Le portail g-INFO pour surveiller la grippe Influenza A

Trung-Tung Doan, Hong Quang Nguyen, Quang Minh Dao, Duc Hung Le, Trong Hieu Vu, Vincent Breton, Yannick Legre

Le portail g-INFO pour surveiller la grippe Influenza A

T. Tung DOAN (1), T. Hieu VU (2), D. Hung LE, Q. Minh DAO (3), H. Quang NGUYEN (1), Vincent BRETON (4), Yannick LEGRE (5)

(1) {dttung, nhquang}@ifi.hut.edu.vn, Institut de la Francophonie pour l'Informatique – 42 Ta Quang Buu, Hanoi, Vietnam
(2) vuhieu@ioit.ac.vn, Institute of Information Technology, Vietnamese Academic of Science and Technology, 18 Hoang Quoc Viet, Hanoi
(3) {ldhung,dqminh}@hpcc.hut.edu.vn, High Performance Computing Center, Hanoi University of Technology, 1 Dai Co Viet, Hanoi, Vietnam
(4) breton@clermont.in2p3.fr, Laboratoire de Physique Corpusculaire, CNRS/IN2P3, 24 avenue des Landais, BP 10448, F-63000 Clermont-Ferrand, France
(5) yannick.legre@healthgrid.org, HealthGrid association, 36 rue Charles de Montesquieu, 63430 Pont-du-Château, France

Overview

In this paper, we present the g-INFO (Grid-based International Network for Flu Observation) project that aims at running and connecting various bioinformatics programs, recognized for their accuracy and speed, to continuously reconstruct a robust phylogenetic tree from a set of sequences publicly available and daily updated for the sake of molecular epidemiology. We implemented a dynamic bioinformatics workflow in g-INFO so that an expert can choose which components he wants as well as the order of the components in the workflow for their specific analysis. Workflows are deployed on grid resources to take advantage of its high security, heterogeneity and large-scale computation. Finally, a portal was developed on the top of g-INFO to help user without previous knowledge of grid to build and run bioinformatics workflows easily.

Enjeux scientifiques, besoin de la grille

Les maladies émergentes sont caractérisées par le fait qu'elles sont par définition peu connues. Il est donc essentiel de pouvoir accumuler rapidement un ensemble d'informations biologiques, épidémiologiques et géographiques sur les foyers identifiés. Ces informations sont souvent très éparpillées géographiquement mais elles doivent être diffusées le plus rapidement possible au sein de la communauté internationale pour alimenter les travaux des chercheurs et pour permettre aux États de prendre des mesures concertées de prévention et de surveillance. Depuis 10 ans, la technologie des grilles informatiques a permis le développement de véritables infrastructures distribuées fournissant de vastes ressources de calculs et de stockage aux communautés scientifiques. Plus récemment, elle met aussi à la disposition de ses utilisateurs des outils de gestion de données distribuées permettant la création de véritables fédérations de bases de données d'accès sécurisé dont les informations peuvent être exploitées à des fins d'analyses statistiques. Grâce à ces progrès récents, les grilles constituent aujourd'hui des infrastructures très prometteuses pour la mise en place de systèmes de surveillance épidémiologique à l'échelle internationale.

Développements, déploiement sur la grille

g-INFO a été mis en œuvre et déployé sur l'infrastructure de grille EGI (European Grid Infrastructure) [1]. L'infrastructure de EGEE est basée sur gLite - un intergiciel de grille [2]. Outre gLite, un déploiement à grande échelle de pipeline phylogénétique nécessite l'utilisation d'un environnement pour la soumission des tâches et la collection de données: l'environnement de production WISDOM (WISDOM Production Environment) [3]. Le WPE est composé de 4 éléments principaux :

- Le Task Manager interagit avec le client et stocke les tâches à accomplir ;
- Le Job Manager soumet les jobs aux éléments de calcul (CE), où les tâches gérées par le Task Manager seront exécutées;
- Le Data Manager interagit avec le client pour gérer les données en mode batch;
- Le système d'information WISDOM utilise AMGA [4] (Application ARDA grille de métadonnées) pour stocker toutes les méta-données nécessaires pour le Data Manager et le Job Manager.

Le Task Manager de g-INFO contient 4 services généraux pour l'analyse phylogénétique:

- BLAST pour chercher des régions de similarité entre séquences [5],
- Muscle pour aligner les séquences [6],
- Gblocks pour corriger les séquences alignées [7]
- et PhyML pour construire les arbres phylogénétiques [8].

