PROCEEDINGS OF SCIENCE

A new flexible workflow on the Grid for monitoring H5N1

Trung-Tung DOAN, Hong-Quang NGUYEN
Institut de la Francophonie pour l'Informatique, UMI UMMISCO 209 (IRD/UPMC)
42, Ta Quang Buu, Hanoi, Vietnam
E-mail: dttung@ifi.edu.vn

Ana Lucia DA-COSTA, Yannick LEGRE
HealthGrid association,
36 rue Charles de Montesquieu, 63430 Pont-du-Château, France

Aurélien BERNARD, Lydia Maigne, Jean SALZEMANN, David Sarramia, Vincent BRETON
Laboratoire de Physique Corpusculaire, CNRS/IN2P3,
24 avenue des Landais, BP 10448, F-63000 Clermont-Ferrand, France

Thanh-Hoa LE
Institute of Biotechnology, Vietnam Academy of Sciences and Technology
18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

Duc-Hung LE
High Performance Computing Center, Hanoi University of Technologies
1, Dai Co Viet, Hanoi, Vietnam

Johan MONTAGNAT
Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis
I3S - UMR6070 - UNSA CNRS, 2000, route des Lucioles - Les Algorithmes - bât. Euclide B - BP 121 - 06903 Sophia Antipolis Cedex - France

In this paper, we introduce a flexible workflow to monitor the evolution of the avian flu virus. Indeed, avian flu remains a major threat to public health worldwide if it acquires the capacity for human to human transmission. The monitoring of H5N1 requires many steps of data analysis. We present here a workflow that can automate this procedure or run it in batch mode. The proposed workflow builds upon existing developments for the g-INFO project and processes molecular biology data from public and private databases using the MOTEUR workflow engine. Some tests show promising results on avian flu observation.

Keywords: grid; workflow; public health; flu; epidemiology

The International Symposium on Grids and Clouds and the Open Grid Forum
Academia Sinica, Taipei, Taiwan
March 19 - 25, 2011

PoS(ISGC 2011 & OGF 31)118

1. Introduction

While the worldwide scientific community is reevaluating international health regulations from the lessons of the pandemic (H1N1) 2009 which has started in April 2009, adequacy of the global response is under scrutiny [1]. Epidemiology is the basis for the practice of preventive medicine and public health by detecting, identifying and analyzing health problems. Molecular epidemiology is using molecular biology to define the distribution of disease and its etiologic determinants. The major issues concerning analysis of genomes, sequences and structures are helped by a comprehensive range of bioinformatics tools permitting to face the dramatic increase in the amount of data available but are requiring adequate computing resources [2]. The 2009 outbreak has demonstrated that continuing vigilance, planning, and strong public health research capability are essential defenses against emerging health threats. This is the reason why more emphasis has been put on global influenza monitoring; indeed monitoring seasonal influenza viruses by sequence analysis provides important and timely information on the appearance of strains with epidemiologic significance [3].

The Grid-based International Network for Flu Observation project (http://g-info.healthgrid.org/) 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. The sequences, extracted from existing data sources populated by the scientific community, are processed dynamically based on Service Oriented Architecture principles (SOA) and Grid technologies. The first prototype of the surveillance network uses flu virus sequences collected from the public international database "Influenza Virus Resource" [4] created and maintained by the National Center for biotechnology Information (NCBI). This prototype executes a pipeline of three well known algorithms commonly used for phylogeny analysis. This pipeline is the same as the one used in the "one click mode" of the phylogeny dedicated website Phylogeny.fr [5].

