E-Biothon : Une plate-forme pour accélérer les recherches en biologie, santé et environnement

Nicolas Bard, Sylvie Boin, François Bothorel, Philippe Chaumeil, Philippe Collinet, Michel Daydé, Benjamin Depardon, Frédéric Desprez, Marie Flé, Alain Franc, et al.

E-Biothon : Une plate-forme pour accélérer les recherches en biologie, santé et environnement

N. Bard (1), S. Boin (2), F. Bothorel (2), P. Chaumeil (9), P. Collinet (3), M. Daydé (4), B. Depardon(5), F. Desprez (6), M. Flé (3), A. Franc (9), J.-M. Frigerio (9), O. Gascuel (7), S. Guindon (7,10), J.-F. Gibrat (7), D. Girou (3), P.-F. Lavallée (3), V. Lefort, (8), G. Lesage (2), M. Rugeri (3), É. Ruinet (3), C. Séguin (5), S. Thérond (3)

(1) CNRS LIP UMR 5668, ENS Lyon, 46 Allée d'Italie, Lyon, France, Nicolas.Bard@ens-lyon.fr

(2) IBM France, Gilles.Lesage@fr.ibm.com

(3) IDRIS-CNRS, Rue John von Neumann, B.P. 167, Orsay, prenom.nom@idris.fr

(4) Université de Toulouse / INPT - CNRS / IRIT, Michel.Dayde@cnrs-dir.fr

(5) SysFera, 13 av. Albert Einstein, Villeurbanne, France, {Cyril.Seguin, Benjamin.Depardon}@sysfera.com

(6) INRIA LIP UMR 5668, ENS Lyon, 46 Allée d'Italie, Lyon, France, Frederic.Desprez@inria.fr

(7) Mathématique Informatique et Génome, INRA, Jean-Francois.Gibrat@jouy.inra.fr

(8) Univ Montpellier 2, CNRS, UMR 5506, LIRMM, Montpellier, France, Vincent.Lefort@lirmm.fr

(9) UMR BioGeCo, INRA & Université Bordeaux, Alain.Franc@pierroton.inra.fr

(10) Department of Statistics, University of Auckland, New Zealand.

Overview

Thanks to the availability of computational grids, Clouds, and their middleware, a seamless access to computation and storage resources is provided to scientists and application developers. Following the success of the Décrypthon project, the E-Biothon project (between CNRS, IBM, Institut Français de Bioinformatique, INRIA, and SysFera) is one example of such a high performance platform. The E-Biothon Cloud aims at promoting access to computational resources to foster the development of large-scale research programs in the field of biology, health, and environmental projects.

In this paper, we present the architecture of the platform, the middleware developed to facilitate access to the Blue Gene/P machines managed and hosted by IDRIS, and the first applications ported on it.

Enjeux scientifiques, besoin en calcul, stockage et visualisation

L'outil informatique (simulation, Big Data, …) est au cœur de la plupart des avancées de la recherche actuelle. Plusieurs acteurs majeurs dans leur domaine, le CNRS, IBM, l'Institut Français de Bioinformatique, INRIA et SysFera se sont alliés pour offrir à l'ensemble de la communauté scientifique la plate-forme expérimentale E-Biothon afin de mettre au point les logiciels et applications permettant d'accélérer les recherches en biologie (génomique, protéomique, …), en santé (épidémiologie, compréhension des maladies génétiques, …) et en écologie-environnement (biodiversité, …).

La France a toujours été à la pointe dans le domaine de la recherche médicale, notamment concernant les grandes pathologies de notre temps (SIDA, cancer ou encore diabète). Il apparaît de plus en plus clairement que les différentes analyses « omiques » effectuées sur les pathogènes et leurs hôtes sont d'une aide précieuse dans la découverte de nouveaux traitements. Les avancées technologiques récentes, telles que les séquenceurs haut-débit, permettent aux chercheurs des sciences de la vie d'avoir accès à des quantités gigantesques de données brutes (des pétaoctets de données sont générés par an) sur les propriétés biologiques des organismes d'intérêt, virus, bactéries, espèce humaine, etc. Analyser ces données brutes pour en extraire des connaissances biologiques est une tâche ardue qui nécessite des traitements informatiques lourds et coûteux au travers d'applications bien conçues.

