MetaMatch : un algorithme pour l'assignation taxonomique en métagénomique

Jean-Marc Frigerio (1), Philippe Chaumeil (1), Pierre Gay (2), Lenaïg Kermarrec (3,4), Frédéric Rimet (3), Agnès Bouchez (3) et Alain Franc (1)
(1) {Jean-Marc.Frigerio, Alain.Franc, Philippe.Chaumeil}@pierroton.inra.fr, INRA, UMR1202 BIOGECO, F-33610 Cestas, France & Univ. Bordeaux, BIOGECO, UMR 1202, F-33400 Talence, France
(2) Pierre.Gay@u-bordeaux1.fr, Univ. Bordeaux, MCIA, F-33400 Talence, France
(3) {Agnes.Bouchez, Frederic.Rimet@thonon.inra.fr}@thonon.inra.fr, INRA, UMR CARRTEL, 74200 Thonon-les-Bains et Université de Savoie, Chambéry, France
(4) lenaig.kermarrec@asconit.com, Asconit Consultants, 66350 Toulouges, France

Overview

Community ecology faces a new challenge as the next-generation sequencing approaches can yield data from hundreds of microbial community samples. This way, combined with accurate and reliable taxonomic assessment, yields hundreds of new data that will contribute to a better understanding of community assemblies formed under various environmental and historical conditions. Algorithms classifying sequences by comparison to a reference library are the most widely used tools for assessing community composition of environmental samples. However, as they are computationally intensive, almost all these algorithms (most standard being BLAST and similar offsprings) use heuristics designed to speed up the database exploration phase, at the cost of being less strict with the quality of the match between a query and a reference. This problem is naturally distributable, as all comparisons (query, reference) are independent. Here, we present a tool enabling comparisons between queries (say, one million reads) and reference sequences (say, several thousands), and its implementation on two infrastructures: a cluster in MCIA (Mésocentre de Calcul Intensif en Aquitaine) and a production grid EGI. We show how tracking the large number of jobs generated was nearly impossible with gLite, and how this problem could be solved using Dirac. We compare time and quality between a run on Avakas and on the grid EGI. As a perspective, we will develop a user friendly interface enabling this tool to be used routinely on the grid as a diagnostic for a user not acquainted with computing subtleties of the grid.

Enjeux scientifiques

Les techniques de séquençage ont prodigieusement évolué ces dernières années, déversant ce qu'il est convenu d'appeler des avalanches de données. Les outils de modélisation statistique et traitement qui étaient dimensionnés pour des productions bas débit sont dépassés, et il existe une forte demande à la fois de transferts vers des machines performantes (clusters, grille) et vers le développement de nouveaux algorithmes adaptés à ces nouvelles dimensions. Ici, nous nous intéressons aux comparaisons de séquences. Actuellement, le programme d'alignement local de séquences le plus largement utilisé, BLAST (Basic Local Alignment Search Tool, Altschul et al., 1990) se décline sous des dizaines de variantes développées depuis 1990, améliorant tel ou tel point de performance. La liste est trop longue pour la mentionner ici. Il effectue des comparaisons de séquences via une heuristique proche d'une distance. Ici, nous présentons *metaMatch,* un outil dans la filiation de BLAST, avec une logique légèrement différente (sans entrer dans les détails, les ancres pour alignement sont calculées dans *metaMatch* en début de processus, alors que dans BLAST, elles sont étendues en cours de processus). Une séquence est une chaîne de caractère dans un alphabet à quatre lettres. Il existe plusieurs façons de calculer une distance entre deux séquences, avec un temps en général proportionnel au produit des deux longueurs. L'enjeu est, connaissant un ensemble de reads (séquences de quelques centaines de bases produites par un séquenceur de nouvelle génération) et une librairie de référence (des séquences d'organismes connus taxonomiquement) de calculer la distance de chaque read à chaque séquence de référence, et de ne retenir pour assignation taxonomique que les plus proches si les informations qu'elles infèrent sont cohérentes et concordantes. Nous avons mis au point un algorithme d'alignement sur ancres (parties identiques maximales), de calcul de distance entre les ancres selon un alignement global, et, le plus délicat, de gestion (imparfaite ..) des extrémités. Pour 10^6 reads de longueur 5×10^3 et 2×10^3 références de longueur 1.5×10^3, cela représente de l'ordre de 1.5×10^{15} opérations élémentaires. La distribution des calculs est alors obligatoire.

