Détermination de sondes oligonucléotidiques pour biopuces phylogénétiques en environnement grille de calcul.

Faouzi Jaziri(1)*, Sébastien Cipière(2)*, Mohieddine Missaoui(3)*, Jérémie Denonfoux(4), Eric Dugat-Bony (5), Nicolas Parisot (6), Antoine Mahul (7), Sébastien Rimour (8), Eric Peyretaillade(9), Pierre Peyret (10)*, David Hill (11)*.

(1) jaziri@isima.fr, Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND, CNRS, UMR 6158, ISIMA / LIMOS, F-63173 AUBIERE.
(2) cipiere@isima.fr, Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND, CNRS, UMR 6158, ISIMA / LIMOS, F-63173 AUBIERE.
(3) missaoui@isima.fr, Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND, CNRS, UMR 6158, ISIMA / LIMOS, F-63173 AUBIERE.
(4) jeremie.denonfoux@univ-bpclermont.fr, CNRS, UMR 6023, LMGE, F-63173 AUBIERE.
(5) eric.dugat@univ-bpclermont.fr, CNRS, UMR 6023, LMGE, F-63173 AUBIERE.
(6) nicolas.parisot@insa-lyon.fr, CNRS, UMR 6023, LMGE, F-63173 AUBIERE.
(7) antoine.mahul@clermont-universite.fr, Clermont Université, Centre Régional de Ressources Informatiques (CRRI).
(8) sebastien.rimour@iut.u-clermont1.fr, CNRS, UMR 6023, LMGE, F-63173 AUBIERE.
(9) peyretai@iut.u-clermont1.fr, CNRS, UMR 6023, LMGE, F-63173 AUBIERE.
(10) pierre.peyret@univ-bpclermont.fr, CNRS, UMR 6023, LMGE, F-63173 AUBIERE.
(11) david.hill@univ-bpclermont.fr, Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND, CNRS, UMR 6158, ISIMA / LIMOS, F-63173 AUBIERE.

Overview

Microorganisms represent the largest diversity of the living beings and they are still largely unknown. Understanding the roles of microorganisms in complex environments such as soils, lakes or oceans is a great challenge in microbial ecology. Fortunately, the development of new genomic approaches allows us to understand the complexity of those systems. In this context, DNA microarrays represent high-throughput tools able to show the expression or the presence of several thousands of genes on a single experiment. However, one of the hardest steps in microarray utilization is the probe selection because of its complexity and the huge amount of data to manipulate. Obtained probes should be specific and sensitive.

To reach the goals of performance and quality, we have developed an algorithm for phylogenetic microarrays and it has been deployed on the European Grid Infrastructure (EGI). Our algorithm aims to design oligonucleotides at large scale for phylogenetic microarrays dedicated to microbial ecology. It is intended to be used massively and it allows bioinformaticians and biologists to design several thousands of probes simultaneously.

Enjeux scientifiques, besoin de la grille

Les microorganismes constituent la plus grande diversité du monde vivant grâce à leurs capacités d'adaptation exceptionnelles, mais restent encore largement méconnus. Ces micro-organismes sont des acteurs essentiels au bon fonctionnement des divers écosystèmes. La compréhension du fonctionnement des écosystèmes et les rôles joués par les microorganismes reste un enjeu majeur de l'écologie microbienne. Le développement de nouvelles approches de génomique permet d'appréhender la complexité de ces systèmes. Ainsi, les biopuces ADN (Schena et al., 1995) représentent des outils à haut débit de choix capables de suivre l'expression ou la présence de plusieurs milliers de gènes en une seule expérience. En effet, cette approche permet de contribuer à une meilleure connaissance du fonctionnement des écosystèmes et d'étudier la dynamique des communautés bactériennes dans des environnements complexes tels que le sol.

