Analyse statistique des mutations somatiques des régions variables des immunoglobulines à partir des résultats d'IMGT/HighV-QUEST

Safa Aouinti[1,2], Patrice Duroux[1],Véronique Giudicelli[1], Dhafer Malouche[2], Sofia Kossida[1], Marie-Paule Lefranc[1]

[1]*IMGT®, the international ImMunoGeneTics information system®, Institut de Génétique Humaine, UMR 9002 CNRS et Université de Montpellier, Montpellier, France, (safa.aouinti@igh.cnrs.fr, patrice.duroux@igh.cnrs.fr, veronique.giudicelli@igh.cnrs.fr, sofia.kossida@igh.cnrs.fr, marie-paule.lefranc@igh.cnrs.fr)*

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

Résumé. Les réponses immunitaires adaptatives de l'espèce humaine et des autres espèces de vertébrés à mâchoires (*gnathostomata*) sont caractérisées par les cellules B et T et leurs récepteurs d'antigènes spécifiques, les immunoglobulines (IG) ou anticorps et les récepteurs de cellules T (TR) (jusqu'à 2.10^{12} différent IG et TR par individu). IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) basé sur IMGT-ONTOLOGY a été créé pour gérer cette diversité. IMGT/HighV-QUEST est l'unique portail web pour l'analyse des séquences IG et TR obtenues par le séquençage haut débit (*'big data from next generation sequencing, NGS'*). L'une de ses caractéristiques majeures est l'identification des clonotypes IMGT (AA) et en particulier l'analyse de leur diversité et expression. Nous présentons une approche statistique basée sur un modèle suivant la loi multinomiale pour l'analyse des mutations somatiques des résultats issus d'IMGT/HighV-QUEST, spécifiques aux gènes variables réarrangés des IG, ainsi que les méthodes de visualisation appropriées. Nous utilisons la numérotation unique IMGT pour une description standardisée des mutations.

Mots-clés. IMGT, IMGT/HighV-QUEST, IMGT-ONTOLOGY, immunoglobuline, anticorps, récepteur T, 'big data', 'next generation sequencing' (NGS), mutation somatique, distribution multinomiale.

Abstract. The adaptive immune responses of humans and other jawed vertebrate species (*gnathostomata*) are characterized by the B and T cells and their specific antigen receptors, the immunoglobulins (IG) or antibodies and the T cell receptors (TR) (up to 2.10^{12} different IG and TR per individual). IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) built on IMGT-ONTOLOGY was created to manage this huge diversity. IMGT/HighV-QUEST is the first web portal, and so far the only one, for the next generation sequencing (NGS) analysis of IG and TR big data sequences. One of its main features is the identification of IMGT clonotypes (AA) and in particular the analysis of their diversity and expression.

We present the statistical approach based on multinomial distribution model for the analysis of somatic mutations of the IG rearranged variable genes, from the results of IMGT/HighV-QUEST, and appropriate visualization methods. The IMGT unique numbering allows a standardized IMGT description of mutations.

Keywords. IMGT, IMGT/HighV-QUEST, IMGT-ONTOLOGY, immunoglobulin, antibody, receptor T, big data, next generation sequencing (NGS), somatic mutation, multinomial distribution model.

1 Introduction

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) créé par Marie-Paule Lefranc en 1989, est l'unique système d'information intégré en immunogénétique et immunoinformatique [Lefranc et al. (2015)]. IMGT®, basé sur IMGT-ONTOLOGY [Giudicelli and Lefranc(2012)], est à l'origine d'une nouvelle science, l'immunoinformatique [Lefranc (2014)].

Les réponses immunitaires adaptatives de l'espèce humaine et des autres espèces de vertébrés à mâchoires (*gnathostomata*) sont caractérisées par les cellules B et T et leurs récepteurs d'antigènes spécifiques, les immunoglobulines (IG) ou anticorps [Lefranc and Lefranc (2001a)] et les récepteurs de cellules T (TR) [Lefranc and Lefranc. (2001b)] (jusqu'à 2.10^{12} différents IG et TR par individu). Les IG sont constituées de deux chaînes lourdes (H) et deux chaînes légères (L) où la chaîne L est une chaîne kappa ou lambda. Chaque chaîne comprend un domaine variable V et un ou plusieurs domaines constants C. Le domaine V est constitué de trois régions hypervariables ou régions de complémentarité à l'antigène (CDR pour *complementarity determining region*) qui déterminent le site de reconnaissance et de liaison à l'antigène et de quatre régions dites charpentes ou FR (*framework region*).

