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; [2]Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France; [3]Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM, UMR 5506 CNRS Université Montpellier II, 161 rue Ada, 34392 Montpellier, France

RESUME

Motivations: Les protéines du complexe majeur d'histocompatibilité (MHC) assurent une fonction essentielle au sein du système immunitaire, en présentant des peptides du soi ou des peptides antigéniques aux récepteurs T. La liaison non covalente de la beta2-microglobuline aux protéines du MHC de classe I est nécessaire à leur expression à la surface cellulaire et à la présentation des peptides. La superfamille du MHC regroupe les protéines de structure homologue aux protéines MHC-I. Ces protéines sont impliquées dans une grande variété de processus biologiques et interagissent ou non avec la beta2-microglobuline. La prédiction de la liaison (ou de l'absence de liaison) à la beta2-microglobuline, pour des protéines de la superfamille du MHC 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 et la méthode d'alignement (pour partie structurale) mises en place au sein d'IMGT s'appliquent avec succès aux protéines de la superfamille du MHC malgré leur faible similarité de séquence, et fournissent une numérotation unique des résidus qui favorise le développement d'une telle approche prédictive.

Résultats: La méthode proposée combine un classifieur Bayesien dit naïf et la numérotation unique IMGT. Elle est composée de deux étapes : un ensemble de descripteurs binaires discriminants (associant une position dans l'alignement et un groupe d'acides aminés) est tout d'abord extrait des données ; les fréquences de ces descripteurs sont ensuite estimées conditionnellement aux classes que l'on cherche à séparer, pour construire le classifieur. Nous appliquons cette approche à un jeu de données composé de 807 séquences alignées, correspondant aux allèles de 47 gènes de la superfamille du MHC. 18 descripteurs sont sélectionnés pour leur capacité à discriminer les protéines selon qu'elles se lient ou non à la beta2-microglobuline. L'analyse structurale des protéines du jeu de données montre que ces descripteurs correspondent à des sites potentiels de contact à la beta2-microglobuline 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 un nouveau gène, une espèce non référencée au sein des données ou un nouveau type de récepteur, avec respectivement 98%, 93% et 79% 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.

Données complémentaires: Les séquences alignées de la superfamille du MHC qui composent le jeu de données sont accessibles dans les sections MHC et RPI d'IMGT Repertoire ; les structures actuellement résolues sont accessibles sous forme de fichiers de coordonnées annotés dans IMGT/3Dstructure-DB (http://imgt.cines.fr).

[*] Auteur à contacter : duprat@ligm.igh.cnrs.fr

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 des peptides antigéniques 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 de structure homologue aux protéines MHC-I. Ces protéines MHC-I-like sont impliquées dans une grande variété de processus biologiques correspondant à différents sites d'interaction protéine-ligand sur la chaîne lourde. Les protéines MHC-I-like correspondant à 34 gènes de mammifères ont été identifiées comme membres de la MhcSF, dont 12 structures 3D sont actuellement disponibles. Les données expérimentales concernant les protéines issues de l'expression de chacun de ces gènes indiquent que 17 correspondent à des protéines liées à la B2M, tandis que 17 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.