L'une des étapes les plus difficiles du développement des biopuces ADN, de par sa complexité et la quantité de données à traiter, est la sélection de sondes qui doivent être à la fois sensibles et spécifiques (Zhou et al, 2002). En effet, pour considérer une sonde, elle doit respecter des critères bien précis et être capable de

* Co premier auteur / Co directeur

détecter spécifiquement sa cible et ce avec la meilleure sensibilité possible. La conception des sondes est donc une étape cruciale dans la conception de biopuces à ADN (Kreil et al, 2006). Quelques logiciels de conception de sondes ADN ont été proposés ces dernières années pour répondre aux exigences des biopuces ADN (Lemoine et al., 2009). Cependant, la conception de sondes pour biopuces ADN en écologie microbienne doit répondre à des spécificités particulières du fait de la complexité des environnements étudiés et doit faire face à l'évolution rapide des formats des biopuces qui dépassent deux millions de sondes actuellement, sans oublier la croissance exponentielle du nombre de séquences déposées dans les bases de données internationales. Cette augmentation du nombre de séquences est clairement visible sur la figure 1 représentant la base de données de l'EMBL (European Molecular Biology Laboratory). La recherche de sondes spécifiques nécessite de faire appel à des comparaisons de séquences pour la recherche de similarité, le nombre de comparaisons étant considérable, cela reste une tâche très lourde en temps de calcul (Wu and Tseng, 2005 ; Chaudhary et al., 2005 ; Datta and Ebedes, 2005). De plus, ces besoins en termes de calcul liés à la conception de sondes ne cessent de croire. Les temps de calcul des logiciels (CommOligo (Li et al., 2005) ; ROSO (Reymond et al., 2004) ou PhylArray (Militon et al., 2007)) peuvent nécessiter plusieurs heures pour assurer la sélection des sondes pour un seul gène.

Travailler à l'échelle des génomes suppose la capacité à répondre à ces besoins en termes de performance et de qualité. Dans cette optique, nous avons développé un algorithme de conceptions de sondes pour biopuces phylogénétiques que nous avons entièrement déployé sur la grille de calcul européenne EGI (European Grid Infrastructure). En effet, nous nous sommes basés sur nos développements antérieurs (Militon et al., 2007) pour proposer des solutions permettant une meilleure sélection des sondes en terme de sensibilité et de spécificité tout en augmentant la performance du logiciel notamment avec l'utilisation de la grille de calcul pour la distribution du calcul à grande échelle.

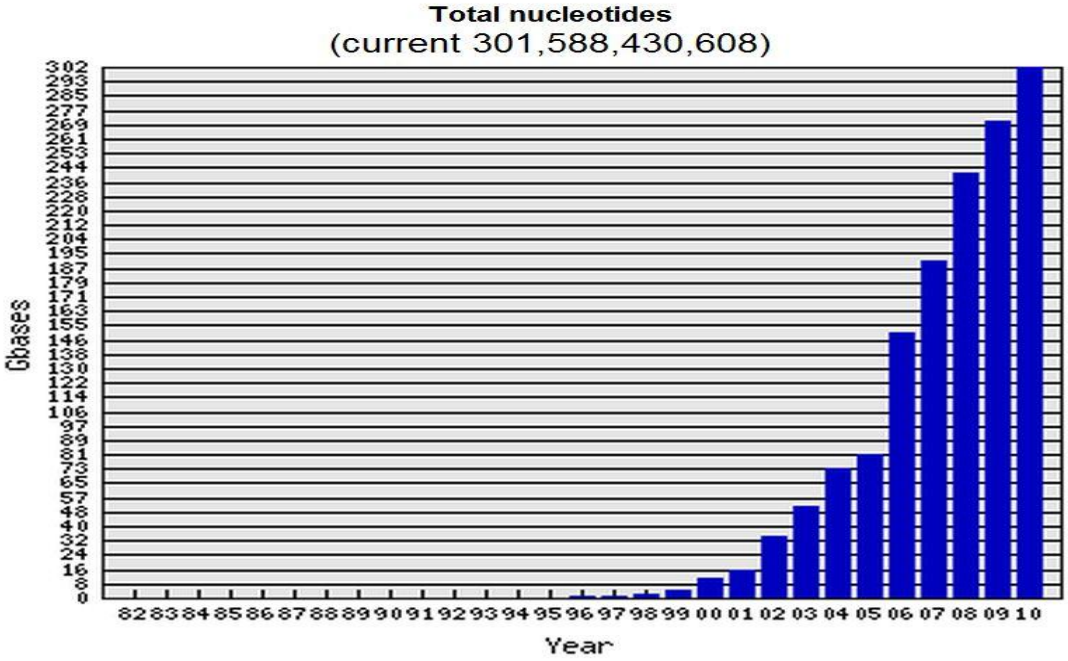

Figure 1. Nombre de nucléotides dans la base EMBL en 2010.
(Source: http://www.ebi.ac.uk/embl/Services/DBStats/)