En 2001, lors du médiatique Téléthon, une première opération « Décrypthon » avait permis la réalisation d'une base de données permettant de lister l'ensemble des protéines du monde vivant. L'objectif était de cartographier le protéome. Entre décembre 2001 et mai 2002, 75 000 internautes ont permis la comparaison de 560 000 protéines connues provenant de diverses espèces. En 2004, fruit d'une collaboration tripartite de l'AFM (Association Française contre les Myopathies), du CNRS et d'IBM, le programme Décrypthon a été lancé avec des objectifs à plus long terme. Les projets sélectionnés par l'AFM avaient tous des problématiques liées aux objectifs scientifiques de l'AFM: guérir les maladies neuromusculaires et les maladies rares, pour la plupart génétiques. Six universités partenaires (Bordeaux I, Lille I, l'ENS de Lyon, Paris IV, Orsay et l'UPMC) où des supercalculateurs avaient été installés par IBM, ont formé une grille de calcul reliée par le réseau RENATER et gérée par le logiciel SysFera-DS de SysFera [1]. Le CRIHAN participait également au programme, en hébergeant les données volumineuses des projets scientifiques. Le programme Décrypthon avait par ailleurs soumis l'application MaxDO à la grille d'internautes (Desktop grid) World Community Grid, sous le nom de Help Cure Muscular Dystrophy (HCMD) [2,3]. Cette application nécessitait en effet, par opposition aux autres applications mises en avant, un temps de calcul bien supérieur à ce qu'une grille dédiée pouvait alors fournir.

Afin d'accélérer drastiquement la découverte de nouveaux traitements de rupture ou de mieux comprendre notre environnement, IBM, le CNRS (via l'IDRIS et l'Institut des Grilles et du Cloud avec le soutien de France Grilles), l'Institut Français de Bioinformatique, INRIA et SysFera se sont associés pour mettre à disposition des chercheurs la plate-forme de

Cloud E-Biothon fournissant le portail applicatif et la puissance de calcul permettant d'aborder le traitement des données complexes de la biologie d'aujourd'hui et de mettre au point les logiciels applicatifs de demain.

La plate-forme E-Biothon propose aux scientifiques une solution unique et indispensable pour réaliser leurs analyses de phylogénie, d'épidémiologie ou encore de génomique. Dans cet article, nous présentons la plate-forme E-Biothon et les premières applications portées sur celle-ci.

Plate-forme de calcul et logiciel de gestion du Cloud

La plate-forme E-Biothon est déployée à l'IDRIS (Institut du développement et des ressources en informatique scientifique), un des trois grands centres de calcul nationaux.

La puissance de calcul allouée au projet E-Biothon consiste en deux racks de Blue Gene/P. Ces deux racks offrent une puissance en crête de 28 téraflops et chaque rack compte mille vingt quatre nœuds, de quatre cœurs chacun. Chaque nœud a une quantité de mémoire RAM partagée de 2 gigaoctets (ainsi nous avons un peu moins de 500 mégaoctets de mémoire par cœur, quand on prend en compte la mémoire occupée par le système d'exploitation). La Blue Gene est aussi dotée d'une capacité de stockage de 200 téraoctets. Pour accéder à cette puissance de calcul, un serveur dédié est utilisé comme machine frontale. Les utilisateurs en ligne de commande peuvent soumettre des jobs via le gestionnaire local de ressources, LoadLeveler.

Portail Web

Le projet E-Biothon requiert également un serveur Web, pour héberger le portail web qui permet aux utilisateurs d'avoir accès à une interface graphique pour gérer leurs simulations. Nous avons décidé d'utiliser une machine virtuelle sur le « CNRS Cloud Recherche ». Il s'agit pour l'instant d'une machine virtuelle légère avec simplement deux processeurs virtuels, 4 gigaoctets de mémoire vive et un disque dur virtuel de 40 gigaoctets.