In this paper, we present an extension of the existing service to meet the specific requirements related to the monitoring of avian influenza. Indeed, while H1N1 virus responsible for influenza A pandemics received great attention from public health authorities and media, H5N1 virus, responsible for the avian flu, has continued to evolve and cause outbreaks. H5N1 is a subtype of the species Influenza A virus, member of Orthomyxoviridae family. Influenza viruses are subtyped based on the antigenicity of their two surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA). The influenza A virus genome consists of eight gene segments encoding 11 viral proteins (gene products) including HA (16 subtypes), NA (9 subtypes), polymerase proteins (PB1, PB2, PA, and PB1-F2), NP, nonstructural proteins (NS1 and NS2), and M1 and M2 proteins. These genes and/or gene products have various basic functions ranging from viral RNA synthesis to receptor binding. Hemagglutinin is a surface protein that acts as a receptor-binding site and is the target of infectivity-neutralizing antibodies [6]. Studies have demonstrated that the level of cleavability of HA determines the virulence of avian influenza viruses in poultry [7]. HA plays an important role in determining the tissue tropism, systemic spread, and pathogenicity of avian influenza viruses. Studies on neuraminidase are

also important. It has been show that most H5N1 viruses are sensitive to neuraminidase inhibitors [8]. Besides HA and NA, studies on other segments such as those of the polymerase (PB2, PB1, and PA) and NP protein show an important effect on the virulence and adaptation of H5N1 virus in the hosts [9]. The Vietnam Ministry of Health has reported two new confirmed human cases of (H5N1) avian influenza infection on 6th and 9th April, 010. Comparing existing information on H1N1 and H5N1, it is very clear that most of the existing data on H5N1 are not publicly available. In response to the request from inter-governmental meetings, WHO has developed an electronic system (https://apps.who.int/fluvirus_tracker) to track influenza A(H5) and other sub-type viruses potential of a pandemic. As of May 25th 2010, there is not one single entry to the database in 2010. At the same time, for every new outbreak of H5N1 virus in Vietnam, the virus is isolated on the dead poultry is sequenced at the Institute of Biotechnology. Analysis of the virus genome and particularly of some key regions related to its pathogenicity and capacity for human-to-human transmission is of utmost importance for the preparedness and the evaluation of pandemic's risk.

For this purpose, a dynamic bioinformatics workflow was implemented in g-INFO. The workflow is considered dynamic because it is configurable so that an expert can choose which components he wants as well as the order of the components in the workflow for his specific analysis. The workflow input data can also be configured. Considering the new H5N1 case in Vietnam, when sequences data are not yet published to a public resource like NCBI, the expert can still use the g-INFO workflow on his own data together with other available public data.

The rest of the paper is organized as follow:

- In section 2, we introduce the existing g-INFO system in which the new workflow will be implemented
- In section 3, we introduce MOTEUR, a workflow engine we use to implement the new workflow in g-INFO
- In section 4, we discuss the tests and the results of the workflow on selected H5N1 data.
- Finally, section 5 will give some conclusions and perspectives.

2. g-INFO system

2.1 General description about g-INFO

The project g-INFO (Grid-based International Network for Flu Observation) is focused on influenza virus with the goal to integrate influenza virus data sources into a federation of databases that can be queried on demand to process selected data through analysis pipelines. g-INFO is implemented and deployed on the EGEE (Enabling Grids for E-sciencE) infrastructure, which is based on a Grid Middleware stack called gLite [10]. Besides gLite, a large-scale deployment of the phylogenetic pipeline requires the use of an environment for job submission and output data collection: the WISDOM Production Environment (WPE) [11]. Initially designed to deploy docking jobs on the grid, the WPE has evolved to use most of the grid services in order to run any software by handling grid jobs in batch mode: automated job submission, status check and report as well as error recovery. WPE has been developed within the EMBRACE project (European Model for Bioinformatics Research and Community

Education) [12]. In the following sub sections, the WPE and its exploitation to handle the g-INFO jobs will be introduced briefly.

2.2 WISDOM Production Environment

The WPE is a middleware designed as an environment that can settle on grid systems or more generally on computing resources, like clusters to handle data and jobs and share the workload on all the integrated resources even if they adopt different technology standards. Based on this middleware, it is possible to build web-services that interact with the system. The middleware is considered as a set of generic services acting as an abstraction level for the specific resources and therefore providing a generic management of data and jobs so that the application services can use any of the underlying systems in a very transparent way (Figure 1). Users are not interacting directly with the grid resources and they are not expected to know how it works since they are just interacting with the top services just like with any other web service.