Le domaine V résulte du réarrangement de trois gènes V, D et J (V-D-J-region codant le domaine VH des chaînes lourdes) ou de deux gènes V et J (V-J-region codant le domaine VL des chaînes légères, kappa ou lambda). Ce sont ces réarrangements qui créent la diversité combinatoire lors de la synthèse des chaînes d'IG (à laquelle s'ajoutent la N-diversité de la jonction et pour les IG, les mutations somatiques).

IMGT/HighV-QUEST est le portail de référence pour l'analyse des séquences d'IG et de TR obtenues par les technologies de séquençage next generation sequencing (NGS) [Alamyar et al. (2012)]. IMGT/HighV-QUEST permet l'analyse des répertoires d'anticorps ou de récepteurs T dans les réponses immunitaires en situation normale (vaccination) ou pathologique (infections, maladies autoimmunes ou cancer). IMGT/HighV-QUEST permet la caractérisation des clonotypes IMGT (AA) (AA pour acide aminé) et l'analyse de leur diversité et expression. Un 'IMGT clonotype (AA)' est défini par un réarrangement V-(D)-J unique, des ancres conservées (C104, W ou F118) et une jonction CDR3-IMGT (AA) unique [Alamyar et al. (2012)]. Chaque IMGT clonotype (AA) est caractérisé par une séquence représentative unique définie par IMGT/HighV-QUEST et l'identification de toutes les séquences qui peuvent lui être rattachées. Ainsi, pour la première fois dans l'analyse des données NGS des récepteurs d'antigènes, l'approche standardisée d'IMGT permet une distinction claire entre la diversité des clonotypes (nombre des clonotypes IMGT (AA) par V, D ou J gène), et l'expression des clonotypes (nombre de séquences assignées à un clonotype IMGT (AA) donné, sans ambiguïté.) [Li et al. (2013)].

Dans ce travail nous analysons les résultats d'IMGT/HighV-QUEST pour étudier les mutations somatiques qui sont spécifiques aux gènes variables réarrangés des IG avec les méthodes de visualisation appropriées. Cette étude est importante pour analyser les modifications de la spécificité et de l'affinité des anticorps au cours des réponses immunitaires normales (lors de vaccinations, infections ou cancers) ou pathologiques (maladies autoimmunes). La numérotation unique IMGT [Lefranc et al. (2015)] a permis une description standardisée des mutations.

2 Contexte et approche méthodologique

La maturation d'affinité est un processus qui consiste à produire des anticorps, caractérisés par les clonotypes des lymphocytes B, suite à un processus de mutations somatiques et de sélection par les antigènes.

En l'absence de toute pression de sélection, des mutations ponctuelles peuvent apparaître aléatoirement dans le gène de la séquence de référence. L'on distingue deux types de mutations selon le résultat : soit des mutations non silencieuses (R pour '*replacement*') où la substitution d'un ou plusieurs nucléotide(s) d'un codon entraîne le changement de l'AA, soit des mutations silencieuses (S) où les mutations ponctuelles des nucléotides d'un codon n'entraînent pas le changement de l'AA associé. De manière intéressante, il a été montré que les anticorps sélectionnés par un antigène comprennent une fréquence plus élevée de mutations R dans les CDR que dans les FR du gène V de l'IG.

2.1 Modèle basé sur la distribution binomiale

Dans un premier temps de l'étude, la probabilité d'excès pour les CDR-IMGT ou de rareté pour les FR-IMGT (*excess or scarcity*) en fréquences des mutations non silencieuses (R), résultant uniquement par chance, est calculée suivant un modèle basé sur la distribution binomiale [Chang et Casali (1994)].

Ce modèle considère que les mutations peuvent avoir lieu de façon aléatoire dans n'importe quelle partie de la région variable du domaine V de l'anticorps avec une chance égale. Un ratio de mutations R/S supérieur à la valeur théorique 2.925 dans le cas des CDR a été considéré comme la signature de la sélection antigénique.

2.1.1 Calcul du nombre de mutations non silencieuses attendu dans les FR-IMGT et CDR-IMGT des gènes V des IG

Pour chaque V gène germline référencé dans IMGT/GENE-DB, le nombre de mutations R attendu a été calculé tel que :

$$\text{Nombre attendu de mutations R} = n \times (Rf_{\texttt{FR-IMGT ou CDR-IMGT}}) \times (Rl_{\texttt{FR-IMGT ou CDR-IMGT}})$$

avec

n : le nombre total des mutations observées

Rf : la fréquence des mutations R par remplacement dans les CDR-IMGT et FR-IMGT calculée comme suit :

$$Rf_{\texttt{FR-IMGT}} = \frac{\sum_{i=1}^{3} R_{FR_i-IMGT}}{\sum_{i=1}^{3} R_{FR_i-IMGT} + \sum_{i=1}^{3} S_{FR_i-IMGT}}$$

$$\text{et } Rf_{\texttt{CDR-IMGT}} = \frac{\sum_{i=1}^{2} R_{CDR_i-IMGT}}{\sum_{i=1}^{2} R_{CDR_i-IMGT} + \sum_{i=1}^{2} S_{CDR_i-IMGT}}$$