Grâce à ces services disponibles dans le WPE, nous pouvons mettre en œuvre des pipelines phylogénétiques statiques. Toutefois, pour implémenter des workflows dans g-INFO, nous utilisons le moteur de workflow MOTEUR [9] développé par les laboratoires I3S et CREATIS. MOTEUR est interfacé avec la grille utilisant gLite et gère des services applicatifs en mode asynchrone. Le portail g-INFO a été développé avec plusieurs technologies web. Ce portail interagit avec le système g-INFO via des services web et aide l'utilisateur à créer des modèles de workflow, à exécuter un workflow existant et à visualiser les résultats avec une interface conviviale sur un navigateur web.

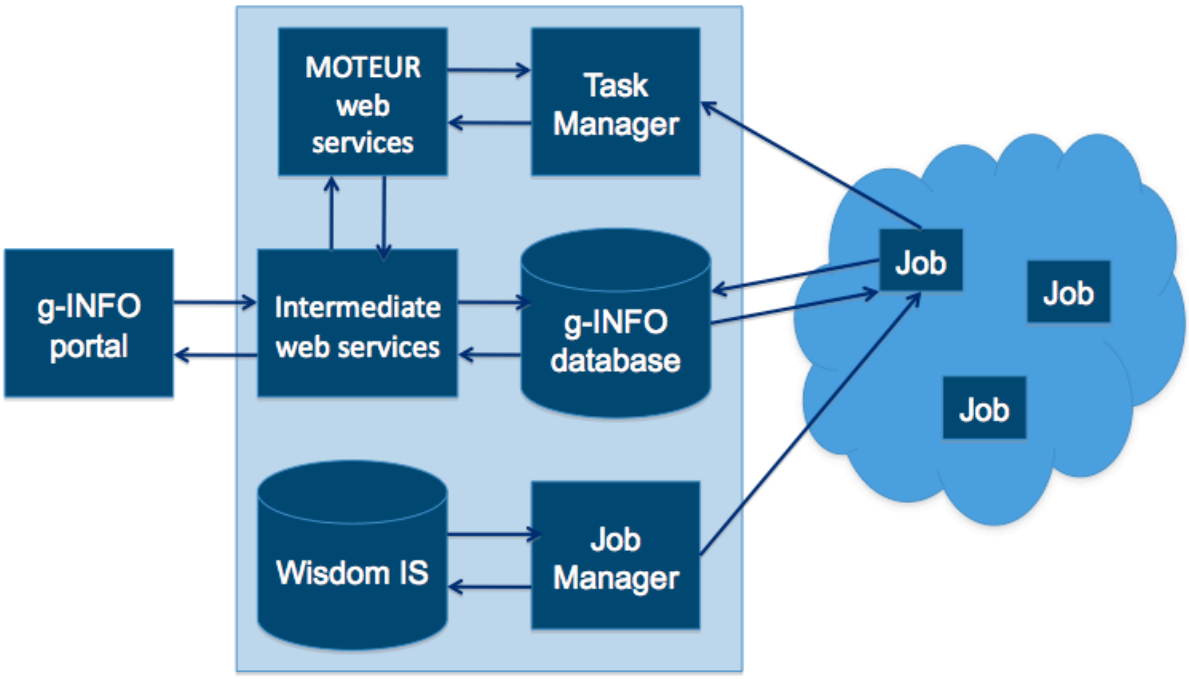

Figure 1 – L'architecture de système g-INFO

Résultats scientifiques

Nous avons fait quelques tests avec les workflows créés par le portail g-INFO. Une des études teste la différence de segments HA / NA de souches de virus H5N1 enregistrés entre l'année 2009 et 2010 et il montre que la plupart des séquences de virus 2010 sont regroupées.

