Université de Montpellier

Master Science et Numérique pour la Santé

Parcours : Bioinformatique, Connaissance et Données

Projet Bibliographique

Classification métagénomique : utilisation de k-mers espacés à la place des k-mers contigüs

Mamy Andrianteranagna
Promotion 2015-2016

janvier 2016

Sommaire

1 Introduction

1.1 Définition et buts de la métagénomique

La métagénomique est l'étude des populations microbiennes via le « métagénome ». Le métagénome est l'ensemble des génomes de tous les espèces présentes dans un environnement donné. La métagénomique consiste à l'identification, à la classification, à la quantification et à la caractérisation des espèces microbiennes qui peuplent un échantillon d'un milieu donné. Elle permet l'exploration des populations microbiennes des milieux naturels tels que les océans, le sol, les tubes digestifs des animaux, etc. sans culture préalable.

Ainsi, la métagénomique permet l'étude des espèces microbiennes non cultivables c'est-à-dire qu'elles ne peuvent pas être cultivées en milieu artificiel ou qu'elles ne sont jamais été l'objet d'une culture justement parce qu'elles ne sont pas encore été identifiées.

En outre, sur le plan génomique et génétique, la métagénomique permet l'assemblage de novo des génomes de ces espèces non cultivables ainsi la prédiction et l'annotation fonctionnelle des gènes. Sur le plan biochimique, elle a permi la découverte de fonction de nombreuses protéines et enzymes.

Bref, la métagénomique regroupe tout ce qui est étude qu'on peut réaliser à partir de l'extraction de l'ensemble de génomes de tous les espèces microbiennes d'un milieu donnée sans avoir à les connaîtres *a priori*.

1.2 Historique et évolution de la métagénomique

Vers le milieux des années 80, les microbiologistes commencent à prendre conscience de l'importance écrasante des espèces microbiennes non cultivables et qu'il n'est plus raisonnable de les ignorer [12]. A titre d'information, seul 1 % de la totalité des espèces microbiennes sont cultivables. De cette prise de conscience se découle alors le besoin de classifier phylogénétiquement des espèces présentes dans un milieux sauvage donné. Ceci a été possible grace à l'avancé de la biologie moléculaire et à l'utilisation de plus en plus facile des séquences d'ARNr 16S comme marqueur phylogénétique [7].

Une autre grande avancée de la biologie moléculaire est survenues au début des années 90. C'est la possibilité de cloner et d'amplifier un fragment d'ADN à partir de n'importe quel milieu grace à la réaction en chaîne des polymérases connue sous le nom de PCR. Cela a permi de de cloner entièrement le gène d'ARNr 16S directement à partir du milieu naturel sans passer par des techniques laborieuses. Cette avancée technologique accélère la détermination et classification des nouvelles espèces microbiennes et, encore une fois, démontre l'importance des espèces non cultivables [11].

Si tout cela s'agit déjà de la métagénomique, ce n'est que vers la fin des années 90 que le terme « métagénomique »a été utilisé pour la première fois pour désigner ce genre d'étude [5].

Actuellement, l'apparition du NGS vers la deuxième moitié du XXI$^{\text{ème}}$ siècle ont totalement bouleversé la métagénomique.

1.3 Les différentes approches de la métagénomique

Comme nous avons dit précédemment, l'étude métagénomique commence toujours par l'extraction du métagénome du milieu. A partir de là, deux approches peuvent être suivies : l'approche fonctionnelle et l'approche par séquençage.

L'*approche fonctionnelle* ou métagénomique fonctionnelle consiste à exprimer les différents fragments provenant du métagénomes dans des organismes d'expréssion hétérologue tel que l'*Escherichia coli* et d'en identifier des fonctions grace à des techniques enzymatiques.

L'*approche par séquençage*, que nous appelons métagénomique par séqueçage consiste à séquencer les fragements du métagénome et d'analyser ces séquences notamment pour quantifier et classifier les espèces qui sont présentes dans le milieu étudié. Pour cela, le clonage et le séquençage uniquement des marqueurs de phylogénétiques, tel que l'ARNr16S, a été très à la mode à partir du début des années 90 grace à la

découverte de la PCR. A l'heure actuelle, le séquençage de tous les fragments métagénomiques ou shotgun sequencing est largement utilisé après l'apparition de la NGS [**?**].

1.4 La métagénomique par séquençage et la classification métagénomique

Le séquençage massif à haut débit (NGS) a permis à l'approche par séquençage d'être l'approche la plus répandue dans les études métagénomiques actuelles. La classification métagénomique (*binning* en anglais) est le fait de classer les fragments de séquences du métagénome. Elle peut être dirigée ou pas selon le type de fragment à classifier.

