BCCM GEN-ERA
BCCM collections in the genomic era

Contract - B2/191/P2/BCCM GEN-ERA

RESUME

Contexte

Le consortium BCCM (Belgian Coordinated Collections of Microorganisms) est un programme du gouvernement fédéral belge et de la Politique scientifique fédérale (Belspo). BCCM est un centre de ressources biologiques (CRB) qui préserve et fournit du matériel microbiologique et génétique pour soutenir les sciences de la vie et le secteur biotechnologique dans le domaine de la recherche fondamentale et appliquée.

Les microorganismes procaryotes et eucaryotes représentent la plupart de la biodiversité présente sur Terre et se retrouvent dans virtuellement tous les environnements. Ils jouent un rôle majeur dans de très nombreuses fonctions, depuis les écosystèmes jusqu'à l'être humain, et contribuent à d'innombrables applications. Cependant, la connaissance actuelle en microbiologie représente à peine la pointe de l'iceberg et la recherche microbiologique doit encore découvrir et comprendre les multiples capacités microbiennes encore cachées. Les CRBs occupent une position essentielle dans cette tâche en isolant, cultivant, identifiant, préservant et distribuant la diversité qui peut être mise en culture.

Depuis sa création en 1983, BCCM a développé et maintenu un leadership au sein des BRCs européens à travers l'implémentation entre autres d'un site web, une certification ISO 9001, un catalogue en ligne, un système de gestion des données de laboratoire, et sa reconnaissance en tant qu'autorité dépositaire internationale. L'étude, la valorisation et la documentation des ressources microbiennes nécessitent également de rester au fait des derniers développements en microbiologie et des technologies modernes afin de pouvoir analyser les microorganismes de manière efficace. De nos jours, le recherche en microbiologie est largement facilitée par les nouvelles techniques de génomique, y compris le séquençage des génomes entiers. Ce dernier fournit l'entièreté de l'information génétique d'un organisme et est de plus en plus requis dans de nombreuses disciplines. Acquérir les connaissances en génomique est donc primordial pour BCCM afin de rester un CRB majeur, pour de futures collaborations nationales et internationales, ainsi que pour répondre aux questions de recherches à venir.

Objectifs

Le premier objectif du projet BCCM GEN-ERA était d'implémenter l'expertise en génomique au sein des collections BCCM, pour lesquelles le véritable challenge était la manipulation et l'analyse des données génomiques (big data). Cela a nécessité l'installation de structures bio-informatiques spécifiques et de softwares pour lesquels les scientifiques de BCCM ont dû être formé afin d'assurer une implémentation à long terme. Lors de ce travail, l'accent a été mis sur le séquençage des génomes entiers car leur détermination contribue grandement à l'expertise des collections dans le domaine de la taxonomie et de la phylogénie, tout en permettant de potentielles analyses fonctionnelles. En outre, mettre à disposition des souches avec un génome séquencé est nécessaire pour rencontrer les besoins des utilisateurs des CRBs et constitue donc une valeur ajoutée importante pour la visibilité et l'attractivité de BCCM.

Le projet BCCM GEN-ERA visait également à répondre à des questions de recherche spécifiques couvrant la biodiversité microbienne des collections BCCM (i.e., bactéries, mycobactéries, cyanobactéries, levures et moisissures) et en particulier sur des microorganismes ayant un impact sociétal (i.e., associés à des insectes pollinisateurs, agents pathogènes, produisant des composés bioactifs). Le projet a impliqué cinq des sept

collections BCCM, en collaboration avec le laboratoire de Phylogénomique des Eucaryotes de l'Université de Liège, ce dernier apportant l'expertise en bio-informatique et (phylo)génomique.

Conclusions

Deux infrastructures bio-informatiques différentes ont été envisages pour la manipulation et l'analyse des données de séquençage, Galaxy et Nextflow. Celles-ci ont été testées et comparées pour leur performance, pertinence, facilité d'utilisation et conformité aux principes FAIR (i.e., Findable, Accessible, Interoperable, Reusable).

