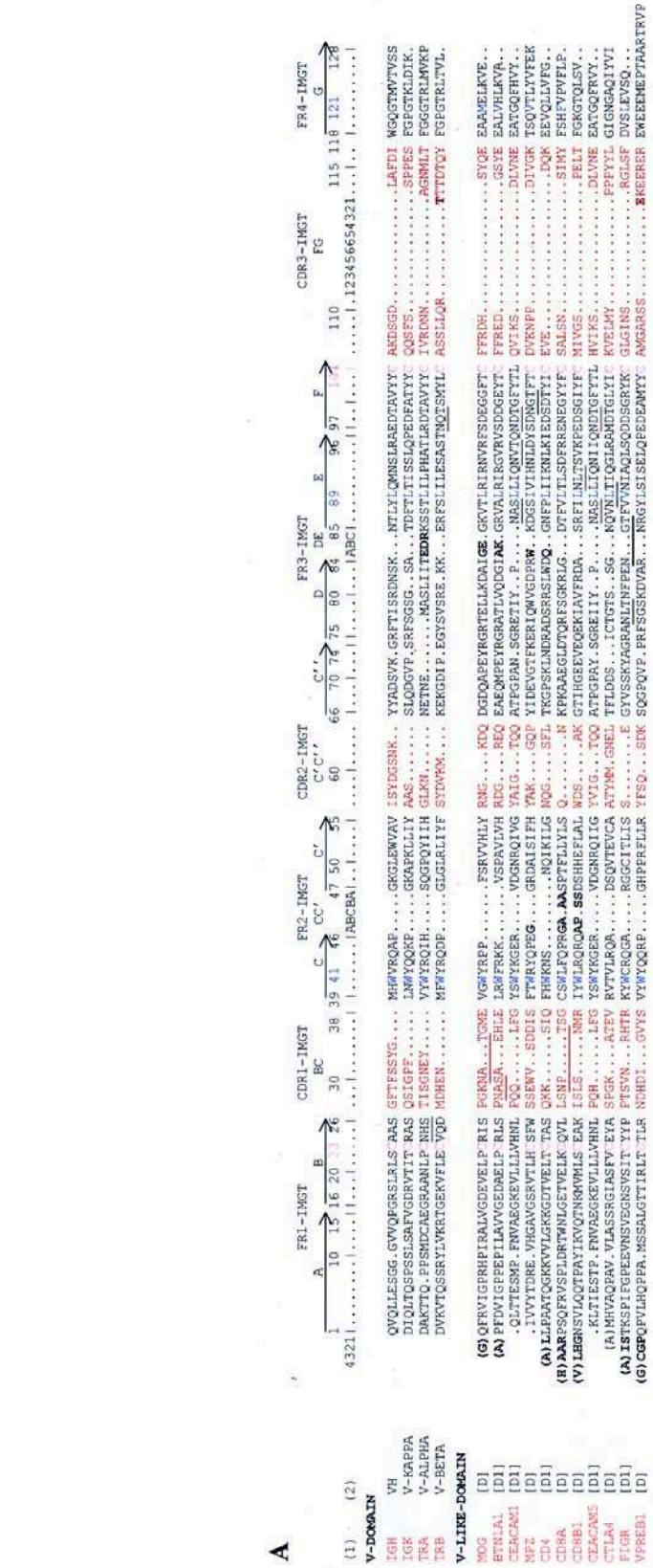

A
FR1-IMGT
CDR1-IMGT
FR2-IMGT
CDR2-IMGT
FR3-IMGT
CDR3-IMGT
FR4-IMGT
V-DOMAIN
IGH VH
IGK V-KAPPA
TRA V-ALPHA
TRB V-BETA
V-LIKE-DOMAIN
MOG [D]
BTN1A1 [D1]
CEACAM1 [D1]
NPS [D]
CD4 [D1]
CD8A [D]
CD8B1 [D]
CEACAM5 [D1]
CTLA4 [D]
FIGR [D1]
VPREB1 [D]
B
C-DOMAIN
IGHG1 CH1
IGLC1 C-LAMBDA1
TRAC C-ALPHA
TRBC2 C-BETA2
C-LIKE-DOMAIN
KIR2DL2 [D1]
KIR2DL1 [D1]
CEACAM1 [D3]
VCAM1 [D1]
CD1A [D3]
CD3E [D]
CD4 [D2]
CEACAM5 [D3]
FCER1A [D1]
FCGR1A [D1]
FCGR2A [D1]

Figure 4. IMGT Protein display. (A) Examples of V-DOMAINs and V-LIKE-DOMAINs (V-set). (B) Examples of C-DOMAINs and C-LIKE-DOMAINs (C-set). #c: rearranged cDNA, g: genomic DNA. Amino acids resulting from a splicing with a preceding exon are shown between parentheses. (A) IG and TR V-DOMAINs: VH (AB027433, #c IGHV3-30-IGHD4-17-IGHJ3), V-KAPPA (AB022654, #c IGKV1-39-IGKJ2), V-ALPHA (AK026255, #c TRAV26-1-TRAJ39), V-BETA (AF043183, #c TRBV28-TRBD1-TRBJ2-3). V-LIKE-DOMAINs: MOG [D] (Z48051, g), BTN1A1 [D1] (U39576, #c, delimitated by homology with MOG [D]), CEACAM1 [D1] (AC004785, g, leader delimited according to [43]), MPZ [D] (D14720, g, MPZ [D] encoded by EX2 and EX3 with (I)53 resulting from the splicing), CD4 [D1] (NT_009759, g, CD4 [D1] encoded by EX2 and EX3 with (G)68 splicing site), CD8A [D] (M27161, g), CD8B1 [D] (M17514, partial g, [44]), CEACAM5 [D1] (M59255, g, leader delimited according to [45]), CTLA4 [D] (AF411058, g), PIGR [D1] (S43441, partial g, limited to EX2; (A) is deduced from S43437), VPREB1 [D] (D86992, g). (B) IG and TR C-DOMAINs: CH1 (J00228, g), C-LAMBDA1 (X51755, g), C-ALPHA (X02883, g), C-BETA2 (M12888, g). C-LIKE-DOMAINs: KIR2DL2 [D1] (AL133414, g), KIR2DL1 [D1] (L41267, #c, delimitated by homology with KIR2DL2 [D1]), CEACAM1 [D3] (AC004785, g), VCAM1 [D1] (M73255, g), CD1A [D3] (M22165, partial g, limited to EX4; (V) is deduced from M22164), CD3E [D] (M23319, M23320 and M23321, partial g, limited to EX4, EX5 and EX6, respectively; (G) is deduced from M23318), CD4 [D2] (NT_009759, g), CEACAM5 (M59257, partial g, limited to EX4; (Y) is deduced from M59256), FCER1A [D1] (L14075, g), FCGR1A [D1] (M63832, partial g, limited to EX3; (D) is deduced from M63831), FCGR2A [D1] (M90723, partial g, limited to EX3; (A) is deduced from M90722). The accession numbers are from IMGT/LIGM-DB (http://imgt.cines.fr) [11,46] for IG and TR, and from EMBL/GenBank/DDBJ [47-49] for IgSF other than IG and TR. Beta strands are shown by arrows. Dots indicate missing amino acids according to the IMGT unique numbering. Putative N-glycosylation sites (N-X-S/T) are underlined.

(1) Gene names (symbols) for IG and TR are according to the IMGT Nomenclature committee (IMGT-NC) [18,19] and the HUGO Nomenclature Committee (HGNC) [50]. Full gene designations are the following: MOG: myelin oligodendrocyte glycoprotein; BTN1A1: butyrophilin, subfamily 1, member A1; CEACAM1: Carcinoembryonic antigen-related cell adhesion molecule 1; MPZ: myelin protein zero (Charcot-Marie-Tooth neuropathy 1B); CD4: CD4 antigen (p55); CD8A: CD8 antigen, alpha polypeptide (p32); CD8B1: CD8 antigen, beta polypeptide 1 (p37); CEACAM5: Carcinoembryonic antigen-related cell adhesion molecule 5; CTLA4: cytotoxic T-lymphocyte-associated protein 4; PIGR: polymeric immunoglobulin receptor; VPREB1: pre-B lymphocyte gene 1; IGHG1: Immunoglobulin heavy constant gamma 1; IGLC1: immunoglobulin lambda constant 1; TRAC: T cell receptor alpha constant; TRBC2: T cell receptor beta constant 2; KIR2DL2: killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 2; VCAM1: vascular cell adhesion molecule 1; CD1A: CD1A antigen, a polypeptide; CD3E: CD3E antigen, epsilon polypeptide (TiT3 complex); CD4: CD4 antigen (p55); FCER1A: Fc fragment of IgE, high affinity I, receptor for alpha polypeptide; FCGR1A: Fc fragment of IgG, high affinity Ia, receptor for (CD64); FCGR2A: Fc fragment of IgG, low affinity IIa, receptor for (CD32).

(2) Domain name. The C-DOMAINs are designated with the IMGT labels (IMGT Scientific chart, http://imgt.cines.fr) The C-LIKE-DOMAINs are designated by the

Figure legend continued

letter D between brackets with a number, corresponding to the position of the domain from the N-terminal end of the protein, and relative to the other domains. There is no number if there is a unique C-LIKE-DOMAIN in the chain.
Amino acid one-letter abbreviation: A (Ala), alanine; C (Cys), cysteine; D (Asp), aspartic acid; E (Glu), glutamic acid; F (Phe), phenylalanine; G (Gly), glycine; H (His), histidine; I (Ileu), isoleucine; K (Lys), lysine; L (Leu), leucine; M (Met), methionine; N (Asn), asparagine; P (Pro), proline; Q (Gln), glutamine; R (Arg), arginine; S (Ser), serine; T (Thr), threonine; V (Val), valine; W (Trp), tryptophan; Y (Tyr), tyrosine.

acids. For C'C" loops shorter than 10 amino acids, gaps are created (missing positions, hatched in IMGT Collier de Perles, or not shown in structural data representations). As an example, in a C'C" loop with 6 amino acids (MOG and MPZ in Figure 3), positions 56 to 58, 63 to 65 are present, whereas positions 59 to 62 are missing. The FG loop (FG-LOOP) comprises position 105 to 117. These positions correspond to a FG loop of 13 amino acids. For FG loops shorter than 13 amino acids, gaps are created from the apex of the loop, in the following order 111, 112, 110, 113, 109, 114, etc (Table 1). For FG loops longer than 13 amino acids (which is rare), additional positions are created, between positions 111 and 112 at the top of the FG loop (Table 1).

1.1.3 Loop length

The loop length (number of codons or amino acids, that is number of occupied positions) is a crucial and original concept of IMGT-ONTOLOGY [13]. The lengths of the BC, C'C" and FG loops characterize the V-LIKE-DOMAINs, as the lengths of the CDR1-IMGT, CDR2-IMGT and CDR3-IMGT characterize the IG and TR V-DOMAINs. Thus, the length of the three loops BC, C'C" and FG is shown, in number of codons (or amino acids), into brackets and separated by dots (Table 2). For examples: *Homo sapiens* MOG [D] [9.6.9] means that in the human MOG [D] domain, the BC, C'C" and FG loops have a length of 9, 6 and 9 codons, respectively; *Homo sapiens* CEACAM1 [D1] [6.7.10] means that in the human CEACAM1 [D1] domain, the BC, C'C" and FG loops have a length of 6, 7 and 10 codons, respectively; *Homo sapiens* MPZ [D] [10.6.11] means that in the human MPZ [D] domain, the BC, C'C" and FG loops have a length of 10, 6 and 11 amino acids, respectively (Figure 3, Table 2).

1.2 Characteristics of the MOG, CEACAM1 and MPZ V-LIKE-DOMAINs

1.2.1 Myelin oligodendrocyte glycoprotein

The myelin oligodendrocyte glycoprotein (MOG) is a specific component of the central nervous system (CNS) localized on the outermost lamellae of

Table 2. Delimitation of the strands and loops for V-LIKE-DOMAINs

		V-LIKE-DOMAIN strands and loops		MOG	BTN1A1	CEACAM1	MPZ	CD4	CD8A	CD8B1	CEACAM5	CTLA4	PIGR	VPREB1
		Numbering	Length	[D] [9.6.9]	[D1] [9.6.9]	[D1] [6.7.10]	[D] [10.6.11]	[D1] [6.6.6]	[D] [7.2.9]	[D] [7.5.9]	[D1] [6.7.10]	[D] [8.9.12]	[D1] [9.2.11]	[D] [9.7.14]
A-STRAND	FR1-IMGT	1.4-1.1	+4	+1	+1			+2	+4	+4			+3	+4
		1-15	15	15	15	13	13	15	15	15	13	14	15	14
B-STRAND		16-26	11	11	11	11	11	11	11	11	11	11	11	11
BC-LOOP	CDR1-IMGT	27-38	12	9	9	6	10	6	7	7	6	8	9	9
C-STRAND		39-46	8	7	7	8	8	6	8	8	8	8	8	8
C'C"-TURN	FR2-IMGT	46A,46B,46C 47B,47A	+5				+1		+4	+4				
C"-STRAND		47-55	9	8	8	9	9	7	9	9	9	9	9	9
C'C"-LOOP	CDR2-IMGT	56-65	10	6	6	7	6	6	2	5	7	9	2	7
C"-STRAND		66-74	9	9	9	8	9	9	9	9	8	6	9	8
D-STRAND		75-84	10	10	10	7	10	10	10	10	7	8	10	10
DE-TURN	FR3-IMGT	84A,84B,84C	+3	+2	+2		+1	+1						
E-STRAND		85-96	12	12	12	11	12	12	12	12	11	12	12	12
F-STRAND		97-104	8	8	8	8	8	8	8	8	8	8	8	8
FG-LOOP	CDR3-IMGT	105-117	13	9	9	10	11	6	9	9	10	12	11	13
		111.1-111.6, 112.6-112.1	+12											+1
G-STRAND	FR4-IMGT	118-128	11	9	9	9	11	9	10	9	9	11	8	16
Total length			128 (+24)	116	116	107	120	108	118	120	107	116	115	126

The delimitations of the strands and loops for the V-LIKE-DOMAINs are identical to those of the IG and TR V-DOMAINs. For more details, see [16]. Lengths of the BC, C'C" and FG loops are shown within brackets. Blank cells indicate no amino acids. A plus sign indicates additional amino acids. Amino acid sequences are shown in Figure 4.

mature myelin [51], and may contribute to myelin maturation and maintenance [52] or as a signal to arrest further myelination [53]. MOG may have an unforeseen immunological status within the central nervous system, providing for instance a rudimentary molecular framework for presentation of pathogens to the immune sytem [54]; MOG is a candidate target antigen for autoimmune-mediated demyelination, and seems to be implicated in pathogenesis of encephalomyelitis and multiple sclerosis, an inflammatory disease of the central nervous system [55,56]. The human MOG gene contains 8 exons [57]. The N-terminal extracellular V-LIKE-DOMAIN of MOG encoded by the exon 2 (EX2) has significant sequence homologies with three non-myelin proteins: *Homo sapiens* butyrophilin (BTN1A1; subfamily 1, member A1) expressed in the mammary gland during lactation and facilitating the interaction between cytoplasmic lipid droplets and the apical membrane [58], *Gallus gallus* B-G antigen encoded by a gene mapping to the major histocompatibility complex [59], and *Gallus gallus* BEN adhesion glycoprotein expressed on the epithelial cells of the bursa of Fabricius and on various neuronal subsets during chicken embryonic development [60]. The IMGT unique numbering allows us to compare *H. sapiens* MOG [D] and *H. sapiens* BTN1A1 [D1] and to describe divergent positions (Figure 3). The MOG and BTN1A1 genes are colocalized near

the human MHC on chromosome 6p21.3-p22 [61]. Two motifs highly homologous to consensus sequences found in glial promoters of proteolipid protein (PLP), protein zero (MPZ), myelin basic protein (MBP), and mouse MOG were also found in human MOG [62]. The *H. sapiens* MOG V-LIKE-DOMAIN has not yet been crystallized, but interestingly PDB has an entry for the *Rattus norvegicus* MOG V-LIKE-DOMAIN in complex with Fab fragment of *Mus musculus* antibody anti-MOG 8-18C5 (1pkq), that allows to identify the MOG FG loop as an important epitope [63,64], and to confirm the N-glycosylated site in N31 (BC-LOOP) (conserved in *H. sapiens* MOG sequence, Figure 3); the MOG FG loop may be implicated in autoimmune recognition [63].

1.2.2 Carcinoembryonic antigen-related cell adhesion molecule 1

Members of the CEA family consist of a single N-terminal V-LIKE-DOMAIN, followed by a variable number of C-LIKE-DOMAINs. Based on sequence similarity and functional characteristics, the CEA family has been subdivided into the CEA subfamily and the pregnancy-specific glycoprotein (PSG) subfamily [65]. Members of the CEA subfamily are anchored in the cell membrane, whereas all of the PSGs appear to be secreted; however, the genes in the CEA and PSG subfamilies have a similar gene structure and organization. The carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1, or biliary glycoprotein BGP) consists of a single N-terminal V-LIKE-DOMAIN followed by 3 C-LIKE-DOMAINs, and is expressed in cells of epithelial and myeloid origin [43]. In granulocytes, CEACAM1 is a main antigen of the CD66 cluster of differentiation antigens that mediate the binding to endothelial E-selectin. The loss or reduced expression of the CEACAM1 adhesion molecule is a major event in colorectal carcinogenesis [66].

1.2.3 Myelin protein zero

The myelin protein zero (MPZ or P0) gene is localized at 1q21.3-q23, about 130 kb of the FCGR2A gene [67]. MPZ is the major structural protein of peripheral myelin, accounting for more than 50% of the protein present in the sheath of peripheral nerves. Expression of the MPZ gene is restricted to Schwann cells; MPZ is not found in the CNS. MPZ corresponds to an integral membrane glycoprotein of 28 kD, and is thought to link adjacent lamellae and thereby stabilize the myelin assembly. The V-LIKE-DOMAIN [D] that is encoded by EX2 and EX3, plays a significant role in myelin membrane adhesion. Several mutations in the MPZ V-LIKE-DOMAIN are associated with the autosomal dominant form of Charcot-Marie-Tooth disease type 1, which is characterized by progressive slowing of nerve conduction and hypertrophy of Schwann cells: the amino acid changes D68>E (C’’-STRAND)

and K74>E (C''-STRAND, near the C''D-TURN) are independently implicated in Charcot-Marie-Tooth disease type 1B (CMT1B) [68], whereas all affected members of another CTM1B family had a 3-bp deletion in EX2 causing loss of the S38 (S38>del; BC-LOOP) [69]. S38>C was found in a 7-year-old boy (heterozygous for the mutation, which was absent in the parents and in 100 unrelated healthy controls) with delayed motor development, hypotonia, muscle weakness, and sensory disturbance thought to be typical of Dejerine-Sottas syndrome, or hereditary motor and sensory neuropathy type III (HMSN3) [68,70]. Partial symptom relief with corticosteroid treatment was reported [71] in a patient with demyelinating CMT1B and a heterozygous R76>H mutation (D-STRAND, near the C''D-TURN). Although this response is rare in such patients, poor myelin compaction by the MPZ protein, caused by the mutation, may have allowed circulating immune elements access to normally sequestered endoneurial components, thus accounting for the response to corticosteroid treatment [71] (OMIM: 159440).