Développement, utilisation des infrastructures

La parallélisation ou distribution d'outils tels que BLAST ou *metaMatch* est naturelle et, pour BLAST notamment, a été largement étudiée (Carvalho & al., 2005, Missaoui, 2009), notamment la distribution sur des plate-formes hétérogènes comme une grille de calcul (Afgan & al., 2006), avec équilibrage dynamique (dynamic-scheduler) sur grille comme pour l'alignement (Chen et Schmidt, 2005). Le portage de *metaMatch* sur la grille EGI fait partie de ce mouvement très général et naturel de transferts d'outils en biologie vers des plate-formes distribuées de calcul (Talbi et Zomaya, 2008). Dans cette même veine, la V.O. GRISBI facilite l'accès à de nombreux outils de bioinformatique (Blanchet & al., 2006).

Le développement s'est déroulé en trois phases. Une première phase a été de concevoir l'algorithme, de l'implémenter en langage C, et de le tester sur des petits jeux de données, à portée d'un ordinateur personnel type PC. Une seconde étape a été de tester son efficacité sur des jeux de données réels par distribution sur les nœuds d'un cluster. Notre choix s'est porté sur le cluster Avakas du mésocentre de calcul MCIA (3168 cœurs Intel Xeon X5675 à 3.06GHz accessibles via l'ordonnanceur TORQUE/MAUI). Cet essai a

servi de benchmark pour le scheduler d'Avakas, lors de la phase de Vérification de Service Régulier. Cette utilisation simultanée de l'ensemble des cœurs pour une seule tâche (fortement déconseillée dans le cahier des bonnes pratiques) ne pouvant être qu'exceptionnelle, la version stabilisée de cet algorithme a été écrite pour une distribution sur un milliers de CPU environ. La troisième phase a été la mise en œuvre sur la grille EGI, via la V.O. MCIA. Une premier essai a été réalisé via l'utilisation de lignes de commandes gLite. Le nombre important de jobs ainsi "lancés" sur plusieurs machines a rendu délicate l'opération de suivi de chacun, et ces essais n'ont pas abouti. Finalement, nous avons utilisé Dirac sur la V.O. France-Grille pour distribuer ces calculs sur la grille. DIRAC est à la fois une interface très conviviale de gestion de jobs lancés sur la grille et une API permettant cette même gestion de façon plus versatile encore, via des scripts python. La stratégie la plus efficace fut de lancer les 1200 jobs via un petit script utilisant l'API et de surveiller l'évolution des jobs au travers de l'interface Web. La récupération des résultats s'est faite naturellement via l'interface.

Résultat Scientifique

Afin de tester notre algorithme *metaMatch*, nous l'avons mis en œuvre en concurrence avec BLAST pour étudier un problème de bioindication de la qualité des eaux après inventaire de communautés de diatomées (Kermarrec, 2012, Kermarrec & al., 2012). Les communautés de diatomées sont un bioindicateur reconnu de la qualité des eaux : il existe un indice construit à partir d'un inventaire taxonomique qui permet de classer les rivières ou lacs selon la qualité de leurs eaux. Plusieurs échantillons ont été pyroséquencés (technologie Roche 454) pour une région d'intérêt taxonomique (SSU rDNA, Medlin et al., 1988). Trois échantillons ont été construits artificiellement en associant des cultures de diatomées, afin de vérifier si notre algorithme permettait de retrouver cet assemblage connu, et quatre échantillons naturels ont été prélevés dans différents écosystèmes. Le résultat présenté ici concerne les échantillons artificiels. Les temps de calcul de BLAST et *metaMatch* sont du même ordre de grandeur, mais fiabilité et précision sont supérieures pour *metaMatch*, qui fournit moins de faux positifs que BLAST. Sur 21 espèces dont les cultures ont été assemblées, BLAST et *metaMatch* en ont retrouvé 20, *metaMatch* a ajouté un faux positif, assez proche d'une espèce présente, alors que BLAST dans ses différentes implémentations a ajouté environ dix faux positifs, dont plusieurs éloignés taxonomiquement de l'échantillon connu. La qualité de l'inventaire avec *metaMatch* et deux choix pour BLAST (taille des ancres) est indiquée dans la figure 1. Chaque procédure pour chaque échantillon produit un inventaire (il y a eu trois procédures différentes de construction de l'assemblage avec les mêmes espèces), et *metaMatch* fournit des inventaires plus proches du réel (KNOWN) que BLAST ou UBLAST.

