Impact de la recherche d'amorces mutées sur les résultats d'analyses métagénomiques

Aymeric Antoine-Lorquin, Frédéric Mahé, Micah Dunthorn, Catherine Belleannée

Impact de la recherche d'amorces mutées sur les résultats d'analyses métagénomiques

Aymeric Antoine-Lorquin [*] [†1], Frédéric Mahé[2], Micah Dunthorn[3], Catherine Belleannée [‡1]

[1] DYLISS (INRIA - IRISA) – INRIA, Université de Rennes 1, CNRS : UMR6074 – Campus de Beaulieu, F-35 042 Rennes Cedex, France
[2] CIRAD – Centre de coopération internationale en recherche agronomique pour le développement : UPR39 – France
[3] Technical University of Kaiserslautern (TU Kaiserslautern) – PO Box 3 049 67.653 Kaiserslautern, Allemagne

Session séquences nucléiques
mardi 28 16h50
Salle des thèses

Introduction

En métagénomique ciblée, une problématique récurrente concerne la quantité de lectures réellement exploitables en sortie d'un séquenceur à haut débit. L'une des étapes influençant cette quantité est la détection dans chaque séquence des amorces utilisées pour amplifier le gène ciblé. Huse et al. (2010) ont émis l'hypothèse que ne retenir que les séquences disposant d'amorces parfaites (non-mutées par une erreur de séquençage) permettrait d'améliorer la qualité globale d'un échantillon. Cette approche, basée sur l'idée que les séquences des amorces doivent dominer numériquement à l'issue de l'étape de PCR, présente l'avantage de ne pas nécessiter l'utilisation d'outils complexes, puisqu'il est possible de rechercher les amorces parfaites à l'aide de simples expressions régulières, naturellement prises en charge par de nombreux langages de programmation (python, perl, ruby, etc.). Néanmoins, cette stratégie élimine probablement aussi des séquences correctes. C'est pourquoi, maintenant qu'il existe des outils capables d'éliminer *a posteriori* les séquences les moins fiables (tel que SWARM de Mahé et al. (2015b)), nous avons voulu savoir si le fait de rechercher des amorces potentiellement mutées permettait d'augmenter le nombre de séquences exploitables par échantillon et si cette augmentation impactait les résultats obtenus au terme de l'analyse métagénomique.

Matériel et méthodes

Jeux de données. Nous avons travaillé sur 9 échantillons biologiques, dans le cadre d'une étude caractérisant la biodiversité des sols tropicaux chez les eucaryotes unicellulaires (Mahé et al. (2015a)). Les échantillons ont été séquencés à la fois en Roche/454 et en Illumina MiSeq, afin de pouvoir constater l'impact de la technologie sur les résultats (i.e. obtient-on les mêmes séquences avec les deux technologies ? Les résultats sont-ils les mêmes en Roche/454 et en Illumina Miseq ?). Le séquençage a ciblé la région V4 de la sous-unité 18S de l'ARN ribosomique (Stoeck et al. (2010)). En effet, cette région possède une portion hypervariable spécifique de chaque espèce et encadrée par deux segments de séquences extrêmement conservés utilisables en tant qu'amorces universelles (ci-après amorces *Forward* et *Reverse*). La détection et l'élimination des amorces dans les échantillons séquencés permet donc d'isoler la partie spécifique, appelée amplicon.

Pour l'ensemble des 9 échantillons, le séquençage Roche/454 a produit 310 375 séquences, contre 5 223 138 séquences sous Illumina MiSeq.

*. Intervenant
†. Corresponding author : aymeric.antoine-lorquin@irisa.fr
‡. Corresponding author : catherine.belleannee@irisa.fr

Recherche des amorces. La recherche des amorces parfaites s'est faite à l'aide d'expressions régulières recherchées en Python (CCAGCA[CG]C[CT]GCGGTAATTCC pour V4F et T[CT][AG]ATCAAGAACGAAAGT pour V4R) ; les séquences obtenues forment l'ensemble des *amplicons Regex* (nombre total d'amplicons Regex Roche/454 : 39 917, Illumina Miseq : 274 801). La recherche des amorces mutées parmi les séquences non regex s'est faite à l'aide de l'outil de *pattern matching* grammatical Logol (Belleannée et al. (2014)) ; les séquences obtenues forment l'ensemble des *amplicons Logol* (nombre total d'amplicons Logol Roche/454 : 2 520, Illumina Miseq : 20 558).