La solution SysFera-DS [1] [4] offre ce portail web d'accès aux applications et, in fine, aux ressources de calcul. À travers ce portail, les chercheurs ont accès à tout un environnement de travail leur permettant d'exécuter simplement les traitements informatiques en lien avec les analyses « omiques » à réaliser, puis de gérer les données générées, tout cela à partir d'un simple navigateur web, sans installation locale et avec une sécurité des données garantie. Ainsi, ils peuvent interagir avec une seule interface conviviale plutôt qu'avec des gestionnaires de ressources de Calcul Haute Performance. SysFera-DS s'occupe ainsi d'abstraire et de rendre transparente l'infrastructure de calcul aux utilisateurs afin de leur permettre de se concentrer sur leurs recherches. Les différents acteurs du projet E-Biothon tirent partie de cette solution :

- Pour les administrateurs : cela permet d'abstraire l'infrastructure et de simplifier son utilisation pour les utilisateurs, ils disposent également d'une interface de suivi de l'utilisation pour suivre la consommation en heures de calcul des différents projets et des différents utilisateurs grâce au module de statistiques intégré.
- Pour les chefs de projet : il est possible de définir des projets au sein du portail et ainsi de faire collaborer des chercheurs autour d'applications communes. Au sein de ces projets, les responsables peuvent définir des rôles pour les différents utilisateurs, gérer les applications qui sont rendues disponibles et suivre la consommation des heures de calcul et ainsi gérer au mieux leurs quotas d'heures allouées.
- Pour les chercheurs : c'est l'accès à toute la puissance de calcul de l'infrastructure E-Biothon en quelques clics. Chaque application expose uniquement ses paramètres propres, permettant ainsi d'exécuter simplement de nombreuses simulations. Les fichiers d'entrée et de résultats peuvent être gérés à travers le portail, aussi simplement que des fichiers locaux.

Applications portées sur la plate-forme

Nous décrivons maintenant les trois applications pilotes qui ont été sélectionnées dans la première partie du projet.

L'application en Barcoding développée par Alain Franc de l'INRA à Bordeaux (en lien avec un consortium associant Thonon, Jouy et l'université de Clermont) permet d'étudier la diversité biologique et la structure des différentes communautés d'organismes. Elle repose sur un outil d'inventaire rapide de communautés de microorganismes à partir de données de séquençages issues d'outils de nouvelle génération (Roche, Illumina). À partir du séquençage massif d'un échantillon environnemental (eau, sol, etc.), tous organismes confondus, les outils déployés sur le E-Biothon permettront de comparer ces données aux bases disponibles, de dresser un inventaire taxonomique et de proposer des indices de diversité de la communauté prélevée.

L'application de phylogénie développée au sein du LIRMM Montpellier par Olivier Gascuel, Stéphane Guindon et Vincent Lefort, permet de réaliser des analyses phylogénétiques dont l'objectif est de reconstruire l'histoire évolutive d'objets biologiques divers, depuis les molécules du vivant jusqu'aux espèces et populations. Ces analyses sont aujourd'hui incontournables dans tous les domaines liés à la biologie. Leurs applications sont très variées, par exemple en génomique

[1] http://www.sysfera.fr/sysfera-ds

pour prédire la fonction des gènes et identifier des cibles thérapeutiques; en médecine pour retracer l'origine des épidémies et prévenir les nouvelles pandémies [7], en écologie pour faire l'inventaire de la biodiversité et préserver l'environnement. PhyML [6] permet de réaliser des analyses phylogénétiques utilisant des modèles d'évolution probabilistes en un temps de calcul raisonnable, ce qui n'était pas possible avant sa publication. Il implémente des algorithmes très rapides qui permettent de traiter avec une grande précision (critère statistique du maximum de vraisemblance) des données inexploitables auparavant avec ce type d'approche. La publication princeps de PhyML (*Systematic Biology* 2003, [6]) est la plus citée au monde en environnement-écologie depuis 2007, avec plus de 6000 citations dans des revues scientifiques (cf. *Science Watch* et *Web of Science*) [2] [3].