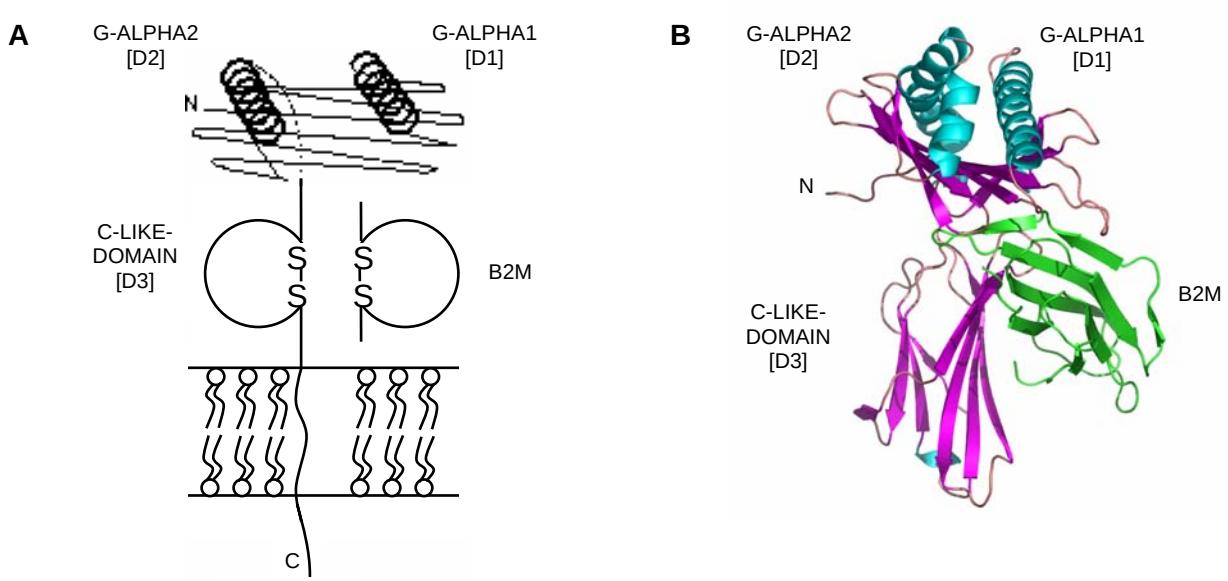

Figure 1. Représentation schématique (A) et structure 3D (B) des protéines MHC-I. (A) Représentation schématique des protéines MHC-I. La protéine MHC-I est représentée comme une protéine transmembranaire, à la surface d'une cellule cible. Les chaînes MHC-I complètes comprennent les 3 domaines extracellulaires G-ALPHA1, G-ALPHA2 et C-LIKE, et les 3 régions intermédiaire, transmembranaire et cytoplasmique (absentes dans la structure 3D) [10,15]. La B2M est composée d'un unique domaine extracellulaire, lié de façon non covalente à la chaîne lourde MHC-I. (B) Les structures secondaires de la chaîne lourde indiquées dans le fichier de coordonnées (annoté 1a1o issu de IMGT/3Dstructure-DB [20]) sont représentées en cyan (hélices), magenta (brins) et saumon (boucles). La B2M est représentée en vert. [D1], [D2] et [D3] indiquent les domaines et leur position en partant de l'extrémité N-terminale.

L'ontologie IMGT [13] établit les règles de description des domaines protéiques de la superfamille du MHC [10,14,15]. Deux G-DOMAINs (G-ALPHA1 [D1] et G-ALPHA2 [D2]) et deux G-LIKE-DOMAINs (G-ALPHA1-LIKE [D1] et G-ALPHA2-LIKE [D2]) correspondent respectivement aux domaines extracellulaires N-terminaux de la chaîne lourde des protéines MHC-I et MHC-I-like (Figure 1). Les G-DOMAINs et G-LIKE-DOMAINs sont des homologues structuraux, chacun constitué d'un feuillet de quatre brins beta antiparallèles (notés A, B, C et D) et d'une longue hélice [10,15]. Certains G-DOMAINs et G-LIKE-DOMAINs sont caractérisés par une faible similarité de séquence. 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 la chaîne lourde des protéines MHC-I et MHC-I-like liées à la B2M est