Calcul de la proximité de deux ensembles de séquences. Le test de dissimilarité de Bray-Curtis permet de visualiser sur un graphe la similarité de deux ensembles de séquences. Il est basé sur les comparaisons 2 à 2 des profils de séquences de deux ensembles de taille identique. Les séquences Regex étant bien plus nombreuses que les séquences Logol, la valeur finale de dissimilarité a été obtenue en faisant la moyenne de 10 000 calculs de dissimilarité de sous-échantillons aléatoires de 1 000 séquences Regex ou Logol, et ce indépendamment pour chaque échantillon biologique.

Regroupement des séquences par similarité. Une clustérisation a été faite avec l'outil SWARM (Mahé et al. (2015b)) sur la totalité des amplicons Regex et Logol et ce séparément pour chaque technologie. Chaque cluster constitué par SWARM contient des séquences proches les unes des autres à quelques substitions près. La règle de validation d'un cluster d'amplicons a été la suivante : pour être conservé, un cluster doit regrouper un minimum de 3 séquences ou au moins 2 séquences provenant de 2 échantillons différents. Chaque cluster valide forme un OTU (*Operationnal taxonomic unit*, unité taxonomique opérationnelle) souvent considéré comme représentant une espèce biologique.

Résultats

Modèle d'amorces mutées. Le modèle de mutation défini par l'expert est le suivant : les amorces Forward et Reverse peuvent posséder jusqu'à 2 substitutions OU jusqu'à 1 insertion-délétion. Par ailleurs, l'amorce Reverse peut être partiellement tronquée sur ses 2 nucléotides terminaux sans que cela compte comme une délétion (i.e. on autorise l'amorce Reverse a être légèrement incomplète, avant de considérer la présence de mutations).

Il n'est pas simple de définir de tels modèles uniquement avec les expressions régulières. En effet, les expressions régulières nécessitent de définir explicitement chaque possibilité recherchée. Par exemple, autoriser deux substitutions sans apriori de position sur un mot de taille n crorrespond à n x (n-1)/2 variants ; soit 190 mots différents pour une séquence de taille 20, donc un nombre d'expression régulière explosif en fonction de la taille du mot. C'est pour cette raison que nous avons utilisé un outil de *pattern matching* permettant d'exprimer des expressions régulières approchées, Logol, qui permet de rechercher facilement des variants de séquences, notamment en autorisant l'ajout de propriétés aux modèles, tel que le nombre d'erreurs autorisées par rapport à une référence. Ainsi, Logol peut couvrir les 190 modèles d'expressions régulières de l'exemple précédent avec un unique modèle d'expression.

Détection des séquences avec amorces mutées. La recherche des amorces parfaites permet de récupérer respectivement 90,2 % et 82,7 % des séquences totales. La recherche d'amorces mutées permet de capturer respectivement 8,3 % et 7,1 % de séquences additionnelles (pour un total respectif de 98,5 % et 89,8 %). La recherche des amorces mutées permet donc d'ajouter une quantité non-négligeable de séquences (+25 619 séquences en Roche/454 et +368 260 séquences en Illumina MiSeq).

Validation des séquences avec amorces mutées

Deux situations peuvent se présenter pour unamplicon Logol : soit il est déjà connu (typiquement, il est identique à unamplicon Regex) et donc tout autant biologiquement envisageable ; soit il est nouveau et la question se pose de savoir s'il ne s'agit pas d'un amplicon trop lourdement muté

ou chimérique, ne reflétant donc pas la réalité biologique. Pour répondre à cette interrogation, nous avons regardé si l'ensemble des amplicons Logol formait une population comparable à la population Regex, puis nous avons observé le devenir individuel de chaque amplicon.

Comparaison globale des amplicons Logol et Regex. L'utilisation du test de dissimilarité de Bray-Curtis a permis de vérifier que les amplicons Logol sont relativement similaires aux amplicons Regex. Les résultats montrent d'une part que les amplicons Logol sont très proches des amplicons Regex et d'autre part, que les amplicons Logol et Regex obtenus via la même technologie de séquençage sont plus proches entre eux qu'avec leurs homologues de la technologie de séquençage alternative pour un échantillon donné (par exemple, les amplicons Regex/Illumina sont plus proches des amplicons Logol/Illumina que des amplcions Regex/454). Les amplicons Logol sont donc comparables aux amplicons Regex, malgré leurs amorces mutées.