Autour du calcul et de la visualisation de relations de synténie dans des génomes bactériens, l'unité Mathématique, Informatique et Génome de l'INRA à Jouy-en-Josas a développé un pipeline permettant de calculer des relations d'homologie entre des génomes bactériens séquencés et de les stocker dans une base de données relationnelle. Elle a aussi développé un outil Web se connectant à la base pour en extraire ces relations et permettre ainsi aux biologistes de les visualiser et de les utiliser pour compléter l'annotation des génomes. Le calcul des relations de synténie nécessite la comparaison deux à deux (avec BLAST [5]) de toutes les séquences protéiques des plus de 2000 génomes bactériens complets (soit environ sept millions) et le calcul de la conservation de l'ordre des gènes orthologues le long des génomes par une technique de programmation dynamique.

Conclusion et perspectives

Démarré sur la base et sur l'expérience acquise pendant le projet Décrypthon, cette nouvelle plate-forme se veut ouverte à de nombreuses applications pouvant être portées sur les Blue Gene/P. Grâce à un portail web simple et dédié aux applications et à la gestion de la plate-forme, la transparence et la simplicité d'utilisation permettent aux scientifiques d'autres disciplines d'avoir accès à leurs applications de manière performante, au sein d'un Cloud, et en mode SaaS [4].

Après cette phase initiale de déploiement, l'objectif est d'ouvrir cette plate-forme à l'ensemble de la communauté scientifique des sciences de la vie à compter de 2014. IBM assure la maintenance des Blue Gene/P et le support matériel autour de ces équipements, le CNRS met à disposition, héberge et administre les calculateurs à l'IDRIS et est en charge, avec INRIA, du support utilisateurs. Enfin, le portail d'accès aux ressources déployé est la solution SysFera-DS développée par SysFera.

Références

[1] Bard, N., Bolze, R., Caron, E., Desprez, F., Heymann, M., Friedrich, A., Moulinier, L., Nguyen, N.-H., Poch, O., Toursel, T., **Décrypthon Grid - Grid Resources Dedicated to Neuromuscular Disorders**, The 8th HealthGrid conference proceedings, June, 2010.

[2] Bard, N., Bertis, V., Bolze, R., Desprez, F., **A Volunteer Computing Platform Experience for Neuromuscular Diseases Problems, Desktop Grid Computing**, Cerin, C., Feday, G. Eds, Chapman & Hall/CRC Numerical Analysis and Scientific Computation Series, Chapman & Hall/CRC, pp. 125-146, May, 2012.

[3] Bertis, V., Bolze, R., Desprez, F., Reed, K., **From Dedicated Grid to Volunteer Grid: Large Scale Execution of a Bioinformatics Application**, Journal of Grid Computing, Vol. 7, N. 4, pp. 463-478, Dec. 2009.

[4] Depardon, B., Kortas, S., Daix, B. and Barate, R., **SysFera-DS : Un portail d'accès unifié aux ressources des centres de calcul**. Mise en application à EDF R&D. In journées scientifiques mésocentres et France Grilles 2012, Paris, France, October 2012.

[5] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., **Basic Local Alignment Search Tool,** J. Mol. Biol., 215: 403-10, 1990.

[6] Guindon, S., Gascuel, O., **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**, Systematic Biology, Oct;52(5):696-704, 2003.

[7] Jung M., Leye N., Vidal N., Fargette D., Diop H., Toure Kane C., Gascuel O., Peeters M., **The origin and evolutionary history of HIV-1 subtype C in Senegal**. PLoS One, 7(3):e33579, 2012.

[8] Kermarrec, L., Franc, A., Rimet, F., Chaumeil, Ph., Humbert J.-F. & Bouchez, A. **Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Molecular Ecology Resources**, 13(4):607-619, 2013.

---

2 http://www2.cnrs.fr/en/454.htm

3 http://www.esi-topics.com/fbp/2005/february05-Guindon-Gascuel.html

4 Software as a Service.