un C-LIKE-DOMAIN [D3] (Figure 1). Les chaînes lourdes de MHC-I-like non liées à la B2M comprennent ou non un C-LIKE-DOMAIN [D3]. La déletion *in vitro* du C-LIKE-DOMAIN d'une chaîne lourde de MHC-I n'affecte ni sa structure, ni sa liaison à B2M et au peptide [16]. La présence ou non d'un C-LIKE-DOMAIN au sein de la chaîne lourde d'une protéine de la MhcSF ne semble donc pas un critère efficace de discrimination entre les protéines liées ou non à la B2M, et les résultats présentés ici se basent 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 de la MhcSF nouvellement identifiées. Nous nous appuyons sur l'alignement 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 [17] a été appliqué avec succès à la prédiction de ligands classe-spécifiques, les caractéristiques fonctionnelles des classes étant connues [18,19]. 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. Ce classifieur se base sur un ensemble de descripteurs binaires (associant 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 (Partie 2). Nous décrivons (Partie 3) la méthode d'extraction des descripteurs discriminants, le classifieur Bayesien employé, et les procédures mises en place pour évaluer les performances de ce classifieur. Les résultats sont présentés en Partie 4, où nous donnons une interprétation structurale ainsi que l'analyse de mutants et les prédictions sur des protéines dont la liaison à la B2M est inconnue à ce jour.

2 LES DONNEES

La superfamille du MHC (MhcSF) est constituée de protéines MHC-I liées à la B2M et de protéines MHC-I-like liées ou non à la B2M. Les types de récepteurs de MHC-I les plus étudiés sont le HLA d'*Homo sapiens*, H2 de *Mus musculus* et RT1 de *Rattus norvegicus*. Les récepteurs HLA et H2 correspondent chacun à 6 gènes: 3 gènes MHC-Ia classiques (HLA-A, HLA-B, HLA-Cw et H2-D, H2-K, H2-Q) et 3 gènes MHC-Ib non classiques (HLA-E, HLA-F, HLA-G et H2-L, H2-M, H2-T) [10] ; seul le gène RT1-AA est pris en compte ici pour le MHC-I de *Rattus norvegicus*. 9 types de récepteurs de MHC-I-like sont actuellement répertoriés : FCGRT, MR1, HFE, CD1, AZGP1, MIC, EPCR, RAE1 et RAET1. Chacun de ces récepteurs a été identifié au sein du génome de nombreux mammifères. Nous avons sélectionné pour notre étude 4 espèces représentatives : *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* et *Bos taurus*. Chaque récepteur est caractérisé par une fonction et une composition en domaines protéiques similaire, quelle que soit l'espèce. Toutes les protéines d'un même type de récepteur interagissent (FCGRT, MR1, HFE, CD1) ou non (CD1, AZGP1, MIC, EPCR, RAE1, RAET1) avec la B2M. La structure d'au moins un récepteur de chaque type a été résolue, et est disponible dans la base de données de structure IMGT/3DStructure-DB [20]. Les 47 gènes MHC-I et MHC-I-like étudiés correspondent donc à 12 types de récepteurs, et sont répertoriés dans notre étude parmi 4 espèces. Le polymorphisme des gènes est pris en compte par la pondération des séquences alléliques décrites au sein d'IMGT ; pour un gène possédant e formes alléliques, un poids de $1/e$ est attribué à chacune de ces séquences. Le faible nombre de gènes (13 MHC-I et 34 MHC-I-like) est donc en partie compensé par le nombre élevé d'allèles pris en compte (768 MHC-I et 39 MHC-I-like). Dans la méthode présentée en Partie 3, chaque étape est détaillée dans un premier temps dans le cas classique, puis dans le cas de gènes polymorphes. Les 807 séquences alléliques correspondant aux 47 gènes MHC-I et MHC-I-like étudiés sont accessibles dans les sections MHC et RPI d'IMGT Répertoire [21], qui

présentent les données expertisées par IMGT concernant les MHC et les protéines apparentées au système immunitaire.