Développements et déploiement sur la grille

Le logiciel que nous proposons est un logiciel de conception de sondes oligonucléotidiques pour biopuces phylogénétiques. Il permet de déterminer des centaines d'oligonucléotides spécifiques pour différents niveaux taxonomiques simultanément. Les oligonucléotides sélectionnés vérifient plusieurs critères de choix implémentés dans le logiciel tels que la spécificité, la taille, le taux de dégénérescence et le nombre maximum d'hybridations croisées autorisées. Notre logiciel est développé avec les langages C++ et Perl, ce dernier étant particulièrement répandu dans le secteur de la bioinformatique appliquée à la génomique. Une interface Web le rendra prochainement accessible à toute la communauté en autorisant la soumission de travaux au travers d'un portail sur Internet.

Notre approche consiste en plusieurs étapes. La première est celle de la construction d'une base de

données de séquences procaryotes et fongiques de petites sous unités de l'ARN Ribosomique (SSU – Small Sub-Units) ; 16S pour les procaryotes et 18S pour les champignons. La construction d'une base de haute qualité composée de sous groupes homogènes de séquences, a nécessité le développement d'un nouvel algorithme en utilisant des outils bioinformatiques tels que blast (Altschul et al., 1990), blasclust et clustalw (Thompson et al., 1994). La deuxième étape est basée sur la construction de séquence consensus à partir d'un alignement de séquences par CLUSTALW-MPI (Li et al., 2003).

La partie la plus coûteuse en termes de temps de calcul est celle du test de spécificité de toutes les sondes possibles pour déterminer leurs hybridations croisées potentielles en utilisant les paramètres choisis par l'utilisateur. Cette partie est entièrement déployée sur la grille de calcul européenne. Le déploiement est réalisé en se basant sur le découpage de la séquence consensus en plusieurs fragments. Le nombre de fragments est variable d'un design à un autre. Ce nombre est déterminé automatiquement en fonction d'une estimation préalable du nombre d'oligonucléotides qui peuvent être générés. Afin de profiter de la taille de la grille européenne, plusieurs conceptions peuvent être lancées de manière simultanée. Notre algorithme permet de superviser l'évolution de chaque job correspondant à une partie du calcul. Seuls les jobs terminés avec des résultats disponibles sont considérés, les autres jobs sont automatiquement resoumis à la grille.

Outils et difficultés rencontrées

Comme tout outil distribué sur grille, ses performances sont fortement dépendantes de l'état de la grille de calcul. Même avec un algorithme de gestion et de supervision des jobs, la grille pose encore des problèmes de fiabilité qui se répercutent directement sur notre logiciel. Notre logiciel reste très fiable lorsque la grille est disponible. D'après les statistiques obtenues, la re-soumission de jobs est nécessaire et la sur-soumission peut-être une technique pour éviter la perte de résultats. Il faut également noter qu'il est également nécessaire de gérer les pannes durant l'exécution de conceptions incluant l'indisponibilité de certaines ressources de la grille (que ce soit des Computing Elements ou des Storage Elements). Les problèmes peuvent également provenir du blocage d'un job au niveau d'un CE. En effet, nous avons remarqué qu'il est possible que la ressource soit affectée à un job sans que celui-ci s'exécute. Parfois des incidents au niveau du réseau peuvent survenir au sein de la grille comme par exemple la panne d'un CE ou la coupure du réseau entre deux ou plusieurs machines. Pour passer outre, une réplication de la base ou un contrôle des données transférées est effectué.

Résultats scientifiques

La conception de sondes à ADN pour biopuce phylogénétique demande une attention particulière car les séquences manipulées sont des séquences hautement conservées entre microorganismes du même genre ou de la même famille. Il est donc primordial de proposer un outil adapté à cette problématique. Nous proposons ici un algorithme permettant de concevoir des oligonucléotides à la fois spécifiques et exploratoires (découverte d'éventuelles nouvelles espèces) pour les biopuces phylogénétiques. Notre logiciel permet une meilleure sélection des sondes en termes de sensibilité et de spécificité. Il profite de la puissance de calcul offerte par la grille européenne EGI pour la parallélisation des conceptions à grande échelle. Il nous a permis de concevoir des oligonucléotides spécifiques et de haute qualité pour un nombre total de 2100 genres procaryotes et 1441 genres fongiques.