2. IMGT Colliers de Perles for C-LIKE-DOMAIN

The IMGT Colliers de Perles for C-LIKE-DOMAIN is based on the IMGT unique numbering for C-DOMAIN [17]. This numbering is itself derived from the IMGT unique numbering for V-DOMAIN [14-16]. Indeed, the sandwich beta sheet of the C-set (C-DOMAIN and C-LIKE-DOMAIN) has the same topology and 3D structure than the V-set (V-DOMAIN and V-LIKE-DOMAIN), but they differ by the number of strands (Figure 1). The C-LIKE-DOMAIN, as the IG and TR C-DOMAIN, is made of seven beta strands linked by beta turns or loops, and arranged so that four strands form one sheet and three strands form a second sheet. A characteristic CD transversal strand links the two sheets; depending from the CD length, the D strand is in the first or second sheet (shown by an arrow in Figure 1). As shown in Table 3, the IMGT unique numbering for the C-LIKE-DOMAIN follows the same rules as those of the C-DOMAIN [17].

2.1. Strands, loops and turns of the C-LIKE-DOMAINs

Three examples of C-LIKE-DOMAIN are shown in Figure 5: the killer cell immunoglobulin-like receptor KIR2DL2 (two domains, long cytoplasmic tail, 2) [D1], the carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1) [D3] and the vascular cell adhesion molecule 1 (VCAM1) [D1].

2.1.1 Strands

The A strand (A-STRAND, positions 1 to 15) and the B strand (B-STRAND, positions 16 to 26, with the 1st-CYS at position 23) are similar to those of the V-DOMAIN and V-LIKE-DOMAIN [16,17]. The C strand (C-STRAND, positions 39 to 45, with the CONSERVED-TRP at position 41) and the D strand (D-STRAND, positions 77 to 84) are shorter of one position and

Table 3. Delimitation of the strands and loops for C-LIKE-DOMAINs

	C-LIKE-DOMAIN strands and loops		KIR2DL2	KIR2DL1	CEACAM1	VCAM1	CD1A	CD3E	CD4	CEACAM5	FCER1A	FCGR1A	FCGR2A
	Numbering	Length	[D1]	[D1]	[D3]	[D1]	[D3]	[D]	[D2]	[D3]	[D1]	[D1]	[D1]
A-STRAND	1.4-1.1	+4	+2	+2	+2	+2		+3	+5	+2	+3	+1	+2
	1-15	15	15	15	15	15	14	15	15	15	15	15	15
AB-TURN	15.1-15.3	+3	+1	+1		+1			+3			+2	+2
B-STRAND	16-26	11	11	11	11	11	11	11	11	11	11	11	11
BC-LOOP	27-36	10	5	5	6	6	8	4	1	6	7	7	7
C-STRAND	39-45	7	7	7	7	7	7	7	7	7	7	7	7
CD-STRAND	45.1-45.7	+7	+5	+5		+4	+5	+4			+2	+2	+2
D-STRAND	77-84	8	8	8	6	6	8	8		6	5	5	6
DE-TURN	84.1-84.7 85.7-85.1	+14	+3	+3		+4	+7	+4					
E-STRAND	85-96	12	12	12	10	11	10	11	10	10	10	10	9
EF-TURN	96.1, 96.2	+2											
F-STRAND	97-104	8	8	8	8	7	7	8	8	8	8	8	8
FG-LOOP	105-117	13	13	13	13	9	11	13	11	13	7	7	7
	111.1-111.6	+12	+1	+1									
G-STRAND	118-128	11	9	9	7	9	8	6	7	6	10	10	10
Total length		95 (+42)	100	100	85	92	96	94	78	84	85	85	86

Delimitations of the strands and loops for C-LIKE-DOMAINs are identical to those of the IG and TR C-DOMAINs. For more details, see [17]. A plus sign indicates additional amino acids. KIR2DL2 [D1] has one additional position in 15.1, whereas 15.2 and 15.3 are unoccupied. VCAM1 [D1] has four additional positions 84.1, 84.2, 85.2 and 85.1. Amino acid sequences are shown in Figure 4.

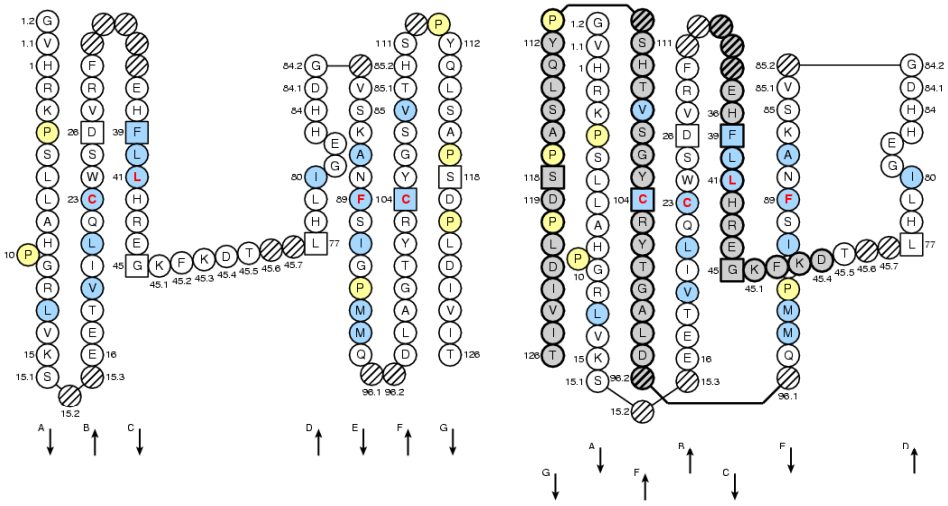
two positions, respectively, compared to the V-DOMAIN and V-LIKE-DOMAIN. As previously described [17], the C' and C" strands are missing and are replaced by the characteristic transversal CD strand (CD-STRAND, positions 45.1 to 45.7). The E strand (E-STRAND, positions 85 to 96, with a conserved hydrophobic amino acid at position 89), the F strand (F-STRAND, positions 97 to 104, with the 2nd-CYS at position 104) and the G strand (G-STRAND, positions 118 to 128, with a conserved hydrophobic amino acid at position 121) are similar to those of the V-DOMAIN and V-LIKE-DOMAIN.

2.1.2 Loops

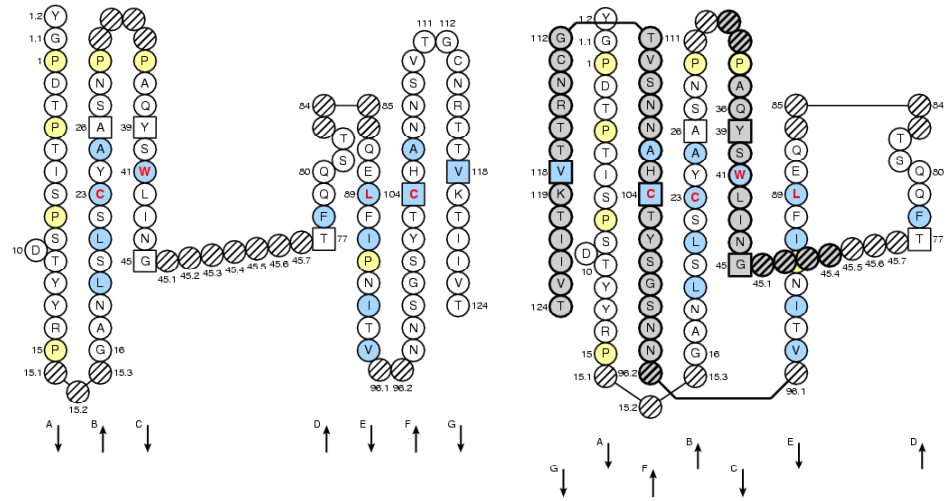
The BC loop (BC-LOOP) comprises positions 27 to 36; the longest BC loops have 10 amino acids, instead of 12 amino acids in the V-DOMAIN and V-LIKE-DOMAIN. For BC loops shorter than 10 amino acids, gaps are created from the apex in the following order 32, 31, 33, 30, 34, etc. As an example, in a BC loop with 5 amino acids (KIR2DL2 [D1] in Figure 5), positions 27 to 29 and 35 and 36 are present, whereas positions 30 to 34 are missing. The FG loop (FG-LOOP) comprises positions 105 to 117 and is similar to that of the V-DOMAIN and V-LIKE-DOMAIN. These positions correspond to a FG loop of 13 amino acids. Gaps for FG loops shorter than 13 amino acids and additional positions for FG loops longer than 13 amino acids, are created following the same rules as those of the V-DOMAIN and V-LIKE-DOMAIN (Table 1). As examples, CEACAM1 [D3] has a FG loop of 13 amino acids, VCAM1 [D] has a FG loop of

Figure 5

A. *Homo sapiens* KIR2DL2 [D1] C-LIKE-DOMAIN



B. *Homo sapiens* CEACAM1 [D3] C-LIKE-DOMAIN



C. *Homo sapiens* VCAM1 [D1] C-LIKE-DOMAIN

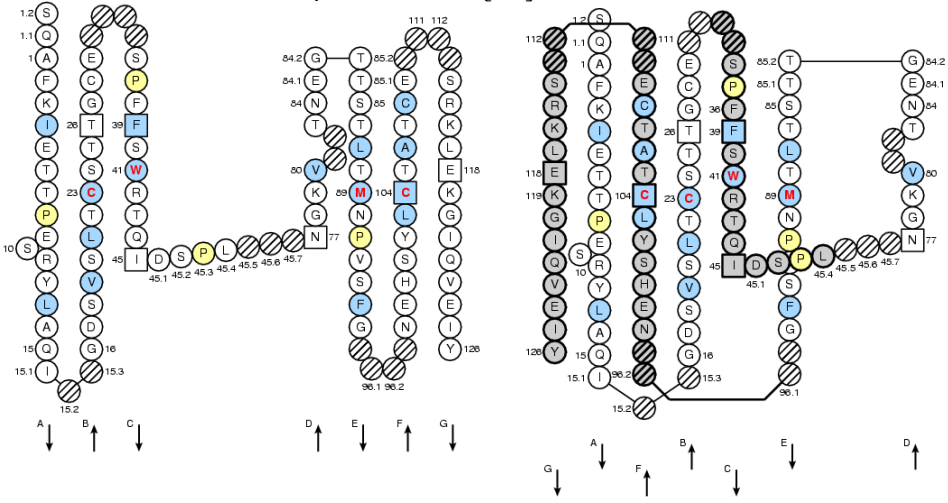


Figure 5. IMGT Colliers de Perles of C-LIKE-DOMAINs on one and two layers. (A) *Homo sapiens* KIR2DL2 [D1], (B) *Homo sapiens* CEACAM1 [D3], (C) *Homo sapiens* VCAM1 [D1]. Amino acids are shown in the one-letter abbreviation. Position at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and tryptophan (W) are found in more than 50% of analysed sequences are shown in blue. All proline (P) are shown in yellow. The positions 26, 39 and 104 are shown in squares by homology with the corresponding positions in the V-set (V-DOMAINs and V-LIKE-DOMAINs). Positions 45 and 77 which delimit the characteristic CD strand of the C-set (C-DOMAINs and C-LIKE-DOMAINs), and position 118 which corresponds structurally to J-PHE or J-TRP of the IG and TR J-REGION [16,17], are also shown in squares. Hatched circles correspond to missing positions according to the IMGT unique numbering for C-DOMAINs and C-LIKE-DOMAINs. Arrows indicate the direction of the beta strands and their different designations in 3D structures (from IMGT Repertoire, http://imgt.cines.fr). The IMGT Colliers de Perles on two layers (on the right hand) show, on the forefront, the GFC strands and, on the back, the ABE strands. Swiss-Prot accession numbers: P43627 for the *H. sapiens* KIR2DL2 protein, P13688 for the *H. sapiens* CEACAM1 protein, and P19320 for the *H. sapiens* VCAM1 protein. The IMGT Colliers de Perles were checked with the following PDB [42] entries: 1efx_D (*H. sapiens* KIR2DL2 [D1]), 1l6z_A (*Mus musculus* CEACAM1 [D4], sequence similar to the *H. sapiens* CEACAM1 [D3]; no available 3D structure of the *H. sapiens* CEACAM1 [D3]), 1vsc_A (*H. sapiens* VCAM1 [D1]).

9 amino acids (with four gaps 110 to 113), whereas KIR2DL2 [D1] has a FG loop of 14 amino acids with the additional position 112.1 (Figure 5).

2.1.3 Turns

The AB turn (AB-TURN) corresponds to additional positions 15.1, 15.2 and 15.3; the longest AB turns have 3 amino acids. For AB turns shorter than 3 amino acids, gaps are created (missing positions, hatched in the IMGT Colliers de Perles (Fig. 5), or not shown in structural data representations) in an ordinal manner. As an example, KIR2DL2 [D1] has one additional position in 15.1, whereas 15.2 and 15.3 are unoccupied. The DE turn (DE-TURN) comprises positions 84.1 to 84.7 and 85.7 to 85.1, corresponding to 14 amino acids. For DE turns shorter than 14 amino acids, gaps are created in the following order: 85.1, 84.1, 85.2, 84.2, 85.3, 84.3, etc. As an example, VCAM1 [D1] has four additional positions 84.1, 84.2, 85.2 and 85.1. The EF turn (EF-TURN) corresponds to additional positions 96.1 and 96.2 when present, corresponding to 2 amino acids. For EF turns shorter than 2 amino acids, gaps are created in the following order: 96.2, 96.1.

2.2 Characteristics of the KIR2DL2 and VCAM1 C-LIKE-DOMAINs

2.2.1 KIR2DL2

The second killer cell immunoglobulin-like receptor with two domains and a long cytoplasmic tail (KIR2DL2) corresponds to a 348 amino acid type I

transmembrane protein [72,73]. Sequence analysis revealed a structure similar to that described for KIR2DL1, with two extracellular C-LIKE-DOMAINs ([D1] and [D2]), a transmembrane domain, and a long cytoplasmic tail with two immunoreceptor tyrosine-based inhibitory motifs (ITIMs). According to the IMGT unique numbering, divergent positions between KIR2DL2 and KIR2DL1 are localized in the A-STRAND (R12>P), BC-LOOP (R28>M), CD-STRAND (K45.1>M, K45.3>N), D-STRAND (H78>R), E-STRAND (G92>S, P93>R, M95>T), FG-LOOP (L114>V) and G-STRAND (T126>I). KIR2D receptors are divided into two families based on their specificities for different HLA-C allotypes: the KIR2DL1 is specific for HLA-Cw2, 4, 6 and 15, whereas KIR2DL2 is specific for HLA-Cw1, 3, 7 and 8. KIR2DL2/HLA-Cw3 and KIR2DL1/HLA-Cw4 share a common binding mode [74], and a single K45.1>M amino acid change between KIR2DL2 and KIR2DL1 is sufficient to switch allotype specificity [75].

2.2.2 VCAM1

The vascular cell adhesion molecule 1 (VCAM1) is expressed by cytokine-activated endothelium, binds leukocyte integrins and is involved in inflammatory and immune functions [76,77]. This type I membrane protein mediates leukocyte-endothelial cell adhesion and signal transduction, and may play a role in the development of arteriosclerosis and rheumatoid arthritis. VCAM1 is present in single copy in the human genome and contains 9 exons spanning about 25 kb of DNA. At least 2 different VCAM1 precursors can be generated from the human gene as a result of alternative mRNA splicing events, which include or exclude exon 5 [76]. The major form is composed of seven C-LIKE-DOMAINs, of which domains [D1], [D2] and [D3] are strikingly homologous in both structure and function to [D4], [D5] and [D6] domains [77,78]. The functionally important [D1] domain essential for binding to the integrin ligand contains two rather than one pair of cysteine residues; 1st-CYS 23 and 2nd-CYS 104 correspond to the core disulfide bond of the domain between the B and the F strands, whereas C28 and C108 form an additional disulfide bond between the BC and the FG loops. Mutagenesis studies directed at [D1] have identified two sets of residues involved in binding [79,80]. According to the IMGT unique numbering, these amino acids are D45.1 and P45.3 (CD-STRAND), and G95 (EF-LOOP). P45.3 appears to be particularly important, since its limited conformational freedom brings the Cα atoms of T43 (C-STRAND) and L45.4 (CD-STRAND) within 7 Å of each other [81]. G95 is located in the EF loop, in close proximity of the CD loop (Figure 5). It might interact directly with the integrin ligand, or it might play an indirect role by stabilizing the structure of the CD loop. There is an extensive network of hydrogen bonds between the CD and EF loops, some of which involve the side chain of H100. A cyclic peptide that mimics the CD loop inhibits binding of a4b1 integrin-bearing cells to VCAM1. [D2] may have a role in ligand binding [81].

3. IMGT Protein displays

A comparison between the V-set (V-DOMAIN and V-LIKE-DOMAIN) and C-set (C-DOMAIN and C-LIKE-DOMAIN) domains is shown in the IMGT Protein display (Figure 4). Amino acid positions shown on the upper line in the IMGT Protein displays correspond to equivalent positions in both sets, whereas amino acid positions on the lower lines are characteristic of each set. Sixty-five positions are structurally equivalent between the V-set and the C-set, when the strands A to G are compared. They comprise: positions 1-15 (A strand), 16-26 (B strand), 39-45 (C strand), 77-84 (D strand), 85-96 (E strand), 97-104 (F strand), 118 up to at least 121 (G strand). Thirty-five positions are characteristic of the C-set numbering. The positions of these additional positions compared to the V-set numbering are designated by a number followed by a dot and a number: 1.1 to 1.9 (at the N-terminal end), 15.1 to 15.3 (at the AB turn, for example 15.1 in KIR2DL2 [D1], KIR2DL1 [D1] and VCAM1 [D1]), 45.1 to 45.7 (that represent the CD transversal strand), 84.1 to 84.7, 85.7 to 85.1 (at the DE loop; these positions correspond to longer antiparallel D and E strands in the C-set), 96.1 and 96.2 (at the EF turn). Thirty-three positions are missing in the C-set, compared to the V-set. Two of these positions (37 and 38) are missing in the BC loop. The thirty-one other positions (46 to 76) correspond to the two C' and C" strands and to the C'C" loop (CDR2-IMGT), present in the V-set but absent in the C-set. Positions 45.1 to 45.7 are structurally different between the C-set and the V-set [17]. Indeed, as described above, these positions represent a transversal strand between C and D in the C-set, whereas there are two additional C' and C" strands in the V-set.