Les G-DOMAINs et G-LIKE-DOMAINs sont observés au sein de toutes les protéines MHC-I et MHC-I-like, contrairement au C-LIKE-DOMAIN qui est absent dans certaines protéines MHC-I-like. Comme indiqué précédemment, la présence ou non d'un C-LIKE-DOMAIN ne semble pas corrélée avec la liaison ou non à la B2M, et seuls les G-DOMAINs et G-LIKE-DOMAINs des protéines MHC-I et MHC-I-like sont pris en compte dans notre étude. La numérotation unique IMGT pour les G-DOMAINs et G-LIKE-DOMAINs [10] a été établie par alignements successifs des séquences (issues d'IMGT Repertoire) et des structures 3D (issues d'IMGT/3Dstructure-DB) de chaînes lourdes de MHC-I et MHC-I-like. Alors que les G-DOMAINs et G-LIKE-DOMAINs sont des homologues structuraux, seules les séquences de G-DOMAINs sont très similaires (plus de 90% d'identité). La similarité des séquences de MHC-I-like est forte (60-90% d'identité) au sein d'un même type de récepteur, mais faible entre eux (20-40% d'identité) ; ainsi, les séquences des G-LIKE-DOMAINs de deux protéines appartenant à une même espèce mais à deux types de récepteurs différents (ex: FCGRT et RAE1 de *Mus musculus*) ne peuvent pas être alignées correctement. Après avoir aligné les séquences de G-DOMAINs des protéines MHC-I, deux étapes sont donc nécessaires pour aligner les protéines MHC-I-like entre elles et avec les protéines MHC-I. Nous réalisons dans un premier temps un alignement structural multiple (avec COMPARER [22]) à partir d'une structure 3D de chacun des 12 types de récepteur MHC-I-like et de 6 structures 3D de MHC-I. Pour chaque type de récepteur, nous alignons alors les séquences restantes avec la séquence de la protéine précédemment alignée en structure. L'alignement multiple ainsi constitué comprend 807 séquences, qui correspondent aux allèles de 47 gènes de la MhcSF. Chacune de ces séquences est composée d'un domaine [D1] (G-ALPHA1 ou G-ALPHA1-LIKE) et d'un domaine [D2] (G-ALPHA2 ou G-ALPHA2-LIKE). La cohérence de cet alignement est validée en terme de séquence par NorMD [23] et en terme de structure par Profit (disponible à l'adresse http://bioinf.org.uk/software/profit). Une nouvelle protéine MHC-I-like sera donc alignée en séquence si elle correspond à un récepteur connu, ou en structure si elle correspond à un nouveau type de récepteur.

L'arbre phylogénétique obtenu à partir des ces séquences 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. Cette observation est cohérente avec les taux de similarité donnés plus haut entre G-DOMAINs et G-LIKE-DOMAINs. 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. Par contre, 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 la Partie 4).

3 LE CLASSIFIEUR BAYESIEN "NAÏF"

Le classifieur Bayesien dit naïf permet d'estimer les probabilités qu'une nouvelle séquence *s* de la superfamille du MHC interagisse ou non avec la B2M. Deux étapes préalables au classement sont nécessaires : (1) un ensemble de descripteurs binaires discriminants (associant une position dans l'alignement et un groupe

d'acides aminés) est extrait des données ; (2) les fréquences de ces descripteurs sont estimées conditionnellement aux classes de séquences que l'on souhaite séparer, pour construire le classifieur. 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 un nouveau gène, une espèce non référencée au sein des données ou un nouveau type de récepteur.

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 [24] et d'après [25]. 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 gènes polymorphes, la table de contingence établie pour une position i et un groupe d'acides aminés g est basée sur les poids des allèles. Un gène de C_β représenté sous 10 formes alléliques 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 superfamille du MHC 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 [26]. 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 gènes 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 [27], 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 de C_X sachant $D(s)$ définie en (4) 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 [28]. 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 gènes 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 gènes polymorphes, 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 formes alléliques d'un même gène. Afin de réduire le temps de calcul, nous considérons donc chaque gène comme un profil p composé d'une ou plusieurs séquences alléliques. Alors que la position i_k de la séquence s d'un gène 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 du gène, 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 superfamille du MHC concernent un nombre restreint de gènes, 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" [29] 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 un nouveau gène, une espèce on référencée au sein des données ou un nouveau type de récepteur. Les séquences de chacun des 12 types de récepteurs, de chacune des 4 espèces et de chacun des 47 gènes 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 RESULTATS ET DISCUSSION

