Diversité et spécificité des récepteurs immunitaires

Comment notre corps fabrique des protéines aléatoires spécialisées pour combattre les pathogènes

Thierry Mora

Résumé Le système immunitaire adaptatif repose sur la diversité des protéines réceptrices de lymphocytes, dont les anticorps sont un exemple. Chaque protéine réceptrice est créée par un processus stochastique d'édition de l'ADN et se lie à un nombre réduit de molécules venant d'agents pathogènes. Cet article décrit comment les progrès récents de séquençage d'ADN et d'analyse statistique avancée permettent de mieux comprendre la genèse et la biochimie de ces protéines, avec des conséquences importantes sur leur fonction immunologique, en se concentrant sur deux questions : quelles sont les lois d'édition de l'ADN qui sous-tendent l'énorme diversité des anticorps produits, et quelle est la relation entre la séquence protéique des anticorps et leur fonction de liaison ?

Mots-clés **Anticorps, répertoire immunitaire, interactions protéine-protéine, inférence statistique.**

Abstract **Diversity and specificity of immune receptors**
The adaptive immune system relies on the diversity of lymphocyte receptor proteins, of which antibodies are an example. Each receptor protein is created by a stochastic DNA editing process, and binds to a limited number of molecules from pathogenic agents. This article describes how recent progress in DNA sequencing and advanced statistical analysis allows us to better understand the generation and biochemistry of these proteins, with important consequences for their immune function, focusing on two questions: what are the laws of DNA editing underlying the enormous diversity of the produced antibodies, and what is the relationship between antibody protein sequence and binding function?

Keywords **Antibodies, immune repertoire, protein-protein interactions, statistical inference.**

On dit souvent que toutes les cellules de notre corps contiennent le même ADN. C'est largement vrai, mais à une exception notable près : les récepteurs de globules blancs, ou plus précisément des lymphocytes B et T, impliqués dans notre système immunitaire. Chaque cellule exprime sa version unique d'un gène codant une protéine réceptrice appartenant à la famille des immunoglobulines, dont font partie par exemple les anticorps, et créée par l'édition aléatoire de l'ADN. Chacune de ces protéines peut se lier spécifiquement à un nombre restreint de bouts d'autres protéines, appelées antigènes. Lorsque ces antigènes appartiennent à des agents pathogènes, la liaison entre la protéine réceptrice et l'antigène conduit au déclenchement de la réponse immunitaire. Afin de nous protéger contre la très grande variété de pathogènes connus ou potentiels, nos lymphocytes expriment donc une très grande diversité de récepteurs, dont seule une petite fraction est spécialisée dans chaque pathogène. L'ensemble des récepteurs exprimés par un individu définit son « répertoire » immunitaire. La diversité de ce répertoire assure la fonction du système immunitaire dit « adaptatif » [1] et est essentielle à notre maintien en bonne santé.

Alors que le principe de la vaccination est connu depuis le XIX^{e} siècle et les expériences de Louis Pasteur, les bases moléculaires de l'immunité n'ont été élucidées que récemment. Notamment, le mécanisme d'amplification des anticorps spécifiques lors d'une infection est resté mystérieux jusqu'à ce que l'immunologiste australien F. Burnet propose la théorie de la sélection clonale en 1957 [2]. D'après cette théorie, fortement spéculative à son origine mais depuis vérifiée dans tous ses aspects et expliquée au niveau moléculaire, la reconnaissance des antigènes par les protéines réceptrices de chaque lymphocyte spécialisé conduit à la prolifération sélective de ces lymphocytes (*figure 1A*). Cette prolifération s'accompagne d'une production de masse d'anticorps, qui ne sont autres que la version soluble des protéines réceptrices. La liaison des anticorps aux protéines de pathogènes va inhiber leur fonction ou ordonner leur destruction. Ce mécanisme constitue la base de la mémoire immunitaire et donc de l'immunisation. Afin d'assurer la tolérance du « soi », les lymphocytes ciblant les protéines de l'individu sont éliminés au terme d'un processus de sélection dans la moelle épinière et le thymus.