In Figure 4, amino acids which result from the splicing with the preceding exon are shown within parentheses. Indeed, the exact delimitations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs can be usually identified when genomic sequences are available [35]. Each domain shown in Figure 4, except CD4 [D1] and MPZ [D], is encoded by a unique exon.

4. IMGT Alignments of alleles

The IMGT unique numbering allows a standardized description of allele polymorphisms and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs. Alleles from the human FCGR3B have recently been described based on these criteria [37]. The mutations and allelic polymorphisms are described per domain and by comparison to the allele *01 from the IMGT reference directory. Based on these criteria, IMGT 'Alignments of alleles' [11,12] allow a standardized display according to the IMGT unique numbering and with the strand and loop delimitations [35]. Other features of the amino acid sequences, such as positions of the N-glycosylation site and amino acids involved in ligand-receptor interactions can easily be visualized.

Conclusion

The IMGT unique numbering gives insight in the structural configuration of the V-LIKE-DOMAINs and C-LIKE-DOMAINs belonging to the human IgSF proteins, but also opens interesting views on the evolution of the sequences of the V-set and C-set [16,17]. Indeed, the IMGT unique numbering can be applied to any IgSF V-set or C-set domain. In the absence of available 3D structures, the V-LIKE-DOMAIN and C-LIKE-DOMAIN IMGT Colliers de Perles are particulary useful for comparison with domains of known 3D structures.

The IMGT unique numbering has many advantages. It allows an easy comparison between sequences coding the V-LIKE-DOMAINs and C-LIKE-DOMAINs, whatever the IgSF protein, the chain type or the species. It has allowed to show that the distinction between C1 and C2 becomes unnecessary (discussed in [17]). It allows, by comparison with genomic sequences, to delimit precisely the V-LIKE-DOMAINs and C-LIKE-DOMAINs. Moreover, it allows to determine the lengths of the BC, C'C" and FG loops of the V-LIKE-DOMAINs and those of the BC loop, CD transversal strand and FG loop of the C-LIKE-DOMAINs. The strand and loop lengths (number of codons or of amino acids, that is number of occupied positions) become crucial information characterizing the V-set and C-set domains, and the corresponding genes, cDNAs and proteins [16,17]. IMGT quality assessment of the data is performed at both the sequence level and 3D structure level. Indeed, the delimitations of the domains are based on the location of the splicing sites in genomic sequences. The IMGT unique numbering has allowed standardized analysis and representations of nucleotide and amino acid sequences (Tables of FR and CDR lengths, Tables of alleles, Alignments of alleles, IMGT Protein displays, IMGT Colliers de Perles, 3D structures). The IMGT unique numbering represents, therefore, a major step forward in analysing and comparing the structure and evolution of the proteins belonging to the immunoglobulin superfamily.

Acknowledgements

We are grateful to Géraldine Folch, Chantal Ginestoux, Joumana Jabado-Michaloud, Dominique Scaviner, Vincent Nègre, Oliver Clément and Denys Chaume for their helpful discussions. E.D. is holder of a doctoral grant from the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR). K.Q. was the recipient of a doctoral grant from the MENESR and is currently supported by a grant from the Association pour la Recherche sur le Cancer (ARC). IMGT is a registered CNRS mark. IMGT is a RIO platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287) programmes of the European Union and received subventions from ARC and from the Génopole-Montpellier-

Languedoc-Roussillon. IMGT is currently supported by the Centre National de la Recherche Scientifique (CNRS), the MENESR (Université Montpellier II Plan Pluri-Formation, BIOSTIC-LR2004 Région Languedoc-Roussillon and ACI-IMPBIO IMP82-2004). Part of this work was carried out in the frame of the European Science Foundation Scientific Network Myelin Structure and its role in autoimmunity (MARIE).

References

1. Lefranc M-P, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, Scaviner D, Ginestoux C, Clément O, Chaume D and Lefranc G. IMGT, the international ImMunoGeneTics information system ®. Nucl. Acids Res. 2005, 33: D593-D597.
2. Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, Bontrop R, Lemaître M, Malik A and Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl. Acids Res. 1997; 25:206-211.
3. Lefranc M-P, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, Lemaître M, Malik A and Chaume D. IMGT, the international ImMunoGeneTics database. Nucl. Acids Res. 1998;26: 297-303.
4. Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, Lemaître M, Malik A, Barbié V and Chaume D. IMGT, the international ImMunoGeneTics database. Nucl. Acids Res. 1999;27: 209-212.
5. Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, Marsh S, Bontrop R, Lemaître M, Lefranc G, Chaume D and Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl. Acids Res., 2000;28: 219-221.
6. Lefranc M-P. IMGT ImMunoGeneTics Database. International BIOforum, 2000, 4, 98-100.
7. Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl. Acids Res. 2001; 29:207-209.
8. Lefranc M-P. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev. Comp. Immunol., 2002, 26: 697-705.
9. 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.
10. Lefranc M-P. IMGT-ONTOLOGY and IMGT databases, tools and web resources for immunogenetics and immunoinformatics. Mol. Immunol. 2004; 40:647-659.
11. Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviener 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, http://imgt.cines.fr. Epub 2003; 4:0004 http://www.bioinfo.de/isb/2003/04/0004/, 22 November 2003. *In Silico* Biol. 2004; 4:17-29.
12. Lefranc M-P, Clément 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. 2005, 0006. http://www.bioinfo.de/isb/2004/05/0006.

13. Giudicelli V and Lefranc M-P. Ontology for Immunogenetics: IMGT-ONTOLOGY. Bioinformatics 1999;12:1047-1054.
14. Lefranc M-P. Unique database numbering system for immunogenetic analysis. Immunol. Today 1997;18:509.
15. Lefranc M-P. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 1999;7:132-136.
16. Lefranc M-P, Pommié 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., 2002, 27: 55-77.
17. Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piédade 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. 2005, 29:185-203.
18. Lefranc M-P and Lefranc G. The Immunoglobulin FactsBook. London, UK: Academic Press, 458 pages, ISBN:012441351X, 2001.
19. Lefranc M-P and Lefranc G. The T cell receptor FactsBook. London, UK: Academic Press, 398 pages, ISBN:0124413528, 2001.
20. Pallarès N, Frippiat JP, Giudicelli V and Lefranc M-P. The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp. Clin. Immunogenet. 1998;15:8-18.
21. Barbié V and Lefranc M-P. The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp. Clin. Immunogenet. 1998;15:171-183.
22. Pallarès N, Lefebvre S, Contet V, Matsuda F and Lefranc M-P. The human immunoglobulin heavy variable (IGHV) genes. Exp. Clin. Immunogenet. 1999;16:36-60.
23. 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. 1999;16:234-240.
24. Folch G, Scaviner D, Contet V and Lefranc M-P. Protein displays of the human T cell Receptor alpha, beta, gamma and delta variable and joining regions. Exp. Clin. Immunogenet. 2000; 17:205-215.
25. Folch G and Lefranc M-P. The human T cell receptor beta variable (TRBV) genes. Exp. Clin. Immunogenet. 2000;17:42-54.
26. Scaviner D and Lefranc M-P. The human T cell receptor alpha variable (TRAV) genes. Exp. Clin. Immunogenet. 2000;17:83-96.
27. 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. 2000;17:216-228.
28. Artero S and Lefranc M-P. The Teleostei immunoglobulin heavy IGH genes. Exp. Clin. Immunogenet. 2000;17:148-161.
29. Artero S and Lefranc M-P: The Teleostei immunoglobulin light IGL1 and IGL2 V, J and C genes. Exp. Clin. Immunogenet. 2000;17:162-172.
30. Bosc N, Contet V and Lefranc M-P. The mouse (*Mus musculus*) T cell Receptor Delta Variable (TRDV), Diversity (TRDD), and Joining (TRDJ) Genes. Exp. Clin. Immunogenet. 2001;18:51-58.

31. Lefranc M-P. Nomenclature of the human immunoglobulin heavy (IGH) genes. Exp. Clin. Immunogenet. 2001; 18:100-116.
32. Lefranc M-P. Nomenclature of the human immunoglobulin kappa (IGK) genes. Exp. Clin. Immunogenet. 2001;18:161-174.
33. Lefranc M-P. Nomenclature of the human immunoglobulin lambda (IGL) genes. Exp. Clin. Immunogenet. 2001;18:242-254.
34. Ruiz M and Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulin with known 3D structures. Immunogenetics 2002;53:857-883.
35. Bosc N and Lefranc M-P. The mouse (Mus musculus) T cell receptor alpha (TRA) and delta (TRD) variable genes. Dev. Comp. Immunol. 2003; 27:465-97.
36. Pommié 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. Recognition 2004; 17: 17-32.
37. Bertrand G, Duprat E, Lefranc M-P, Marti J and Coste J. Characterization of human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. Tissue Antigens 2004; 64:119-31.
38. Williams AF and Barclay AN. The immunoglobulin superfamily-domains for cell surface recognition. Annu Rev Immunol 1988;6:381-405.
39. Bork P, Holm L and Sander C. The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol 1994;242:309-320.
40. Hunkapiller T and Hood L. Diversity of the immunoglobulin gene superfamily. Adv. Immunol. 1989, 44:1-63.
41. Jones EY. The immunoglobulin superfamily. Curr. Opin. Struct. Biol. 1993; 3:846-852.
42. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE. The Protein Data Bank. Nucl. Acids Res. 2000; 28:235-242.
43. Hinoda Y, Neumaier M, Hefta SA, Drzeniek Z, Wagener C, Shively L, Hefta LJ, Shively JE and Paxton RJ. Molecular cloning of a cDNA coding biliary glycoprotein I: primary structure of a glycoprotein immunologically crossreactive with carcinoembryonic antigen. Proc. Natl. Acad. Sci. 1988; 85(18):6959-6963.
44. DiSanto JP, Smith D, De Bruin D, Lacy E and Flomenberg N. Transcriptional diversity at the duplicated human CD8 beta loci. Eur. J. Immunol. 1993; 23(2):320-326.
45. Schrewe H, Thompson J, Bona M, Hefta LJ, Maruya A, Hassauer M, Shively JE, von Kleist S and Zimmermann W. Cloning of the complete gene for carcinoembryonic antigen: analysis of its promoter indicates a region conveying cell type-specific expression. Mol. Cell Biol. 1990; 10(6):2738-2748.
46. Chaume D, Giudicelli V and Lefranc M-P. IMGT/LIGM-DB. In: The Molecular Biology Database Collection. Nucl Acids Res 2004;32. http://www3.oup.co.uk/ nar/ database/summary/504
47. Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, Van den Broek A, Cochrane G, Duggan K, Eberhardt R, Faruque N, Garcia-Pastor M, Harte N, Kanz C, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Stoehr P, Stoesser G, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W and Apweiler R. The EMBL Nucleotide Sequence Database. Nucl Acids Res 2004;32:D27-D30.

48. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL. GenBank: update. Nucl Acids Res 2004;32:D23-D26.
49. Miyazaki S, Sugawara H, Ikeo K, Gojobori T and Tateno Y. DDBJ in the stream of various biological data. Nucl Acids Res 2004;32:D31-D34.
50. Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW and Povey S. Guidelines for human gene nomenclature. Genomics 2002;79:464-470.
51. Brunner C, Lassmann H, Waehneldt TV, Matthieu JM and Linington C. Differential ultrastructural localization of myelin basic protein, myelin/oligodendroglial glycoprotein, and 2',3'-cyclic nucleotide 3'-phosphodiesterase in the CNS of adult rats. J. Neurochem. 1989; 52:296-304.
52. Birling MC, Roussel G, Nussbaum F and Nussbaum JL. Biochemical and immunohistochemical studies with specific polyclonal antibodies directed against bovine myelin/oligodendrocyte glycoprotein. Neurochem. Res. 1993; 18:937-45.
53. Gardinier MV, Amiguet P, Linington C and Matthieu JM. Myelin/Oligodendrocyte glycoprotein is a unique member of the immunoglobulin superfamily. J. Neurosci. Res. 1992; 33:177-187.
54. Steinman L. Connections between the immune system and the nervous system. Proc. Nat. Acad. Sci. 1993; 90:7912-7914.
55. Sun J, Link H, Olsson T, Xiao BG, Andersson G, Ekre HP, Linington C and Diener P. T and B cell responses to myelin-oligodendrocyte glycoprotein in multiple sclerosis. J. Immunol. 1991; 146:1490-1495.
56. Kerlero de Rosbo N, Milo R, Lees MB, Burger D, Bernard CCA and Ben-Nun A. Reactivity to myelin antigens in multiple sclerosis: peripheral blood lymphocytes respond predominantly to myelin oligodendrocyte glycoprotein. J. Clin. Invest. 1993; 92:2602-2608.
57. Pham-Dinh D, Della Gaspera B, Kerlero de Rosbo N and Dautigny A. Structure of the human myelin/oligodendrocyte glycoprotein gene and multiple alternative spliced isoforms. Genomics 1995; 29:345-352.
58. Jack LJW and Mather IH. Cloning and analysis of cDNA encoding bovine butyrophilin, an apical glycoprotein expressed in mammary tissue and secreted in association with the milk-fat globule membrane during lactation. J. Biol. Chem. 1990; 265:14481-14486.
59. Kaufman J and Salomonsen J. B-G: We know what it is, but what does it do? Immunol. Today 1992; 13:1-3.
60. Pourquié O, Corbel C, Le Caer JP, Rossier J and Le Douarin N. BEN, a surface glycoprotein of the immunoglobulin superfamily, is expressed in a variety of developing systems. Proc. Nat. Acad. Sci. 1992; 89:5261-5265.
61. Pham-Dinh D, Mattei M-G, Nussbaum J-L, Roussel G, Pontarotti P, Roeckel N, Mather IH, Artzt K, Lindahl KF and Dautigny A. Proc. Nat. Acad. Sci. 1993; 90:7990-7994.
62. Roth M-P, Malfroy L, Offer C, Sevin J, Enault G, Borot N, Pontarroti P and Coppin H. The human myelin oligodendrocyte glycoprotein (MOG) gene: complete nucleotide sequence and structural characterization. Genomics 1995; 28:241-250.
63. Breithaupt C, Schubart A, Zander H, Skerra A, Huber R, Linington C and Jacob U. Structural insights into the antigenicity of myelin oligodendrocyte glycoprotein. Proc. Natl. Acad. Sci. 2003; 100(16):9446-9451.

64. Clements CS, Reid HH, Beddoe T, Tynan FE, Perugini MA, Johns TG, Bernard CC and Rossjohn J. The crystal structure of myelin oligodendrocyte glycoprotein, a key autoantigen in multiple sclerosis. Proc. Natl. Acad. Sci. 2003; 100(19):11059-11064.
65. Thompson JA, Grunert F and Zimmermann W. Carcinoembryonic antigen gene family: molecular biology and clinical perspectives. J. Clin. Lab. Anal. 1991; 5:344-366.
66. Neumaier M, Paululat S, Chan A, Matthaes P and Wagener C. Biliary glycoprotein, a potential human cell adhesion molecule, is down-regulated in colorectal carcinomas. Proc. Nat. Acad. Sci. 1993; 90:10744-10748.
67. Pham-Dinh D, Fourbil Y, Blanquet F, Mattei M-G, Roeckel N, Latour P, Chazot G, Vandenberghe A and Dautigny A. The major peripheral myelin protein zero gene: structure and localization in the cluster of Fc-gamma receptor genes on human chromosome 1q21.3-q23. Hum. Molec. Genet. 1993; 2:2051-2054.
68. Hayasaka K, Himuro M, Wang Y, Takata M, Minoshima S, Shimizu N, Miura M, Uyemura K and Takada G. Structure and chromosomal localization of the gene encoding the human myelin protein zero (MPZ). Genomics 1993; 17:755-758.
69. Kulkens T, Bolhuis PA, Wolterman RA, Kemp S, te Nijenhuis S, Valentijn LJ, Hensels GW, Jennekens FGI, de Visser M, Hoogendijk JE and Baas F. Deletion of the serine 34 codon from the major peripheral myelin protein P(0) gene in Charcot-Marie-Tooth disease type 1B. Nature Genet. 1993; 5:35-39.
70. Tachi N, Ishikawa Y and Minami R. Two cases of congenital hypomyelination neuropathy. Brain Dev. 1984; 6:560-565.
71. Watanabe M, Yamamoto N, Ohkoshi N, Nagata H, Kohno Y, Hayashi A, Tamaoka A and Shoji S. Corticosteroid-responsive asymmetric neuropathy with a myelin protein zero gene mutation. Neurology 2002; 59:767-769.
72. Wagtmann N, Biassoni R, Cantoni C, Verdiani S, Malnati MS, Vitale M, Bottino C, Moretta L, Moretta A and Long EO. Molecular clones of the p58 NK cell receptor reveal immunoglobulin-related molecules with diversity in both the extra- and intracellular domains. Immunity 1995; 2:439-449.
73. Dohring C, Samaridis J and Colonna M. Alternatively spliced forms of human killer inhibitory receptors. Immunogenetics 1996; 44:227-230.
74. Boyington JC and Sun PD. A structural perspective on MHC class I recognition by killer cell immunoglobulin-like receptors. Mol. Immunol. 2001; 38:1007-1021.
75. Winter CC and Long EO. A single amino acid in the p58 killer cell inhibitory receptor controls the ability of natural killer cells to discriminate between the two groups of HLA-C allotypes. J. Immunol. 1997; 158:4026-4028.
76. Cybulsky MI, Fries JWU, Williams AJ, Sultan P, Eddy R, Byers M, Shows T, Gimbrone MA and Collins T. Gene structure, chromosomal location, and basis for alternative mRNA splicing of the human VCAM1 gene. Proc. Nat. Acad. Sci. 1991; 88:7859-7863.
77. Osborn L, Hession C, Tizard R, Vassallo C, Luhowskyj S, Chi-Rosso G and Lobb R. Direct expression cloning of vascular cell adhesion molecule 1, a cytokine-induced endothelial protein that binds to lymphocytes. Cell 1989; 59:1203-1211.
78. Hession C, Tizard R, Vassallo C, Schiffer SB, Goff D, Moy P, Chi-Rosso G, Luhowskyj S, Lobb R and Osborn L. Cloning of an alternate form of vascular cell adhesion molecule-1 (VCAM1). J. Biol. Chem. 1991; 266:6682-6685.

