Analyse Statistique des Propriétés Physicochimiques des Acides Aminés des Régions Hypervariables des Anticorps

Safa Aouinti[1,2,3], Dhafer Malouche[2,3], Marie-Paule Lefranc[1]

[1] *IMGT®, the international ImMunoGeneTics information system®, Institut de Génétique Humaine, UPR CNRS 1142, Université Montpellier 2, Montpellier, France, safa.aouinti@igh.cnrs.fr, marie-paule.lefranc@igh.cnrs.fr*

[2] *École Supérieure de la Statistique et de l'Analyse de l'Information de Tunis, Tunisie, dhafer.malouche@me.com*

[3] *École Nationale d'Ingénieurs de Tunis-(U2S), Tunisie*

Résumé. L'anticorps est une protéine qui détecte et neutralise spécifiquement les antigènes qui représentent toute substance étrangère à l'organisme déclenchant une réponse immunitaire pour l'éliminer. Il est constituée de deux chaînes lourdes (H) et deux légères (L) où la chaîne L est une chaîne kappa ou lambda. Chaque chaîne est constituée par un domaine variable V et un constant C. Le domaine V est constitué des régions hypervariables ou CDR (*complementarity determining region*) qui déterminent le site de reconnaissance et de liaison à l'antigène et des régions dites charpentes ou FR (*framework region*). Une étude statistique a été menée sur les propriétés physicochimiques des séquences des acides aminés des trois régions hypervariables (CDR-IMGT) du domaine V des anticorps. Pour les CDR1-IMGT et CDR2-IMGT l'analyse a été effectuée sur 36811 séquences (23174 IGHV, 7460 IGKV et 6177 IGLV) et 7961 séquences pour les CDR3-IMGT (2178 IGHV, 3084 IGKV et 2699 IGLV). Cette étude a permis de localiser les positions importantes pour la diversité et celles conservées ainsi que l'extraction des acides aminés importants pour le maintient de la structure de l'anticorps.

Mots-clés. Analyses Multivariées, méthodes de classification, réseaux graphiques, anticorps, IMGT, CDR-IMGT, propriétés physicochimiques, IMGT Colliers de Perles, acides aminés.

Abstract. The antibody is a protein to detect and neutralize specifically the antigens that are foreign substances to the organism able to trigger an immune response to remove it. It comprises two heavy (H) chains and two light (L) chain, where L is a kappa chain or a lambda chain. Each chain consists of a variable V and a constant C domain. An antibody variable domain comprises the hypervariable regions or CDR (complementarity determining region) which determine the site of recognition as well as antigen-binding and regions called framework or FR (framework region). A statistical study was conducted on the physicochemical properties of the amino acid sequences of the three V domain hypervariable regions (CDR-IMGT). Analysis for the CDR1-IMGT and CDR2-IMGT was performed on 36811 sequences (23174 IGHV, 7460 IGKV et 6177 IGLV) and 7961 sequences for the CDR3-IMGT (2178 IGHV, 3084 IGKV et 2699 IGLV). The study allowed the localization of important diversity positions and those preserved as well as extracting the important amino acids to maintain the antibody structure.

Keywords. Multivariate analysis, classification methods, antibodies, graphics networks, IMGT, CDR-IMGT, physicochemical properties, IMGT Colliers de Perles, amino acids.

1 Introduction

IMGT® (`http://www.imgt.org`), the international ImMunoGeneTics information system®, crée par Marie-Paule Lefranc en 1989, est l'unique système d'information intégré en immunogénétique et immunoinformatique. IMGT® est spécialisé dans les séquences, structures et données génétiques des immunoglobulines (IG) ou anticorps, récepteurs T (TR) et proteines majeures d'histocompatibilité (MH) des vertébrés ainsi que des protéines des superfamilles IgSF et MhSF du système immunitaire (SI) des vertébrés et invertébrés [Lefranc M-P. (2009)].

Pour comprendre les caractéristiques structurales des domaines des anticorps en relation avec les propriétés physicochimiques des acides aminés qui constituent leurs chaînes protéiques, IMGT® a effectué une première analyse statistique sur les régions charpentes (FR-IMGT) délimitées de façon standardisée grâce à la numérotation unique IMGT. Ces résultats ont fait l'objet d'une publication par [Pommié et al. 2004] qui sert de référence pour la comparaison entre les anticorps nouvellement étudiés et le répertoire exprimé chez l'homme. Le but de ce papier est d'analyser et voir s'il existe des conservations entre les positions des régions hypervariables ou CDR-IMGT des gènes V des chaînes lourdes (IGH) et légères (IGK ou IGL) et les classes physicochimiques.