4.1 Performance du classifieur et nombre de descripteurs

Les taux de bon classement sont représentés en Figure 2, pour toutes les valeurs de $f = 1, 2...40$ et les trois procédures de leave-one-out. La performance la plus faible est obtenue lorsque tous les gènes d'un même type de récepteur constituent l'échantillon de test, et ce quel que soit le nombre de descripteurs. Ce résultat était prévisible. En effet, 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 79% des séquences (37 gènes parmi les 47 gènes du jeu de données) dont le type de récepteur n'est pas représenté au sein de 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 gènes mal classés (sur 10) 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 beaucoup plus hydrophobes que ceux des autres protéines de la MhcSF. Pour les 2 autres leave-one-out, 93% et 98% des séquences du jeu de données 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'un même gène.

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. 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 79%, 93% 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.

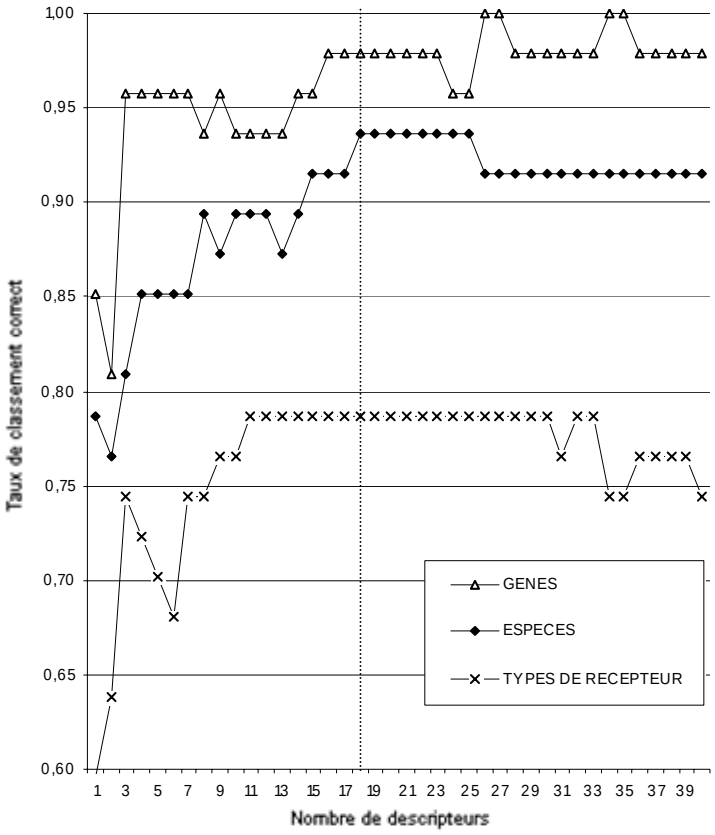

Figure 2. Performance des classifieurs en fonction du nombre de descripteurs, pour les 3 procédures de leave-one-out. La performance la plus élevée pour les 3 procédures est obtenue pour 18 descripteurs.

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

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.

Afin d'identifier les sites potentiels de liaison à la B2M au sein des chaînes lourdes de protéines MHC-I et MHC-I-like, nous avons réalisé une analyse exhaustive des contacts pour les structures 3D de 29 protéines de la MhcSF liées à la B2M (26 protéines MHC-I et 6 protéines MHC-I-like). Les sites potentiels de contact correspondent aux positions de [D1] et [D2] identifiées en contact avec la B2M dans au moins une de ces structures. La zone de contact potentiel que nous avons identifiée correspond aux brins et boucles du domaine [D1] (G-ALPHA1 ou G-ALPHA1-LIKE) et aux brins A, B et C et à la boucle BC du domaine [D2] (G-ALPHA2 ou G-ALPHA2-LIKE).

Table 1. Les 18 descripteurs sélectionnés.

Domaine IMGT	Position IMGT	Groupe discriminant	Type de descripteur
[D1]	8	CDPNT	3
	11	ILV	3
	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	1
	27	AG	1
	32	DE	1
	39	EVQH	4
	83	DE	2
	85	G	2

Les descripteurs de type 1 et 2 sont favorables à la liaison à la B2M et représentés respectivement en jaune et bleu en Figure 3A. Les types 3 et 4 sont défavorables et représentés respectivement en vert et orange en Figure 3B. Les types 1 et 3 sont dans la zone de contact à la B2M.