79. Osborn L, Vassallo C, Browning BG, Tizard R, Haskard DO, Benjamin CD, Dougas I and Kirchhausen T. Arrangement of domains, and amino acid residues required for binding of vascular cell adhesion molecule-1 to its counter-receptor VLA-4 (alpha 4 beta 1). J. Cell Biol. 1994; 124:601-608.
80. Vonderheide RH, Tedder TF, Springer TA and Staunton DE. Residues within a conserved amino acid motif of domains 1 and 4 of VCAM-1 are required for binding to VLA-4. J. Cell Biol. 1994; 125:215-22.
81. Wang JH, Pepinsky RB, Stehle T, Liu JH, Karpusas M, Browning B and Osborn L. The crystal structure of an N-terminal two-domain fragment of vascular cell adhesion molecule 1 (VCAM-1): a cyclic peptide based on the domain 1 C-D loop can inhibit VCAM-1-alpha 4 integrin interaction. Proc. Nat. Acad. Sci. 1995; 92:5714-5718.

Publication 4

BERTRAND, G.[*], **DUPRAT, E.**[*], LEFRANC, M.-P., MARTI, J. and COSTE, J. (2004) Human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. *Tissue Antigens*, 64, 2, 119-131.

[*] contribution équivalente

Publication 5

KAAS, Q., **DUPRAT, E.**, TOURNEUR, G. and LEFRANC, M.-P. (2005) IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: *Immunoinformatics* (Brusic V, Schoenbach C eds.) Springer, The Netherlands (in press).

Publication 6

DUPRAT, E., LEFRANC, M.-P. and GASCUEL, O. (2005) Prédire l'interaction des protéines de la superfamille du MHC avec la beta2-microglobuline en combinant classifieur Bayesien « naïf » et alignement multiple IMGT. *Actes des Journées Ouvertes Biologie Informatique Mathématiques 2005*.

Prédire l'interaction des protéines de la superfamille du MHC avec la beta2-microglobuline en combinant classifieur Bayesien "naïf" et alignement multiple IMGT

Elodie Duprat[1], Marie-Paule Lefranc[1,2] et Olivier Gascuel[3]

[1] IMGT, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier II, Institut de Génétique Humaine, IGH, UPR CNRS 1142, 141 rue de la Cardonille, 34396 MONTPELLIER Cedex 5, France
duprat@ligm.igh.cnrs.fr
[2] Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 PARIS, France
lefranc@ligm.igh.cnrs.fr
[3] Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM, UMR CNRS 5506, Université Montpellier II, 161 rue Ada, 34392 MONTPELLIER, France
gascuel@lirmm.fr

Résumé: *Les protéines du complexe majeur d'histocompatibilité (MHC) assurent une fonction essentielle au sein du système immunitaire, en présentant des peptides du soi ou du non soi aux récepteurs T. La liaison non covalente de la beta2-microglobuline (B2M) aux protéines du MHC de classe I (MHC-I) est nécessaire à leur expression à la surface cellulaire et à la présentation des peptides. La superfamille du MHC (MhcSF) regroupe les protéines du MHC mais aussi les protéines de structure similaire aux protéines MHC-I. Ces protéines MHC-I-like sont impliquées dans une grande variété de processus biologiques et interagissent ou non avec la B2M. La prédiction de la liaison (ou de l'absence de liaison) à la B2M, pour des protéines de la MhcSF nouvellement identifiées, permettrait d'une part d'indiquer leur mécanisme de reconnaissance moléculaire, et d'autre part de détecter des mutants pathologiques dont l'expression à la surface cellulaire est affectée. La description standardisée des domaines protéiques et la méthode d'alignement (pour partie structurale) mises en place au sein d'IMGT pour le MHC, s'appliquent avec succès aux protéines MHC-I-like malgré leur faible similarité de séquence. La standardisation IMGT fournit ainsi une numérotation unique des résidus qui favorise le développement d'une telle approche prédictive. La méthode proposée combine un classifieur Bayesien dit naïf et la numérotation unique IMGT pour les G-DOMAINs et G-LIKE-DOMAINs, observés pour toutes les protéines de la MhcSF. Cette méthode comprend deux étapes: la sélection d'un ensemble de descripteurs binaires discriminants (qui associent une position dans l'alignement et un groupe d'acides aminés) à partir des données, et la construction du classifieur par estimation des fréquences de ces descripteurs conditionnellement aux classes que l'on cherche à séparer. Le jeu de données est composé de 806 séquences alignées, qui correspondent à des formes alléliques de 47 protéines de la MhcSF appartenant à 9 types de récepteurs et à 4 espèces. 18 descripteurs sont sélectionnés pour leur capacité à discriminer les protéines selon qu'elles se lient ou non à la B2M. L'analyse structurale des protéines de la MhcSF montre que ces descripteurs correspondent à des sites potentiels de contact à la B2M ou à des sites impliqués dans le maintien d'une conformation favorable au contact. La performance du classifieur est évaluée par la procédure de "leave-one-out", déclinée en 3 modes qui distinguent les cas où la prédiction concerne une nouvelle protéine, une espèce non référencée au sein des données ou un nouveau type de récepteur, avec respectivement 98%, 94% et 70% de succès. Ces taux élevés de bonne prédiction mettent en évidence l'efficacité de l'approche proposée, qui devrait trouver des applications dans d'autres problématiques biologiques. Les séquences alignées des protéines de la MhcSF sont accessibles dans les sections MHC et RPI d'IMGT Repertoire; les structures 3D analysées sont accessibles sous forme de fichiers de coordonnées annotés dans IMGT/3Dstructure-DB (http://imgt.cines.fr).*

Mots clés: Classifieur Bayesien naïf, numérotation unique IMGT, validation croisée, complexe majeur d'histocompatibilité, superfamille du MHC, beta2-microglobuline, G-DOMAIN, G-LIKE-DOMAIN

1 Introduction

Les protéines du complexe majeur d'histocompatibilité (MHC) assurent une fonction essentielle au sein du système immunitaire, en présentant des peptides du soi ou du non soi aux récepteurs T. La liaison non covalente de la beta2-microglobuline (B2M) à la chaîne lourde transmembranaire (I-ALPHA) des protéines du MHC de classe I (MHC-I) est nécessaire à la présentation des peptides [1-3], à la stabilisation de la

conformation du complexe [4-6] et à son expression à la surface cellulaire [7-9]. La superfamille du MHC (MhcSF) [10] regroupe les protéines du MHC mais aussi les protéines de structure similaire aux protéines MHC-I. Ces protéines MHC-I-like sont impliquées dans une grande variété de processus biologiques qui correspondent à différents sites d'interaction protéine-ligand sur la chaîne lourde (I-ALPHA-LIKE). D'après la littérature, 34 protéines MHC-I-like de mammifères ont été identifiées à ce jour; les structures 3D sont connues pour 12 d'entre elles. Les données expérimentales montrent que 17 de ces protéines se lient à la B2M, tandis que les 17 autres ne s'y lient pas. La prédiction automatique de la liaison (ou de l'absence de liaison) à la B2M, pour des protéines de la MhcSF nouvellement identifiées, permettrait d'une part d'indiquer leur mécanisme de reconnaissance moléculaire [11,12], et d'autre part de détecter des mutants pathologiques dont l'expression à la surface cellulaire est affectée.

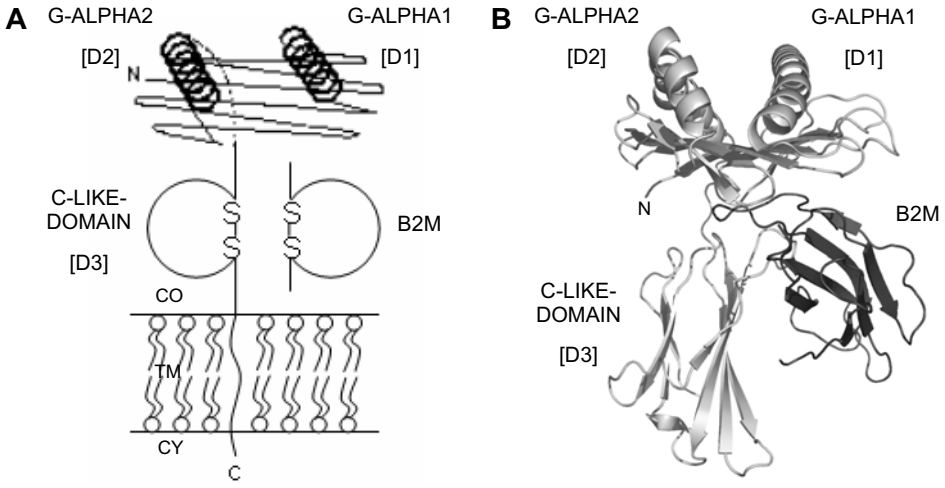


Figure 1. Représentation et structure 3D d'une protéine MHC-I. (A) Représentation d'une protéine MHC-I à la surface d'une cellule cible. (B) Structure 3D d'une protéine MHC-I (code 1bii, IMGT/3Dstructure-DB [15]). La chaîne lourde I-ALPHA comprend les 3 domaines extracellulaires G-ALPHA1, G-ALPHA2 et C-LIKE, et les 3 régions de connexion (CO), transmembranaire (TM) et cytoplasmique (CY) absentes dans la structure 3D [10,16]. La B2M est composée d'un unique domaine extracellulaire (lié de manière non covalente à I-ALPHA). [D1], [D2] et [D3] indiquent les domaines et leur position en partant de l'extrémité N-terminale de la chaîne I-ALPHA.

Les règles de description des domaines protéiques de la MhcSF [10] sont définies dans IMGT-ONTOLOGY [13]. Les deux domaines extracellulaires N-terminaux de la chaîne lourde des protéines MHC-I (Fig. 1) et MHC-I-like sont respectivement des G-DOMAINs (G-ALPHA1 [D1] et G-ALPHA2 [D2]) et G-LIKE-DOMAINs (G-ALPHA1-LIKE [D1] et G-ALPHA2-LIKE [D2]). Chaque chaîne lourde de protéine MHC-II est constituée d'un unique G-DOMAIN (G-ALPHA [D1] pour les chaînes I-ALPHA et G-BETA [D1] pour les chaînes I-BETA). Les G-DOMAINs et G-LIKE-DOMAINs ont des structures très similaires, chacune constituée d'un feuillet de quatre brins beta antiparallèles (notés A, B, C et D) et d'une longue hélice [10,14,16]. Cette grande similarité des structures 3D est d'autant plus frappante qu'il n'existe qu'une faible similarité de séquence entre les G-DOMAINs et G-LIKE-DOMAINs. La méthode d'alignement des G-DOMAINs et G-LIKE-DOMAINs développée au sein d'IMGT combine par conséquent leurs informations de séquence et de structure. Le domaine extracellulaire C-terminal de chacune des deux chaînes lourdes de MHC-II (liées de manière non covalente) est un C-LIKE-DOMAIN [D2]. Cette organisation modulaire empêche toute interaction des protéines MHC-II avec la B2M. Les séquences et structures des G-DOMAINs de protéines MHC-II sont prises en compte pour définir la numérotation unique IMGT, mais ne sont pas concernées par notre problématique de classification. Le domaine extracellulaire C-terminal de la chaîne lourde des protéines MHC-I et MHC-I-like liées à la B2M est un C-LIKE-DOMAIN [D3] (Fig. 1). Les chaînes lourdes de protéines MHC-I-like non liées à la B2M comprennent ou non un C-LIKE-DOMAIN [D3]. La délétion in vitro du C-LIKE-DOMAIN d'une chaîne lourde de protéine MHC-I

n'affecte ni sa structure, ni sa liaison à la B2M et au peptide [17]. La présence ou non d'un C-LIKE-DOMAIN ne semble donc pas être un critère efficace de discrimination entre les protéines de la MhcSF liées ou non à la B2M, et les résultats présentés ici se basent exclusivement sur l'analyse des G-DOMAINs et G-LIKE-DOMAINs.

L'objectif de notre étude est de prédire la liaison (ou l'absence de liaison) à la B2M pour des protéines MHC-I et MHC-I-like nouvellement identifiées. Nous nous appuyons sur l'alignement multiple IMGT, construit à partir de protéines dont la classe (liaison ou non à la B2M) a été déterminée expérimentalement. Parmi les nombreuses approches de classification supervisée, c'est-à-dire qui utilisent une connaissance a priori du découpage des données en classes, le classifieur Bayesien dit naïf [18] a été appliqué avec succès à la prédiction de ligands classe-spécifiques, les caractéristiques fonctionnelles des classes étant connues [19,20]. Outre la simplicité de sa mise en œuvre, l'avantage majeur de ce classifieur est de s'accommoder d'un jeu de données de taille restreinte, ce qui est le cas ici. L'approche proposée combine par conséquent un classifieur Bayesien naïf et la numérotation unique IMGT pour les G-DOMAINs et G-LIKE-DOMAINs. Ce classifieur se base sur un ensemble de descripteurs binaires (qui associent une position dans l'alignement et un groupe d'acides aminés), préalablement extraits des données pour leur capacité à discriminer les deux classes de séquences.

Dans la suite, nous présentons plus en détail les séquences des protéines de la MhcSF, leur alignement et les caractéristiques du problème posé. Nous décrivons ensuite la méthode de sélection des descripteurs discriminants, le classifieur Bayesien employé et les procédures mises en place pour évaluer sa performance. Les résultats sont présentés conjointement à une interprétation structurale, à l'analyse de mutants et aux prédictions sur des protéines dont la liaison à la B2M est inconnue à ce jour.

2 Les données

Dans cette étude, 806 séquences protéiques de la MhcSF ont été regroupées en 9 types de récepteurs d'après leurs caractéristiques biologiques et leur composition en domaines protéiques. Ces séquences correspondent à des formes alléliques de 47 protéines MHC-I et MHC-I-like de 4 espèces de mammifères: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* et *Bos taurus*. Les séquences alléliques (767 pour les 13 protéines MHC-I et 39 pour les 34 protéines MHC-I-like, décrites au sein d'IMGT) sont nommées "allèles" dans ce contexte et dans la suite du texte, et diffèrent par au moins un acide aminé, pour une protéine codée par un gène donné, dans une espèce donnée.

Les protéines MHC-I comprennent les protéines HLA d'*Homo sapiens* et H2 de *Mus musculus*, qui correspondent à 6 types de gènes: 3 MHC-Ia classiques (HLA-A, HLA-B, HLA-Cw chez l'homme, H2-D, H2-K, H2-L chez la souris) et 3 MHC-Ib non classiques (HLA-E, HLA-F, HLA-G chez l'homme, H2-M, H2-Q, H2-T chez la souris) [10]; seule la protéine RT1-AA est prise en compte ici pour le MHC-I de *Rattus norvegicus*. Les protéines MHC-I constituent un type de récepteur et sont liées à la B2M. Huit types de récepteurs MHC-I-like ont été définis dans cette étude: AZGP1, CD1, EPCR, FCGRT, HFE, MIC, MR1 et RAE. Chacun de ces récepteurs a été identifié chez de nombreux mammifères. Toutes les protéines d'un même type de récepteur interagissent (CD1, FCGRT, HFE, MR1) ou non (AZGP1, EPCR, MIC, RAE) avec la B2M. A l'exception des récepteurs MR1, la structure d'au moins un récepteur de chaque type a été résolue et est disponible dans IMGT/3DStructure-DB [15], la base de données de structures 3D d'IMGT.

La numérotation unique IMGT pour les G-DOMAINs des protéines MHC-I et MHC-II, et les G-LIKE-DOMAINs des protéines MHC-I-like [10] a été établie par alignements successifs de séquences et de structures 3D: 767 séquences et 155 structures de MHC-I, 504 séquences et 33 structures de MHC-II, 39 séquences et 27 structures de MHC-I-like. La localisation des exons au niveau des séquences génomiques nous permet de délimiter les domaines protéiques. Tous les G-DOMAINs et G-LIKE-DOMAINs sont alors alignés ensemble, en séquence ou structure selon leur similarité. La similarité des séquences de la MhcSF qui correspondent à un même type de récepteur est forte (60-90% d'identité) tandis que celle des séquences qui correspondent à des types de récepteurs différents est faible (15-40%); l'unité structurale de ces domaines protéiques nous permet néanmoins de nous affranchir de cette faible similarité de séquence. Un numéro est

attribué à chaque position de l'alignement final; les positions rarement occupées dans l'alignement sont identifiées avec une lettre (par exemple la position 7A). Ce mode de gestion des positions insérées nous permet de conserver les identifiants numériques de l'alignement multiple, quelle que soit la longueur des séquences des G-DOMAINs ou G-LIKE-DOMAINs ajoutées. Nous obtenons ainsi la numérotation unique pour les G-DOMAINs et G-LIKE-DOMAINs [10]; les séquences et les structures numérotées sont ajoutées respectivement à IMGT Repertoire [21] et IMGT/3DstructureDB [15]. Chaque nouvelle protéine de la MhcSF est dans un premier temps décrite en terme de domaines. La numérotation unique IMGT est alors appliquée à chaque G-DOMAIN ou G-LIKE-DOMAIN identifié, par alignement avec la séquence numérotée la plus similaire ou par alignement de structure dans le cas d'une similarité de séquence insuffisante. Les alignements multiples de séquences et de structures sont respectivement effectués avec MUSCLE [22] et COMPARER [23]. La cohérence de l'alignement final est validée en terme de séquence par NorMD [24] et en terme de structure par Profit (http://bioinf.org.uk/software/profit).

Les G-DOMAINs et G-LIKE-DOMAINs sont observés au sein de toutes les protéines MHC-I et MHC-I-like; comme indiqué précédemment, chaque séquence de l'alignement multiple utilisé pour la classification est donc composée uniquement de deux G-DOMAINs ou G-LIKE-DOMAINs. L'arbre phylogénétique obtenu à partir de ces séquences protéiques alignées indique l'antériorité évolutive de la spécialisation des protéines MHC-I-like sur la spéciation. En effet, chaque type de récepteur définit un clade regroupant les séquences de plusieurs espèces, ce qui laisse supposer que ces différentes fonctions sont apparues avant l'origine des espèces étudiées. La deuxième caractéristique mise en évidence par la phylogénie est directement liée à notre problématique de classification supervisée. Les deux classes de séquences (liaison ou non à la B2M) ne constituent pas en effet deux clades distincts, mais plusieurs clades non corrélés à la classification. Ainsi, le plus proche voisin des 3 séquences de type EPCR correspond aux 7 séquences de type CD1, les séquences de ces deux types de récepteurs n'appartenant pas à la même classe (CD1 se lie à B2M tandis qu'EPCR ne s'y lie pas). L'analyse des plus proches voisins dans l'arbre phylogénétique ne permettrait donc pas le classement d'une séquence correspondant à un nouveau type de récepteur MHC-I-like. Lorsque le récepteur est déjà connu, le problème de classification semble plus simple puisque toutes les séquences d'un même récepteur ont le même comportement vis-à-vis de la B2M (à moins qu'il ne s'agisse d'un mutant pathologique, comme nous le verrons dans l'analyse des résultats).