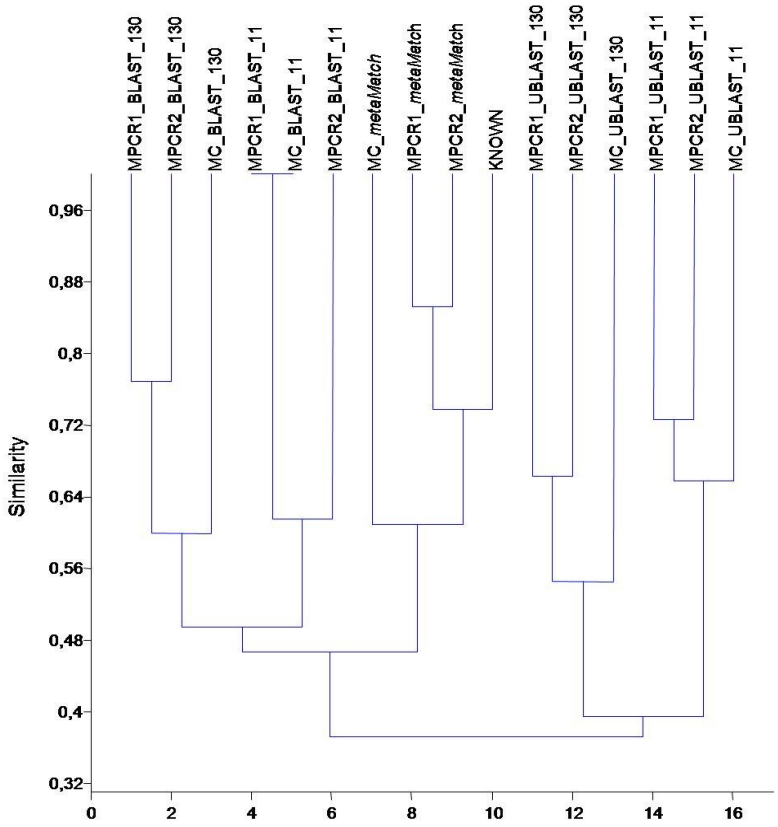

Figure 1: *Dendrogramme de classification ascendante hiérarchique sur distances de Jaccard entre 16 inventaires d'un même assemblage d'espèces de diatomées*. KNOWN est l'inventaire connu (espèces utilisées pour créer l'échantillon artificiel); Mx*_metaMatch* sont les inventaires obtenus avec *metaMatch*; Mx_BLAST_.. sont les inventaires obtenus avec BLAST; Mx_UBLAST_.. sont les inventaires obtenus avec UBLAST. Les inventaires réalisés avec *metaMatch* sont systématiquement plus proche de l'inventaire connu (LK & al., submittted).