La Figure 3 représente ces deux ensembles de 9 descripteurs au sein des structures 3D de FCGRT de *Rattus norvegicus* et de RAE1B de *Mus musculus*. Parmi les 9 descripteurs qui semblent favorables à l'interaction à la B2M (Figure 3A), 5 correspondent à une position localisée dans la zone de contact potentiel (en jaune sur la Figure 3A). C'est également le cas de 6 descripteurs (en vert sur la Figure 3B) parmi les 9 défavorables à l'interaction à la B2M (Figure 3B).

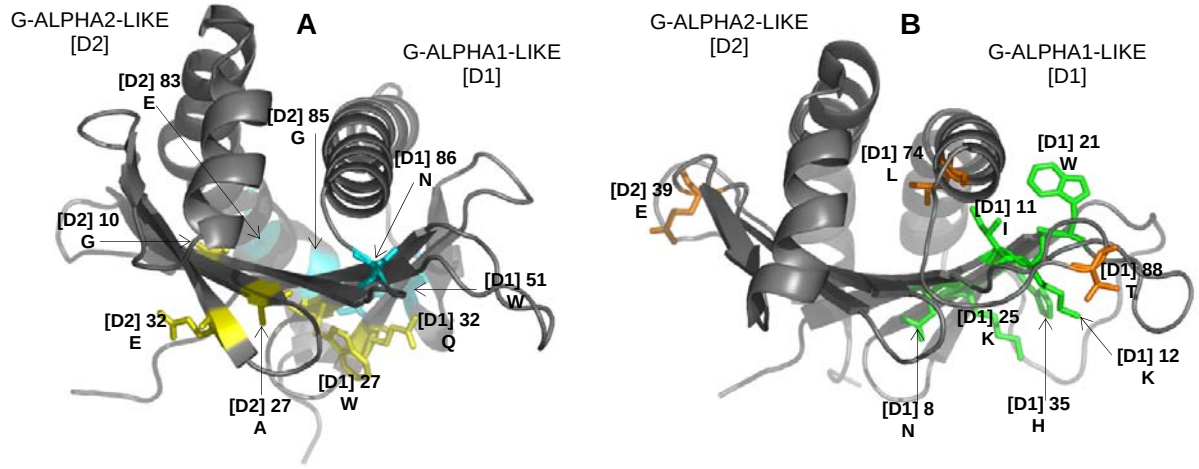

Figure 3. Structure 3D des deux G-LIKE-DOMAINs des protéines FCGRT de *Rattus norvegicus* (A) et RAE1B de *Mus musculus* (B), et représentation des 4 types de descripteurs. (A) Fichier de coordonnées 3fru, chaîne A ; les chaînes latérales correspondant aux positions des 9 descripteurs favorables à la liaison à la B2M sont représentées en jaune pour la zone de contact potentiel, ou en bleu sinon. (B) Fichier de coordonnées 1jfm, chaîne A ; les chaînes latérales correspondant aux positions des 9 descripteurs défavorables à la liaison à la B2M sont représentées en vert pour la zone de contact potentiel, ou en orange sinon. Le domaine, la position et l'acide aminé observé dans la structure de référence sont indiqués pour chaque descripteur. Visualisation par le logiciel Pymol, librement accessible sur http://pymol.sourceforge.net.