3 Le classifieur Bayesien naïf

Le classifieur Bayesien dit naïf permet d'estimer les probabilités qu'une nouvelle séquence s de la MhcSF interagisse ou non avec la B2M. Deux étapes préalables au classement sont nécessaires: (1) la sélection d'un ensemble de descripteurs binaires discriminants (qui associent une position dans l'alignement et un groupe d'acides aminés) d'après les données; (2) la construction du classifieur, par estimation des fréquences de ces descripteurs conditionnellement aux classes de séquences que l'on souhaite séparer. Le classement de la séquence s est effectué d'après la description de s pour l'ensemble des descripteurs binaires, et de leurs fréquences au sein des classes. Nous proposons 3 modes d'application de la procédure de "leave-one-out", destinés à évaluer la performance du classifieur dans les cas où la prédiction concerne une nouvelle protéine, une espèce non référencée au sein des données ou un nouveau type de récepteur. Le polymorphisme des protéines de la MhcSF est pris en compte par la pondération des allèles: pour une protéine qui possède e allèles, un poids de $1/e$ est attribué à chacune de ces séquences. Cette approche permet de compenser en partie le faible nombre de protéines par le nombre élevé d'allèles pris en compte. Dans la suite, chaque étape est détaillée dans un premier temps dans le cas classique, puis dans le cas de protéines polymorphes.

3.1 Sélection des descripteurs

L'objectif est de sélectionner un ensemble de descripteurs pour leur capacité à discriminer les 2 classes de séquences C_{β} et $C_{\neg\beta}$ du jeu de données (protéines liées ou non à la B2M). Chaque descripteur associe une position i dans l'alignement multiple et un groupe d'acides aminés g. Les acides aminés sont regroupés d'après leur homologie fonctionnelle au sein des V-REGIONs des immunoglobulines [25] et d'après [26]. Le

gap est considéré comme un acide aminé additionnel. L'ensemble des groupes d'acides aminés utilisé est:

$g = \{DNEQKR\},\{IVLFCMAW\},\{GTSYPH\},\{GAS\},\{CDPNT\},\{EVQH\},\{MILKR\},\{FWY\},\{DE\},\{NQ\},$
$\{RHK\},\{ST\},\{AGILPV\},\{CM\},\{ILV\},\{AG\},\{P\},\{AVIL\},\{F\},\{G\},\{W\},\{Y\},\{A\},\{R\},\{N\},\{D\},\{C\},\{Q\},$
$\{E\},\{H\},\{I\},\{L\},\{K\},\{M\},\{S\},\{T\},\{V\},\{-\}$.

Les descripteurs ainsi définis sont binaires, un acide aminé du groupe g pouvant être observé ou non à la position i d'une séquence. Pour un groupe g, $\neg g$ représente l'ensemble des acides aminés non inclus dans g. Par exemple, pour le groupe $g = \{IVLFCMAW\}$, on a $\neg g = \{DNEQKRGTSYPH-\}$. La capacité discriminante de chacun des groupes d'acides aminés est évaluée à chaque position de l'alignement, afin de sélectionner les couples (i, g) les plus discriminants. Les nombres d'occurrence d'acides aminés de g et de $\neg g$ à la position i des séquences des classes C_β et $C_{\neg\beta}$ sont présentés sous forme d'une table de contingence :

$$CT = \begin{array}{c|c|c} & C_\beta & C_{\neg\beta} \\ \hline g & a & b \\ \hline \neg g & c & d \\ \hline \end{array} \qquad (1)$$

Dans le cas de protéines exprimées sous différentes formes alléliques, la table de contingence établie pour une position i et un groupe d'acides aminés g est basée sur les poids des allèles. Une protéine de C_β représenté par 10 allèles aura une contribution de 2/10 pour a et 8/10 pour c dans (1), dans le cas où 2 allèles ont un acide aminé de g à la position i et les 8 autres un acide aminé de $\neg g$.

La capacité de discrimination d'un groupe d'acides aminés g à la position i, est estimée par la mesure du χ^2 à partir de la table de contingence (1):

$$\chi^2(CT) = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(a + c)(c + d)(b + d)}. \qquad (2)$$

Cette mesure prend une valeur d'autant plus grande que la différence entre les deux diagonales ad et bc est importante en valeur absolue. Pour une position i, le groupe d'acides aminés g associé à la valeur de χ^2 la plus élevée est sélectionné; dans le cas où plusieurs groupes sont caractérisés par la même valeur de χ^2, nous sélectionnons le groupe comprenant le moins d'acides aminés. Les couples (i, g) ainsi générés sont classés par ordre décroissant selon leur valeur de χ^2. Les f premiers couples de cette liste constituent l'ensemble des descripteurs, où f est un paramètre ajusté suivant une procédure décrite en 3.4. L'ensemble $D = (d_1, d_2, ..., d_k, ..., d_f)$ est ainsi constitué des f descripteurs d_k les plus discriminants, combinant une position i_k dans l'alignement et un groupe d'acides aminés g_k .

3.2 Apprentissage du classifieur de Bayes

La probabilité qu'une nouvelle séquence s de la MhcSF appartienne à la classe C_X sachant la description de s pour l'ensemble des descripteurs binaires préalablement définis, est estimée selon la règle de Bayes par:

$$P(C_X / D(s)) = \frac{P(C_X) P(D(s) / C_X)}{P(D(s))}, \qquad (3)$$

avec:

$X \in \{\beta, \neg\beta\}$,

$D(s) = (d_1(s), d_2(s), ..., d_k(s), ..., d_f(s))$.

La classe C_X prédite pour la séquence s correspond à celle dont la probabilité sachant $D(s)$ est la plus élevée. Dans (3), $P(C_\beta / D(s))$ et $P(C_{\neg\beta} / D(s))$ ont le même dénominateur $P(D(s))$; la probabilité $P(C_X / D(s))$ la plus élevée est donc identifiée par comparaison des valeurs de numérateur de $P(C_\beta / D(s))$ et $P(C_{\neg\beta} / D(s))$. Par ailleurs, le classifieur de Bayes dit naïf s'appuie sur l'hypothèse d'indépendance des descripteurs conditionnellement à la classe. Cette hypothèse est simplificatrice, mais elle s'est avérée efficace

pour de très nombreux jeux de données réels, même avec des descripteurs fortement corrélés; cette propriété est expliquée par des arguments théoriques dans [27]. La probabilité que la séquence s appartienne à la classe C_X est donc exprimée par:

$$P(C_X/D(s)) \propto P(C_X)\prod_k^f P(d_k(s)/C_X)\,. \qquad (4)$$

Les probabilités $P(C_\beta)$ et $P(C_{\neg\beta})$ sont estimées a priori, par les proportions de protéines dans le jeu de données qui correspondent respectivement à des protéines liant ou non la B2M. Les probabilités $P(d_k(s)/C_X)$ sont estimées au cours de la phase d'apprentissage du classifieur, par les fréquences d'occurrence du descripteur d_k (présence ou absence de g_k à la position i_k) pour la classe C_X. Ces fréquences sont corrigées par le facteur de Lidstone [28], afin de remédier au problème posé par les fréquences nulles. En effet, dans le cas d'un descripteur d_k pour lequel toutes les séquences s' de la classe C_X sont telles que $d_k(s) \neq d_k(s')$, la probabilité (4) de C_X sachant $D(s)$ est nulle, quelle que soit la contribution des autres descripteurs. L'utilisation de fréquences non corrigées est donc susceptible d'aboutir à la dominance d'un unique descripteur pour le classement d'une nouvelle séquence s. Les fréquences corrigées sont définies par:

$$P(d_k(s)/C_X) = \frac{N(d_k(s)/C_X) + \lambda}{|C_X| + 2\lambda}\,, \qquad (5)$$

où $N(d_k(s)/C_X)$ est le nombre de séquences s' de C_X qui vérifient $d_k(s) = d_k(s')$. Nous avons choisi $\lambda = 1/|C_X|$ d'après des analyses préliminaires destinées à ajuster λ, et d'après [29]. Du fait de la binarité des descripteurs, le facteur de λ. au dénominateur est égal à 2 pour qu'un descripteur soit vrai ou faux avec une probabilité totale de 1.

L'estimation et la correction des fréquences dans le cas de protéines polymorphes est traitée de manière analogue à l'établissement des tables de contingence, en prenant en compte la somme des poids des séquences s' de C_X telles que $d_k(s) = d_k(s')$.

3.3 Performance du classifieur et nombre de descripteurs

Afin d'évaluer la performance d'un classifieur, le jeu de données est généralement divisé en un jeu d'apprentissage et un jeu de test. Les étapes de sélection de l'ensemble des descripteurs et de construction du classifieur (détaillées en 3.1 et 3.2) sont effectuées pour l'échantillon d'apprentissage, le classifieur ainsi construit étant ensuite appliqué aux séquences de l'échantillon de test pour prédire leur classe d'appartenance. La performance du classifieur est alors évaluée par le nombre d'observations test dont la classe prédite est égale à la classe réelle. Dans le cas de protéines exprimées sous différentes formes alléliques, une approche simple consiste à classer itérativement toutes les séquences alléliques du test, puis à pondérer les succès et erreurs par l'inverse du nombre d'allèles, comme nous l'avons vu dans la phase d'apprentissage. Sur notre jeu de données, des études préliminaires montrent que la classe prédite est identique pour toutes les allèles d'une même protéine. Afin de réduire le temps de calcul, nous considérons donc chaque protéine codée par un gène donné comme un profil p composé d'une ou plusieurs séquences alléliques. Alors que la position i_k de la séquence s d'une protéine non polymorphe présente un acide aminé de g_k ou de $\neg g_k$, des acides aminés de g_k et de $\neg g_k$ peuvent être observés conjointement à la position i_k d'un profil p. Nous estimons alors $P(C_X/D(p))$ en remplaçant l'expression $P(d_k(s)/C_X)$ dans (4) par la moyenne algébrique, au sein de l'ensemble des allèles de la protéine, des probabilités conditionnelles correspondant à chacun des 2 cas ($i_k(s) = g_k$ et $i_k(s) = \neg g_k$).

Les données actuelles sur la MhcSF concernent un nombre restreint de protéines, et ne peuvent pas être divisées en un échantillon d'apprentissage et un échantillon de test de taille suffisante. Nous utilisons donc une approche de type "leave-one-out" [30] pour définir ces échantillons. Lorsque l'on dispose de n observations, le principe de base est d'apprendre sur $n-1$ observations, de tester sur l'observation restante, et d'itérer le processus n fois. La performance est évaluée par la moyenne des n tests. Nous déclinons ici cette procédure selon trois modes, destinés à évaluer la performance du classifieur dans les cas où la

prédiction concerne une nouvelle protéine, une espèce non référencée au sein des données ou un nouveau type de récepteur. Les séquences de chacun des 9 types de récepteurs, de chacune des 4 espèces et de chacune des 47 protéines constituent itérativement l'échantillon de test.

Afin d'ajuster le nombre f de descripteurs à prendre en compte dans le classifieur, nous construisons un classifieur successivement pour chaque valeur de f comprise entre 1 et 40. Pour $f = 1$, l'unique descripteur pris en compte est donc le premier de la liste, c'est-à-dire celui qui présente la meilleure discrimination au sens de la mesure du χ^2 (2). En augmentant le nombre de descripteurs, on s'attend à une augmentation de performance (évaluée par leave-one-out, comme décrit ci-dessus), jusqu'à atteindre un plateau correspondant à la taille f optimale. Le petit nombre d'observations disponibles ici rend cependant cette procédure difficile, et il est plus juste de parler de "bonne taille" que de taille optimale, comme nous le verrons dans la partie suivante.

4 Résultats et discussion

4.1 Performance du classifieur et nombre de descripteurs

Les taux de bon classement sont représentés pour toutes les valeurs de $f = 1, 2...40$ et les trois procédures de leave-one-out (Fig. 2). La performance la plus faible est obtenue lorsque toutes les protéines d'un même type de récepteur constituent l'échantillon de test, et ce quel que soit le nombre de descripteurs. Ce résultat était prévisible car cette procédure de leave-one-out correspond au classement de séquences de test dont le pourcentage d'identité avec les séquences d'apprentissage ne dépasse pas 40%. La performance la plus élevée, quelle que soit la procédure de leave-one-out, est obtenue par un classifieur constitué de 18 descripteurs. Un tel classifieur classe correctement 70% des séquences (33 protéines parmi les 47 du jeu de données) dont le type de récepteur n'est pas représenté dans l'échantillon d'apprentissage. Ce résultat est satisfaisant, du fait de la faible similarité de ces séquences de test avec les séquences d'apprentissage. De plus, 7 protéines mal classées (sur 14) correspondent au récepteur CD1, dont les deux G-LIKE-DOMAINs lient de petits lipides, à la place des peptides habituellement présentés par les protéines MHC-I et certaines protéines MHC-I-like; les domaines de CD1 sont par conséquent plus hydrophobes que ceux des autres protéines de la MhcSF. 94% et 98% des séquences sont classées correctement par un classifieur constitué de 18 descripteurs et appris à partir d'échantillons d'apprentissage qui excluent respectivement les séquences d'une même espèce et d'une même protéine.

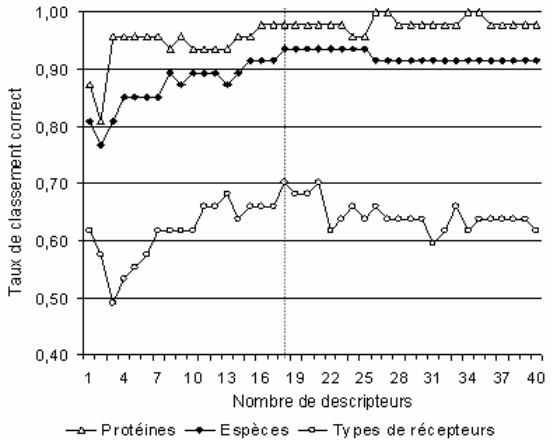


Figure 2. Performance des classifieurs en fonction du nombre de descripteurs et de la procédure de leave-one-out. La performance la plus élevée pour les 3 procédures est obtenue pour 18 descripteurs.

Afin de vérifier la signification statistique des performances des classifieurs mesurées par les 3 procédures de leave-one-out, nous avons testé notre approche sur 100 jeux de données engendrés par remaniements aléatoires des séquences du jeu de données initial. Les positions de chaque séquence ont ainsi été "mélangées" aléatoirement pour générer une séquence de même composition en acides aminés. Les performances sont alors proches de 50% de bon classement pour les 3 procédures d'évaluation. Les performances évaluées à 70%, 94% et 98% par leave-one-out sur le jeu de données initial sont donc bien significatives. De plus, cette expérience montre que ce n'est pas la composition globale des séquences en acides aminés qui détermine leur liaison ou non à la B2M. L'apprentissage du jeu de données initial (sans re-échantillonnage) par le classifieur Bayesien est donc réalisé d'après (4) pour les 18 descripteurs les plus discriminants dans la liste ordonnée suivant la mesure du χ^2 (2). L'ensemble de ces 18 descripteurs est présenté en Table 1.

Domaine IMGT	Position IMGT	Groupe discriminant	Type de descripteurs
[D1]	8	CDPNT	3
	11	ILV	4
	12	MILKR	3
	21	W	3
	25	DNEQKR	3
	27	FYW	1
	32	EVQH	1
	35	EVQH	3
	51	W	2
	74	MILKR	4
	86	NQ	2
	88	CDPNT	4
[D2]	10	G	2
	27	AG	1
	32	DE	1
	39	EVQH	4
	83	DE	2
	85	G	2

Table 1. Les 18 descripteurs sélectionnés. Les descripteurs de type 1 et 2 sont favorables à la liaison à la B2M (Fig. 3A); les types 3 et 4 sont défavorables à la liaison à la B2M (Fig. 3B). Les types 1 et 3 sont localisés dans la zone potentielle de contact à la B2M.

Nous avons également évalué les performances de 2 classifieurs construits, respectivement, à partir de 9 descripteurs localisés dans la zone de contact, et à partir de 5 descripteurs localisés hors de cette zone (dont la définition est donnée plus loin). Le nombre de descripteurs de chacun de ces classifieurs a été optimisé comme décrit précédemment, à partir de l'alignement multiple initial respectivement restreint aux sites de contact potentiel ou excluant ces sites. Chacun de ces classifieurs s'avère être aussi performant que le classifieur établi pour 18 descripteurs (qui incluent les 9 et 5). Cette expérience met donc en évidence une certaine redondance statistique de nos 18 descripteurs, chacun de ces 2 classifieurs étant suffisant pour classer correctement les nouvelles séquences. Mais comme nous allons le voir dans la partie suivante, les 18 descripteurs sélectionnés admettent une interprétation biologique et structurale.

4.2 Analyse du contexte structural des descripteurs

Afin d'identifier les sites potentiels de liaison à la B2M au sein des G-DOMAINs des protéines MHC-I et G-LIKE-DOMAINs des protéines MHC-I-like, nous avons réalisé une analyse exhaustive des contacts pour 165 structures 3D de protéines de la MhcSF liées à la B2M (156 structures 3D de 12 protéines MHC-I et 9 structures 3D de 6 protéines MHC-I-like). Les sites potentiels de contact correspondent aux positions des domaines [D1] (G-ALPHA1 ou G-ALPHA1-LIKE) et [D2] (G-ALPHA2 ou G-ALPHA2-LIKE) en contact avec la B2M dans au moins une de ces structures. Ces positions sont localisées dans les brins A, B, C et D, dans les boucles AB et BC, et dans l'hélice de [D1], et dans les brins A, B et C, et dans la boucle BC de [D2].

Les descripteurs peuvent être classés en quatre types, selon qu'ils sont favorables ou défavorables à la liaison à la B2M, et qu'ils correspondent ou non à des sites potentiels d'interaction à la B2M. La première distinction (favorable/défavorable) est issue de l'analyse des diagonales de la table de contingence (1). Pour un descripteur donné (i_k, g_k), une table de contingence (1) dont la diagonale ad est majoritaire indique qu'un acide aminé du groupe g_k à la position i_k d'une protéine est plutôt favorable à son interaction à la B2M; de manière analogue, une table de contingence dont la diagonale bc est majoritaire indique qu'un acide aminé du groupe g_k à la position i_k d'une protéine est plutôt défavorable à son interaction à la B2M. L'interprétation du contexte structural des 18 descripteurs doit par conséquent être réalisée indépendamment pour les descripteurs de chaque type.