La classification dirigée utilise un seul fragment ou amplicon, dans la plupart des cas, des marqueurs phylogénétiques connus comme le gène d'ARNr 16S. Cette méthode de classification est aussi appelée, dans la littérature, *amplicon-based classification* ou métagénomique dirigée. La classification non dirigée, quant à elle, utilise l'ensemble des fragments du métagénome, c'est à dire tous les *reads*. Elle est aussi appelée *shotgun sequencing-based classification* ou métagénomique non dirigée [10] et, actuellement, la plus utilisée.

Selon l'utilisation ou pas des génomes de référence, la classification peut être à taxonomie dépendante ou à taxonomie indépendante. La classification à taxonomie dépendante utilise des génomes de références phylogénétiquement connues. Il s'agit alors d'une classification supervisée et c'est la méthode de classification la plus utilisée. La classification à taxonomie indépendant n'utilise pas de génomes de références et s'agit d'une classification non supervisée.

De nombreux outils sont actuellement disponibles en ligne pour l'analyse et la classification métagénomique. Les plus connus d'entre eux sont présentés par [3].

1.5 Problématique et objectif de l'analyse bibliographique

Pour cette étude bibliographique, ce qui nous intéresse est la classification non dirigée et à taxonomie dépendante qui utilise des génomes de références afin classifier l'ensemble des reads. Dans ce contexte, deux méthodes sont actuellement les plus utilisées : la méthode basée sur l'alignement des séquences et la méthode non basée sur l'alignement [10]. Parmi la méthode non basée sur l'alignement, il y a le comptage des k-mers communs entre le read et le génome de référence. Cette méthode est utiisée par de nombreux outils publiés récemment tels que LMAT [1], Kraken [14]. Afin d'améliorer la sensibilité/spécificité de la classification utilisant la méthode des k-mers commun, [2] suggère l'utilisation des k-mers espacés à la place des k-mers contigus.

Dans ce travail bibliographique, nous essayons d'analyser cette nouvelle approche de classification proposée par [2]. Dans un premier temps, nous allons présenter le principe de k-mers espacées. Dans un second temps, nous allons analyser les résultats obtenus en comparant les deux méthodes. Et enfin, nous allons discuter de l'intérêt de cette nouvelle approche de classification utilisant les k-mers communs dans le contexte actuel de la métagénomique.

2 Les algorithmes de classification utilisant des k-mers communs

Le k-mer est définit comme un sous-mot de longueur k contenu dans un mot. Comme un read est un fragment de séquence génomique issus du séquençage à haut débit, il est constitué de k-mers. Un read de longueur L contient L-k+1 k-mers de longueur k décalé à la fois d'une seule base. Ceci est appelé des k-mers contigüs qui sont utilisés dans de nombreux analyses de séquence issue de la NGS.

2.1 Les k-mers contigüs et la classification métagénomique

Dans la classification métagénomique, les k-mers contigüs peuvent être utilisés soit pour l'alignement des reads aux génomes de références (mapping), soit pour comparer ceux-ci aux génomes de références sans tenir compte des positions. Dans le dernier cas, qui est plus économique en temps et en espace, le nombre

d'occurence de chaque k-mer d'un read donnée est enregistré pour chaque génome de référence. Afin d'illustrer comment ça marche cette algorithme, nous allons nous baser avec ce que fait l'outil Kraken pour classer les reads métagénomiques.

Pour classifier un read, Kraken [14] cherche chaque k-mer de ce read dans chaque génome de référence. Il associe alors le k-mer donné au noeud représentant le plus ancien ancêtre commun des génomes de références dans lesquelles il apparaît. Un poids correspondant au nombre de k-mers associé (nombre de hit) est alors attribué à chaque noeud. Le reads est alors classé au niveau de la taxa dont la somme des poids de la racine à la feuille est la plus élevée.

2.2 Les k-mers espacées, une alternative aux k-mers contigüs dans la classification métagénomique

Le concept de k-mer espacé vient des graines espacées. L'alignement par extension de graine est notamment utilisé pour comparer une séquence donnée avec des séquences dans une base de donnée (par exemple le cas de BLAST). Cependant, la taille de la base de donnée qui ne cesse d'augmenter ralentisse la comparaison. Afin de palier à cela, Ma *et al.*[9] ont proposé pour la première fois, dans son algo PatternHunter, l'utilisation de graine espacée tout en augmentant la sensibilité des résultats. Des nombreuses études ultérieures ont repris ce concepte dans différents type d'analyse de séquence génomique.