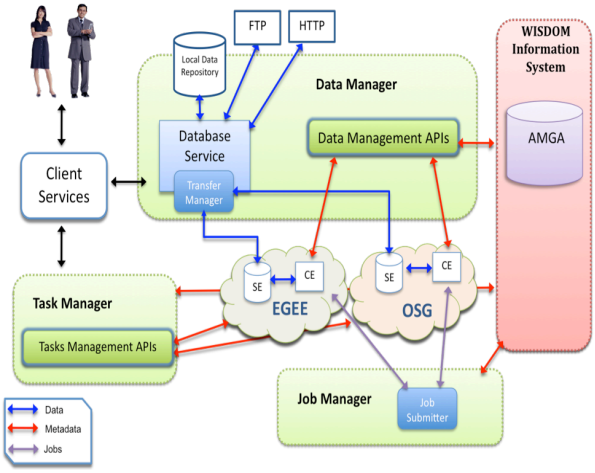

Figure 1. WISDOM Production Environment

As presented in the Figure 1, the WPE is composed of 4 principal components:

- The Task Manager interacts with the client and hosts the tasks to be done;
- The Job Manager submits the jobs to the Computing Elements (CEs) where the tasks managed by the Task Manager will be executed;
- The Data Manager interacts with the client to handle data in batch mode;
- The WISDOM Information System uses AMGA (ARDA Metadata Grid Application) [13] to store all meta-data needed by the Data Manager and the Job Manager.

2.3 Integration of g-INFO within WPE

As seen on the Figure 1, the WPE is composed of 4 principal components but g-INFO uses only 3 of them: the Task Manager to manage g-INFO's tasks, the Job Manager to submit g-INFO's jobs and the WISDOM Information Service as required by the Data Manager and the Job Manager. The project g-INFO does not use the WPE Data Manager. Instead, it uses AMGA directly to store influenza virus data and meta-data. Because of the small size of the sequence data for each virus (several text lines in FASTA format), it is not necessary to use a powerful

tool like the Data Manager, which is used to deploy automatically large data files on the grid and manage their replica. Indeed, it is easier and faster to access directly an AMGA server rather than accessing a Storage Element (SE) on the Grid for such small files.

The Data Collection is used to collect and integrate influenza data from other data source (public or private). NCBI has been selected as the first data provider for the g-INFO. The Data Service provides access to influenza data stored in an AMGA server for other component in g-INFO such as the services in the Task Manager or the g-INFO portal. In this first implementation, the data is centralized in one single AMGA server located in the CNRS/IN2P3/LPC laboratory but it is possible to distribute the data on several AMGA servers to obtain a load-balanced multiple servers. From a specific AMGA server, AMGA APIs enable the access to the data stored in other servers.

In the Task Manager, we implemented 4 general tasks for phylogenetic analysis:

- BLAST's task is for running BLAST [14] on sequences;
- MUSCLE's task is for sequences' alignment [15];
- Gblocks's task is for sequences' curation [16];
- PhyML's task is for constructing phylogenetic trees [17].

The g-INFO's jobs are submitted by the Job Manager. Once a g-INFO job is submitted successfully on the Grid, it will look for a g-INFO's task in the Task Manager. If a job finds a task, it grabs and executes it. The WPE keeps a job alive as long as possible so that during its runtime, a job can execute many tasks. When a job dies (normally because the associated proxy is expired), WPE will submit another job to replace it. In this way, there are always available jobs to process tasks. In consequence, a task can be executed immediately after it is created.

With the g-INFO's tasks available in WPE, we can implement phylogenetic pipelines such as the one described in a previous work [18] to monitor specific features of the influenza virus strains. This pipeline is just a starting point. It is aimed at providing a complementary service to the public health research community by producing common interest epidemiologic indicators on all available data. In the case of H5N1, virologists need flexible workflows that can be configured and run automatically / manually. The next section will present how we provide these new services in g-INFO.