Créer des protéines aléatoires par édition de l'ADN

La théorie de Burnet laissait en suspens la question de savoir comment cette diversité de protéines pouvait être engendrée. En effet, la couverture par notre système immunitaire de pratiquement n'importe quelle molécule suggère que le système immunitaire doit comporter un très grand nombre de récepteurs. Ces récepteurs sont-ils tous encodés « en dur » par une multitude de gènes présents dans le

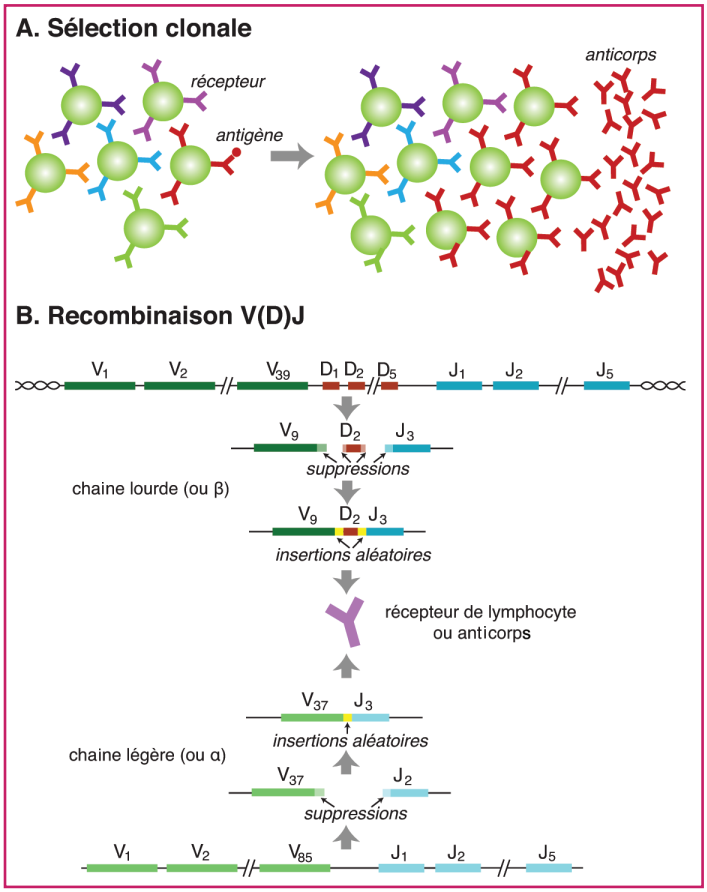

Figure 1 - **Création et sélection de protéines aléatoires**. A) Chaque lymphocyte exprime un récepteur unique, qui peut se lier spécifiquement à une fraction de morceaux de protéines étrangères, ou « antigènes ». Les lymphocytes reconnaissant des agents pathogènes prolifèrent, fournissant au corps un grand nombre de cellules et anticorps spécialisés dans la lutte contre l'infection et forment la base de l'immunité adaptative. B) Chaque récepteur unique résulte de l'édition aléatoire de l'ADN lors du développement des lymphocytes. Trois segments appelés V, D et J sont pris parmi une liste de modèles encodés dans la souche de l'ADN et collés ensemble. Des éliminations et ajouts de paires de bases aléatoires viennent compléter le processus.

génome de chaque cellule, ou bien résultent-ils d'une modification « somatique » de l'ADN (c'est-à-dire unique à chaque cellule et non transmise aux progénitures) ? S. Tonegawa démontra en 1978 la validité de la deuxième hypothèse [3] : chaque protéine réceptrice résulte de l'édition aléatoire de l'ADN « souche » lors du développement des lymphocytes.

