Biologie Aujourd'hui **211 (3)**, 229-231 (2017)

https://doi.org/10.1051/jbio/2017033

Disponible en ligne :
www.biologie-journal.org

ARTICLE ORIGINAL

IGoR : un outil pour apprendre et simuler la génération aléatoire de récepteurs d'antigènes

Thierry Mora*

Laboratoire de physique statistique, École Normale Supérieure, CNRS, UPMC etUPD, 24 rue Lhomond, 75005 Paris, France

Reçu le 2 décembre 2017

Résumé – Les récepteurs d'antigènes, qui sont à la base du système immunitaire adaptatif, sont créés de manière stochastique par un processus d'édition de l'ADN appelé recombinaison V(D)J. Alors que le séquençage à haut débit permet maintenant d'étudier le répertoire de ces récepteurs, il devient possible d'apprendre les lois probabilistes de ce processus aléatoire, et de les utiliser afin d'analyser des récepteurs particuliers, d'engendrer des répertoires synthétiques à des fins de contrôle, ou bien d'aider à l'identification de récepteurs spécifiques de certaines maladies, avec des applications possibles pour le diagnostic médical. Cet article décrit comment ceci peut être réalisé à l'aide du logiciel *IGoR*, qui apprend des modèles statistiques à partir des données de séquençage, et permet d'annoter des séquences existantes ou bien d'en engendrer de nouvelles, synthétiques, en suivant les lois du processus de recombinaison.

Mots clés : récepteurs d'antigènes, répertoire immunitaire, recombinaison V(D)J, inférence statistique, analyse de séquences

Abstract – *IGoR*: a tool for learning and simulating the random generation of antigen receptors. Antigen receptors, which form the base of the adaptive immune system, are created stochastically by a DNA editing process called V(D)J recombination. As high-throughput sequencing enables to study the repertoire of these receptors, it is now possible to learn the probabilistic laws of this random process, and to use them to analyse receptors of interest, generate synthetic repertoires to create controls, or aid the identification of receptors that are specific to diseases, with possible applications for medical diagnostics. This article describes how these tasks can be performed using the *IGoR* software, which can learn statistical models from data, annotate existing sequences, or generate new synthetic ones with the same laws as the recombination process.

Keywords: antigen receptors, immune repertoires, V(D)J recombination, statistical inference, sequence analysis

Les récepteurs d'antigènes, exprimés par les lymphocytes B et T de notre système immunitaire adaptatif, jouent un rôle essentiel dans la reconnaissance et l'élimination des agents pathogènes et des infections. La diversité de ces récepteurs immunitaires est générée par un processus d'édition de l'ADN appelé recombinaison VDJ (Tonegawa, 1988), qui se produit dans chaque lymphocyte. Par ce processus, l'ADN natif est réarrangé aléatoirement à partir de segments génomiques, auxquels s'ajoutent des suppressions et insertions aléatoires de paires de bases. Chaque récepteur est formé de deux chaînes (légère et lourde pour les immunoglobulines, alpha et bêta ou gamma et delta pour les récepteurs de cellules T). La première chaîne (légère, alpha, gamma) se compose de deux segments appelés V comme « *variable* » et J comme « *joining* » séparés par un joint de paires de bases aléatoires, tandis que la seconde chaîne comporte en plus un segment D comme « *diversity* » entre les segments V et J, avec deux joints VD et DJ. À la diversité combinatoire résultant du choix de chacun de ces segments parmi plusieurs dizaines dans le génome, s'ajoute la diversité jonctionnelle provenant de la variété d'options dans le nombre de paires rabotées ou ajoutées et sur la nature des paires ajoutées. Comme chacune des deux chaînes se forme de manière essentiellement indépendante, la diversité des récepteurs s'en trouve démultipliée, donnant lieu à un nombre démesuré de combinaisons, et encore augmentée par des hypermutations induites de manière active pendant la maturation d'affinité des lymphocytes B (Cobey *et al.*, 2015). Nous abritons de

*Auteur correspondant : tmora@lps.ens.fr

l'ordre du trillion de lymphocytes B ou T, dont une grande partie exprime un récepteur unique. Ce nombre reste toutefois infime comparé à la diversité combinatoire suggérée par le processus de recombinaison V(D)J (Mora & Walczak, 2016), de la même manière qu'une bibliothèque, même gigantesque, ne saurait épuiser la diversité des livres qui pourraient être écrits.