Caractérisation individuelle des amplicons Logol. Dans le cadre de notre analyse, chaque amplicons Logol correspond à l'une des 4 situations suivantes après la clustérisation par SWARM :

— Cas 1) L'amplicon Logol appartient à un cluster non-valide, i.e. est isolé. Il n'a alors pas d'impact sur les résultats finaux car il en sera éliminé. Les amplicons chimériques font par exemple partie de cette catégorie
— Cas 2) L'amplicon Logol s'ajoute à un cluster Regex valide. Il augmente ainsi l'abondance totale du cluster mais ne modifie pas le nombre d'OTU identifiés. Cela valide tout de même l'amplicon.
— Cas 3) L'amplicon Logol s'ajoute à un cluster Regex non-valide et le rend valide. Il permet l'identification d'un nouvel OTU.
— Cas 4) L'amplicon Logol appartient à un cluster valide purement Logol. Il permet l'identification d'un nouvel OTU.

Les deux derniers cas sont particulièrement intéressants, puisqu'ils modifient concrètement le résultat de l'analyse, en ajoutant de nouveaux OTU.

En Roche/454, 9 % des séquences Logol appartiennent à un cluster non-valide (contre 5 % pour les séquences Regex) et sont donc rejetées. La clustérisation aboutit à 4 432 OTU, dont 2 059 contiennent des séquences Logol (46,5 %). Pour 1 632 OTU, il s'agit d'une augmentation de l'abondance totale d'OTU déjà détectés avec les séquences Regex. Les 427 autres OTU sont de nouveaux OTU (dont 205 uniquement constitués de séquences Logol).

En Illumina, 24 % des séquences Logol appartiennent à un cluster non-valide (contre 20 % pour les séquences Regex) et sont donc rejetées. La clusterisation aboutit à 28 377 OTU, dont 12 693 contiennent des séquences Logol (44,7%). Pour 10 835 OTU, il s'agit d'une augmentation de l'abondance totale d'OTU déjà détectés avec les séquences Regex. Les 1 858 autres OTU sont de nouveaux OTU (dont 937 uniquement constitués de séquences Logol).

La récupération des séquences disposant d'amorces mutées a donc permis l'identification de 10 à 7 % de nouveaux OTU dans les échantillons (respectivement en 454/Roche et Illumina).

Conclusion

La recherche des amorces mutées permet d'améliorer de façon non-négligeable la sensibilité d'une analyse métagénomique en augmentant le rappel parmi les séquences d'un échantillon.

Bien sûr, cette recherche nécessite l'utilisation de moyens adaptés. Il existe un certain nombre d'outils disponibles pour mettre en œuvre la recherche des amorces mutées, tels que CutAdapt (Martin, 2011), simple et très rapide mais assez peu flexible, ou Logol, beaucoup plus lent, qui permet un contrôle complet sur les spécificités du modèle. En elle-même, cette recherche est simple à inclure dans les pipelines métagénomiques existants.

Un post-filtrage par clustérisation permet d'éliminer, parmi les nouveaux candidats, les séquences isolées et de ne conserver que les séquences similaires aux séquences à amorces parfaites. Ces séquences supplémentaires permettent d'accroître la sensibilité des analyses métagénomiques,

en permettant la détection de nouveaux OTU (+7 à +10 %, dans notre étude, en fonction de la technologie de séquençage), que ce soit en augmentant l'abondance de séquences détectées en nombre insuffisant pour être validées ou que ce soit en détectant des séquences totalement nouvelles qui n'étaient pas visibles auparavant.

Références

Belleannée, C., Sallou, O., and Nicolas, J. (2014). "Logol: Expressive Pattern Matching in Sequences. Application to Ribosomal Frameshift Modeling". *Pattern Recognition in Bioinformatics*, number 8626 in Lecture Notes in Computer Science, pp 34–47. Springer International Publishing. `https://hal.inria.fr/hal-01059506v1`

Bray, J. Roger, and J. T. Curtis. "An Ordination of the Upland Forest Communities of Southern Wisconsin". *Ecological Monographs* 27.4 (1957): 326–349. Web.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). "Ironing out the wrinkles in the rare biosphere through improved OTU clustering". *Environmental Microbiology*, 12(7):1889–1898.

Mahé, F., Mayor, J., Bunge, J., Chi, J., Siemensmeyer, T., Stoeck, T., Wahl, B., Paprotka, T., Filker, S., and Dunthorn, M. (2015a). "Comparing High-throughput Platforms for Sequencing the V4 Region of SSU-rDNA in Environmental Microbial Eukaryotic Diversity Surveys". *Journal of Eukaryotic Microbiology*, 62(3):338–345.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015b). "Swarm v2: highly-scalable and high-resolution amplicon clustering". *PeerJ*, 3:e1420.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". *EMBnet.journal*, 17(1):pp. 10–12.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., and Richards, T. A. (2010). "Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water". *Molecular Ecology*, 19 Suppl 1:21–31.

Mots clefs : pattern matching, Illumina, Roche/454, 18S ribosomal RNA