Il existe deux types principaux de lymphocytes impliqués dans le système immunitaire adaptatif : les cellules B et les cellules T, chacune présente en très grand nombre dans notre corps (de l'ordre d'un millier de milliards de cellules). Chaque cellule exprime sur sa surface une protéine réceptrice de la famille immunoglobuline formée de manière semblable dans les cellules B et T : cette protéine est composée de deux chaines, l'une « légère » et l'autre « lourde », codées par deux gènes distincts. La chaine lourde (appelée chaine β dans le contexte des cellules T) est assemblée à partir de l'ADN souche par un processus appelé recombinaison VDJ, par l'action d'enzymes et formation de complexes macromoléculaires. Trois segments, appelés V (« variable »), D (« diversity ») et J (« joining ») pour des raisons historiques, sont choisis de manière aléatoire parmi un catalogue de modèles contenus dans l'ADN souche. Ces segments sont ensuite « collés » les uns aux autres (*figure 1B*). Ce collage n'est pas parfait : à chacune des deux jonctions, les segments d'ADN modèle sont rabotés d'un nombre aléatoire de paires de bases et de nouvelles paires de bases aléatoires sont ajoutées entre les segments. Cette diversité, dite « de jonction », s'ajoute à la diversité combinatoire résultant du choix des trois segments. La chaine légère (appelée chaine α pour les lymphocytes T) subit un processus très similaire, mais avec les segments V et J seulement. Par ce processus, un nombre astronomique de combinaisons et donc de séquences codantes de protéines peut être potentiellement engendré. Dans le cas des lymphocytes B, cette diversité se trouve encore augmentée par l'induction d'« hypermutations » activées par la reconnaissance de pathogènes et le déclenchement de la réponse immunitaire.

Vue la nature aléatoire des séquences ainsi produites, la question de leur fonction et de leurs propriétés se pose naturellement. Par exemple, près des trois quarts de ces séquences contiennent un codon d'arrêt prématuré, dû le plus souvent à un décalage du cadre de lecture, qui rend le gène non codant et empêche la production de la protéine. Plus généralement, on peut s'attendre à ce que de nombreuses protéines résultant de la recombinaison VDJ ne se replient pas correctement, ou n'aient pas de bonnes propriétés de liaison entre les deux chaines ou avec les antigènes. Le système immunitaire comporte un mécanisme de nettoyage des cellules exprimant de telles protéines défectueuses, mais les règles précises de cette sélection, et comment en deviner le résultat à partir de la seule séquence protéique, restent des questions largement ouvertes.

Inférence des règles de tirage au sort des récepteurs immunitaires

Le séquençage massif d'ADN rendu possible par les technologies à haut débit permet, depuis les premières démonstrations en 2009 [4-5], d'établir une liste des récepteurs exprimés par un petit échantillon de lymphocytes (de l'ordre du million). Cette nouvelle technologie est en train de révolutionner la manière d'étudier le système immunitaire adaptatif, et promet de nouvelles méthodes de diagnostic ou de médecine personnalisée [6]. En effet, le répertoire de récepteurs d'un individu reflète son histoire immunologique et contient en principe une mine d'informations sur les maladies qu'il a ou a pu avoir, sa propension à développer des maladies auto-immunes ou des allergies, etc.

Ces données de séquençage permettent également d'apprendre les lois probabilistes de génération aléatoire des séquences. Ces lois permettent à leur tour de contraindre les modèles de fonctionnement des macromolécules et enzymes impliquées dans ce processus complexe. L'apprentissage de ces lois se heurte à une propriété fondamentale de la recombinaison VDJ : sa dégénérescence. Une même séquence finale d'ADN recombiné peut avoir été produite d'un grand nombre de manières différentes. Il est simplement impossible de remonter à la chaine d'événements d'édition de l'ADN (choix des segments V, D et J, nombre de paires de bases supprimées à chaque bout, nombre et nature des paires de bases ajoutées) à partir d'une séquence donnée. De manière surprenante, cette limitation peut être surmontée en analysant un grand nombre de séquences. À l'aide de méthodes d'inférence statistique basées sur le principe de « vraisemblance maximale » (une méthode consistant à trouver le jeu de paramètres qui maximise la probabilité que les données aient été engendrées sous l'hypothèse du modèle), il est possible d'apprendre les lois de probabilité de chaque événement

d'édition avec une grande précision à partir de jeux de données de quelques milliers de séquences [7-9]. Cet apprentissage statistique permet, entre autres, d'obtenir la loi de probabilité du nombre de paires de bases insérées à chaque jonction (*figure 2A*). Cette loi se distingue pour les cellules T par un pic à quatre paires de bases, suivi par une décroissance exponentielle à grands nombres d'insertions, et ne dépend pas de la jonction considérée (VD, DJ ou VD) ou de l'individu échantillonné. Ce résultat éclaire le fonctionnement de la polymérase « *terminal deoxynucleotidyl transferase* » (TdT) impliquée dans le processus d'insertion, et a par la suite inspiré l'hypothèse d'un mécanisme en deux étapes fondé sur l'étude de la structure cristalline de la TdT liée à de l'ADN [10].