Dans la classification métagénomique que nous étudions ici, l'utilisation de k-mers espacées est proposée à la place des k-mers contigü que nous venons de voir dans la sous-section précédente. L'objet est d'introduire des trous au niveau de chaque k-mer. Ces trous, ou joker, représenté par un « - », peuvent prendre n'importe quelle base lors de la comparaison. Les autres bases autre que les jokers sont représenté par un « # »dont le nombre total représente le poids du k-mer. Deux métriques peuvent alors être mésurés pour un read : le nombre de hit (nombre de k-mers retrouvés dans le génome de référence) et la couverture ou le nombre total de positions couvertes par tous les k-mers matchés. Ce sont ces métriques qui seront utilisé pour classifier les reads.

Le k-mer espacé est aussi proposés par différents auteurs pour améliorer différentes méthodes d'alignement de reads dans différents contexte d'analyse NGS : Chip-Seq [4], reconstruction phylogénétique [8], etc.

3 Comparaison des résultats de classification utilisant les k-mers contigus versus k-mers espacés

3.1 comparaison sur la sensibilité-spécificité des résultats données par Kraken et Seed-Kraken

Afin de comparer les résultats de classification en utilisant les k-mers contigüs versus k-mers espacées, les auteurs ont choisis d'utiliser Kraken [14] que nous avons décrit auparavant. Ils ont adapté cet outil afin qu'il puisse intégrer une autre option capable de traiter les k-mers espacées. Cette option, ils l'ont appelé seed-Kraken.

Quatre données métagénomiques ont été utilisées pour la comparaison entre les deux options Kraken (k-mers contigüs) et Seed-Kraken (k-mers espacés). Trois de ces données sont des données simulés (HiSeq, MiSeq, Simba-5) dont chacun contient 10000 séquences et 1 est une donnée réelle (HMPtongue) construite à partir de 50000 séquences sélectionnées au hasard dans les séquences contenant dans SRS011086 de l'Human Microbiome Project. HiSeq et MiSeq ont été chacun construites à partir des génomes microbiens réels [14] et ont déjà été les objets de validation de Kraken [**?**]. Pareil pour Simba-5 qui a été construit à partir des données génomiques présentes dans RNA-Seq [**?**]. Tous les reads ont une taille de 100 paires de base.

Concernant la base de données utilisée comme référence afin de comparer les k-mers, elle est composée de 915 genomes bactériens. Ces génomes sont constituées de l'ensemble des génomes bactériennes ajouté à tous les génomes à partir desquelles les données simulées ont été construites.