Galaxy est une plateforme bio-informatique en ligne dont l'objectif est notamment de mettre des analyses génomiques à disposition de tous les chercheurs, y compris ceux sans grande compétences informatiques, grâce à un interface graphique facile d'utilisation. L'installation d'un site GALAXY « BCCM » a ainsi été testée mais a rencontré des problèmes de sécurité qui ont compliqué son déploiement. Un administrateur système s'est donc révélé nécessaire pour maintenir l'infrastructure. Ces difficultés ont également limité l'interopérabilité et la réutilisation des outils bio-informatiques. De plus, certains programmes requis dans les analyses génomiques modernes n'étaient pas disponibles. Pour ces raisons, Nextflow, conçu pour réaliser des analyses bio-informatiques en utilisant des lignes de commande, a finalement été choisi. Au total, 14 workflows Nextflow, supportés par 11 containers Singularity ont été implémentés. Ils couvrent les besoins courants des collections telles que BCCM en termes de génomique, en particulier pour la taxonomie et la modélisation des métabolismes microbiens. Ils peuvent être utilisés sur des procaryotes et des petits eucaryotes d'une manière totalement reproductible. Les workflows sont fournis sous la forme de programmes qui peuvent être lancés avec une seule ligne de commande, assurant ainsi la reproductibilité des analyses. Cette « boîte à outils GEN-ERA » est accessible librement à partir du GitHub https://github.com/Lcornet/GENERA qui inclut en outre une documentation détaillée à destination des utilisateurs. Nextflow répond ainsi aux critères d'un usage à long terme et FAIR de l'infrastructure bio-informatique à BCCM. Le seul inconvénient, comparé à Galaxy, est la convivialité d'utilisation. Travailler avec des lignes de commandes est en effet moins intuitif et a nécessité des formations spécifiques, mais a pu être réduit à un minimum grâce aux containers Singularity.

L'infrastructure Nextflow mise en place à BCCM, en collaboration avec l'Université de Liège, a été utilisée pour investiguer des questions de recherche pouvant être résolues grâce à des analyses génomiques. Les microorganismes suivants ont notamment été étudiés : des pathogènes fongiques causant des infections cutanées, des mycobactéries infectieuses, des bactéries et des levures isolées du tube digestif d'insectes pollinisateurs (abeilles et bourdons) ainsi que des cyanobactéries montrant des activités biologiques. Ces analyses ont permis des percées dans leur domaine respectif et ont ouvert de nouvelles perspectives pour des recherches futures.

Le projet BCCM GEN-ERA a établi une expertise génomique au sein de BCCM en mettant en place une structure bio-informatique, en fournissant des outils d'analyses génomiques et en développant les compétences en génomique de ses scientifiques. Ces investissements ont été réalisés en visant une implémentation à long terme afin que la génomique devienne une activité continue des collections BCCM. Dans ce cadre, une étape majeure a été franchie avec la création du GitHub BCCM GEN-ERA qui peut être considéré comme un portail pour l'utilisation, la réutilisation et l'apprentissage des outils bio-informatiques d'analyse génomique. Il a été élaboré pour offrir un accès libre aux programmes d'analyse de données génomique provenant de microorganismes procaryotes et eucaryotes. Il centralise des logiciels pour l'assemblage et l'annotation des génomes, la phylogénomique ou la modélisation métabolique des génomes. Il fonctionne sous le forme d'une plateforme en ligne commune qui peut être utilisée par toutes les collections et les scientifiques de BCCM ainsi que par d'autres institutions (e.g. CRBs, laboratoires de microbiologie) intéressées par des thématiques similaires.

Mots-clés
Génomique; BCCM; microorganismes; génomes; collections; bio-informatique; taxonomie; phylogénie; évolution moléculaire; biodiversité.