Les 9 descripteurs dont l'observation est un critère favorable à l'interaction à la B2M sont analysés pour la structure 3D de FCGRT de *Rattus norvegicus*. Parmi les protéines de la MhcSF qui se lient à la B2M et dont la structure a été résolue, la protéine FCGRT de *Rattus norvegicus* présente en effet un acide aminé du groupe caractéristique de la classe C_β pour chacune de ces 9 positions. De même, le contexte structural des 9 descripteurs dont l'observation est un critère défavorable à l'interaction à la B2M doit être interprété pour une protéine non liée à la B2M, et pour laquelle ces 9 positions sont constituées d'un acide aminé du groupe caractéristique de la classe $C_{\neg\beta}$: la protéine RAE1B de *Mus musculus* correspond à ce critère. Parmi les 9 descripteurs favorables à l'interaction à la B2M (Fig. 3A), 4 correspondent à une position localisée dans la zone potentielle de contact. C'est également le cas de 5 descripteurs parmi les 9 défavorables à l'interaction (Fig. 3B).

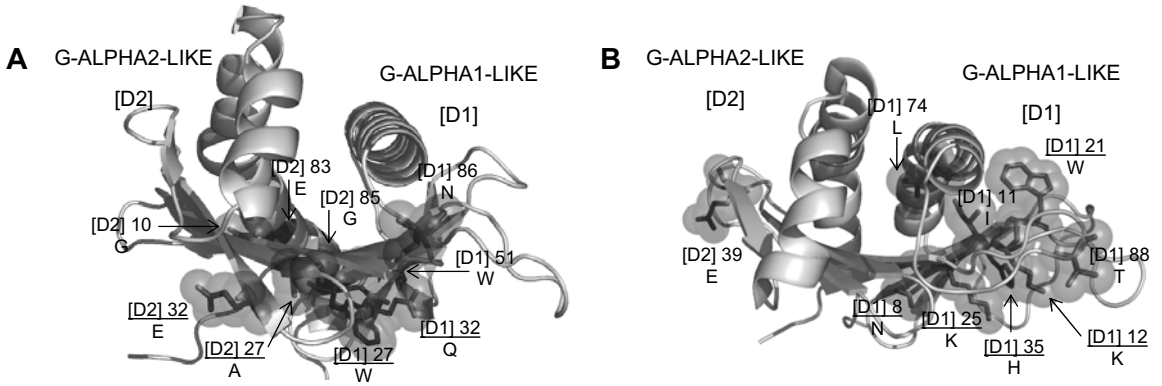


Figure 3. Structure 3D et représentation des 4 types de descripteurs pour les domaines G-ALPHA1-LIKE et G-ALPHA2-LIKE de protéines MHC-I-like se liant ou non à la B2M. Le domaine, la position et l'acide aminé observé sont indiqués pour chaque descripteur; les chaînes latérales de ces acides aminés sont mises en évidence par des sphères. (A) FCGRT de *Rattus norvegicus* (MHC-I-like lié à la B2M, code 3fru_A IMGT/3Dstructure-DB [15]). Les descripteurs favorables à la liaison à la B2M sont localisés dans la zone potentielle de contact à la B2M (type 1, soulignés: [D1] 27, 32, [D2] 27, 32) ou en dehors (type 2: [D1] 51, 86, [D2] 10, 83, 85). (B) RAE1B de *Mus musculus* (MHC-I-like non lié à la B2M, 1jfm_A). Les descripteurs défavorables à la liaison à la B2M sont dans la zone potentielle de contact (type 3, soulignés: [D1] 8, 12, 21, 25, 35) ou en dehors (type 4: [D1] 11, 74, 88, [D2] 39).

Globalement, les descripteurs favorables à l'interaction localisés dans la zone potentielle de contact semblent correspondre à une orientation de chaîne latérale ou à une propriété physico-chimique favorable au contact direct avec la B2M, tel qu'un résidu aromatique F, W ou Y en position [D1] 27 (W pour FCGRT de *Rattus norvegicus*, Fig. 3A). Les descripteurs favorables localisés hors de cette zone semblent maintenir une conformation adéquate au contact. Les résidus [D1] 51 et [D2] 85 pourraient en effet maintenir la fermeture du sillon (par le rapprochement des deux hélices) à une extrémité (Fig. 3A). Au contraire, les descripteurs défavorables situés dans cette zone semblent empêcher le contact direct par gêne stérique, tels que les résidus K en positions [D1] 12 et 25 de RAE1B de *Mus musculus* (Fig. 3B). La déstabilisation de la conformation propice à l'interaction par des résidus tels que E, V, Q ou H en [D2] 39 serait à analyser en détail.

La définition des descripteurs en terme de position et de groupes d'acides aminés permet donc d'identifier les propriétés physico-chimiques dont l'observation à une position semble être favorable ou non au contact direct (pour les positions localisées dans la zone potentielle de contact), ou stabiliser ou non la conformation moléculaire globale (pour les positions localisées hors de cette zone). La détermination de ces 4 types de descripteurs sur les chaînes lourdes de protéines MHC-I et MHC-I-like donne des indications qui devraient être précieuses pour des expériences futures de mutagenèse dirigée.

4.3 Mutagenèse dirigée et classement de nouvelles séquences MHC-I

D'après l'analyse bibliographique concernant les résultats de mutagenèse dirigée des gènes de MHC-I, la mutation du résidu N à la position [D1] 86 de la protéine HLA-A empêche la liaison de HLA-A à la B2M [31,32]. Cela conforte notre étude, car nous avons trouvé (Table 1) que la position [D1] 86 associée au groupe amide NQ est favorable à la liaison à la B2M. Ceci est sans doute lié à une implication dans le maintien de la conformation globale de la protéine, car cette position est en dehors de la zone potentielle de contact à la B2M. Les résultats expérimentaux indiquent qu'il s'agit d'un site de N-glycosylation: un oligosaccharide, relié par une liaison N-glycosidique au groupement amide du résidu N en [D1] 86, favorise l'interaction entre la protéine HLA-A et la B2M. Sur la base de nos résultats, nous pouvons étendre ce résultat expérimental en prédisant que la liaison d'un oligosaccharide en [D1] 86 est caractéristique des protéines de la MhcSF liées à la B2M.

Nous avons également classé 3 protéines MHC-I de vertébrés inférieurs: *Ambystoma mexicanum* (P79458), *Oncorhynchus kisutch* (Onki-UAA, Q9GJB4) et *Salmon trutta* (Satr-UBA, Q9GJJ8). Alors que la structure et l'origine évolutive des gènes du MHC des amphibiens [33] et des poissons téléostéens [34] sont largement étudiées, peu de données expérimentales indiquent la voie d'expression cellulaire impliquée et la liaison ou non de ces MHC-I à la B2M [35]. Le classifieur rattache les protéines de *Ambystoma mexicanum* et *Salmon trutta* à la classe des protéines de la MhcSF qui se lient à la B2M, avec un rapport élevé entre les probabilités conditionnelles de liaison ou non à la B2M (respectivement de 7.10^3 et 2.10^3); ce rapport est seulement de 9 pour la protéine d'*Oncorhynchus kisutch*, ce qui est favorable à la liaison mais avec une certitude moindre. Cette étude donne donc à penser que ces protéines MHC-I se lient à la B2M. Elles devraient par conséquent être exprimés à la surface cellulaire par le même processus que les protéines MHC-I et MHC-I-like de mammifères liés à la B2M.

5 Conclusion

Notre étude met en évidence l'efficacité de la combinaison du classifieur Bayesien naïf et de la numérotation unique IMGT pour les G-DOMAINs et G-LIKE-DOMAINs, pour prédire l'interaction des protéines de la MhcSF avec la B2M. Cette approche est en effet performante quel que soit le type de nouvelle séquence à classer, et malgré la faible similarité de séquence des protéines appartenant à un nouveau type de récepteur. Nous identifions 4 types de descripteurs, correspondant à des propriétés physico-chimiques discriminantes: favorable ou non au contact direct à la B2M, et favorable ou non au maintien d'une conformation propice à l'interaction. L'analyse structurale de ces descripteurs et la confrontation de nos résultats à ceux de mutagenèse dirigée mettent en évidence la cohérence biologique de l'approche. En indiquant si une nouvelle protéine se lie ou non à la B2M, le classifieur ainsi construit donne des informations concernant ses ligands potentiels et la voie impliquée pour son expression à la surface cellulaire. Cette approche performante devrait trouver des applications dans d'autres problématiques biologiques pour lesquelles on dispose d'un alignement multiple de qualité, et de classes de séquences connues a priori, qu'il s'agisse de classes fonctionnelles, structurales ou d'interaction.

Remerciements

E.D. est allocataire de recherche du Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR). Ce projet est soutenu par le CNRS, le MENESR (ACI IMPBio), le PPF (UMII, IUF).

References

[1] L.F. Boyd, S. Kozlowski and D.H. Margulies, Solution binding of an antigenic peptide to a major histocompatibility complex class I molecule and the role of B2-microglobulin. Proc. Natl. Acad. Sci. USA, 89:2242-2246, 1992.

[2] B. Ortmann, M.J. Androlewicz and P. Cresswell, MHC class I/beta2-microglobulin complexes associate with TAP transporters before peptide binding. Nature, 368:864-867, 1994.

[3] M.J. Shields, R. Kubota, W. Hodgson, S. Jacobson, W.E. Biddison and R.K Ribaudo, The effect of human b2-microglobulin on major histocompatibility complex I peptide loading and the engineering of a high affinity variant. Implications for peptide-based vaccines. J. Biol. Chem., 273:28010-28018, 1998.

[4] A. Townsend, T. Elliot, V. Cerundolo, L. Foster, B. Barber and A. Tse, Assembly of MHC class I molecules analyzed in vitro. Cell, 62:285-295, 1990.

[5] J.C. Solheim, J.R. Cook and T.H. Hansen, Conformational changes induced in the MHC class I molecule by peptide and beta 2-microglobulin. Immunol. Res., 14:200-217, 1995.

[6] D.M. Hill, T. Kasliwal, E. Schwarz, A.M. Hebert, T. Chen, E. Gubina, L. Zhang and S. Kozlowski, A dominant negative mutant B2-microglobulin blocks the extracellular folding of a major histocompatibility complex class I heavy chain. J. Biol. Chem., 278:5630-5638, 2003.

[7] D.B. Williams, B.H. Barber, R.A. Flavell and H. Allen, Role of beta2-microglobulin in the intracellular transport and surface expression of murine class I histocompatibility molecules. J. Immunol., 142:2796-2806, 1989.

[8] C.M. D'Urso, Z.G. Wang, Y. Cao, R. Tatake, R.A. Zeff and S. Ferrone, Lack of HLA class I antigen expression by cultured melanoma cells FO-1 due to a defect in B2m gene expression. J. Clin. Invest., 87, 284-292, 1991.

[9] Z. Wang, Y. Cao, A.P. Albino, R.A. Zeff, A. Houghton and S. Ferrone, Lack of HLA class I antigen expression by melanoma cells SK-MEL-33 caused by a reading frameshift in Beta2-microglobulin messenger RNA. J. Clin. Invest., 91:684-692, 1993.

[10] M.-P. Lefranc, E. Duprat, Q. Kaas, M. Tranne, A. Thiriot and G. Lefranc, IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev. Comp. Immunol., 2005, in press.

[11] A. Porgador, O. Mandelboim, N.P. Restifo and J.L. Strominger, Natural killer cell lines kill autologous B2-microglobulin-deficient melanoma cells: implications for cancer immunotherapy. Proc. Natl. Acad. Sci. USA, 94:13140-13145, 1997.

[12] A.L. Cook, Beta 2 microglobulin and resistance to murine respiratory mycoplasmosis. Contemp. Top. Lab. Anim. Sci., 43:18-24, 2004.

[13] V. Giudicelli and M.-P. Lefranc, Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics, 15:1047-1054, 1999.

[14] E. Duprat, Q. Kaas, V. Garelle and M.-P. Lefranc, IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In: *Recent Research and Developments in Human Genetics*, 2, pp. 111-136, 2004.

[15] Q. Kaas, M. Ruiz and M.-P. Lefranc, IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucl. Acids Res., 32:D208-D210, 2004.

[16] Q. Kaas, E. Duprat, G. Tourneur and M.-P. Lefranc, IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: *Immunoinformatics: Opportunities and challenges of bridging immunology with computer and information sciences*, in press.

[17] E.J. Collins, D.N. Garboczi, M.N. Karpusas and D.C. Wiley, The three-dimensional structure of a class I major histocompatibility complex molecule missing the alpha3 domain of the heavy chain. Proc. Natl. Acad. Sci. USA, 92:1218-1221, 1995.

[18] I.J. Good, *The estimation of probabilities: an essay on modern Bayesian methods*, Research Monograph 30, MIT Press, Cambridge, 1965.

[19] R. Bandyopadhyay, X.X. Tan, K.S. Matthews and D. Subramanian, Predicting protein-ligand interactions from primary structure. Technical Report, Rice University, TR02-398, 2002.

[20] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido and S. Ahmad, A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. Bioinformatics, 19:234-240, 2003.

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

[22] E.C. Robert, MUSCLE: multiple sequence alignement with high accuracy and high throughput. Nucl. Acids Res., 32:1792-1797, 2004.

[23] A. Sali and T.L. Blundell, Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J. Mol. Biol., 212:403-428, 1990.

[24] J.D. Thompson, F. Plewniak, R. Ripp, J.C. Thierry and O. Poch, Towards a reliable objective function for multiple sequence alignments. J. Mol. Biol., 314:937-951, 2001.

[25] C. Pommié, S. Levadoux, R. Sabatier, G. Lefranc and M.-P. Lefranc, IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acids properties. J. Mol. Recognition, 17:17-32, 2003.

[26] T.D. Wu and D.L. Brutlag, Identification of protein motifs using conserved amino acids properties and partitioning techniques, *Proceedings of the Thirteenth International Conference on Intelligent Systems for Molecular Biology*, pp. 402-410, 1995.

[27] P. Domingos and M. Pazzani, Beyond independence: conditions for the optimality of the simple Bayesian classifier, *Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, pp. 105-112, 1996.
[28] G. Lidstone, Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. Transactions of the Faculty of Actuaries, 8:182-192, 1920.
[29] R. Kohavi, B. Becker and D. Sommerfield, Improving simple bayes, *Proceedings of the Ninth European Conference on Machine Learning, Poster Papers*, Springer-Verlag, New York, 1997.
[30] D.J. Hand, Recent advances in error rate stimation. Pattern Recognition Letters, 4:335-346, 1986.
[31] J.A. Barbosa, J. Santos-Aguado, S.J. Mentzer, J.L. Strominger, S.J. Burakoff and A.P. Biro, Site-directed mutagenesis of class I HLA genes. Role of glycosylation in surface expression and functional recognition. J. Exp. Med., 166:1329-1350, 1987.
[32] J. Santos-Aguado, A.P. Biro, U. Fuhrmann, J.L. Strominger and J.A. Barbosa, Amino acid sequences in the alpha1 domain and not glycosylation are important in HLA-A2/beta2-microglobulin association and cell surface expression. Mol. Cell Biol., 7:982-990, 1987.
[33] B. Sammut, L. Du Pasquier, P. Ducoroy, V. Laurens, A. Marcuz and A. Tournefier, Axolotl MHC architecture and polymorphism. Eur. J. Immunol., 29:2897-2907, 1999.
[34] J.D. Hansen, P. Strassburger, G.H. Thorgaard, W.P. Young and L. Du Pasquier, Expression, linkage, and polymorphism of MHC-related genes in Rainbow trout, Oncorhynchus mykiss. J. Immunol., 163:774-786, 1999.
[35] A.B. Antao, V.G. Chinchar, T.J. McConnell, N.W. Miller, L.W. Clem and M.R. Wilson, MHC class I genes of the channel catfish: sequence analysis and expression. Immunogenetics, 49:303-311, 1999.

Publication 7

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 4 2006, pages 453–459
doi:10.1093/bioinformatics/bti826

Structural bioinformatics

A simple method to predict protein-binding from aligned sequences—application to MHC superfamily and β2-microglobulin

Elodie Duprat[1], Marie-Paule Lefranc[1,2] and Olivier Gascuel[3,*]

[1]Laboratoire d'ImmunoGénétique Moléculaire, IGH (UPR CNRS 1142), 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France, [2]Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France and [3]Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM (UMR CNRS-UM2 5506), 161 rue Ada, 34392 Montpellier Cedex 5, France

Received on August 3, 2005; revised and accepted on December 7, 2005
Advance Access publication December 13, 2005
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The MHC superfamily (MhcSF) consists of immune system MHC class I (MHC-I) proteins, along with proteins with a MHC-I-like structure that are involved in a large variety of biological processes. β2-Microglobulin (B2M) non-covalent binding to MHC-I proteins is required for their surface expression and function, whereas MHC-I-like proteins interact, or not, with B2M. This study was designed to predict B2M binding (or non-binding) of newly identified MhcSF proteins, in order to decipher their function, understand the molecular recognition mechanisms and identify deleterious mutations. IMGT standardization of MhcSF protein domains provides a unique numbering of the multiple alignment positions, and conditions to develop such predictive tool.

Method: We combine a simple-Bayes classifier with IMGT unique numbering. Our method involves two steps: (1) selection of discriminant binary features, which associate an alignment position with an amino acid group; and (2) learning of the classifier by estimating the frequencies of selected features, conditionally to the B2M binding property.

Results: Our dataset contains aligned sequences of 806 allelic forms of 47 MhcSF proteins, corresponding to 9 receptor types and 4 mammalian species. Eighteen discriminant features are selected, belonging to B2M contact sites, or stabilizing the molecular structure required for this contact. Three leave-one-out procedures are used to assess classifier performance, which corresponds to B2M binding prediction for: (1) new proteins, (2) species not represented in the dataset and (3) new receptor types. The prediction accuracy is high, i.e. 98, 94 and 70%, respectively. Application of our classifier to lower vertebrate MHC-I proteins indicates that these proteins bind to B2M and should then be expressed on the cellular surface by a process similar to that of mammalian MHC-I proteins. These results demonstrate the usefulness and accuracy of our (simple) approach, which should apply to other function or interaction prediction problems.

Availability: Data and MhcSF multiple alignments are available on the IMGT website (http://imgt.cines.fr).

Contact: gascuel@lirmm.fr, duprat@ligm.igh.cnrs.fr, lefranc@ligm.igh.cnrs.fr

Supplementary information: Supplementary material is downloadable at http://imgt.igh.cnrs.fr/MhcSF-B2M.html.

*To whom correspondence should be addressed.

1 INTRODUCTION

Major histocompatibility complex (MHC) proteins play a key role in the immune system, by displaying self and non-self peptides for recognition by T cell receptors. MHC class I (MHC-I) proteins have a transmembrane heavy chain (I-ALPHA) non-covalently linked to β2-microglobulin (B2M). The interaction between I-ALPHA and B2M is required for peptide display, stabilization of the molecular structure and cell surface expression of the complex (D'Urso *et al.*, 1991; Hill *et al.*, 2003).