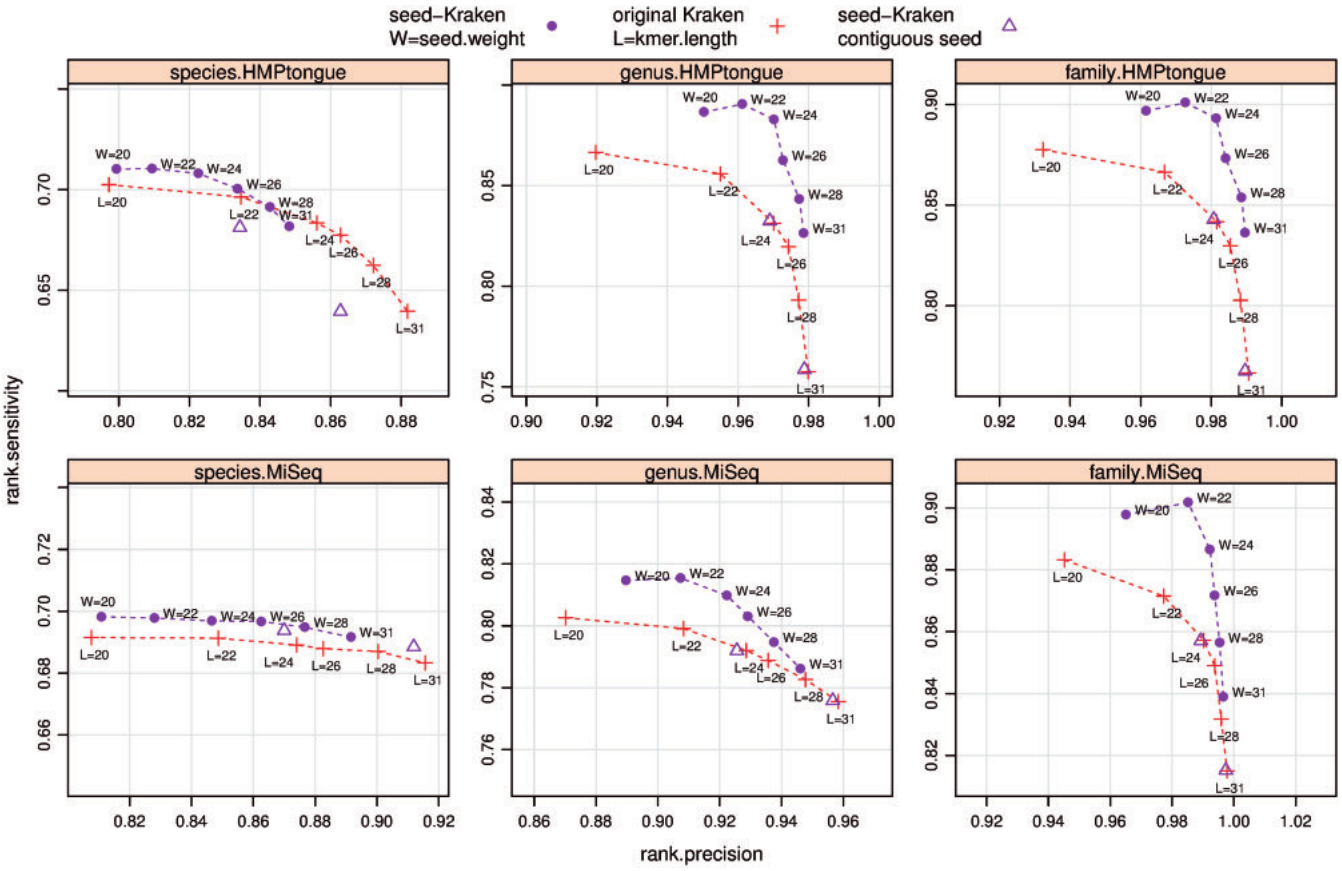

FIGURE 1 – Sensibilité/précision de la classification obtenue avec kraken (point) et de seed-kraken (+) sur les données Miseq et de HMPtongue sur trois niveaux taxonomique (famille, genre, espèce). Les deux triangles en bleu représentent, respectivement, les k-mers contigüs de longueur 24 et 31 lancés avec seed-kraken afin d'évaluer l'effet de l'adaptation apporté à celui-ci

Deux critères ont été utilisés pour comparer les résultats par kraken et par seed-kraken : la sensibilité et précision de la classification. La sensibilité de la classification est le pourcentage de reads correctement classifiés parmi la totalité des reads (1). La précision de la classification est la fraction correcte de toute la classification (2). Ces critères ont été évalués sur trois niveaux taxonomiques différents qui sont la famille, le genre et l'espèce.

$$sensibilite = \frac{nombre\ \ de\ \ reads\ \ correctement\ \ classifies}{nombre\ \ total\ \ de\ \ reads} \tag{1}$$

$$precision = \frac{nombre\ \ de\ \ classifications\ \ correctes}{nombre\ \ total\ \ de\ \ classifications} \tag{2}$$

La longueur de k-mers (pour les k-mers contigüs) et le poids (pour les k-mers esapés) a été variée de 20, 22, 24, 26, 28 et 31.

D'après la figure 1 de [2], au niveau de classification par famille et par genre, on observe, en général, des gains en sensibilité et en précision avec les k-mers espacés. Le gain en sensibilité est observé sur n'importe quelle valeur de poids des k-mers espacés. Il est, cependant, beaucoup plus importante pour les poids élevées (à partir de 24). Inversement, le gain en précision est beaucoup plus marqué pour les poids faible (20 et parfois 22) et qu'il est moindre voir inexistant pour les poids élevé (à partir de 24).

Au niveau classification par espèce, un léger gain en sensibilité est accompagné d'une forte perte en précision. Les résultats montrent que plus le poids des k-mers est élevé plus cette perte en précision est importante.

Bref, le k-mers espacés apporte des gains significatifs en sensibilité et en précision au niveau famille et genre. Cependant, au niveau classification par espèce, il n'apporte qu'une faible amélioration de la sensibilité et diminue la précision de la classification.

3.2 Le coût en temps de calcul et en espace mémoire de l'utilisation de k-mers espacés

Kraken n'indexe pas le complémentaire des k-mers. Afin de tenir compte ces derniers, seed-kaken dois indexer aussi les complémentaires des k-mers espacées étant donné que ceux-ci ne peut pas être obtenu directement. Par conséquent, ce rajout d'indexation augugmente l'espace mémoire utilisée ainsi que le temps de recherche de k-mers. Le temps de calcul augmente alors de l'ordre de 3 à 5 fois avec seed-kraken comparé à kraken.