Pour tester la performance de notre approche en termes de temps de calcul, nous avons lancé des calculs pour tous les genres fongiques obtenus après la création de la base ARN ribosomique. Les calculs ont été lancés par paquet de 20 conceptions à la fois. Seuls les jobs terminés avec succès dont les résultats finaux ont été obtenus sont comptabilisés dans le nombre de jobs terminés avec succès. Nous avons estimé le temps de calcul total pour les champignons à un peu plus d'un an-CPU en prenant une durée moyenne de 6 heures et demi par conception. Après 7 semaines de calcul sur la grille européenne, nous avons obtenu les résultats pour 1356 genres des champignons ce qui représente environ 95% des données. Les 5% restants ont été resoumis dans un deuxième temps.

Bilan et perspectives

Nous présentons dans ce papier une approche de détermination de sondes oligonucléotidiques pour biopuces phylogénétiques procaryotes et fongiques. Notre approche a été entièrement déployée sur la grille de calcul européenne. L'algorithme développé permet de concevoir simultanément des milliers de sondes spécifiques et exploratoires pour l'étude de la présence et/ou l'évolution des procaryotes et champignons dans

différents milieux complexes.

Le but est de créer une plateforme accessible depuis internet. Cette interface permettra à un utilisateur d'utiliser l'ensemble des logiciels de bioinformatique développés ces dernières années par les différentes équipes du LMGE et du LIMOS. Notre objectif est alors de rendre ce service le plus rapide et le plus complet possible. Les traitements seront soient envoyés sur nos clusters ou pour les calculs plus importants sur la grille européenne EGI. Les résultats ainsi obtenus seront accessibles au travers de la même interface web, après l'envoi d'un mail de confirmation d'exécution correcte. Les outils que nous développons permettront à des utilisateurs ne possédant pas de certificat (qui est la clef qui ouvre les portes de la grille) de pouvoir eux aussi profiter des formidables ressources de calcul qu'offre la grille au monde de la recherche et à la connaissance humaine en général. La seule contrepartie sera une inscription gratuite dans notre base de données d'utilisateurs pour que nous puissions générer un couple identifiant et mot de passe, garantissant la confidentialité des données et de leurs traitements.

Références

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

Chaudhary, V., Liu, F., Matta, V. and Yang, L.T. (2005) Parallel Implementations of Local Sequence Alignment: Hardware and Software. In Sons, J.W. (ed), *Parallel Computing for Bioinformatics and Computational Biology*.

Datta, A. and Ebedes, J. (2005) Multiple Sequence Alignment in Parallel on a Cluster of Workstations. In Sons, J.W. (ed), *Parallel Computing for Bioinformatics and Computational Biology*.

Kreil, D.P., Russell, R.R. and Russell, S. (2006) Microarray oligonucleotide probes, *Methods in enzymology*, **410**, 73-98.

Lemoine, S., Combes, F. and Crom, S.L. (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge, *Nucleic Acids Research*, **37**, 1726–1739.

Li, K.-B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing, Bioinformatics, **19**, 1585-1586.

Li, X., He, Z. and Zhou, J. (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation, *Nucleic Acids Research*, **33**, 6114-6123.

Militon, C., Rimour, S., Missaoui, M., Biderre, C., Barra, V., Hill, D., Mone, A., Gagne, G., Meier, H., Peyretaillade, E. and Peyret, P. (2007) PhylArray: phylogenetic probe design algorithm for microarray, *Bioinformatics*, **23**, 2550-2557.

Reymond, N., Duret, H.C.L., Calevro, F., Beslon, G. and Fayard, J.-M. (2004) ROSO: optimizing oligonucleotide probes for microarrays, *Bioinformatics*, **20**, 271-273.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Research, **22**, 4673–4680.

Wu, X. and Tseng, C.-W. (2005) Searching Sequence Databases Using High Performance BLASTs In Sons, J.W. (ed), *Parallel Computing for Bioinformatics and Computational Biology*.

Zhou, J. and Thompson, D.K. (2002) Challenges in applying microarrays to environmental studies *Current Opinion in Biotechnology*, **13**, 204-207.