De novo comparison of metagenomic data.

A new tool: Compareads

Nicolas MAILLET[1], Claire LEMAITRE[1], Rayan CHIKHI[2], Dominique LAVENIER[1] and Pierre PETERLONGO[1]

[1] INRIA Rennes - Bretagne Atlantique/IRISA, EPI GenScale (Symbiose), Campus universitaire de Beaulieu, 35042, RENNES Cedex, France
{nicolas.maillet, claire.lemaitre, dominique.lavenier, pierre.peterlongo}@inria.fr

[2] ENS Cachan/IRISA, EPI GenScale (Symbiose), UMR 6074 CNRS, Campus universitaire de Beaulieu, 35042, RENNES Cedex, France
rayan.chikhi@irisa.fr

Keywords Comparative metagenomics, Next-generation sequencing, Bloom filter

Comparaison *de novo* de données métagénomique

Un nouvel outil : Compareads

Mots-clés Métagénomique comparative, Séquençage de nouvelle génération, filtre de Bloom

La métagénomique étudie l'ensemble des génomes, ainsi que leurs interactions, au sein d'un même biotope. Une très grande proportion des micro-organismes n'est pas cultivable en milieu contrôlé (plus de 99,9% dans l'eau de mer[1]). C'est pourquoi la métagénomique va consister à séquencer tous les génomes présents dans un échantillon sans passer par une phase de culture ou de différenciation des espèces en amont. À l'issu d'un séquençage, on récupère des centaines de millions de fragments (des *lectures*, ou *reads*) qui représentent chacun une fraction de l'ADN des divers organismes présents dans cet échantillon. Cette discipline récente, supportée par l'évolution rapide des technologies de séquençage à haut débit, amène naturellement de nouvelles problématiques comme, par exemple, la mesure de similarité entre deux métagénomes. Une manière d'évaluer cette similarité peut être de compter le nombre de *reads* communs entre deux métagénomes. Deux échantillons contenant des espèces différentes partageront peu de *reads* en commun. À l'inverse, deux métagénomes de compositions identiques auront un grand nombre de *reads* similaires. Cette mesure approximative ne reflète probablement pas une réalité effective, elle permet simplement d'établir une mesure de similarité relative entre plusieurs métagénomes. La mise en œuvre de cette mesure de similarité métagénomique n'est cependant pas immédiate. Une comparaison classique de tous les *reads* d'un échantillon A contre ceux d'un échantillon B représente un volume de calcul très conséquent. À titre d'exemple, deux *runs* de type Illumina de 2×10^8 *reads* impliqueraient 4×10^{16} comparaisons individuelles de *reads*. Avec une hypothèse (optimiste) de 100ns pour comparer deux *reads*, le calcul prendrait 4×10^9 secondes, soit 15 ans sur un processeur actuel (8 cœurs).

Notre méthode pour comparer deux métagénomes repose d'abord sur une manière simple et rapide d'exprimer une similarité entre deux *reads*. Deux *reads* sont considérés comme similaires s'ils partagent au moins P mots non chevauchant de K caractères. Le principe de la comparaison de deux métagénomes A et B s'effectue alors de la manière suivante :

– 1. Tous les mots de K caractères du métagénome A sont indexés. La structure de données utilisée stocke en mémoire tous les mots différents de K caractères présents dans le jeu de données A. L'interrogation de cette structure de données avec un mot quelconque de K caractères indique rapidement si ce mot y est présent ou non.

– 2. Pour chaque *read* du métagénome B, on test la présence de tous ses mots non chevauchant de K caractères dans la structure de données (*i. e.* le métagénome A). Si au moins P mots sont trouvés, le *read*

est retenu et stocké dans un ensemble $B*$. À la fin de cette étape, l'ensemble $B*$ représente donc tous les *reads* de B qui ont au moins une occurrence dans A.

– 3. On répète l'étape 1., en utilisant cette fois-ci le métagénome B.

– 4. On répète l'étape 2., en utilisant cette fois-ci le métagénome A. Les *reads* retenus sont mis dans l'ensemble $A*$.

Pour éviter des comparaisons inutiles, optimiser le processus et réduire encore les temps de calcul, l'étape 3. indexe uniquement l'ensemble $B*$. En effet, si un *read* de B n'est pas présent dans A, il est inutile d'effectuer des tests de comparaison avec ce *read* à l'étape 4..
Le résultat de la comparaison est un couple de nombres (X, Y) où X est le cardinal de $A*$ et Y le cardinal de $B*$. X représente donc le nombre de *reads* de A ayant au moins une occurrence dans B, et Y, le nombre de *reads* de B ayant au moins une occurrence dans A. Dans la structure d'indexation mise en place, l'information qui relie mots et *reads* est perdue. Ainsi, aux étapes 2. et 4., on teste simplement l'appartenance d'un mot à un métagénome, ce qui génère des faux-positifs. Si pour un *read* donné de A, au moins P mots qui le composent appartiennent au métagénome B, rien n'indique que ces mots appartiennent au **même** *read* de B. En pratique, pour des mots suffisamment longs, le nombre de faux-positifs reste faible.
Dans cette approche, la structure d'indexation est centrale. On doit représenter en un minimum d'espace mémoire un très grand nombre de mots de K caractères (plusieurs milliards). De plus, cette structure est extrêmement sollicitée et doit donc être rapide. Notre implémentation repose sur un index probabiliste basé sur le concept de filtre de Bloom[2]. L'inconvénient de cette structure probabiliste est qu'elle génère également des faux-positifs.

Le logiciel Compareads est une première implémentation (mono-cœur) réalisée en C pour valider cette approche. La comparaison de deux métagénomes comprenant 10^8 *reads* chacun dure environ 5 heures. De nombreuses optimisations, dont de la parallélisation sur multi-cœurs, sont encore possibles pour réduire significativement ce temps.
Une validation fonctionnelle portant sur 15 métagénomes, représentant 3 conditions différentes d'un même milieu, a été conduite afin de vérifier que cette mesure de similarité permet rapidement de positionner des métagénomes les uns par rapport aux autres. Le calcul des similarités deux à deux des 15 métagénomes (105 comparaisons) a ainsi permis de retrouver aisément la hiérarchisation relative aux 3 conditions initiales.
Une seconde validation a consisté à filtrer 4 jeux de données métagénomique fortement contaminés par des séquences d'ADN humain. Chaque échantillon a été comparé au génome humain afin d'en éliminer les séquences similaires. Les résultats des 4 intersections sont venus confirmer de précédentes études réalisées à l'aide de BLAST[3] et de MG-RAST[4] en terme de pourcentage de contamination.

La méthode mise en pratique dans le logiciel Compareads a plusieurs avantages. Premièrement, cette méthode permet de donner un score de similarité entre plusieurs métagénomes. Deuxièmement, cette méthode peut servir de filtre numérique sur des données métagénomique, afin par exemple de retirer une espèce d'un métagénome. Enfin, la consommation mémoire n'est pas dépendante de la quantité de données à traiter, quantité qui peut être très fluctuante suivant les technologies de séquençage utilisées.

Références

[1] R. I. Amann, W. Ludwig and K. H. Schleifer, Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59 :143-169, 1995.

[2] B. H. Bloom, Space/time trade-offs in hash coding with allowable errors *Communications of the ACM*, 13 :422-426, 1970.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, Basic local alignment search tool *J. Mol. Biol.*, 215 :403-410, 1990.

[4] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. A. Edwards, The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes *BMC Bioinformatics*, 9 :386, 2008.