2 Méthodologie

Trois études ont été réalisées sur les propriétés physicochimiques des séquences des acides aminés des trois régions hypervariables (CDR-IMGT) du domaine V des anticorps. Pour les CDR1-IMGT et CDR2-IMGT, l'analyse a été effectuée sur 36811 séquences (23174 IGHV, 7460 IGKV et 6177 IGLV) et 7961 séquences pour les CDR3-IMGT (2178 IGHV, 3084 IGKV et 2699 IGLV). Ces études ont abouti à des résultats complémentaires pour pouvoir répondre aux questions qui concernent la problématique de la diversité des boucles hypervariables des domaines variables des anticorps.

Une analyse préliminaire a été effectuée sur 25877 séquences de gènes qui ont été structurées sous forme de tableaux de contingence. Chaque cellule de ces tableaux indique l'effectif des séquences ayant une certaine classe de propriété physicochimiques IMGT à une position bien déterminée des CDR-IMGT.

Les effectifs des cellules de ces tableaux ont été visualisées à l'aide de diagrammes en mosaïque qui ont permis d'identifier dans un premier temps l'existence de dépendances entre les différentes classes des propriétés physicochimiques les positions des CDR-IMGT ; soit l'hypothèse nulle H_0 ="Les deux variables (positions des acides aminés des CDR-IMGT et classes des propriétés physicochimiques) sont indépendantes", si la p-valeur est inférieure au seuil de signification statistique fixé à α=5%, on aura tendance à rejeter H_0 et opter pour l'hypothèse alternative H_1 ="Les deux variables (positions des acides aminés des CDR-IMGT et classes des propriétés physicochimiques) sont dépendantes". Cependant, la valeur du test $\chi 2$ n'est pas elle-même un bon indice de l'importance de l'association entre les traits mesurés et c'est, pour cette raison, qu'il faut analyser plus finement cette association par les résidus de Pearson [Mayer (2006)].

En passant à une deuxième étude, une analyse factorielle des correspondances (AFC) a permis de représenter la majorité de l'information contenue dans les tableaux de données et de tirer les liens qui existent entre les variables. L'AFC va permettre de révéler les liens et les oppositions qui existent ou non entre les différentes positions des acides aminés des CDR-IMGT et les classes des propriétés physicochimiques en fonction des écarts à

l'indépendance.

2.1 Classification des positions des acides aminés des CDR-IMGT

Pour aller plus loin qu'une AFC, une Classification Ascendante Hiérarchique (CAH) est réalisée en considérant la distance euclidienne calculée à partir des coordonnées sur les deux premiers axes. Ce critère est lié à des calculs d'inertie à chaque étape de la procédure d'agrégation [Confais (2004)] et induit une minimisation de la décroissance de la variance interclasse à chaque étape de regroupement [Ward (1963)]. Cet outil est utile pour établir des typologies et pour bien interpréter les analyses factorielles des correspondances. Il s'agit alors de regrouper les positions les plus proches les unes des autres, puis les groupes de positions et ainsi de suite jusqu'à ce que l'on obtienne un arbre généalogique (un dendrogramme) de ces positions qui met en relief les positions qui sont proches les unes des autres tout en représentant les relations d'inclusion entre les groupes.

2.2 Validation du nombre de classes des positions des acides aminés des CDR-IMGT

Pour trouver un meilleur partitionnement et pour valider le nombre de classes à retenir (niveau d'élagage du dendrogramme), un graphique de la décroissance de la variance interclasses en fonction du nombre de classes était nécessaire dans un premier temps.

Dans un deuxième temps, le recours à l'indice silhouette [Rousseeuw (1987)] était très utile de point de vue interprétation et validation.

En effet, chaque classe est représentée par une silhouette qui montre quelles positions CDR-IMGT sont correctement placées à l'intérieur de la classe et lesquelles n'ont simplement qu'une position intermédiaire.