Rl : la longueur relative des FR-IMGT ou CDR-IMGT telle que :

$$Rl_{\texttt{FR-IMGT}} = \frac{\sum_{i=1}^{3} longueur_{FR_i-IMGT}}{\sum_{i=1}^{3} longueur_{FR_i-IMGT} + \sum_{i=1}^{2} longueur_{CDR_i-IMGT}}$$

$$\text{et } Rl_{\texttt{CDR-IMGT}} = \frac{\sum_{i=1}^{2} longueur_{CDR_i-IMGT}}{\sum_{i=1}^{3} longueur_{FR_i-IMGT} + \sum_{i=1}^{2} longueur_{CDR_i-IMGT}}$$

2.1.2 Calcul de la probabilité d'excès ou de rareté des mutations (R) dans les CDR-IMGT et FR-IMGT des gènes V résultant uniquement par chance

Le calcul de la probabilité p d'excès ou de rareté des mutations (R) est basé sur la distribution de la loi binomiale tel que :

$$p = \frac{n!}{k!(n-k)!} \times q^k \times (1-q)^{n-k} \tag{1}$$

avec
n : le nombre total des mutations observées
k : le nombre de mutations (R) observées dans les CDR-IMGT et FR-IMGT
q : la probabilité qu'une mutation R est localisée dans les CDR-IMGT ou FR-IMGT telle que

$$q = Rl_{\texttt{CDR-IMGT}} \times Rf_{\texttt{CDR-IMGT}} \ \texttt{ou} \ Rl_{\texttt{FR-IMGT}} \times Rf_{\texttt{FR-IMGT}}$$

2.2 Modèle basé sur la distribution multinomiale

Le problème du modèle précédent est qu'il n'est adapté, par définition, qu'aux variables ayant deux possibilités de distribution alors que les mutations des IG ont quatre possibilités de distribution différentes (mutations silencieuses (S) ou non silencieuses (R) dans les CDR-IMGT et/ou FR-IMGT des gènes V). La solution proposée par [Lossos et al. (2000)] est de calculer la probabilité d'excès ou de rareté des mutations (R) résultant par chance en se basant sur un modèle suivant la distribution multinomiale [Hogg (1978)] permettant de détecter la pression de sélection antigénique. Le modèle est appliqué comme suit :

Soit n le nombre de mutations total dans chaque IMGT clonotype (AA) tel que

$$n = r1 + s1 + r2 + s2$$

avec
r_1 et r_2 : les mutations (R) dans les FR-IMGT et les CDR-IMGT respectivement.
s_1 et s_2 : les mutations (S) dans les FR-IMGT et les CDR-IMGT respectivement.

Les probabilités théoriques des mutations r_1, s_1, r_2 et s_2 sont notées respectivement par p_1, q_1, p_2, q_2 et calculées comme suit :

$$\begin{aligned} p_1 &= Rf_{\texttt{FR-IMGT}} \times Rl_{\texttt{FR-IMGT}} \\ q_1 &= (1 - Rf_{\texttt{FR-IMGT}}) \times Rl_{\texttt{FR-IMGT}} \\ p_2 &= Rf_{\texttt{CDR-IMGT}} \times (1 - Rl_{\texttt{FR-IMGT}}) \\ q_2 &= (1 - Rf_{\texttt{CDR-IMGT}}) \times (1 - Rl_{\texttt{FR-IMGT}}) \end{aligned}$$

La probabilité d'observer r_2 ou plus de mutations (R) dans les CDR-IMGT est calculée comme suit [Lossos et al. (2000)] :

$$\mathbb{P}(R_2 \geq r_2) = \sum_{k=r_2 \ldots n,\ R_1+S_1+k+S_2=n} \left(\begin{array}{c} n \\ R_1, S_1, k, S_2 \end{array} \right) p_1^{R_1} q_1^{S_1} p_2^{k} q_2^{S_2}$$

La somme prend des valeurs k allant de r_2 à n et toutes les combinaisons de R_1, S_1, S_2, tel que $R_1 + S_1 + k + S_2 = n$.

