Analyse comparée d'arbres taxonomiques

Clémence REDA, encadrée par Macha NIKOLSKI
Equipe CBIB du CGFB, Bordeaux

21 août 2016

Le contexte général

La métagénomique est l'étude du contenu génétique d'échantillons d'organismes prélevés dans un milieu naturel. En général, ces organismes ne peuvent pas être cultivés en laboratoire, par exemple certaines bactéries dans l'intestin ou dans des écosystèmes marins fragiles. L'ADN présent dans ces échantillons est séquencé : on obtient des reads, c'est-à-dire des morceaux annotés de séquences d'ADN. Le but par la suite est d'assigner ces reads à une espèce de bactérie, c'est-à-dire d'identifier à quels organismes appartiennent ces séquences d'ADN. Il est rare de pouvoir identifier précisément à quelle espèce le read appartient, aussi on se contente généralement d'assigner le read à un noeud de l'arbre phylogénétique (ou arbre taxonomique), permettant d'identifier un read à un rang : une espèce, un genre, une famille, un ordre, une classe, un phylum, un domaine ou un règne, dans l'ordre décroissant de précision sur l'organisme. Tout le travail de la métagénomique consiste alors à extraire le maximum d'informations pertinentes de cet ensemble de données pour répondre à un problème de biologie, de médecine, etc.

Le problème étudié

Il existe d'ores et déjà des programmes qui permettent de préciser l'assignement d'un read à un rang, ou des mesures sur les arbres phylogénétiques qui permettent de quantifier leur pertinence vis-à-vis d'une matrice de distance donnée entre les reads, ou vis-à-vis d'un autre arbre de référence (algorithmes de Neighbor Joining, de Unweigthed Pairs, sur les Tree distance,...), mais il manque des outils pour comparer les arbres obtenus pour les différents échantillons, quantifier leurs similitudes mais aussi leurs différences, et surtout tirer des conclusions sur la présence ou l'absence de certains organismes en fonction des données cliniques (ou métadonnées). Par exemple, pour un groupe de patients sains ou atteints de mucovisidose sur lesquels on a prélevé des échantillons de l'intestin, il peut être intéressant de savoir comment l'âge du patient ou son traitement actuel peuvent influer sur la présence de E.Coli dans son intestin.

Ces questions sont traitées pour le moment par recoupement d'analyses statistiques, en utilisant le test de Wilcoxon ou de Mc Nemar. In fine, l'objectif est de pouvoir clusteriser les patients de manière semi-automatique pour pouvoir par exemple améliorer le traitement ou donner un diagnostic.

La contribution proposée

Trois approches ont été proposées, toutes répondant à des problèmes légèrement différents.

1. Pour faire l'analyse des données issues de l'assignation des noeuds par le logiciel TANGO, on calcule divers scores sur l'arbre de départ (qui quantifient le nombre de noeuds/bactéries en commun entre deux échantillons, la proximité phylogénétique des bactéries présentes dans les échantillons, ...) et sur les valeurs des données cliniques. Puis le programme propose de partitionner les patients en fonction de la valeur d'une certaine donnée clinique, on calcule la distance entre deux classes de patients pour toutes les paires de classes possibles, et on renvoie les paires de classes de patients les plus éloignées. Cela permet de savoir si la donnée clinique choisie est un facteur discriminant pour les populations bactériennes des patients.
2. La deuxième approche tente de comparer les arbres taxonomiques des patients avant assignation des noeuds. Elle utilise un classificateur naïf bayésien, qui tente de classer les patients dans divers groupes de valeurs de données cliniques en fonction de leurs populations bactériennes. La classification obtenue est évaluée par un coefficient de Youden. Si le coefficient a une valeur satisfaisante, la classification selon la population bactérienne est cohérente vis-à-vis de la donnée clinique, ce qui souligne une corrélation entre cette donnée clinique et les bactéries considérées.
3. La troisième approche compare les arbres avant assignation des noeuds, mais évite d'avoir à fournir a priori des hypothèses sur les probabilités d'assignation aux noeuds, ou d'avoir une certaine valeur de donnée clinique. Le programme clusterise les patients par un algorithme des K-moyennes, en utilisant des distances sur les arbres taxonomiques qui correspondent à chaque patient. Ces clusters sont comparés aux clusters des patients selon la valeur des données cliniques. Si ces deux groupes de clusters sont proches (ceci étant quantifié par une distance), cela peut signifier que les données cliniques considérées et les populations bactériennes sont liées. L'algorithme renvoie également les bactéries en commun entre les patients d'un même cluster.

Les arguments en faveur de sa validité

Ces algorithmes ont été testés sur une base de données existante, issues d'une étude de l'hôpital Pellegrin de Bordeaux. Les résultats obtenus par les algorithmes ont été comparés à ceux résultant de l'analyse statistique, et semblent similaires.

Toutefois, pour pouvoir valider véritablement les résultats biologiques obtenus, il faudrait pouvoir avoir accès à plusieurs dizaines d'ensembles d'échantillons et itérer les calculs précédents pour pouvoir les confirmer. Cela sort malheureusement du cadre du stage. Ainsi, les hypothèses utilisées, notamment sur les probabilités d'assignation à un noeud donné de l'arbre taxonomique, ou la probabilité d'avoir une certaine valeur de métadonnée, ne sont pas a priori plus valables que d'autres. D'autre part, selon le traitement des résultats donnés en entrée (normalisation des données numériques, existence d'une procédure standardisée pour la récupération des données...), les résultats numériques obtenus peuvent être incohérents. De plus, l'interprétation des données obtenus ne peut pas se passer de l'avis du praticien.

Le bilan et les perspectives

L'utilisation d'algorithmes de Machine Learning en métagénomique n'est pas nouvelle, mais est ici exploitée pour faire le lien entre des données cliniques, dont le traitement relève en général plus de la statistique, et les arbres. D'une part, pour les arbres étiquetés et ordonnés, beaucoup de problèmes de comparaison sont NP-complets (l'inclusion, ...). Ce qui permet d'échapper à ce problème de complexité dans notre problème est que la comparaison pertinente ici soit essentiellement celle des ensembles de feuilles. Mais la complexité temporelle des deux dernières méthodes reste à améliorer. D'autre part, il serait intéressant dans le futur de pouvoir tester ces algorithmes sur d'autres bases de données.