L'analyse statistique de séquences permet également d'inférer un indice de viabilité ou de taux de survie des récepteurs de lymphocytes, à partir de l'étude des différences entre les séquences recombinées non codantes (et donc produits bruts de la recombinaison, sans influence de la sélection des lymphocytes) et les séquences codantes, dont la majorité a été soumise à des tests de fonctionnalité, tels que le bon repliement de la protéine ou sa capacité à être présentée sur la surface de la cellule [11]. Cet indice de viabilité peut se décomposer en facteurs de sélection agissant sur chaque résidu d'une courte portion de la chaine β appelée « *complementarity determining region 3* » (CDR3) (*figure 2B*). Cette région correspond à une boucle de contact avec l'antigène qui concentre la partie la plus aléatoire de la protéine. L'examen des corrélations avec diverses propriétés biochimiques des résidus [12] révèle une préférence pour les résidus tendant à former des virages ou à se trouver dans le cœur d'un complexe d'interaction. Par ailleurs, l'analyse du modèle tend à suggérer que le processus de recombinaison VDJ est biaisé afin d'augmenter la probabilité de produire des protéines viables, ce qui pourrait s'expliquer par la pression sélective sur le système immunitaire des vertébrés à l'échelle évolutive.

Les modèles statistiques ainsi obtenus permettent également de quantifier la diversité des séquences aléatoires en termes d'entropie. Ainsi, on peut estimer un indice de diversité correspondant à 10^{23} séquences distinctes pour les récepteurs de cellules T, et encore plus pour les cellules B [13].

Le modèle peut également être utilisé afin de prédire avec succès le nombre de chaines de récepteurs que deux individus distincts sont susceptibles d'avoir en commun par « coïncidence », c'est-à-dire par la recombinaison VDJ aléatoire d'exactement la même séquence, et ce de manière indépendante chez les deux individus [11]. La méthode peut être étendue afin de détecter l'excès de ces séquences communes, ce qui a permis par exemple de formuler l'hypothèse que les vrais jumeaux conservent jusqu'à l'âge adulte des lymphocytes échangés par le cordon ombilical pendant la grossesse [14].

Mesure massive de courbes de titrage d'anticorps à l'aide de levure

Afin de pouvoir interpréter les séquences de protéines réceptrices et anticorps exprimés dans le corps d'un individu, il faut pouvoir déterminer ou prédire à quels antigènes ces protéines sont susceptibles de se lier. Cette question, encore très largement ouverte, est très importante d'un point de vue médical car sa réponse permettrait de déterminer l'ensemble des pathogènes auquel le système immunitaire d'un individu est réactif sur simple examen de son répertoire immunitaire. Bien que des travaux très récents aient fait des progrès dans cette direction [15-16], il n'existe toujours pas de formule systématique permettant de prédire la fonction de liaison des récepteurs immunitaires à partir de leur séquence.

D'un point de vue biochimique, la fonction d'un anticorps ou récepteur immunitaire est principalement déterminée par sa constante de dissociation avec sa cible moléculaire : plus la liaison est robuste, meilleure est la réponse immunitaire. La fonction immunitaire est typiquement assurée pour des constantes de dissociation allant de 1 à 100 nmol/L. Il existe des méthodes expérimentales permettant de « cribler » un grand nombre d'anticorps en fonction de leurs propriétés de liaison avec une cible donnée, afin d'en sélectionner les meilleurs. Ces techniques reposent sur l'utilisation d'organismes biologiques, comme les bactériophages [17] ou les levures [18], afin de relier les protéines à leur séquence génétique et d'en sélectionner les meilleures par évolution dirigée.