3. Flexible workflows in g-INFO

The difference between the data processing pipeline previously implemented within g-INFO and the new flexible workflows is that the pipeline is static as it is hardcoded and deployed on the server side in the g-INFO system [18] while the workflow is dynamic and configurable from the user side. Experts without knowledge of grid computing (or even programming) can create their own workflow and run it with a desktop tool or through a portal. To enable workflows in g-INFO, we use the MOTEUR workflow engine described below.

3.1 MOTEUR

MOTEUR is a workflow designer and enactor developed by I3S and CREATIS laboratories that is interfaced with gLite grid middleware and handles application services asynchronously [19]. For this reason it is perfectly suited to handle long makespan workflows

such as g-INFO. MOTEUR provides a very flexible framework to run g-INFO as the workflow can be built from a set of independent services, and can be modified interactively through a graphical interface. Furthermore, it provides advanced data parallelism constructs well adapted to exploit distributed grid resources. MOTEUR can also be run in command line allowing a daily and automatic execution of the g-INFO pipeline. The use of a workflow engine such as MOTEUR is very relevant in the context of a bioinformatics platform with modular services since designing as many pipelines as there are workflows, users or execution conditions would become untraceable. A bioinformatics platform should be a toolbox of independent tools and algorithms, and a workflow engine will be used to handle all those services altogether in a coherent way at runtime without adaptation of the users on the services themselves.

3.2 g-INFO workflow

In order to permit experts to create a g-INFO workflow with MOTEUR, we provide for each task in the Task Manager a corresponding asynchronous web service. Each web service contains at least 4 principal functions:

- submitSequence: creates a corresponding task in the Task Manager, i.e. BLAST asynchronous web service will create BLAST task
- isFinished: to check if an instance task is finished so that the workflow can continue with the next instance task. If there are several instances of a task, the workflow runs asynchronously: it doesn't wait for all instances of a task to be completed before moving to the next task.
- getOutput: returns a LFN of a file on a Storage Element which contains the list of sequences as the output of submitSequence to be used for the input of the next step. For example, BLAST returns a list of hits, MUSCLE returns a list of aligned sequences, etc.
- getOutputArchive: returns a LFN of an archive of other outputs of a task including error or log files.

In order to prepare input and collect result for a workflow running with MOTEUR, g-INFO system provides two web services:

- gINFOSequenceFileBuilder: gets sequence data from g-INFO's AMGA server based on AMGA queries, then put them on the Grid as a LFN (Logical File Name) file. This LFN will be used as input for the first component in a workflow.
- gINFOResultsCollector: collects output on the Grid of a workflow as well as its components.

3.3 Input data for g-INFO workflow

Input data can be provided to a g-INFO workflow using the two following methods:

- The first method consists in using the g-INFO Data Manager. Input data must be prepared in a sequence query file which contains one or several queries to g-INFO's AMGA server. This sequence query file will be parsed by gINFOSequenceFileBuilder. The sequence query file can be created on the g-INFO portal without knowledge of AMGA queries.

- The second method is to copy manually a sequence data file from a local computer to the Grid. This could be done also by submitting the sequence data file to the g-INFO portal.

When an expert wants to study a set data that have not been published yet, he can chose one of the two methods mentioned above. By choosing the first method, he needs to integrate his data with g-INFO data using the component Data Collection in the g-INFO Data Manager. By this way, he can share his data with other selected users of g-INFO system but still keeps the data private from outside.

4. Monitoring H5N1 with g-INFO workflows

Phylogenetic analysis using parsimony with influenza virus sequences from GenBank showed that all eight genes formed a unique branch. The first test we made using a workflow with 3 components (muscle, gblocks and phyml) shows that: H5N1 sequences of the year 2010 are clustered in two branches of HA and NA (Figure 2).