Pour chaque position CDR-IMGT (x_i), les informations suivantes sont fournies :
- le numéro de la classe à laquelle elle appartient,
- le numéro de la classe voisine,
- la valeur de la silhouette s_i,
- l'identificateur à trois caractères de la position CDR-IMGT (x_i),
- une ligne dont la longueur est proportionnelle à s_i.

La valeur s_i est calculée comme suit :

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

avec

$$\begin{cases} a_i : \quad \textit{Distance moyenne de la position } \text{CDR-IMGT } (x_i) \textit{ par rapport à toutes les positions appartenant à la classe } A(i) \textit{ à laquelle appartient la position } \text{CDR-IMGT}(i). \\ \qquad\qquad a_i = \frac{1}{n(A(i))} \sum_{x \in A(i)} d(x_i, x) \\ b_i : \quad \textit{Distance moyenne de la position } \text{CDR-IMGT } (i) \textit{ par rapport à toutes les positions appartenant à la classe } B(i) \textit{ la plus proche (voisine de la position } \text{CDR-IMGT } (i)). \\ \qquad\qquad b_i = \min_{B \setminus A(i)} \frac{1}{n(A(i))} \sum_{x \in A(i)} d(x_i, x) \end{cases}$$

2.3 Identification des classes des propriétés physicochimiques IMGT

Une fois que le nombre de classes a été déterminé, il est judicieux de savoir la nature des groupes obtenus. Les profils lignes des tableaux de contingence ont été considérés comme un moyen d'identification de la classe des propriétés physicochimiques adéquate à chaque groupe en fixant les deux pourcentages 50% et 80% comme seuils jugés importants pour l'analyse. Dans le cas où les profils lignes sont plus faibles que ces deux seuils pour

certaines positions d'un même groupe, l'identification de la propriété physicochimique est appuyée sur les indices silhouette des positions (bien classées) dont la valeur dépasse 0.5, ayant des pourcentages supérieurs aux seuils fixés et donc, autrement dit, ces dernières sont celles qui vont emporter la propriété physicochimique pour toutes les autres positions dans une même classe.

2.4 IMGT Colliers de Perles

Dans le but d'avoir une visualisation rapide, les résultats obtenus seront représentées sur des graphiques en 2D appelés IMGT Colliers de Perles qui permettent de positionner les acides aminés de séquences protéiques des domaines V préalablement gappées selon la numérotation unique IMGT ainsi de délimiter les FR-IMGT et CDR-IMGT standardisés [Ruiz (2002)]. (voir l'exemple de la figure-1).