Les progrès récents dans le séquençage à haut débit (Robins *et al.*, 2009 ; Weinstein *et al.*, 2009) permettent maintenant d'établir une liste assez complète des séquences de récepteurs d'antigène dans un échantillon de lymphocytes, donnant la possibilité d'étudier la composition et la diversité du répertoire immunitaire selon l'individu, son état de santé, l'organe ou le sous-type cellulaire considéré. Ces méthodes promettent de révolutionner la manière d'aborder le répertoire immunitaire et ses implications pour la médecine (Calis & Rosenberg, 2014). Plus fondamentalement, les données ainsi produites nous permettent maintenant de caractériser avec une grande précision les lois probabilistes, ou « grammaire », de la recombinaison V(D)J et donc de la génération aléatoire de la diversité des récepteurs (Murugan *et al.*, 2012 ; Elhanati *et al.*, 2015).

Apprendre cette grammaire n'est pas aisé, car le produit final de la recombinaison V(D)J, c'est-à-dire la séquence ADN chaîne de récepteur, masque en fait la série d'événements aléatoires (choix des segments V, D et J parmi les modèles du génome, nombre de rabotages, nombre et nature des insertions) l'ayant engendrée. Ainsi, une annotation précise et certaine de la séquence, qui donnerait l'origine mécanistique de chaque paire de bases (si celle-ci vient du génome souche et de quel segment, ou bien si elle correspond à une insertion aléatoire) est en principe impossible, même si des logiciels très efficaces en donnent un scénario vraisemblable (Bolotin *et al.*, 2015 ; Duez *et al.*, 2016). Ceci complique la collection de statistiques sur l'usage des segments du génome, de leur rabotage, et du processus d'insertions aléatoires. Ces caractéristiques sont des variables « cachées » qui déterminent la séquence mais auxquelles la séquence seule ne permet pas de remonter de manière univoque. Néanmoins, une astuce statistique, fondée sur une classe d'algorithmes appelés « espérance-maximisation », permet tout de même de découvrir ces statistiques cachées avec une grande précision du moment qu'un grand nombre de séquences est disponible. Cette méthode peut être appliquée aux données de séquençage afin d'inférer la statistique complète des scénarios de recombinaison V(D)J. Une astuce supplémentaire permet d'approcher au plus près du processus pur de recombinaison, avant que les protéines codées par ces séquences n'aient subi de test de viabilité, en ne considérant que les recombinaisons « hors-cadre », c'est-à-dire décalées par rapport au cadre de lecture normal et donc incapables de se replier correctement. Ces séquences résultent d'une recombinaison ratée et ne doivent leur survie qu'au gène de récepteur exprimé sur le second chromosome. Comme la sélection est aveugle à leur présence, elles constituent un laboratoire privilégié afin de comprendre le répertoire primordial.

Plusieurs résultats découlent de cet apprentissage. D'une part, on constate que la distribution de probabilités de scénarios de recombinaison, et par conséquent de génération de séquences, est extrêmement hétérogène. Parmi les séquences effectivement produites par le processus de recombinaison V(D)J, la probabilité de leur génération couvre un nombre considérable d'ordres de grandeur, typiquement de 10^{-20} à 10^{-5} pour la chaîne bêta des récepteurs de lymphocytes T. Cette observation a des conséquences importantes pour la constitution d'un répertoire « public » qui serait partagé par un grand nombre d'individus : les séquences dont la génération serait la plus fréquente tendraient à être partagées par un grand nombre d'individus, car elles seraient susceptibles d'être engendrées chez chacun de manière indépendante. Et de fait, c'est le cas de la plupart des séquences publiques ou partagées (Venturi *et al.*, 2008 ; Elhanati *et al.*, 2014), sauf chez les jumeaux, qui partagent davantage de récepteurs, probablement grâce au fait qu'ils échangent des cellules sanguines pendant la grossesse (Pogorelyy *et al.*, 2017a). Néanmoins, certaines séquences dérogent à cette règle, car elles sont trouvées chez plus d'individus que ne l'aurait prévu le simple biais de recombinaison. Ces séquences sont intéressantes car leur présence répétée suggère une pression sélective qui a favorisé leur expansion, suggérant une cause commune, notamment dans des cohortes d'individus partageant une même condition médicale. Il est ainsi possible d'identifier, à l'aide du modèle de génération, ces séquences « surreprésentées » chez des individus infectés par le cytomégalovirus ou atteints de diabète de type 1 (Seay *et al.*, 2016 ; Emerson *et al.*, 2017 ; Pogorelyy *et al.*, 2017b), parmi de nombreux exemples.

Afin de faciliter la caractérisation statistique du processus de recombinaison V(D)J à partir de données de séquençage, nous avons développé, avec Quentin Marcou et Aleksandra Walczak, *IGoR* (Marcou *et al.*, 2017), un logiciel qui permet de construire et d'utiliser des modèles génératifs de répertoire. L'utilisateur peut apprendre les lois probabilistes de recombinaison V(D)J à partir d'un jeu de données, générer des séquences synthétiques à partir de ces lois, y compris avec des hypermutations aléatoires ciblées, ou bien annoter des séquences d'intérêt particulier, en estimant leur probabilité de génération et en listant les scénarios de recombinaison vraisemblables les ayant engendrées. L'outil permet notamment de créer des contrôles computationnels afin de mieux repérer les anomalies ou des déviations dans des répertoires donnés.