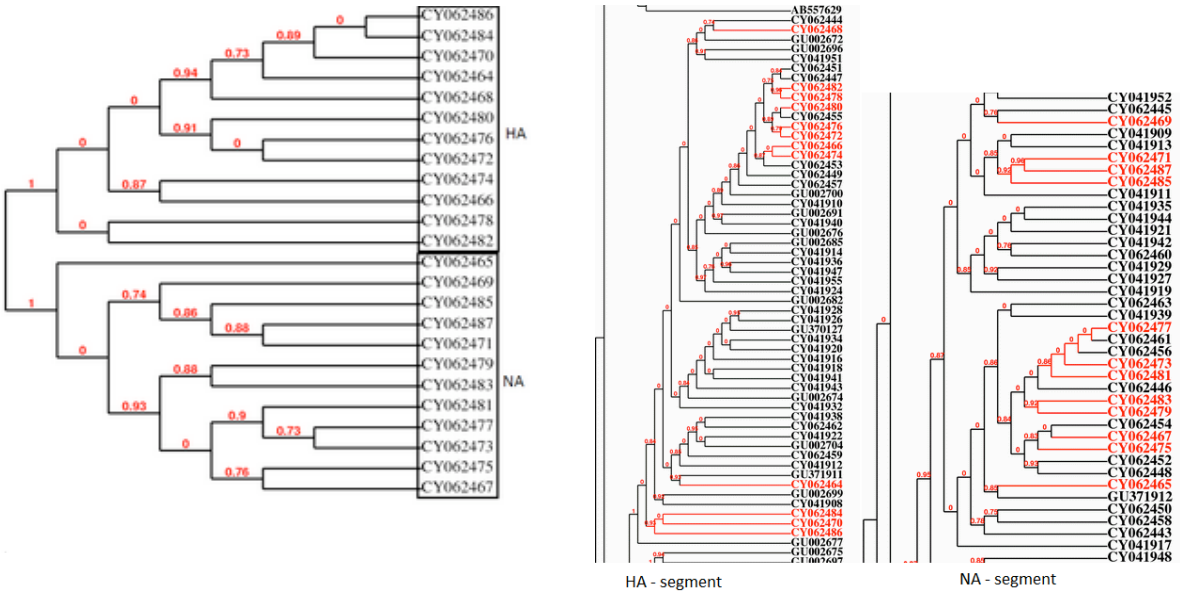

Figure 2. Analysis of sequences of H5N1 (human host) in 2010

In the second test of the g-INFO workflow, we study the difference of HA / NA segment of H5N1 strains between year 2009 and 2010. In Figure 2, we can see that most of 2010 virus sequences are clustered together.

The last test studies the capacity of running, in parallel, several instances of a g-INFO workflow. In this test, we add a BLAST component to the previous workflow. For each sequence in the input file, BLAST will find n sequences in a database that are the more similar to that sequence. The rest of the workflow will construct a phylogenetic tree from these n + 1 sequences as in the first two examples. Supposed that we have an input file of N sequences, we obtain N instances (or pipelines) of the workflow. In g-INFO system, each job can handle a task. If and only if the task has been completed, the job is free to handle another task. In

consequence, we can conclude that we need at least N jobs to run the workflow in parallel. This is illustrated in Figure 3. Most of the time, the workflow requires 3 jobs to run its 3 instances I_1, I_2 and I_3.

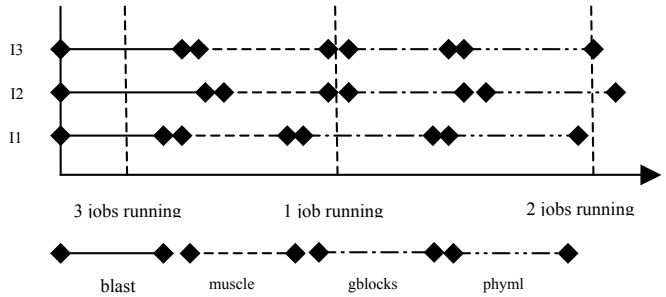

Figure 3. Number of jobs requires to run a workflow

Supposed that we have submitted M jobs and M ≥ N, given T is the time to run the workflow; T_i is the time to run the instance ith then:

$$T \approx \max(T_i), i = \overline{1, N}$$

Table 1 shows the time to run a workflow on 12 sequences of H5N1 (HA segment, 2010, human host). Time to run 12 instances of the workflow is 897s while the time maximum of an instance is 866s. We submitted only 20 jobs for this test. In the real production running mode, WPE has approximate 3000 jobs available on the Grid.