Perspectives

La taxonomie moléculaire, bien qu'ancienne (Hillis & al., 1996) est en plein développement et à la base des études de biodiversité (Hébert & al., 2003) : il faut connaître la diversité des communautés pour analyser leur fonctionnement ou les services qu'elles peuvent rendre (production, conservation, épuration, molécules chimiques, etc …) ou comme bioindicateur de qualité des écosystèmes (comme la diversité des diatomées pour la qualité des eaux). Il existe plusieurs dizaines de millions d'espèces différentes, dont la majorité ne sont pas connues ni décrites. La taxonomie s'oriente vers la taxonomie moléculaire, et les séquenceurs de nouvelle génération permettent un accès à des données sur l'ensemble des organismes d'une communauté (métagénomique), ce qui était impossible auparavant. La comparaison avec des séquences de référence sur organismes connus est une étape essentielle pour asseoir sur des bases solides la taxonomie moléculaire. La plupart des algorithmes et méthodes, comme celle présentée ici, reviennent en fait à comparer des séquences inconnues (les reads), avec des informations bien connues dans une base de référence. Ces opération se prêtent naturellement à la distribution, et cette “stratégie” de portage sur la grille avec un souci d'accès simplifié via un portail comme Dirac a pour perspective de se diversifier fortement. On peut ainsi penser au phylotypage qui, plus que donner une distance entre séquences, permet d'insérer en un temps linéaire un read de nature inconnue sur une phylogénie connue de la base réalisée une fois pour toute (Matsen & al., 2010), et de progresser ainsi vers une utilisation fluide d'une variété d'outils en taxonomie moléculaire sur la grille de calcul EGI.

Références

Afgan, E., Sathyanarayana, P. et Bangalore, P. - 2006 - Dynamic Task Distribution in the Grid for BLAST. *IEEE International Conference on Granular Computing.* 554-557.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipmanl, D.J. - 1990 - Basic Local Alignment Search Tool, J *ournal of Molecular Biology,* **215:** 403-410.

Blanchet, C., & al. - 2006 - Grid deployment of legacy bioinformatics applications with transparent data access, in 7th IEEE/ACM International Conference on Grid Computing (GRID 2006), September 28-29, 2006, Barcelona, Spain, Proceedings.

Carvalho, P. C., Glória, R.V., de Miranda, A. B. et Degrave , W. M. - 2005 - Squid – a simple bioinformatics grid . *BMC Bioinformatics,* **6:**197 - doi:10.1186/1471-2105-6-197 .

Chen, C. et Schmidt, B. - 2005 - An adaptive grid implementation of DNA sequence alignment,
Future Generation Computer Systems, **21:** 988-1003.

Jaziri & al. - 2011 - Détermination de sondes oligonucléotidiques pour biopuces phylogénétiques en environnement grille de calcul. Communication aux Journées Scientifiques de France-Grille, Lyon, 2011.

Hebert P. D. N., Cywinska A., Ball S. L. , et deWaard J. R. - 2003 - Biological identifications through DNA barcodes. *Proc. Natl. Acad. Sci. U.S.A.* **270**:313-321.

Hillis D. M., Moritz C., et Mable B. - 1996 - *Molecular Systematics*. 2e édition. Sunderland, Massachusetts: Sinauer Associates.

Kermarrec, L. - 2012 - Apport des outils de la biologie moléculaire pour l'utilisation des diatomées comme bioindicateurs de la qualité des écosystèmes aquatiques lotiques, et pour l'étude de leur taxonomie. *Thèse de l'Université de Grenoble*.

Kermarrec, L., Chaumeil, Ph., Rimet, F., Frigerio, J.-M., Bouchez, A. & Franc, A. - xxxx – *metaMatch*, a tool for metabarcoding. Soumis.

Matsen, F. A., Kodner, R. B. et Armbrust, E. V. - 2010 – pplacer : linear time maximum likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics,* **11:**538, doi:10.1186/1471-2105-11-538

Medlin, L. K., Elwood, H. J., Stickel, S. & Sogin, M.L. - 1988 - The characterization of enzimatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene*, **71:**491–499.

Missaoui, M. – 2009 - Contributions algorithmiques à la conception de sondes pour biopuces à ADN en environnement parallèle. Thèse, Université Blaise Pascal, 2009.

Talbi, E.-G. Et Zomaya, A.Y. (Ed.) - 2008 - *Grid Computing for Bioinformatics and Computational Biology.* Wiley.