Un des défis principaux au développement de méthodes de diagnostic à partir du répertoire est l'identification des séquences de récepteurs impliquées dans une maladie ou reconnaissant un antigène donné (Dash *et al.*, 2017 ; Glanville *et al.*, 2017). Nous espérons qu'*IGoR* puisse contribuer à l'identification de nouvelles associations entre séquences et pathogenèse, et à terme à de nouveaux diagnostics, par l'analyse de données de séquençage de répertoires.

Références

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V, Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods*, 12, 380-381.

Calis J.J., Rosenberg B.R. (2014). Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol*, 35, 581-590.

Cobey, S., Wilson, P., Matsen, F.A. 4th. (2015). The evolution within us. *Philos Trans R Soc Lond B Biol Sci*, 370, 1676.

Dash P., Fiore-Gartland A.J., Hertz T., Wang G.C., Sharma S., Souquette A., Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547, 89-93.

Duez, M., Giraud, M., Herbert, R., Rocher, T., Salson, M., Thonier, F. (2016). Vidjil: a web platform for analysis of high-throughput repertoire sequencing. *PLoS ONE*, 11, e06166126.

Elhanati Y., Murugan A., Callan C.G., Mora T., Walczak A.M. (2014). Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci USA*, 111, 9875-9880.

Elhanati Y., Sethna Z., Marcou Q., Jr, G.C., Mora T., Walczak A.M. (2015). Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond, B, Biol Sci,* 370, 1676.

Emerson, R.O., DeWitt, W.S., Vignali, M., Gravley, J., Hu, J.K., Osborne, E.J., Desmarais C., Klinger M., Carlson C.S., Hansen J.A., Rieder M., Robins, H.S. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet*, 49, 659-665.

Glanville J., Huang H., Nau A., Hatton O., Wagar L.E., Rubelt F., Ji X., Han A., Krams S.M., Pettus C., Haas N., Arlehamn C.S.L., Sette A., Boyd S.D., Scriba T.J., Martinez O.M., Davis M.M. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547, 94-98.

Marcou Q., Mora T., Walczak A.M. (2017). IGoR: a tool for high-throughput immune repertoire analysis. *BioRxiv preprint*.

Mora T., Walczak A., Quantifying lymphocyte receptor diversity, *in:* J. Das, C. Jayaprakash (Eds.), *Systems immunology: an introduction to modeling methods for scientists*, CRC Press, 2016.

Murugan A., Mora T., Walczak A.M., Callan C.G. (2012). Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA*, 109, 16161-16166.

Pogorelyy M.V, Elhanati Y., Marcou Q., Sycheva A.L., Komech E.A., Nazarov V.I., Britanova O.V., Chudakov D.M., Mamedov I.Z., Lebedev Y.B., Mora T., Walczak, A.M. (2017a). Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput Biol*, 13, e1005572.

Pogorelyy, M.V, Minervina, A.A., Chudakov, D.M., Mamedov, I.Z., Lebedev, Y.B., Mora, T., Walczak, A.M. (2017b). Method for identification of condition-associated public antigen receptor sequences. *BioRxiv preprint*.

Robins H.S., Campregher P.V., Srivastava S.K., Wacher A., Turtle C.J., Kahsai O., Riddell S.R., Warren E.H., Carlson C. S. (2009). Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, 114, 4099-4107.

Seay, H.R., Yusko, E., Rothweiler, S.J., Zhang, L., Posgai, A.L., Campbell-Thompson, M., Vignali M., Emerson R.O., Kaddis J.S., Ko D., Nakayama M., Smith M.J., Cambier J.C., Pugliese A., Atkinson M.A., Robins H.S., Brusko, T.M. (2016). Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight*, 1, 1-19.

Tonegawa S. (1988). Somatic generation of immune diversity. *Biosci Rep*, 8, 3-26.

Venturi, V., Price, D.A., Douek, D.C., Davenport, M.P. (2008). The molecular basis for public T-cell responses? *Nat Rev Immunol*, 8, 231-238.

Weinstein J.A., Jiang N., White R.A., Fisher D.S., Quake S.R. (2009). High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324, 807-810.

Citation de l'article : Mora, T. (2017). *IGoR*: un outil pour apprendre et simuler la génération aléatoire de récepteurs d'antigènes. *Biologie Aujourd'hui*, **211**, 229-231