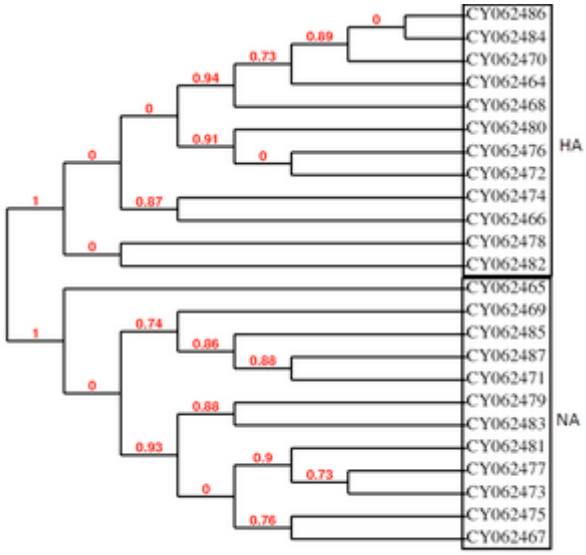

Figure 2 – Test de 24 séquences de H5N1

Un autre test a pour but de confirmer la capacité à exécuter en parallèle plusieurs instances d'un workflow g-INFO. Premièrement, l'outil BLAST trouve n séquences dans une base de données qui sont le plus semblables à la séquence d'entrée. Le reste du workflow va construire un arbre phylogénétique à partir de ces n + 1 séquences. Le test fonctionne exactement sur 12 séquences du virus H5N1 (segment HA, année 2010, hôte humain).

Sequence	**CY062476**	**CY062478**	**CY062480**	**CY062482**	**CY062484**	**CY062486**
T_i (s)	**792**	**709**	**830**	**866**	**727**	**710**
Sequence	**CY062466**	**CY062464**	**CY062468**	**CY062472**	**CY062470**	**CY062470**
T_i (s)	**749**	**851**	**785**	**771**	**708**	**728**
Tmin = 708s Tmax = 866s Tmean = 769s T = 897s						

Tableau 1 – Test de performance de g-INFO

Le temps pour exécuter 12 instances du workflow est de 897s tandis que la durée maximale d'une instance est de 866s. Nous avons utilisé seulement 20 jobs pour ce test. Dans le mode de production réelle, le WPE peut soumettre jusqu'à environ 1000 jobs disponibles sur la grille.

Perspectives

L'analyse du génome du virus de la grippe Influenza est vraiment importante pour comprendre sa pathogénicité, son origine et sa capacité de transmission d'homme à homme afin d'anticiper une éventuelle pandémie. H1N1 a reçu une grande attention ces derniers temps par les autorités de santé publique et les médias, mais le virus H5N1 a également continué à évoluer et à provoquer des foyers épidémiques. Le portail de g-INFO peut aider les experts à créer des workflows bio-informatiques dynamiques et à les exécuter dans gINFO pour surveiller les virus Influenza A. Le workflow phylogénétique actuel est juste un point de

départ. Le travail en perspective comprend l'accès à plusieurs bases de données. Bien qu'ayant la possibilité d'utiliser des données non publiques, un cadre de sécurité doit être développé pour permettre aux propriétaires des données de garder les privilèges sur leurs propres données.

Références

[1] Tiziana Ferrari (2011), *Annual Report on the EGI Production Infrastructure*, EGI Document 413-v8.

[2] Stephen Burke, Simone Campana, Antonio Delgado Peris, Flavia Donno, Patricia Mendez Lorenzo, Rober to Santinelli, Andrea Sciaba (2007), *gLite 3 User Guide Manual Series*, CERN-LCG-GDEIS-722398.

[3] V. Breton, A. L. D. Costa, P. D. Vlieger, L. Maigne, D. Sarramia, Y. Kim, D. Kim, H. Q. Nguyen, T. Solomonides, and Y. Wu (2009), *Innovative in silico approaches to address avian flu using grid technology*, Infectious Disorders Drug Targets, 9(3):358-65.

[4] N. Santos and B. Koblitz (June 2006), *Distributed Metadata with the AMGA Metadata Catalog*, Workshop on Next-Generation Distributed Data Management, HPDC-15, Paris, France.

[5] Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res. 25:3389-3402.

[6] Edgar, Robert C. (2004), *MUSCLE: multiple sequence alignment with high accuracy and high throughput*, Nucleic Acids Research 32(5), 1792-97.

[7] Castresana, J (2000). *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis*, Molecular Biology and Evolution 17, 540-552.

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

[9] T. Glatard, J. Montagnat, D. Lingrand, X. Pennec (2008). *Flexible and efficient workflow deployement of data-intensive applications on grids with MOTEUR*, International Journal of High Performance Computing Applications (IJHPCA), 22 (3), 347-360.