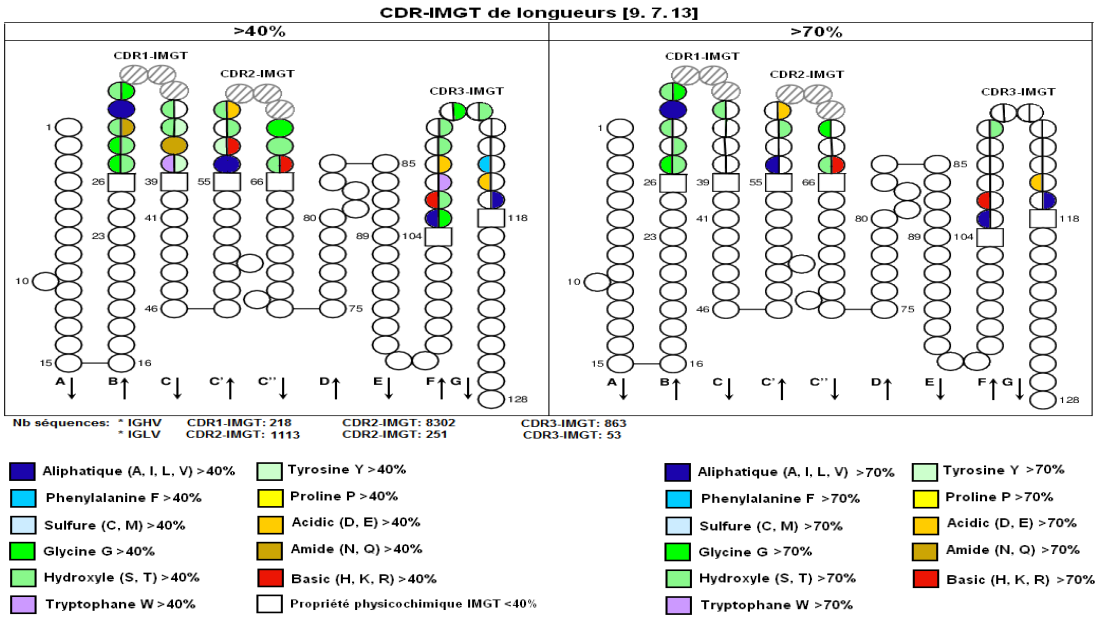

FIGURE 1 – IMGT Collier de Perles du profil statistique des positions des CDR-IMGT des IGHV versus IGLV selon les classes des propriétés physicochimique IMGT. Les demi-cercles correspondent aux cas des IGHV (à gauche) et aux IGLV (à droite). Les propriétés sont visualisées à des seuils $\geqslant 40\%$ et $\geqslant 70\%$.

2.5 Spécification des acides aminés des positions des CDR-IMGT

Une étude complémentaire a été menée pour détailler les résultats obtenus et spécifier quels sont les acides aminés les plus présents à ces classes de propriétés physicochimiques IMGT identifiées. Ce but a été atteint au moyen de cartes de double classification (ou *heatmaps*) [Eisen (1998)] sur les positions et sur les acides aminées qui ont permis de détecter les acides aminées qui ont des profils semblables ou encore ceux qui coïncident avec les groupements des acides aminés selon les propriétés physicochimiques IMGT.

Il s'agit d'une double classification hiérarchique avec la distance euclidienne qui opère en même temps sur les lignes et sur les colonnes d'une matrice X croisant en lignes les positions CDR-IMGT et en colonnes les 20 acides aminés. La matrices X est représentées sous forme d'une mosaïque (ou carte centrale) où chaque entrée de la matrice est colorée en fonction de sa mesure ou de sa dissemblance (distance) selon une échelle indiquant le z-score correspondant (voir l'exemple de la figure-2).

Ces cartes ont segmenté les positions les plus conservées des positions où se focalise la diversité, d'une part, et les AA les plus présents de ceux rarement détectés, d'autre part.

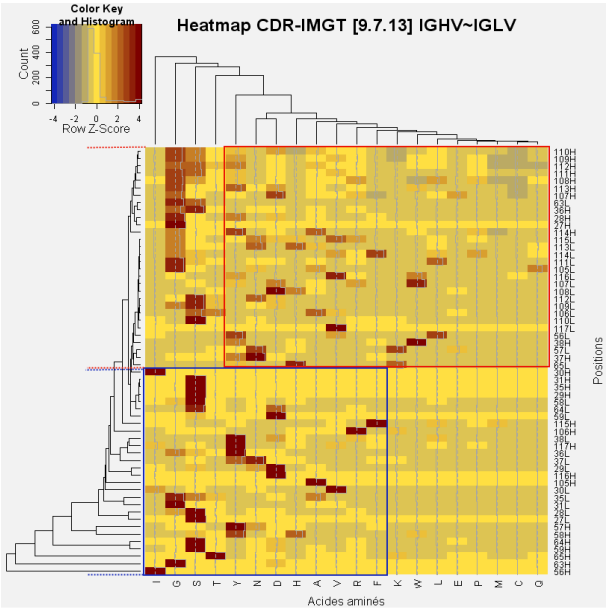

FIGURE 2 – Cartes de double classification pour les positions CDR-IMGT (en lignes) des IGHV versus IGLV et les 20 acides aminés (en colonnes)

Pour une lecture plus rapide des cartes, des réseaux graphiques ont été générés pour visualiser les acides aminés les plus présents à toutes les positions des CDR-IMGT des différentes chaînes. Dans ce graphe, ces acides aminés sont représentés par des rectangles et les positions par des cercles colorés en fonction des CDR-IMGT et du type de chaînes. Les arêtes reliant une position à un AA peuvent avoir 3 types de représentations différentes en fonction de leurs fréquences d'apparitions et les valeurs des z-scores : le trait continu signifie qu'à cette position donnée, l'AA correspondant est trouvé à un seuil $> 50\%$ (z-score$\geqslant 2.58$). Pour un seuil variant entre 25% et 50% ($1.96 \leqslant$z-score<2.58) le trait sera en pointillés et pour un seuil $<25\%$ ($1.65<$z-score<1.96), le trait sera en tiret-point. Si une position a une seule arête et qu'elle est représentée par un trait continu, son AA correspondant est présent à un seuil d'au moins 70%. Ce graphe permet aussi d'afficher les connections possibles entres les acides aminés surtout ceux qui appartiennent à des classes de propriétés physicochimiques IMGT (voir l'exemple de la figure-3).