La probabilité (P) d'observer le nombre r_2 est calculée en se basant sur la formule suivante :

$$P = \mathbb{P}(R_2 > r_2) + 0.5 \times \mathbb{P}(R_2 = r_2) \tag{2}$$

Par convention standard, la fonction densité de probabilité en r_2 de l'équation (2) est multipliée par 0.5 ce qui permet de calculer la P-valeur dans le cas de la sélection positive comme étant 1 moins celle de la sélection négative [Hershberg et al. (2008)].

Si $P \leq 0.05$, ceci indique que les différences en nombres observés et attendus de mutations R dans les CDR-IMGT ne sont pas dues au hasard mais en raison de la sélection antigénique positive qui 'favorise' les mutations R dans les CDR-IMGT.

La probabilité d'observer r_1 ou moins de mutations (R) dans les FR-IMGT est calculée comme suit :

$$\mathbb{P}(R_1 \leq r_1) = \sum_{k=0 \ldots r_1,\ k+S_1+R_2+S_2=n} \left(\begin{array}{c} n \\ k, S_1, R_2, S_2 \end{array} \right) p_1^{k} q_1^{S_1} p_2^{R_2} q_2^{S_2}$$

La somme prend des valeurs k allant de 0 à r_1 et toutes les combinaisons de S_1, R_2, S_2, tel que $k + S_1 + R_2 + S_2 = n$.

Le calcul de la probabilité (P) d'observer le nombre r_1 dans les FR-IMGT est basé sur la formule suivante :

$$P = \mathbb{P}(R_1 < r_1) + 0.5 \times \mathbb{P}(R_1 = r_1) \tag{3}$$

avec $\mathbb{P}(R_1 < r_1) = \mathbb{P}(R_1 \leq r_1) - \mathbb{P}(R_1 = r_1)$
et $\mathbb{P}(R_1 = r_1) = \mathbb{P}(R_1 \leqslant r_1) - \mathbb{P}(R_1 \leqslant r_1 - 1)$

Si $P \leq 0.05$, on est amené à considérer que la sélection agit contre les mutations (R) dans les FR-IMGT (sélection négative qui 'défavorise' les mutations R dans les FR-IMGT).

N.B. Pour les CDR-IMGT et FR-IMGT, les valeurs (P) unilatérales ont été utilisées.

Pour contrôler l'inflation du taux d'erreurs dans le cas des tests multiples, nous avons appliqué la procédure de Benjamini & Hochberg (BH) qui contrôle le FDR (*False discovery rate*).

3 Résultats

Nous présenterons une étude descriptive d'un exemple de jeu de données issu d'un répertoire IG humain avec les méthodes de visualisations appropriées. Les résultats obtenus seront ensuite exposés en appliquant les deux modèles cités ci-dessus à l'analyse des mutations des IMGT clonotypes (AA) du même jeu de données.

4 Conclusion

L'analyse des mutations somatiques de la V-REGION des IG est importante pour étudier la réponse de l'anticorps. Le processus de mutations somatiques a un biais intrinsèque pour accumuler les mutations (R) dans les CDR par rapport aux FR. Effectuée en parallèle avec les analyses d'affinité d'anticorps spécifiques durant la réponse immunitaire, cette approche permettra d'estimer la sélection de l'antigène dans les mutations observées.

Bibliographie

[1] Lefranc, M-P. et al. (2015) IMGT®, the international ImMunoGeneTics information system® 25 years. Nucleic Acids Res., 43 : D413-22.
[2] Giudicelli, V. and Lefranc, M-P. (2012) IMGT-ONTOLOGY 2012. Front. Genet., 3 :79.
[3] Lefranc, M-P. (2014). Immunoglobulin (IG) and T cell receptor genes (TR) : IMGT® and the birth and rise of immunoinformatics. Front Immunol. 5 :22.
[4] Lefranc, M.-P. and Lefranc, G. (2001a). The Immunoglobulin FactsBook. Academic Press, London, UK, (458 pages).
[5] Lefranc, M-P. and Lefranc, G. (2001b). The T cell receptor FactsBook. Academic Press, London, UK (398 pages).
[6] Alamyar, E. et al. (2012) IMGT/HighV-QUEST : the IMGT web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res., 8 :1 :2.
[7] Li, S. et al. (2013) IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nature Comm., 4 :2333.
[8] Chang, B. and Casali, P. (1994), The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement, Immunol. Today, 15(8) :367-373.
[9] Lossos, I.S. et al. (2000), The inference of antigen selection on Ig genes. J. Immunol., 1 ;165(9) :5122-6.
[10] Hogg, R. (1978) Introduction to Mathematical Statistics. Macmillan, New York.
[11] Hershberg, U. et al (2008), Improved methods for detecting selection by mutation analysis of Ig V region sequences. Int. Immunol., 20(5) :683-694.