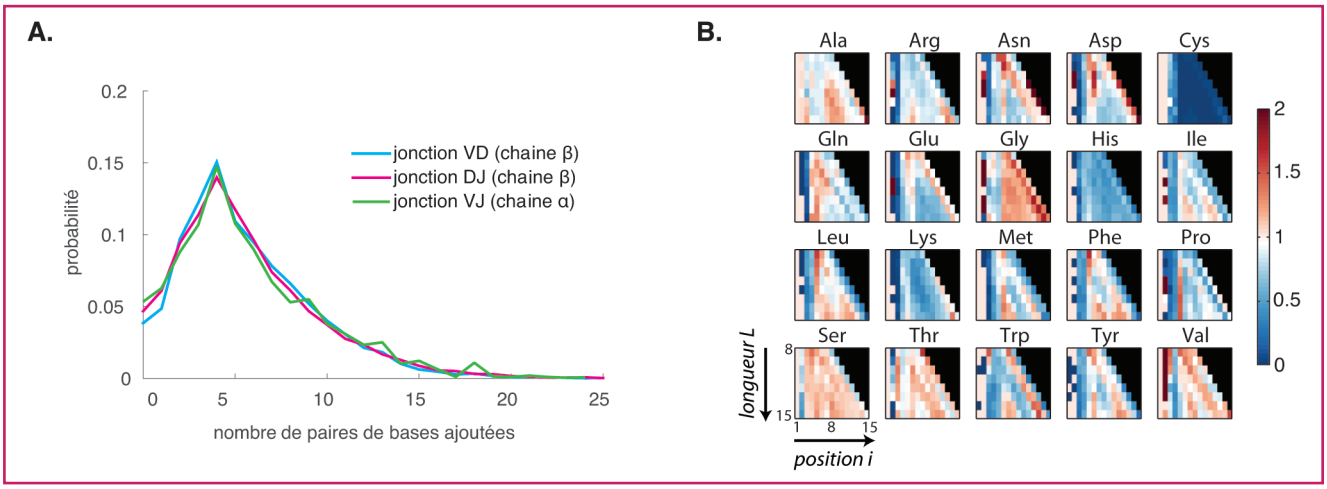

Figure 2 - **Lois statistiques de la genèse et viabilité des récepteurs immunitaires**. A) Distribution de probabilité du nombre de paires ajoutées aux trois jonctions des récepteurs de cellules T, telle qu'inférée par la méthode de vraisemblance maximale à partir de données de séquençage à haut débit. La distribution est universelle et décroit de manière exponentielle à grand nombre. La forme de la distribution impose des contraintes aux modèles d'opération de la polymérase impliquée dans l'ajout d'insertions. B) Facteurs de viabilité agissant sur les résidus de la boucle principale (CDR3) du récepteur en contact avec les antigènes, en fonction de la taille de la boucle, de la position et de la nature du résidu dans cette boucle, tels que révélés par l'analyse statistique des données de séquençage. Le rouge indique que le résidu a tendance à rendre la protéine viable, tandis que le bleu montre une influence délétère.

