

UNIVERSITÉ DE MONTPELLIER

MASTER SCIENCE ET NUMÉRIQUE POUR LA SANTÉ

PARCOURS : BIOINFORMATIQUE, CONNAISSANCE ET DONNÉES

PROJET BIBLIOGRAPHIQUE

**Nouvelle génération de métagénomique :
comparaison de l'utilisation de la méthode des
k-mers contigus avec celle des k-mers espacés**

Mamy ANDRIANTERANAGNA
Promotion 2015-2016

janvier 2016

Sommaire

1	Introduction	1
2	Les algorithmes de classification des reads	2
3	Comparaison des deux algorithmes d'alignement	2
4	Discussions et Conclusion	2

1 Introduction

- définition et buts de la métagénomique
 - métagénomique = étude des populations microbiennes via le métagénome
 - métagénome = ensemble des génomes de tous les espèces présentes dans un environnement donné
 - métagénomique → identification, classification et quantification des espèces microbiennes présentes dans un échantillon d'un milieu donné
 - métagénomique → exploration des populations microbiennes des océans, du sol, des tubes digestifs, etc.
 - métagénomique → étude des espèces microbiennes non cultivables
- historique et évolution de la métagénomique [2]
 - milieux des années 80 : prise de conscience des microbiologistes sur l'importance et le besoin d'étudier les microorganismes non cultivables -> classification (phylogénétique) des espèces présentes dans un milieu sauvage donné, grâce à l'utilisation de plus en plus facile des séquences d'ARNr 16S (phylogenetic stain) rendue facile (cette facilité est due à des progrès techniques telles que publié dans [3])
 - début des années 90 : PCR → possibilité de cloner entièrement le gène d'ARNr 16S à partir du milieu sans passer par des techniques lourdes ⇒ rapidité et efficacité de la détermination et classification des nouvelles espèces microbiennes [4]
 - fin des années 90 : naissance de l'appellation métagénomique [1]
 - au début, la métagénomique sert uniquement à identifier les espèces présentes dans le milieu étudié puis, par la suite, elle permet aussi de caractériser leurs fonctions
- approches de la métagénomique
 - approche par séquençage (sequence-based analysis) : séquençage des marqueurs de phylogénétiques (ex. ARNr16S) ou shotgun sequencing
 - approche fonctionnelle (fonctionnal métagénomics) → hétérologous expression (*E.coli*)
- la nouvelle génération de métagénomique
 - influence de l'apparition de NGS ⇒ shotgun sequencing metagenomics
 - objectif de la classification : comparaison des fragments de séquences à une séquence génomique de référence
 - méthodes de classifications : amplicon-based et shotgun sequencing based [?], taxonomy dependent (→ classification supervisée → utilisation de références phylogénétiquement connues) or taxonomy independent (→ classification non supervisée)
 - méthode de classification la plus répandue pour shotgun sequencing metagenomics → comptage des k-mers communs (entre les reads et chaque génome de référence) → classification
- objectif de l'étude
 - proposition de [?] d'utiliser les k-mers espacés à la place des k-mers contigus pour améliorer la sensibilité/spécificité de la classification
 - objectif : comparaison des résultats de classification (sensibilité/spécificité), comparaison sur l'implémentation et la complexité

2 Les algorithmes de classification des reads

- méthode des k-mers contigus dans la classification métagénomique
 - construction des k-mers de longueur k sur un read de longueur L , décalé à la fois d'une seule base
 - nb de k-mers = $L-k+1$
- méthode des k-mers espacés
 - concept venant directe des graines espacées des alignements par extension de graine
 - graines espacées : proposé par [?] dans son algo PatternHunter pour garder la rapidité de comparaison lors des alignements par graines (cette rapidité diminue dû à l'augmentation des tailles des bases de données) tout en augmentant la sensibilité
 - nombre de hit : nb de k-mers (espacés) retrouvés dans le génome de référence
 - couverture : nb total de positions couvertes par tous les k-mers matchés

3 Comparaison des deux algorithmes d'alignement

- comparaison sur la sensibilité-spécificité des résultats
 - utilisation de Kraken [?] adapté pour traiter les k-mers espacés → seed-Kraken
 - données utilisées pour la comparaison entre les deux méthodes : 3 données métagénomiques simulées (HiSeq, MiSeq, Simba-5) chacun contenant 10000 séquences et 1 donnée réel construite à partir de 50000 séquences sélectionnées de SRS011086 de l'Human Microbiome Project (HMP-tongue)
 - MiSeq ← reads de 10 génomes bactériens (illumina, simulé)
 - HMPtongue ← provenant de reads sélectionnées au hasard (illumina, données réelle)
 - base de données (référence) : composé de 915 génomes bactériens (un par espèce + génomes de toutes les souches bactériennes à partir desquelles les données simulées ont été construites)
 - paramètre de comparaison : sensibilité et précision de classification sur trois niveaux taxonomique : famille, genre et espèce.
 - $sensibilite = \frac{nb\ de\ reads\ correctement\ classifie}{nb\ total\ des\ reads}$
 - $precision = \frac{nb\ de\ classification\ correcte}{nb\ total\ des\ classification}$
 - fig3 de l'article → gain en sensibilité et en précision avec les k-mers espacés au niveau classification par genre et par famille (poids faible *rightarrow* plus de sensibilité et moins de précision, poids élevées *rightarrow* moins de sensibilité mais plus de précision, gain en sensibilité plus marqué au niveau classification par famille), un léger gain en sensibilité et une diminution en précision au niveau classification par espèce
- comparaison sur l'implémentation, la complexité temporelle et spatiale
 - adaptation de Kraken : indexation du complément de chaque k-mers espacés ⇒ augmentation de l'espace mémoire utilisée et du temps de recherche de k-mers
 - temps de calcul lors de classification augmenté de 3 à 5 fois avec seed-kraken comparé à kraken

4 Discussions et Conclusion

- intérêt des deux méthodes aux contextes actuels de l'étude
- perspective (intérêt futur de la nouvelle approche?)

Références

- [1] 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.
- [2] Jo Handelsman. Metagenomics : Application of Genomics to Uncultured Microorganisms Metagenomics : Application of Genomics to Uncultured Microorganisms. 68(4) :669–685, 2004.

- [3] 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.
- [4] 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.