**The analysis of clonal expansions in normal and autoimmune B cell repertoires**

L’objectif de cet article est de discuter de la diversité des définitions possibles d’un clone au sein d’une population de lymphocytes B. Il traite d’une part la définition théorique d’un clone à partir de la différentiation des cellules, et d’autre part d’une définition plus expérimentale de la notion de clone, à partir du séquençage des BCR. Les différentes méthodes expérimentales présentées révèlent que seul le séquençage de masse de la chaîne lourde ou légère des anticorps permet réellement d’étudier un nombre important de clonotypes, ce qui est fondamental pour notre projet. En ce qui concerne l’analyse des séquences obtenues, il est souligné dans l’article que l’ordre de traitement des séquences dans la comparaison de celles-ci a une forte importance. Ainsi, il semble que la méthode de lignée soit la plus robuste. Elle consiste à regrouper les séquences appartenant à un même clone, et à les placer dans une lignée en fonction des mutations qu’elles présentent par rapport à la lignée germinale qui correspond à ce clone (défini par le gène V et le gène J, la longueur de la séquence CDR3 ainsi qu’une certaine identité dans la séquence). La suite de l’article traite de problématiques qui concernent la finesse de l’analyse. Il propose par exemple des pistes pour différencier les mutations des erreurs de séquençage, des méthodes pour identifier le caractère stéréotypé d’un ensemble de BCR, ainsi que des calculs attestant de la rareté de l’identité de CDR3 pour deux cellules n’appartenant pas à un même clone.

**GLaMST: grow lineages along minimum spanning tree for b cell receptor sequencing data**

L’objectif de cet article est de présenter un algorithme de construction d’un arbre de lignée pour une population séquencée de récepteurs de lymphocytes B. L’hypermutation somatique ainsi que la sélection des récepteurs présentant le plus d’affinités avec les antigènes est en effet similaire à une situation d’évolution des espèces, pour laquelle on a l’habitude de construire des arbres phylogénétiques. La problématique est ici d’ordonner les séquences relevées, ainsi que de déterminer les intermédiaires entre la séquence germinale et les séquences observées. Pour ce faire, cet algorithme se base sur le principe du maximum de parcimonie, puisque les méthodes Bayésiennes ainsi que le maximum de vraisemblance nécessitent des connaissances a priori qui ne sont pas disponibles pour la population étudiée. Le principe de l’algorithme est de construire pas à pas un arbre reconstruit, à partir de la racine qu’est la séquence germinale identifiée comme à la base de la population clonale étudiée. La première étape est la comparaison 2 à 2 de l’ensemble des séquences, afin d’identifier les étapes nécessaires pour passer de l’une à l’autre. Ces opérations sont stockées. Ensuite, à chaque étape, l’algorithme construit un Minimum Spanning Tree reliant les clonotypes en attente aux clonotypes de l’arbre reconstruit, et l’ensemble des opérations stockées sont testées. Celle qui rapproche le plus l’arbre reconstruit des autres clonotypes est conservée et appliquée pour agrandir l’arbre reconstruit, jusqu’à avoir intégré tous les clonotypes. Une fois cet arbre construit, il est simplifié au maximum, notamment en enlevant les branches dont la feuille n’est pas une séquence observée.

Cet algorithme est ensuite comparé aux algorithmes existants, et présente tant sur des populations simulées que des prélèvements réels de bien meilleurs résultats, autant sur le plan de la qualité des résultats que sur le plan de la vitesse d’exécution.