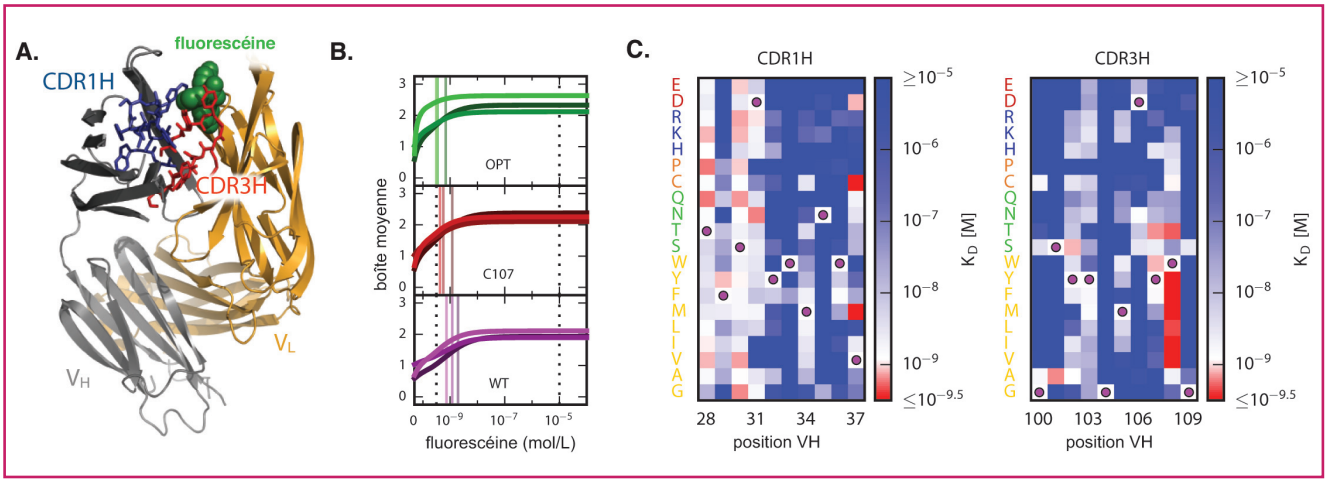

Figure 3 - **Mesure massive et parallèle de courbes de titrages de pairs anticorps-antigène**. A) Structure du complexe de liaison entre un antigène, la fluorescéine, et un anticorps spécifique (code Protein Data Bank 1FLR). Les régions CDR1H et CDR3H, qui correspondent à des boucles de liaisons avec l'antigène sur la chaine lourde (VH) de l'anticorps, sont mutées artificiellement afin de construire une bibliothèque d'anticorps. B) Exemples de courbes de titrage de trois anticorps distincts (parmi plusieurs milliers), inférées par séquençage à haut débit. Chaque niveau de foncé correspond à une réplique de l'expérience. Les lignes verticales indiquent la constante de dissociation K_D inférée en mol/L. C) Paysage de mutations uniques. Chaque carré correspond à l'effet d'une mutation à cette position en ce résidu par rapport à la séquence de référence. Les points indiquent les résidus de la séquence de référence.

Bien qu'elles permettent de concevoir des anticorps spécifiques à n'importe quelle cible et d'en apprendre beaucoup sur les propriétés d'ensemble de liaison de familles d'anticorps [19], ces techniques ne permettent pas directement de mesurer la constante de dissociation d'anticorps donnés en unités physiques. Cependant, l'utilisation de la présentation par levure à diverses concentrations de la cible permet de le faire [20]. Le principe consiste à mesurer, par cytométrie de flux, la fluorescence moyenne de cellules de levure exprimant l'anticorps à leur surface, et ayant été préalablement plongées dans un bain d'antigènes marqués avec une étiquette fluorescente. Cette fluorescence moyenne, tracée en fonction de la concentration d'antigènes dans le bain, donne une courbe de titrage dont la constante de dissociation peut être extraite, en moles par litre.

Cette technique peut être combinée avec le séquençage à haut débit afin d'obtenir les courbes de titrage d'un très grand nombre d'anticorps à la fois [21], ainsi qu'illustré par la *figure 3* pour l'exemple de la fluorescéine comme antigène modèle. L'idée de cette méthode, baptisée « Tite-Seq », est de faire exprimer par une population de levures une « bibliothèque » de plusieurs milliers d'anticorps (différant par un, deux ou trois résidus d'un anticorps de référence), et de les analyser par cytométrie de flux. Au lieu d'enregistrer la fluorescence moyenne, on trie les cellules selon l'intensité de leur fluorescence dans quatre boites, puis on obtient la séquence de tous les anticorps exprimés dans chacune des boites. En général, les anticorps ayant de bonnes propriétés de liaison vont se retrouver préférentiellement dans les boites de haute fluorescence, tandis que les mauvais anticorps seront plutôt dans les boites de basse fluorescence. À l'aide de techniques d'inférence statistique, on peut utiliser ces données de séquençage brutes, prises à plusieurs concentrations de fluorescéines, afin de reconstruire la courbe de titrage et la constante de dissociation de chaque membre de la bibliothèque initiale (*figure 3B*). Cela permet, par exemple, de mesurer le paysage d'affinité des mutations uniques (*figure 3C*).

Ce type de technique ouvre la voie à des mesures systématiques et à haut débit de la constante de dissociation d'un grand nombre de paires anticorps-antigènes, qui sont amenées à jouer un rôle essentiel dans la compréhension de la réponse immunitaire et dans le diagnostic à partir de données de répertoires.

Conclusion

Les anticorps et protéines réceptrices de lymphocytes sont difficiles à étudier du fait de leur grande diversité, à la fois en termes de séquences et de propriétés de liaison. Le développement des techniques de séquençage promet à la fois de mieux comprendre leur diversité et dynamique naturelle dans leur contexte immunologique, avec des applications potentielles révolutionnaires en médecine, et d'autre part de mieux comprendre le paysage d'affinité de ces protéines et mieux concevoir et identifier les récepteurs spécifiques à certains agents pathogènes ou allergènes. Les travaux présentés ici illustrent l'utilité d'approches d'inférence statistique nouvelles afin d'analyser les données de séquençage brutes et d'en extraire les quantités biologiques ou physico-chimiques pertinentes.