Globalement, les descripteurs favorables à l'interaction 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 aromatique F, W ou Y en position [D1] 27 (W pour FCGRT de *Rattus norvegicus* en Figure 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 (Figure 3A). Au contraire, les descripteurs défavorables situés dans la zone de contact potentiel semblent empêcher le contact direct par gêne stérique, tels que les résidus K en position [D1] 12 et [D1] 25 de RAE1B de *Mus musculus* (Figure 3B). La déstabilisation de la conformation propice à l'interaction par des résidus tels que E, V, Q ou H en position [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 de contact potentiel), 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 en position [D1] 86 du gène HLA-A empêche la liaison de la protéine HLA-A à la B2M [30,31]. Cette observation conforte notre étude, puisque 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 localisée hors de la zone de contact potentiel 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 chaîne lourde de HLA-A et la B2M. Sur la base de nos résultats, on peut étendre ce résultat expérimental en prédisant que la liaison d'un oligosaccharide en position [D1] 86 est caractéristique des protéines MHC-I et MHC-I-like qui se lient à la B2M.

Nous avons également classé 3 protéines MHC-I de vertébrés inférieurs, correspondant à un gène MHC-Ia de *Salmon trutta* (Satr-UBA, Q9GJJ8 dans Swiss-prot), MHC-Ia d'*Ambystoma mexicanum* (P79458) et MHC-Ib de *Oncorhynchus kisutch* (Onki-UAA, Q9GJB4). Alors que la structure et l'origine évolutive des gènes du MHC des amphibiens [32] et des poissons téléostéens [33] 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 [34]. Le classifieur rattache les protéines MHC-Ia de *Salmon trutta* et *Ambystoma mexicanum* 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 MHC-Ib d'*Oncorhynchus kisutch*, ce qui est également favorable à la liaison, mais avec une certitude moindre. Cette étude donne donc à penser que les protéines MHC-I étudiées pour ces trois espèces se lient à la B2M. Elles devraient par conséquent être exprimés à la surface cellulaire par le même processus que les MHC-I et MHC-I-like de mammifères liés à la B2M, et reconnaître des ligands homologues à ceux de la B2M chez les mammifères.

5 CONCLUSION

Notre étude met en évidence l'efficacité de la combinaison du classifieur Bayesien naïf et de la numérotation unique IMGT, pour prédire l'interaction des protéines de la superfamille du MHC 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 globale 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.

REFERENCES

[1] Boyd,L.F., Kozlowski,S. and Margulies,D.H. (1992) Solution binding of an antigenic peptide to a major histocompatibility complex class I molecule and the role of B2-microglobulin. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 2242-2246.

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

[3] Shields,M.J., Kubota,R., Hodgson,W., Jacobson,S., Biddison,W.E. and Ribaudo,R.K. (1998) 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. *Journal of Biological Chemistry*, 273, 28010-28018.

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

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

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

[7] Williams DB, Barber BH, Flavell RA and Allen H. (1989) Role of beta2-microglobulin in the intracellular transport and surface expression of murine class I histocompatibility molecules. *The Journal of Immunology*, 142, 2796-2806.

[8] D'Urso,C.M., Wang,Z.G., Cao,Y., Tatake,R., Zeff,R.A. and Ferrone,S. (1991) Lack of HLA class I antigen expression by cultured melanoma cells FO-1 due to a defect in B2m gene expression. *The Journal of Clinical Investigation*, 87, 284-292.

[9] Wang,Z., Cao,Y., Albino,A.P., Zeff,R.A., Houghton,A. and Ferrone,S. (1993) Lack of HLA class I antigen expression by melanoma cells SK-MEL-33 caused by a reading frameshift in Beta2-microglobulin messenger RNA. *The Journal of Clinical Investigation*, 91, 684-692.

[10] 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* (in press)

[11] Porgador,A., Mandelboim,O., Restifo,N.P. and Strominger,J.L. (1997) Natural killer cell lines kill autologous B2-microglobulin-deficient melanoma cells: implications for cancer immunotherapy. *Proceedings of the National Academy of Sciences of the United States of America*, 94, 13140-13145.

[12] Cook,A.L. (2004) Beta 2 microglobulin and resistance to murine respiratory mycoplasmosis. *Contemporary Topics in Laboratory Animal Science*, 43, 18-24.

[13] Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics*, 15:1047-1054.

[14] Duprat,E., Kaas,Q., Garelle,V. 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 and Developments in Human Genetics*, 2, 111-136.