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						

Table 1. A workflow's run time

5. Conclusion

Analysis of the influenza virus genome is of utmost importance to understand its pathogenicity, origin and capacity for human-to-human transmission, and anticipate a potential pandemic. H1N1 received lately great attention from public health authorities and media, but the H5N1 virus has also continued to evolve and cause outbreaks, requiring relevant tracking. Therefore, we presented in this paper a dynamic bioinformatics workflow implemented in g-INFO to monitor avian flu.

The g-INFO system connects various bioinformatics programs and reconstructs a robust phylogenetic tree from a set of sequences, either public or belonging to a researcher. g-INFO is implemented on the EGEE infrastructure and deployed though the WISDOM Production Environment. The WPE brings the possibility to take advantage of the heterogeneity and dynamism of the grid technology.

The MOTEUR workflow engine is used to handle the bioinformatics algorithms as modular services, making the workflow fully configurable by the user. Users can compose a workflow based on the available bioinformatics components of g-INFO. Input data can be chosen from a ported public data source or can be private.

g-INFO workflow is very suitable for batch analysis. Taking advantages of the Grid power, several workflows with many instances can run simultaneously.

The current phylogenetic workflow is just a starting point. The work in perspective includes the access to other influenza databases in addition to NCBI. In order to have the possibility of using non-public data, a security framework must be developed to allow the data owner to keep privileges on his own data. More bioinformatics tools will be added to the g-INFO system with the priority for those experts can use to study H5N1. We need also to improve the g-INFO portal to have a more user-friendly interface with features such as search engine or possibility to create, and execute customized workflow.

References

[1] Report of the First Meeting of the Review Committee on the Functioning of the International Health Regulations (2005) in Relation to Pandemic (H1N1) 2009, 12–14 April 2010, Geneva, Switzerland, WHO

[2] Frederica P. Perera, I. Bernard Weinstein, *Molecular epidemiology: recent advances and future directions, carcinogenesis*, vol. 21, n°3, pp 517-524, 2000.

[3] Stack JC. , *Protocol for sampling viral sequences to study epidemic dynamics*, JR Soc Interface, 2010.

[4] Bao Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. *The Influenza Virus Resource at the National Center for Biotechnology Information*, J. Virol. 2008 Jan;82(2):596-601.

[5] Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.-F., Guindon S., Lefort V., Lescot M., Claverie J.-M., Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist, Nucleic Acids Research. 2008 Jul 1; 36 (Web Server Issue):W465-9. Epub 2008 Apr 19.

[6] Skehel JJ, Wiley DC. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem. 2000;69:531-69.

[7] Horimoto T, Kawaoka Y. Reverse genetics provides direct evidence for a correlation of hemagglutinin cleavability and virulence of an avian influenza A virus. J Virol. 1994 May;68(5):3120-8.

[8] World Health Organization Global Influenza Program Surveillance Network. Evolution of H5N1 avian influenza viruses in Asia. Emerg Infect Dis. 2005 Oct;11(10):1515-21.

[9] Gabriel G, Herwig A, Klenk HD (2008) Interaction of polymerase subunit PB2 and NP with importin alpha1 is a determinant of host range of influenza A virus. PLoS Pathog 4: e11.

[10] gLite middleware, [Online], Available: http://glite.web.cern.ch/glite/default.asp

[11] 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, Innovative in silico approaches to address avian flu using grid technology, Infectious Disorders Drug Targets, Nov. 2008.

[12] EMBRACE Grid, [Online], Available: http://www.embracegrid.info/page.php

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

[14] 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.

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

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

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

[18] Trung-Tung DOAN, , Aurélien BERNARD, Ana Lucia DA-COSTA, Vincent BLOCH, Thanh-Hoa LE, Yannick LEGRE, Lydia MAIGNE, Jean SALZEMANN, David SARRAMIA, Hong-Quang NGUYEN, Vincent BRETON, Grid-based International Network for Flu Observation, HealthGrid 2010 Conference, accepted paper, 2010.

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