The MHC superfamily (MhcSF) (Lefranc *et al.*, 2005a) includes MHC proteins, as well as proteins with a MHC-I-like structure which are involved in a large variety of biological processes. Thirty-four mammalian MHC-I-like proteins have currently been identified, and the 3D structure is available for 12 of them (Kaas and Lefranc, 2004). Among these 34 proteins, only 17 are constitutively bound to B2M, according to the experimental data. This study is designed to predict B2M binding (or non-binding) of newly identified MHC-I or MHC-I-like protein sequences. Such prediction should be useful for deciphering the function of these new sequences, determining their mechanism of molecular recognition, detecting mutations leading to defects in their cell surface expression, or clarifying a number of biological questions, as illustrated below with lower vertebrate MHC.

Description rules for MhcSF protein domains are defined in IMGT, the international ImMunoGeneTics information system® (Lefranc *et al.*, 2005b), and are based on the IMGT-ONTOLOGY concepts (Giudicelli and Lefranc, 1999). The two N-terminal extracellular domains of the heavy chain of MHC-I (Fig. 1) and MHC-I-like proteins are G-DOMAINs and G-LIKE-DOMAINs, respectively. These domains are strikingly similar according to their 3D structure, with each being composed by one sheet of four antiparallel beta strands and one long helical region (Kaas and Lefranc, 2005). This high 3D structure similarity is noticeable as G-DOMAINs and G-LIKE-DOMAINs sequences have low homology (∼30% identity). The third (and last) extracellular heavy chain domain is a C-LIKE-DOMAIN (Duprat *et al.*, 2004; Lefranc *et al.*, 2005c). This C-LIKE-DOMAIN is always present in MHC-I, but absent in some MHC-I-like proteins. The C-LIKE-DOMAIN of a MHC-I heavy chain was experimentally

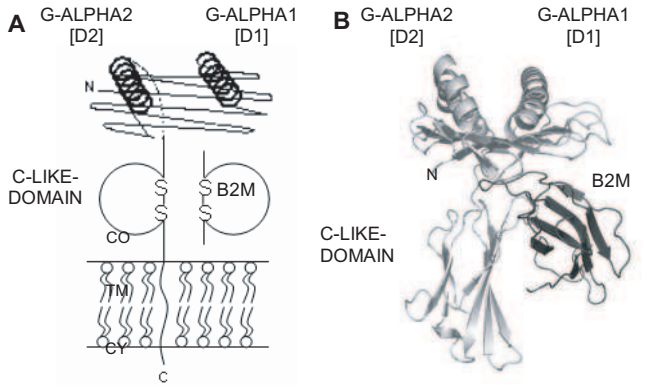


Fig. 1. MHC-I protein representation on the target cell surface (**A**) and 3D structure (**B**). The heavy chain consists of, from the N-terminal to the C-terminal end, the G-ALPHA1 [D1], G-ALPHA2 [D2] and C-LIKE extracellular domains, and (absent in the 3D structure) the connecting (CO), transmembrane (TM) and intracytoplasmic (CY) regions. B2M has a unique extracellular domain, non-covalently bound to the heavy chain. (modified from Lefranc *et al*., 2005a).

deleted, but the protein remained structurally unchanged and the B2M and the peptide were still bound conventionally (Collins *et al*., 1995). The presence of the C-LIKE-DOMAIN thus does not seem to be a valuable criterion for discrimination between B2M bound and unbound MhcSF proteins, and our results are based solely on an analysis of the two G- and G-LIKE-DOMAINs.

Our prediction method combines IMGT multiple alignment for G-DOMAINs and G-LIKE-DOMAINs (Lefranc *et al*., 2005a), along with experimental knowledge on the B2M bound/unbound properties of these proteins. We use a supervised classification approach (Duda *et al*., 2001), where classes are a priori known (here bound/unbound) in the learning set, and the goal is to predict the class of new unknown instances. In this context, the simple-Bayes classifier (Good, 1965) is easy to implement, accurate for small datasets (as is the case here) and its results are easily interpretable. Moreover, it was successfully applied for the prediction of class-specific ligands using functional features (Bandyopadhyay *et al*., 2002; Cao *et al*., 2003). Our classifier is based on binary features, consisting of a multiple alignment position and an amino acid group, and are selected from the dataset for their ability to discriminate between the two sequence classes. Three leave-one-out experiments are used to assess classifier performance—these experiments consider B2M binding prediction for new proteins, species not represented in the dataset or new receptor types.

We next give further details on MhcSF protein sequences, their alignment and the main aspects of our supervised classification problem. We then describe the method for selecting discriminant features, the classifier learning procedure, and experiments to assess its accuracy. The results are analysed in the light of structural interpretation, site-directed mutagenesis literature and B2M binding prediction for lower vertebrate MHC-I proteins.

2 DATA

2.1 MhcSF proteins

In this study, MhcSF consists of 806 protein sequences corresponding to allelic and homologous forms of 47 MHC-I and MHC-I-like proteins from four mammalian species: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Bos taurus* (see Supplementary Data 1 for details). Sequences described as 'allelic' in this study refer to sequences which, for a given protein from a given species, differ by at least one amino acid (see Supplementary Data 1 for allele homogeneity). The high allele number partly compensates for the small amount of proteins. MhcSF proteins each includes two G-DOMAINs or G-LIKE-DOMAINs and are grouped here into nine functional receptor types:

- MHC-I proteins (13 items, 767 alleles, B2M bound, C-LIKE-DOMAIN) are highly polymorphic and display a huge diversity of self and non-self peptides to T cell receptors.
- AZGP1 proteins (3 items, B2M unbound, C-LIKE) regulate fat degradation in adipocytes (Sanchez *et al*., 1999).
- CD1 proteins (7 items, B2M bound, C-LIKE) display phospholipid antigens to T cells and participate in immune defence against microbian pathogens (Zeng *et al*., 1997); these proteins differ from other MhcSF proteins by the high hydrophobicity of their antigen binding sites.
- EPCR proteins (3 items, B2M unbound, not C-LIKE) interact with activated C protein and are involved in the blood coagulation pathway (Simmonds and Lane, 1999).
- FCGRT proteins (4 items, B2M bound, C-LIKE) transport maternal immunoglobulins through placenta and govern neonatal immunity (West and Bjorkman, 2000).
- HFE proteins (3 items, B2M bound, C-LIKE) interact with transferrin receptor and consequently take part in iron homeostasis by regulating iron transport through cellular membranes (Feder *et al*., 1998).
- MIC proteins (2 items, B2M unbound, C-LIKE) are induced by stress and involved in tumor cell detection (Holmes *et al*., 2002).
- MR1 (3 proteins, B2M bound, C-LIKE) function is currently unknown (Miley *et al*., 2003).
- RAE proteins (9 items, 14 alleles, B2M unbound, no C-LIKE) are inducible by retinoic acid and stimulate cytokine/chemokine production and cytotoxic activity of NK cells (Li *et al*., 2002).

2.2 IMGT multiple alignment

The IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN (Lefranc *et al*., 2005a) is built by successive alignments of sequences and 3D structures of MHC-I, MHC-II and MHC-I-like proteins. MhcSF sequences that belong to the same receptor type are close (60-90% of identity), whereas MhcSF sequences from different receptor types are quite different (15-40%). All G-DOMAINs and G-LIKE-DOMAINs are then aligned together using the following strategy: (1) we perform structural alignment of nine—one per receptor type—3D structures; (2) remaining sequences are aligned within each receptor class against the previously structurally aligned protein. Finally, IMGT numbering is obtained by attributing a number to each position of the resulting multiple alignment.

Newly identified MhcSF proteins (e.g. amphibian and teleost MHC-I proteins, whose prediction results are presented hereafter) are first described in terms of domains. IMGT numbering of their G-DOMAINs or G-LIKE-DOMAINs is then obtained by sequence alignment with the numbered sequence with highest similarity in

the learning set, or by structural alignment when sequence similarity is insufficient.

MUSCLE (Edgar, 2004) and COMPARER (Sali and Blundell, 1990) are used for sequence and 3D structure multiple alignments; Fasta2 (Pearson and Lipman, 1988) is used for pairwise sequence alignments. Sequence and structure consistency of the resulting alignment are validated with NorMD (Thompson *et al*., 2001) and ProFit (http://bioinf.org.uk/software/profit), respectively.

2.3 Phylogeny

Evolutionary relationships within the MhcSF are established by phylogenetic analysis (Guindon and Gascuel, 2003) of 47 MHC-I and MHC-I-like protein sequences from the IMGT multiple alignment (see Supplementary Data 2 for details). The resulting phylogeny indicates that specialization occurred before speciation. Indeed, each receptor type corresponds to a clade containing all available sequences for that receptor from the species at hand. This seems to indicate that the various functions of MhcSF proteins appeared before the common ancestor of the studied mammalian species. This is in line with the small sequence similarity of G-DOMAINs and G-LIKE-DOMAINs (see above), and is further supported by the high-bootstrap value obtained for every receptor clade.

The second insight gained through this phylogeny is directly related to our prediction problem. Indeed, the two sequence classes (bound versus unbound to B2M) constitute several clades unrelated to the phylogeny, instead of two monophyletic clades. For example, the nearest neighbour of the EPCR clade is CD1, while CD1 binds B2M and EPCR does not. This indicates that nearest neighbour analysis would be inaccurate for predicting B2M binding of any MHC-I-like protein belonging to a new receptor type. However, classification seems to be easier when sequences of the same receptor type as the sequence to be predicted are already known, as all sequences from the same receptor type have the same behaviour regarding B2M [unless they correspond to a pathogenic mutant, as described in Barbosa *et al*. (1987) and Santos-Aguado *et al*. (1987), and explained in the Results section].

3 SIMPLE-BAYES CLASSIFIER

The simple-Bayes classifier estimates the probability of classes (B2M bound/unbound) for a new sequence s of MhcSF, given its description with a feature set. Two steps are required to infer this classifier from the learning set: (1) selection of discriminant binary features; and (2) learning of the classifier by estimating the frequencies of selected features, conditionally to the classes. Within this process, protein polymorphism is taken into account by weighting alleles: for a protein with e alleles, the weight of each of them is set equally at $1/e$. In the following, each step is first detailed for the standard case and then for polymorphic proteins.

3.1 Feature selection

This step aims at selecting features regarding their ability to discriminate between C_β and $C_{\neg\beta}$ classes (i.e. bound and unbound to B2M, respectively). Each (binary) feature consists of an alignment position i and an amino acid group g, and denotes the presence/absence of an amino acid from g at i in the studied sequence. Amino acids are grouped based on statistical analysis of immunoglobulin sequences (Pommié *et al*., 2004) and using standard physicochemical criteria (Wu and Brutlag, 1995): $\{IVLFCMAW\}$ $\{DNEQKR\}$ $\{GTSYPH\}$ $\{AGILPV\}$ $\{CDPNT\}$ $\{MILKR\}$ $\{EVQH\}$ $\{AILV\}$ $\{GAS\}$ $\{FWY\}$ $\{ILV\}$ $\{RHK\}$ $\{DE\}$ $\{NQ\}$ $\{ST\}$ $\{CM\}$ $\{AG\}$ (see also Supplementary Data 3). The 20 amino acids are also considered as 'groups', leading to a total of 37 possible groups. For a given group g, $\neg g$ represents the amino acids excluded from g plus the gap; e.g. $\neg g = \{DNEQKRGTSYPH—\}$ when $g = \{IVLFCMAW\}$. The g and $\neg g$ groups are dealt with in a symmetrical way and tested simultaneously for each position.

The discrimination capacity of each group is evaluated at every position of the alignment. Occurrences of the amino acids from g and $\neg g$ at position i of the sequences from classes C_β and $C_{\neg\beta}$ are counted in the contingency table:

$$\mathrm{CT} = \begin{array}{c|cc} & C_\beta & C_{\neg\beta} \\ g & a & b \\ \neg g & c & d \end{array} \quad (1)$$

In case of polymorphic proteins, this contingency table is computed according to the allele weights. For example, the contributions for a and c in (1) are set at 2/10 and 8/10, respectively, for a protein from C_β represented by two alleles having an amino acid from g at i and eight alleles where site i belongs to $\neg g$. The discrimination capacity of any (i,g) pair is estimated using the χ^2-measure that is applied to contingency table CT (1):

$$\chi^2(\mathrm{CT}) = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(c+d)(b+d)} \quad (2)$$

The highest the χ^2-value, the highest is the difference between both contingency table diagonals, which are represented by ad and bc terms. For a given position i, the amino acid group g with the highest discrimination capacity is selected—if several groups have same χ^2-value, the one with the smallest size is chosen. Resulting (i,g) pairs are ordered according to their χ^2-value, starting from the best ones. The f first pairs define the selected features, where f is a parameter that is tuned with data (see Section 3.4). The feature set $D = (d_1, d_2, \ldots, d_k, \ldots, d_f)$ consists of the f-most discriminant features d_k, with each combining a multiple alignment position i_k and an amino acid group g_k.

3.2 Simple-bayes classifier and learning procedure

The probability that a new MhcSF sequence s belongs to class C_X given its description with feature set D, is provided by the Bayes formula:

$$P(C_X|D(s)) = \frac{P(C_X)P(D(s)|C_X)}{P(D(s))}, \quad (3)$$

where $X \in \{\beta, -\beta\}$ and $D(s) = (d_1(s), d_2(s), \ldots, d_k(s), \ldots, d_f(s))$.

The class with highest probability is then predicted. Note that both $P(C_X)$ $P(D(s) \mid C_X)$ terms can simply be compared to perform this prediction, and that computing $P(D(s))$ is useless. Moreover, the simple-Bayes classifier is based on the assumption that features are independent conditionally to the classes. This is a simplifying assumption which nevertheless proved reliable for many real datasets, even with strongly correlated features—this property is explained with theoretical arguments in (Domingos and Pazzani, 1996). The probability that a given sequence s belongs to class C_X

is then obtained using:

$$P(C_X|D(s)) \propto P(C_X) \prod_{k=1}^{f} P(d_k(s)|C_X). \quad (4)$$

The probabilities $P(C_\beta)$ and $P(C_{-\beta})$ are a priori estimated by the proportions of proteins in the dataset which bind or not B2M, respectively. The probabilities $P(d_k(s)|C_X)$ are estimated during classifier learning by the frequencies of features d_k (presence or absence of g_k at position i_k) within class C_X. These frequencies are corrected by Lidstone's (1920) factor, in order to overcome the problem arising from null frequencies. Indeed, in case of a feature d_k for which all the sequences s' of the class C_X are such as $d_k(s) \neq d_k(s')$, the probability (4) of C_X knowing $D(s)$ is null, irrespective of the contribution of other features. The use of non-corrected frequencies is thus likely to lead to predominance of a single feature for the classification of a new sequence s. Corrected frequencies are defined by:

$$P(d_k(s)|C_X) = \frac{N(d_k(s)|C_X) + \lambda}{|C_X| + 2\lambda} \quad (5)$$

where $N(d_k(s)|C_X)$ is the number of sequences s' of C_X with $d_k(s) = d_k(s')$. We chose $\lambda = 1/|C_X|$ according to preliminary analyses conducted to adjust λ, and according to (Kohavi *et al*., 1997). Since our features are binary, the factor of λ in denominator is equal to 2, so a feature is true or false with a total probability of 1. Estimation and correction of the frequencies for polymorphic proteins are treated in a way similar to the contingency table calculation, by taking into account the sum of the weights of the sequences s' of C_X such as $d_k(s) = d_k(s')$.

3.3 Classifier performance and number of features

In order to evaluate the performance of a classifier, the dataset at hand is usually divided into a learning sample and a test sample. The feature selection and classifier learning stages (detailed in Sections 3.1 and 3.2) are carried out using the learning sample, and the classifier built in this way is then applied to sequences of the test sample to predict their membership class. Classifier performance is evaluated by the number of test sequences whose predicted class is equal to the real class. In case of proteins expressed in various allelic forms, a simple approach involves classifying all allelic sequences of the test sample and then balancing successes and errors by the inverse of the allele number, as we saw in the learning step. In our dataset, preliminary studies showed that the predicted class is identical for all alleles of the same protein. In order to reduce computing time, we thus consider each protein encoded by a given gene as a profile p made up of one or more allelic sequence. Amino acids of g_k and $\neg g_k$ can be observed jointly at position i_k of a profile p. We then estimate $P(C_X|D(p))$ by replacing, in (4), $P(d_k(s)|C_X)$ terms by the average within all alleles of the conditional probabilities corresponding to each two cases ($i_k(s) = g_k$ and $i_k(s) = \neg g_k$).

Current data on MhcSF is related to a limited number of proteins and cannot be divided into a learning sample and a test sample of sufficient size. We then use the leave-one-out procedure (Hand, 1986) to define these samples. When there are n observations, the guiding principle is to learn on $n - 1$ observations, to test the remaining observation, and to iterate this process n times. The performance is evaluated by the average of the n test results.

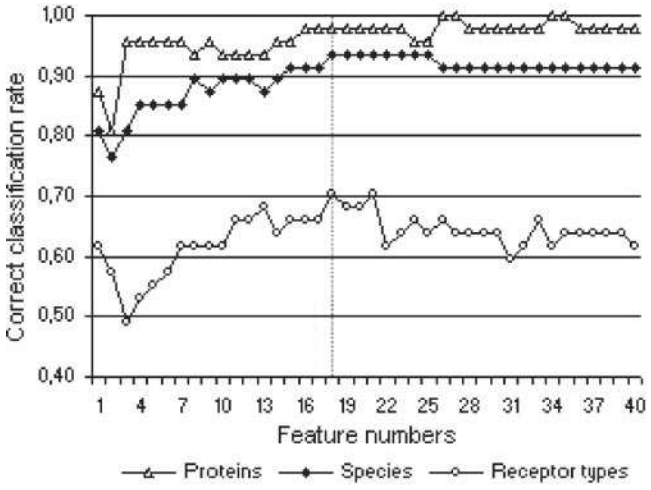


Fig. 2. Classifier accuracy as a function of the feature number and the leave-one-out procedure. Best accuracy is obtained with 18 features.

Here we apply this procedure in three different ways to evaluate the performance of the classifier when the prediction relates to a new protein, a species not represented in the dataset, or a new receptor type. Sequences of each 47 proteins, each 4 species and each 9 receptor types constitute the test sample repeatedly. For example with species leave-one-out, we predict human sequences with a classifier built using rat, mouse and cow sequences, then mouse sequences using a classifier built from human, rat and cow, etc., finally averaging the test results to obtain classifier accuracy in predicting sequences from a species not represented in the database.