4 Discussions et Conclusion

4.1 L'utilisation des k-mers espacés represente-t-elle rééllement des intérêts à la métagénomique actuelle ?

Juste avant l'apparition de la NGS, un projet de métagénomique compte environ 2 millions de reads (1 milliard de bp). Après l'apparition de la NGS, ce nombre devient de l'ordre de milliard de reads [6] (produit en un seul run de plateforme NGS). Ces énormes données produites par la NGS imposent alors depuis un grand défit sur le temps de calcul et l'espace mémoire lors de leur traitement.

En outre, les courtes séquences des reads produites par le NGS (par rapport au séquençage Sanger utilisé auparavant) apporte une grande difficulté d'analyse notamment pour l'assemblage de novo [6]. Face à ces défis qui attends encore la métagénomique actuelle, la question se pose si l'amélioration de la sensibilité apportée par la méthode des k-mers espacés est-elle primordiale. Etant donné que cette amélioration est aussi accompagné de coût additionnel en temps et en espace, la réponse semble plutôt négative. Ce qui ne veut pas dire que ce travail inutile, il est certe intéressant mais devrait être accompagné par des amélioration au niveau facilité de calcul et de stockage. D'ailleurs, beaucoup d'autres travaux vont dans la même direction en essayant d'améliorer la performance de la classification métagénomique.

4.2 Les autres tentatives d'amélioration de la classification métagénomique

Le nombre de travaux effectués sur l'algorithme de classification métagénomique montre l'intérêt dans un futur proche de l'amélioration de cette performance de la classification métagénomique. Parmi ces travaux, on peut citer celui de Kumar et al. [6] et de Wang et al. [13]. Ce dernier propose le text mining appliqué dans la classification métagénomique. Ramazotti et al. propose aussi une nouvelle méthode de métagénomique combinant la classification non dirigée et la classification dirigée pour améliorer la classification à taxonomie indépendante c'est à dire sans utilisation de génomes de références.

Références

[1] Sasha K. Ames, David a. Hysom, Shea N. Gardner, G. Scott Lloyd, Maya B. Gokhale, and Jonathan E. Allen. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, 29(18) :2253–2260, 2013.

[2] Karel Břinda, Maciej Sykulski, and Gregory Kucherov. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, 2015.

[3] Pravin Dudhagara, Sunil Bhavsar, Chintan Bhagat, Anjana Ghelani, Rajesh Patel, S Bhavsar, C Bhagat, A Ghelani, S Bhatt, R Patel, and Metagenomics Studies. Web Resources for Metagenomics Studies. *Genomics, Proteomics & Bioinformatics*, 13(5) :296–303, 2015.

[4] Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology*, 10(7) :e1003711, 2014.

[5] J Handelsman, M R Rondon, S F Brady, J Clardy, and R M Goodman. Molecular biological access to the chemistry of unknown soil microbes : a new frontier for natural products. *Chemistry & biology*, 5(10) :R245–R249, 1998.

[6] Satish Kumar, Kishore Kumar Krishnani, Bharat Bhushan, and Manoj Pandit Brahmane. Metagenomics : Retrospect and Prospects in High Throughput Age. 2015, 2015.

[7] David J Lane, Bernadette Pace, Gary J Olsen, David A Stahl, Mitchell L Sogin, and Norman R Pace. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Nati. Acad. Sci. USA*, 82 :6955–6959, 1985.

[8] Chris-Andre Leimeister, Marcus Boden, Sebastian Horwege, Sebastian Lindner, and Burkhard Morgenstern. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30(14) :1991–1999, 2014.

[9] Bin Ma, John Tromp, and Ming Li. PatternHunter : faster and more sensitive homology search. *Bioinformatics (Oxford, England)*, 18(3) :440–445, 2002.

[10] S. S. Mande, M. H. Mohammed, and T. S. Ghosh. Classification of metagenomic sequences : methods and challenges. *Briefings in Bioinformatics*, 13(6) :669–681, 2012.

[11] T. M. Schmidt, E. F. DeLong, and N. R. Pace. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology*, 173(14) :4371–4378, 1991.

[12] V Torsvik, J Goksøyr, F L Daae, Vigdis Torsvik, Jostein Goksyr, and Frida Lise Daae. High diversity in DNA of soil bacteria. *Applied and environmental microbiology*, 56(3) :782–787, 1990.

[13] Ying Wang, Xiaoye Lei, Shun Wang, Zicheng Wang, Nianfeng Song, Feng Zeng, and Ting Chen. Effect of k-tuple length on sample-comparison with high-throughput sequencing data. *Biochemical and biophysical research communications*, pages 1–7, 2015.

[14] Derrick E Wood and Steven L Salzberg. Kraken : ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3) :R46, 2014.