Références

[1] Murphy K., Travers P., Walport M., Janeway C., *Janeway's Immunobiology*, 8th ed., Garland Science, **2012**.

[2] Burnet F.M., A modification of Jerne's theory of antibody production using the concept of clonal selection, *Austalian Journal of Science*, **1957**, *20*, p. 67.

[3] Tonegawa S., Somatic generation of immune diversity, *Biosci. Rep.*, **1988**, *8*, p. 3.

[4] Weinstein J.A., Jiang N., White R.A., Fisher D.S., Quake S.R., High-throughput sequencing of the zebrafish antibody repertoire, *Science*, **2009**, *324*, p. 807.

[5] Robins H.S. *et al.*, Comprehensive assessment of T-cell receptor β-chain diversity in αβ T-cells, *Blood*, **2009**, *114*, p. 4099.

[6] Calis J.J., Rosenberg B.R., Characterizing immune repertoires by high throughput sequencing: strategies and applications, *Trends Immunol.*, **2014**, *35*, p. 581.

[7] Murugan A., Mora T., Walczak A.M., Callan C.G., Statistical inference of the generation probability of T-cell receptors from sequence repertoires, *Proc. Natl Acad. Sci. USA*, **2012**, *109*, p. 16161.

[8] Elhanati Y., Sethna Z., Marcou Q., Callan G.C. Jr, Mora T., Walczak A.M., Inferring processes underlying B-cell repertoire diversity, *Philos. Trans. R. Soc. Lond. B, Biol. Sci.*, **2015**, *370*, p. 20140243.

[9] Marcou Q., Mora T., Walczak A.M., IGoR: a tool for high-throughput immune repertoire analysis, **2017**, arXiv:1705.08246.
[10] Loc'h J., Rosario S., Delarue M., Structural basis for a new templated activity by terminal deoxynucleotidyl transferase: implications for V(D)J recombination, *Structure*, **2016**, *24*, p. 1.
[11] Elhanati Y., Murugan A., Callan C.G., Mora T., Walczak A.M., Quantifying selection in immune receptor repertoires, *Proc. Natl Acad. Sci. USA*, **2014**, *111*, p. 9875.
[12] Martin J., Lavery R., Arbitrary protein-protein docking targets biologically relevant interfaces, *BMC Biophysics*, **2012**, *1*, p. 7.
[13] Mora T., Walczak A., Quantifying lymphocyte receptor diversity, **2016**, arXiv :1604.00487.
[14] Pogorelyy M.V. *et al.*, Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires, *PLoS Comput. Biol.*, **2017**, *13*, p. e1005572.
[15] Dash P. *et al.*, Quantifiable predictive features define epitope-specific T cell receptor repertoires, *Nature*, **2017**, *547*, p. 89.
[16] Glanville J. *et al.*, Identifying specificity groups in the T cell receptor repertoire, *Nature*, **2017**, *547*, p. 94.
[17] Smith G.P., Petrenko V.A., Phage display, *Chem. Rev.*, **2017**, *97*, p. 391.
[18] Gai S.A., Wittrup K.D., Yeast surface display for protein engineering and characterization, *Curr. Opin. Struct. Biol.*, **2007**, *17*, p. 467.
[19] Boyer S. *et al.*, Hierarchy and extremes in selections from pools of randomized proteins, *Proc. Natl Acad. Sci. USA*, **2016**, *113*, p. 3482.
[20] Van Antwerp J.J., Wittrup K.D., Fine affinity discrimination by yeast surface display and flow cytometry, *Biotechnol. Prog.*, **2000**, *16*, p. 31.
[21] Adams R.M., Kinney J.B., Mora T., Walczak A.M., Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves, *eLife*, **2016**, 5:e23156.

Thierry Mora
est chargé de recherche au CNRS, au Laboratoire de physique statistique*.
Il a reçu la Médaille de bronze du CNRS en 2016.

* Laboratoire de physique statistique, École normale supérieure, CNRS et UPMC, 24 rue Lhomond, F-75005 Paris.
Courriel : tmora@lps.ens.fr