**Using Genotype Abundance to Improve Phylogenetic Inference**

Les algorithmes inférant des arbres de lignée via le principe du maximum de vraisemblance présentent régulièrement des résultats dégénérés, c’est-à-dire plusieurs arbres différents dont la parcimonie est identique. Cet article se propose de donner une méthode permettant de sélectionner l’arbre le plus vraisemblable au sein de cette ‘forêt de parcimonie’. Il est proposé de prendre en compte, pour cette étape du traitement algorithmique, l’abondance des séquences au sein de la population de cellules séquencées, sur le principe intuitif qu’un génotype partagé par un grand nombre de cellules aura plus de chance de présenter des descendants mutants qu’un génotype faiblement présent. Le résultat de cet algorithme est un « Genotype-Collapsed Tree », c’est-à-dire un arbre où chaque génotype observé apparaît de manière unique, annoté avec son abondance dans l’échantillon de séquences analysées. Pour concrétiser cette intuition, ils ont mis en équation un modèle stochastique, dans lequel l’abondance n’intervient que dans un second temps, pour classer les arbres de la forêt de parcimonie. Ils utilisent des estimateurs statistiques pour inférer les paramètres de probabilité de descendance ainsi que de mutation, ce qui permet d’éviter à l’expérimentateur de faire ces estimations lui-même. Une fois ce modèle réalisé, il a d’abord été testé sur des séquences simulées suivant le mieux possible la réalité biologique d’évolution d’une séquence germinale de BCR par SHM. Sur ces lignées, l’arbre exact est disponible puisqu’elles ont été créées de toute part par ordinateur. Cela permet de vérifier que l’arbre obtenu par leur méthode est le plus proche des arbres de la forêt de l’arbre exact, voire l’arbre exact lui-même. Ils ont ensuite effectué ces mêmes tests sur des séquences issues de prélèvements de souris, cette fois en comparant les arbres issus des séquences IGH et IGL puisque la coévolution de celles-ci prévoit des lignées similaires lorsque l’on s’intéresse indépendamment à l’une et à l’autre. C’est de nouveau leur méthode qui minimise la distance entre les arbres issus des séquences IGH et ceux issus des séquences IGL. Ceci est un indicateur de la bonne robustesse de leur méthode. Cependant, il semble que le nombre de séquences sur lequel ils travaillent est relativement faible, et la vitesse de leur algorithme n’est que très rapidement évoqué, ce qui interroge sur la pertinence de celui-ci pour des situations comme les nôtres, où l’on a parfois plus d’un millier de génotypes différents à classer dans un arbre.

**Synthèse des trois articles et ouverture à notre sujet de recherche :**

Le premier article proposé souligne la difficulté de définir exactement un clone, et souligne aussi la pertinence de la construction d’un arbre de lignée pour comprendre le processus évolutif qui tend à sélectionner les lymphocytes B dont le récepteur est le mieux adapté à l’antigène contre lequel il est dirigé. Il s’agit d’ordonner les différents clonotypes d’un même clone (défini par le gène V et le gène J, la longueur de la séquence CDR3 ainsi qu’une certaine identité dans la séquence) par rapport à la séquence germinale de ce clone, donnée disponible sur le site IGMT.

Les deux articles suivants résultent de ce constat, puisque ce sont des méthodes de construction d’arbre de lignée, appliquées au cas des BCR. L’article GLaMST fournit une méthode efficace de construction d’un arbre de lignée à partir du principe du maximum de parcimonie, en intégrant au besoin les séquences intermédiaires dans l’histoire évolutives des BCR, même si celles-ci ne sont pas observées dans l’échantillon analysé. Cet algorithme se révèle précis et surtout efficace, puisqu’il est capable de traiter un nombre important de clonotypes en un temps relativement court. Enfin, le dernier article permet de lever un des problèmes liés à l’utilisation du principe du maximum de parcimonie, puisqu’il se propose de lever l’ambiguïté entre les résultats équivalents de ces algorithmes, en introduisant l’analyse de la donnée de l’abondance des séquences, plutôt que de considérer que chaque séquence unique n’est présente qu’une seule fois dans l’échantillon. Il semblerait donc que la solution la plus efficace et précise serait de combiner GLaMST avec le principe du dernier article afin de répondre à la problématique de notre projet de recherche.