3 Conclusions

Les classes des propriétés physicochimiques IMGT trouvées dans les différentes positions des CDR-IMGT comparées entre les chaînes lourdes versus chaînes légères ou encore entre les IGKV et IGLV des chaînes légères ont été visualisées à des seuils importants sur les IMGT Colliers de Perles pour préciser les positions sur lesquelles on peut intervenir sans risque de modifier la spécificité de reconnaissance de l'antigène. Suite à cette analyse, la diversité a été détectée d'une manière plus importante surtout dans les positions intermédiaires des chaînes lourdes des boucles CDR3-IMGT que dans les chaînes légères et c'est la glycine qui est détectée comme l'acide aminé le plus probable à y apparaître, et ceci, vu qu'il est le moins contraint par sa structure latérale. La conservation de certaines classes de propriétés physicochimiques IMGT dans certaines positions des CDR-IMGT a été détectée quelle que soit leurs longueurs.

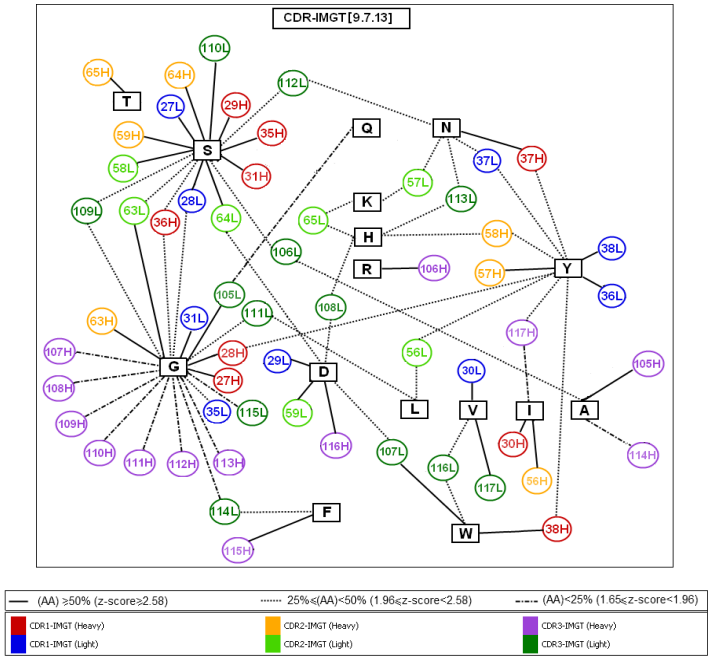

FIGURE 3 – Réseaux graphiques des positions CDR-IMGT des IGHV versus IGLV par la présence d'acides aminés.

Bibliographie

[1] Eisen M. B., Spellman P. T., Brown P. O. and Botstein D. (1998), *Cluster analysis and display of genome-wide expression patterns*. Proceeding of the National Academy of Sciences of the USA, 95 : 14863–14868.

[2] Lefranc M-P., Giudicelli V., Ginestoux C., et al. (2009),*IMGT®, the international ImMunoGeneTics information system®*. Nucleic Acids Research, 37 : D1006–12.

[3] MacQueen J.B. (1967), *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1 : 281-297.

[4] Meyer D., Zeileis A., Hornik K. (2006), *The Strucplot Framework : Visualizing Multiway Contingency Tables with vcd*. Journal of Statistical Software, vol. 17.

[5] Pommié C., Levadoux S. (2004), Sabatier R., Lefranc G., Lefranc MP., *IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties*. Journal of Molecular Recognition, 17 : 17–34.

[6] Rousseeuw P.J.(1987), *Silhouettes : a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20 : 53–65.

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

[8] Ward J.H. (1963), *Hierachical grouping to optimize an objective function*. Journal of American Statistical Association, 58 : 236–244.