In order to adjust the number f of features to be used by the classifier, we iteratively build a classifier for each value of f ranging from 1 to 40. For $f = 1$, the single feature taken into account is thus the first of the list, i.e. that which presents best discrimination accuracy regarding the χ^2-measure (2). By increasing the number of features, an increase in performance is expected (evaluated by leave-one-out, as described above), until reaching a plateau corresponding to the optimal size f.

4 RESULTS

4.1 Classifier performance and number of features

The correct classification rates are shown in Figure 2 for all values of $f = 1, 2, \ldots, 40$ and for the three leave-one-out procedures. The lowest performance is obtained when all proteins of the same receptor type constitute the test sample, regardless of the number of features. This result was expected as this leave-one-out procedure corresponds to the classification of test sequences having a percentage of identity with the learning sequences <40%. The best performance, regardless of the leave-one-out procedure, is obtained by a classifier made up of 18 features. Note that this number is inevitably approximate, due to the small amount of available proteins. Such a classifier correctly classifies 70% of the sequences (33 proteins among 47) belonging to a receptor type not represented within the learning sample. Random prediction is about 50% when predictions are well balanced, i.e. when they satisfy class priors as is the case here. Accuracy of 70% is therefore highly significant from a statistical point of view. Moreover, 7 missclassified proteins

Table 1. The 18 selected features

IMGT domain	IMGT position	Discriminant group	Feature type
[D1]	8	CDPNT	3
	11	ILV	4
	12	MILKR	3
	21	W	3
	25	DNEQKR	3
	27	FYW	1
	32	EVQH	1
	35	EVQH	3
	51	W	2
	74	MILKR	4
	86	NQ	2
	88	CDPNT	4
[D2]	10	G	2
	27	AG	1
	32	DE	1
	39	EVQH	4
	83	DE	2
	85	G	2

Features of types 1 and 2 are favourable to B2M binding, whereas features of types 3 and 4 are unfavourable; features of types 1 and 3 are located in potential B2M contact zone.

(among 14) belong to CD1 whose G-LIKE-DOMAINs bind phospholipids (instead of peptides, see data section) and are much more hydrophobic than those of other MhcSF proteins; missclassification of these proteins was thus expected. Finally, the two other leave-one-out procedures show very high accuracy of 94 and 98%, for test sequences belonging to new species and new proteins, respectively. These results compare favourably with those of the simple score-based approach, which involves outputting the bound/unbound status of the protein that is closest (using FASTA with Blosum62) from the protein to be predicted. Using our three leave-one-out procedures, we found accuracies of 48, 89 and 100%, for new receptor type, species and protein, respectively. These results confirm our phylogenetic analysis (see above) indicating that new receptor prediction can hardly be done using protein neighbourhood, while the two other prediction tasks are relatively easy.

Final (i.e. without re-sampling) learning of our Bayes classifier is thus carried out for the 18 most discriminant features according to the χ^2-measure (2). These features are displayed in Table 1. Note that the same feature set is obtained using mutual information, which is another standard association measure (Shannon, 1948). We also evaluated the performance of two classifiers being built with 9 and 5 features located within and outside of the potential zone of B2M contact, respectively (this zone is defined hereafter). The number of features of each of these classifiers was adjusted as above described, starting from the initial multiple alignment but restricting it to the potential contact sites or excluding these sites, respectively. Each of these two classifiers proves to be as accurate as the classifier built with the 18 feature set (which includes the nine and five features of restricted classifiers). This experiment highlights a certain statistical redundancy of our 18 features. However, we shall see in the next section that all of our 18 selected features can be biologically and/or structurally interpreted and are thus useful for understanding MhcSF/B2M interaction.

4.2 Structural analysis of selected features

In order to identify potential sites of B2M contact on MHC-I and MHC-I-like heavy chains, we carried out an exhaustive contact analysis for the 165 known 3D structures of complexes between a MhcSF protein and B2M (see Supplementary Data 4 for details). Based on this analysis, selected features can be classified in four types, depending on whether they correspond or not to potential sites of B2M contact, and whether they are favourable or not for B2M binding. The latter distinction (favourable/unfavourable) results from an analysis of the diagonals of contingency Table 1 and identifies class-conserved features. For a given feature (i_k, g_k), a contingency Table 1 with dominant *ad* diagonal (>*bc*) indicates that an amino acid of group g_k at position i_k of a protein tends to be favourable for its interaction with B2M. In the same way, dominant *bc* diagonal indicates that an amino acid of g_k at i_k is unfavourable for B2M interaction.

Structural interpretation of selected features must then be carried out independently for each feature type. Nine features are favourable for the interaction with B2M, and are analysed using the 3D structure of *Rattus norvegicus* FCGRT. Indeed, this protein (with known structure) possesses an amino acid belonging to the conserved group of class C_β, for each of the nine positions involved. In the same way, nine features are unfavorable for the interaction with B2M and are analysed using the 3D structure of *Mus musculus* RAE1B (unbound to B2M and representative of $C_{-\beta}$ for the nine positions). Figure 3 displays the structural context of the selected features for these two 3D structures, while 3D coordinate files and PyMOL scripts for dynamic visualization are available in the Supplementary Data 5. Among the nine features which seem favourable for the interaction with B2M, four correspond to a position located in the potential zone of B2M contact. The same holds for five features among the nine that are unfavourable for the interaction with B2M.

Overall, features that are favourable for the interaction with B2M and located in the potential zone of contact seem to correspond to a side chain orientation or a physicochemical property favourable for direct contact with B2M, such as a large and aromatic residue F, W or Y at position [D1] 27 (W for *Rattus norvegicus* FCGRT). The features favourable for the interaction with B2M and located outside of the potential zone of contact seem to maintain a structure suitable for B2M contact; e.g. residues [D1] 51, [D2] 83 and 85 could ensure closure of the groove (by bringing the two helices closer) at one end. On the contrary, the unfavourable features located in the potential zone of contact seem to prevent direct contact by steric hindrance, such as residues N and K at position [D1] 8 and 25 of *Mus musculus* RAE1B, respectively. Destabilization of the conformation favourable for the interaction by residues such as E, V, Q or H at position [D2] 39 should be analysed in detail.

Definition of the features in terms of position and amino acid group thus facilitates determination of the physicochemical properties whose detection at a given position seems to be favourable or not for direct contact (for those located in the potential zone of B2M contact), or for stabilizing or not the molecular structure (for those located outside of this zone). The determination of these 4 types of features on heavy chains of MHC-I and MHC-I-like proteins should thus be valuable for future site-directed mutagenesis experiments.

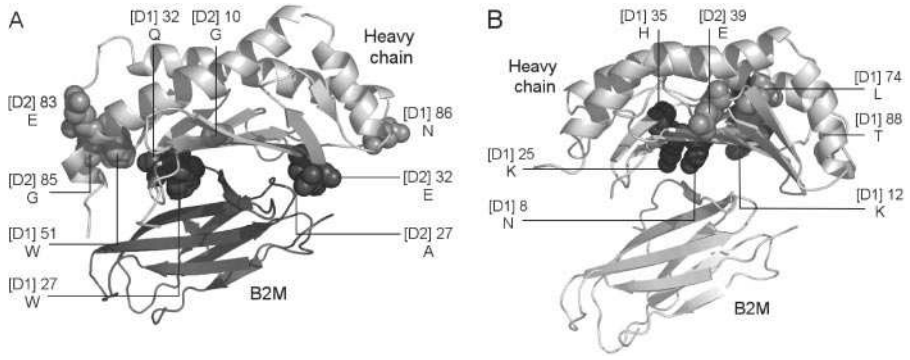


Fig. 3. Structural context of selected features for (**A**) *Rattus norvegicus* FCGRT and (**B**) *Mus musculus* RAE1B proteins. Each MHC-I-like heavy chain consists of the [D1] (in the back) and [D2] (in the front) extracellular domains. B2M is complexed with FCGRT, but virtually placed for RAE1B (see Supplementary Data 5). The C-LIKE extracellular domain of FCGRT heavy chain is not shown. Each feature is labelled with the domain, position and amino acid observed in the 3D structure; the corresponding side chains are represented by spheres. Features located in the potential B2M contact zone are shown in dark grey and the others are in light grey. Coordinate files: (A) 3fru and (B) 1jfm.

4.3 Site-directed mutagenesis

Polymorphism analysis and site-directed mutagenesis on MHC-I genes described in the literature relate mainly to the interaction affinity of MHC-I proteins with proteins required for peptide presentation (Paquet and Williams, 2002). Among them, the two site-directed mutagenesis on asparagine N (to aspartate D and glutamine Q) at position [D1] 86 of HLA-A gene are the only ones described as preventing the interaction between a MHC-I protein and B2M (Barbosa *et al*., 1987; Santos-Aguado *et al*., 1987). This partly supports the findings of our study as we found (Table 1) that position [D1] 86 associated with amino acid group NQ (amide) is favourable for the interaction with B2M. Our classifier highlights the importance of this position for B2M binding, but partly fails to identify the exact amino acid required as it suggests that mutation N>D could be deleterious, while overlooking that N>Q is also deleterious. In fact, all sequences in our dataset corresponding to B2M-bound proteins possess an N at [D1] 86, except those of CD1 which possess a Q at this position. Moreover, Q is totally absent at [D1] 86 in sequences corresponding to B2M unbound proteins. This explains selection (by our classifier) of the NQ group as being the most discriminant one at this position. However, as said earlier, CD1 proteins are atypical and much more hydrophobic than other MhcSF proteins. Thus, careful analysis of our dataset and of selected features also suggests that N>Q could be deleterious. We must keep in mind, however, that our dataset is limited and that the 18 features selected by our classifier only give statistical trends, which can be interpreted at the structural level, but should be validated by site-directed mutagenesis.

4.4 Prediction for lower vertebrate MHC-I sequences

We also classified 8 MHC-I proteins of lower vertebrates: *Salmon trutta* (Satr-UBA, Q9GJJ8 in UniProt/Swiss-Prot), *Ambystoma mexicanum* (P79458), *Oncorhynchus kisutch* (Onki-UA, Q9GJB4) and *Oncorhynchus mykiss* (Onmy-UAA, -UBA, -UCA, -UDA and -UEA; Shiina *et al*., 2005). Although sequences of amphibian and teleost MHC genes are known and their evolutionary origin well studied (Sammut *et al*., 1999; Hansen *et al*., 1999), few experimental data relate to cellular expression and interaction or not of their MHC-I protein with B2M (Antao *et al*., 1999). We thus analysed these 8 proteins by aligning them with IMGT multiple alignment (see above), numbering their positions, and applying our Bayes classifier. The prediction obtained in this way is the same for the 8 MHC-I proteins, which could hardly be due to chance (∼5%, given class priors), and indicates that those proteins very likely bind to B2M. This strongly suggests that they should be expressed on the cellular surface by the same process as that of mammalian MHC-I proteins.

5 DISCUSSION

This paper addresses the problem of predicting the interaction between MhcSF proteins and B2M, by only using sequences and multiple alignment. This problem is difficult, due to low sequence similarity of MhcSF proteins, and constitutes a good feasibility test of function and interaction prediction solely based on sequence information. Our method combines a simple-Bayes classifier with high-quality multiple alignment and unique numbering, as provided by the IMGT information system. Our results show that this method is accurate, even when the sequence to be predicted has low similarity with sequences in the learning set. Moreover, the results of our method are interpretable as it identifies sites associated with physicochemical properties that are well conserved within one class and avoided in the other. In our interaction problem, the conserved sites of both bound/unbound classes belong to the potential contact zone, but also stabilize, or not, the structure required for this contact. Finally, we show that the predictions of our method are confirmed by site-directed mutagenesis, and we illustrate its usefulness by analysing lower vertebrate MHC-I proteins which appear to be similar to mammalian MHC-I proteins. This simple method should thus be readily applicable to numerous other problems, when functions or interactions are to be predicted, and when a learning set of classified and aligned sequences is available. A direction for further research would be to combine our supervised classification approach with other methods, based on unsupervised classification and site conservation (Lichtarge *et al*., 1996; del Sol Mesa *et al*., 2003), using simple models of protein interaction (Gomez *et al*., 2003), or combining both functional and structural attributes of interacting protein sequence pairs (Huang *et al*., 2004).

ACKNOWLEDGEMENTS

This work was supported by CNRS, MENESR (doctoral grant to E.D.), Université Montpellier II Plan Pluri-Formation, ACI-IMPBIO, GIS AGENAE and BIOSTIC-LR.

Conflict of Interest: none declared.

REFERENCES

Antao,A.B. *et al.* (1999) MHC class I genes of the channel catfish: sequence analysis and expression. *Immunogenetics*, **49**, 303–311.

Bandyopadhyay,R., Tan,X.X., Matthews,K.S. and Subramanian,D. (2002) Predicting protein-ligand interactions from primary structure. *Technical Report TR02-398*. Rice University, Houston, TX.

Barbosa,J.A. *et al.* (1987) Site-directed mutagenesis of class I HLA genes. Role of glycosylation in surface expression and functional recognition. *J. Exp. Med.*, **166**, 1329–1350.

Cao,J. *et al.* (2003) A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics*, **19**, 234–240.

Collins,E.J. *et al.* (1995) The three-dimensional structure of a class I major histocompatibility complex molecule missing the alpha 3 domain of the heavy chain. *Proc. Natl Acad. Sci. USA*, **92**, 1218–1221.

Domingos,P. and Pazzani,M. (1996) Beyond independence: conditions for the optimality of the simple Bayesian classifier. *Proceedings of the Thirteenth International Conferences on Machine Learning (ICML)*, Bari, Italy, Morgan Kauffman, San Mateo, CA, pp. 105–112.

Duda,R.O., Hart,P.E. and Stork,D.G. (2001) *Pattern Classification*. 2nd edition. Wiley, New York.

Duprat,E. *et al.* (2004) IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. *Recent Res. Dev. Hum. Genet.*, **2**, 111–136.

D'Urso,C.M. *et al.* (1991) Lack of HLA class I antigen expression by cultured melanoma cells FO-1 due to a defect in B2m gene expression. *J. Clin. Invest.*, **87**, 284–292.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Feder,J.N. *et al.* (1998) The hemochromatosis gene product complexes with the transferrin receptor and lowers its affinity for ligand binding. *Proc. Natl Acad. Sci. USA*, **95**, 1472–1477.

Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics*, **15**, 1047–1054.

Gomez,S.M., Noble,W.S. and Rzhetsky,A. (2003) Learning to predict protein–protein interactions from protein sequences. *Bioinformatics*, **19**, 1875–1881.

Good,I.J. (1965) The estimation of probabilities: an essay on modern Bayesian methods. In *Research Monograph 30*. MIT Press, Cambridge, MA.

Guindon,S. and Gascuel,O. (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Hand,D.J. (1986) Recent advances in error rate estimation. *Pattern Recogn. Lett.*, **4**, 335–346.

Hansen,J.D. *et al.* (1999) Expression, linkage, and polymorphism of MHC-related genes in rainbow trout, *Oncorhynchus mykiss*. *J. Immunol.*, **163**, 774–786.

Hill,D.M. *et al.* (2003) A dominant negative mutant B2-microglobulin blocks the extracellular folding of a major histocompatibility complex class I heavy chain. *J. Biol. Chem.*, **278**, 5630–5638.

Holmes,M.A. *et al.* (2002) Structural studies of allelic diversity of the MHC class I homolog MIC-B, a stress-inducible ligand for the activating immunoreceptor NKG2D. *J. Immunol.*, **169**, 1395–1400.

Huang,Y. *et al.* (2004) Predicting protein–protein interactions by a supervised learning classifier. *Comput. Biol. Chem.*, **28**, 291–301.

Kaas,Q. *et al.* (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.

Kaas,Q. and Lefranc, M.-P. (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biology*, **5(4)**, 0046 (advance access).

Kohavi,R. *et al.* (1997) Improving Simple Bayes. *Proceedings of the Ninth European Conference on Machine Learning (ECML)*, Springer Verlag, Heidelberg, pp. 78–87.

Lefranc,M.-P. *et al.* (2005a) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev. Comp. Immunol.*, **29**, 917–938.

Lefranc,M.-P. *et al.* (2005b) IMGT, the international ImMunoGeneTics information system®. *Nucleic Acids Res.*, **33**, D593–D597.

Lefranc,M.-P. *et al.* (2005c) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.*, **29**, 185–203.

Li,P. *et al.* (2002) Crystal structures of RAE-1beta and its complex with the activating immunoreceptor NKG2D. *Immunity*, **16**, 77–86.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Lidstone,G. (1920) Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Trans. Fac. Act.*, **8**, 182–192.

Miley,M.J. *et al.* (2003) Biochemical features of the MHC-related protein 1 consistent with an immunological function. *J. Immunol.*, **170**, 6090–6098.

Paquet,M.-E. and Williams,D.B. (2002) Mutant MHC class I molecules define interactions between components of the peptide-loading complex. *Int. Immunol.*, **14**, *347–358*.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Pommié,C. *et al.* (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recogn.*, **17**, 17–32.

Sali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.

Sammut,B. *et al.* (1999) Axolotl MHC architecture and polymorphism. *Eur. J. Immunol.*, **29**, 2897–2907.

Sanchez,L.M. *et al.* (1999) Crystal structure of human ZAG, a fat-depleting factor related to MHC molecules. *Science*, **283**, 1914–1919.

Santos-Aguado,J. *et al.* (1987) Amino acid sequences in the alpha 1 domain and not glycosylation are important in HLA-A2/beta2-microglobulin association and cell surface expression. *Mol. Cell. Biol.*, **7**, 982–990.

Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

Shiina,T. *et al.* (2005) Interchromosomal duplication of major histocompatibility complex class I regions in rainbow trout (*Oncorhynchus mykiss*), a species with a presumably recent tetraploid ancestry. *Immunogenetics*, **56**, 878–893.

Simmonds,R.E. and Lane,D.A. (1999) Structural and functional implications of the intron/exon organization of the human endothelial cell protein C/activated protein C receptor (EPCR) gene: comparison with the structure of CD1/major histocompatibility complex alpha1 and alpha2 domains. *Blood*, **94**, 632–641.

del Sol Mesa,A. *et al.* (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.

Thompson,J.D. *et al.* (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.*, **314**, 937–951.

West,A.P.,Jr and Bjorkman,P.J. (2000) Crystal structure and immunoglobulin G binding properties of the human major histocompatibility complex-related Fc receptor. *Biochemistry*, **39**, 9698–9708.

Wu,T.D. and Brutlag,D.L. (1995) Identification of protein motifs using conserved amino acids properties and partitioning techniques. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **19**, 402–410.

Zeng,Z.-H. *et al.* (1997) Crystal structure of mouse CD1: An MHC-like fold with a large hydrophobic binding groove. *Science*, **277**, 339–345.