[15] 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: Opportunities and challenges of bridging immunology with computer and information sciences.* (in press)

[16] Collins,E.J., Garboczi,D.N., Karpusas,M.N. and Wiley,D.C. (1995) The three-dimensional structure of a class I major histocompatibility complex molecule missing the alpha3 domain of the heavy chain. *Proceedings of the National Academy of Sciences of the United States of America*, 92:1218-1221.

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

[18] Bandyopadhyay,R., Tan,X.X., Matthews,K.S. and Subramanian,D. (2002) Predicting protein-ligand interactions from primary structure. *Technical Report, Rice University*, TR02-398.
[19] Cao,J., Panetta,R., Yue,S., Steyaert,A., Young-Bellido,M. and Ahmad,S. (2003) A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics*, 19, 234-240.
[20] Kaas,Q., Ruiz,M., Lefranc,M.-P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Research*, 32, D208-D210.
[21] Lefranc,M.-P., Giudicelli,V., Kaas,Q., Duprat,E., Jabado-Michaloud,J., Scaviner,D., Ginestoux,C., Clément,O., Chaume,D. and Lefranc,G. (2005) IMGT, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, 33, D593-D597.
[22] 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. *Journal of Molecular Biology*, 212, 403-428.
[23] Thompson,J.D., Plewniak,F., Ripp,R., Thierry,J.C. and Poch,O. (2001) Towards a reliable objective function for multiple sequence alignments. *Journal of Molecular Biology*, 314, 937-951.
[24] Pommié,C., Levadoux,S., Sabatier,R., Lefranc,G. and Lefranc M.-P. (2003) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acids properties. *Journal of Molecular Recognition*, 17, 17-32.
[25] Wu,T.D. and Brutlag,D.L. (1995) Identification of protein motifs using conserved amino acids properties and partitioning techniques. *Proceedings of the Thirteenth International Conference on Intelligent Systems for Molecular Biology*, 19, 402-410.
[26] Domingos,P. and Pazzani,M. (1996) Beyond independence: conditions for the optimality of the simple Bayesian classifier. *Proceedings of the Thirteenth International Conference on Machine Learning*, 105-112. Bari, Italy: Morgan Kaufmann.
[27] Lidstone,G. (1920) 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.
[28] Kohavi,R., Becker,B. and Sommerfield,D. (1997) Improving simple bayes. *Proceedings of the Ninth European Conference on Machine Learning.* Poster Paper.
[29] Hand,D.J. (1986) Recent advances in error rate stimation. *Pattern Recognition Letters*, 4, 335-346.
[30] Barbosa,J.A., Santos-Aguado,J., Mentzer,S.J, Strominger,J.L., Burakoff,S.J. and Biro,A.P. (1987) Site-directed mutagenesis of class I HLA genes. Role of glycosylation in surface expression and functional recognition. *The Journal of Experimental Medicine*, 166, 1329-1350.
[31] Santos-Aguado,J., Biro,A.P., Fuhrmann,U., Strominger,J.L. and Barbosa,J.A. (1987) Amino acid sequences in the alpha1 domain and not glycosylation are important in HLA-A2/beta2-microglobulin association and cell surface expression. *Molecular and Cellular Biology*, 7, 982-990.
[32] Sammut,B., Du Pasquier,L., Ducoroy,P., Laurens,V., Marcuz,A. and Tournefier,A. (1999) Axolotl MHC architecture and polymorphism. *European Journal of Immunology*, 29, 2897-2907.
[33] Hansen,J.D., Strassburger,P., Thorgaard,G.H., Young,W.P. and Du Pasquier,L. (1999) Expression, linkage, and polymorphism of MHC-related genes in Rainbow trout, *Oncorhynchus mykiss*. *The Journal of Immunology*, 163:774-786.
[34] Antao,A.B., Chinchar,V.G., McConnell,T.J., Miller,N.W., Clem,L.W. and Wilson,M.R. (1999) MHC class I genes of the channel catfish: sequence analysis and expression. *Immunogenetics*, 49, 303-311.