Grid-based International Network for Flu Observation (g-INFO)

Trung-Tung DOAN[a,1], Aurélien BERNARD[b], Ana Lucia DA-COSTA[c], Vincent BLOCH[b], Thanh-Hoa LE[e], Yannick LEGRE[c, d], Lydia MAIGNE[b], Jean SALZEMANN[b], David SARRAMIA[b], Hong-Quang NGUYEN[a], Vincent BRETON[b]

[a] *Institut de la Francophonie pour l'Informatique*
L'équipe Modélisation et Simulation Informatique
42, Ta Quang Buu, Hanoi, Vietnam

[b] *Laboratoire de Physique Corpusculaire, CNRS/IN2P3,*
24 avenue des Landais, BP 10448, F-63000
Clermont-Ferrand, France

[c] *HealthGrid association, 36 rue Charles de Montesquieu,*
63430 Pont-du-Château, France

[d] *Maat Gknowledge, Méjico, 2. 45004 Toledo, Spain*

[e] *Institute of Biotechnology, Vietnam Academy of Sciences and Technology*
18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

Abstract. The 2009 H1N1 outbreak has demonstrated that continuing vigilance, planning, and strong public health research capability are essential defenses against emerging health threats. Molecular epidemiology of influenza virus strains provides scientists with clues about the temporal and geographic evolution of the virus. In the present paper, researchers from France and Vietnam are proposing a global surveillance network based on grid technology: the goal is to federate influenza data servers and deploy automatically molecular epidemiology studies. A first prototype based on AMGA and the WISDOM Production Environment extracts daily from NCBI influenza H1N1 sequence data which are processed through a phylogenetic analysis pipeline deployed on EGEE and AuverGrid e-infrastructures. The analysis results are displayed on a web portal (http://g-info.healthgrid.org) for epidemiologists to monitor H1N1 pandemics.

Keywords: Grid, public health informatics, flu, surveillance network, phylogeny

Introduction

According to the World Health Organization update in February 2010, more than 212 countries have reported laboratory confirmed cases of pandemic influenza H1N1 in 2009, including at least 15292 deaths [1]. 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

[1] Corresponding Author.

influenza viruses by sequence analysis provides important and timely information on the appearance of strains with epidemiologic significance [2]. Continuous molecular epidemiology analysis of viral genetic data collected provides scientists with clues about the temporal and geographic evolution of the virus as well as clues about which viral genes are associated with virulence.

Since the beginning of the pandemic on 19 April 2009, thousands of clinical samples have been characterized antigenically and genetically in laboratories. A large fraction of the clinical samples, especially in the early days of the pandemics, have been sequenced. The data have been continuously cleaned up and submitted from all over the world to the Influenza Virus Resource of the National Center for Biotechnology Information (NCBI) which collects highly valuable antigenic and molecular information accumulated since 50 years. This information is highly relevant to make prediction about future pandemics [3].

The quality of the response to an emerging disease relies on timely and reliable information, broadly shared and quickly analyzed. In the case of influenza H1N1, all data were collected into one public database at NCBI. In the case of H5N1, the situation is much more complex. Molecular data are stored in databases which remain non-synchronized. Systems are non-interoperable, leading to data and application silos, duplication of work, and costly efforts to integrate data. The grid technology is one of the most promising and dynamic concepts able to address such bottlenecks: grid permits an up-to-date data sharing and access, strengthens transparency and interoperability in order to process dynamically the large amount of molecular data made available by the research community. In the present paper, researchers from France and Vietnam are proposing a global surveillance network based on grid technology: the goal is to federate influenza data servers, in a secure way, and deploy automatically molecular epidemiology studies.

This project is well integrated in the current trend expressed by the global informatics community to move towards collaborative, distributed development and architecture. In 2009, the WHO and the Centers for Disease Control and Prevention's National Center for Public Health Informatics (CDC/NCPHI) have announced their collaboration on a Global Public Health Grid (GPHG) initiative to enable global data exchange and collaborative development of globally shareable and interoperable systems, tools and services [4]. This paper describes one of the pilot projects of GPHG initiative.

1. Background

Molecular data about influenza virus is of high importance to make prediction about emerging strains and therefore develop early response. For this reason, some institutes and laboratories such as Influenza Virus Resource [5] (National Center for Biotechnology Information), BioHealthBase [6] (National Institute of Allergy and Infectious Diseases), Influenza Virus Database [7] (Beijing Institute of Genomics - Chinese Academy of Sciences), Influenza Sequence Database [8] (Los Alamos National Laboratory), have been collecting influenza virus data from countries all around the world and are providing rich services to reinforce influenza surveillance and early warning capabilities. Some of these database resources ([5], [6], [7]) supply phylogenetic tools with web-based interface to process the data, others don’t [8]. In this

case, epidemiologists can use local tools or web-based tools such as Phylogeny.fr [9]. Nevertheless, there are some issues that need to be considered:

- The restriction in terms of sequences length and number of sequences returned due to the performance of the web server is a limiting factor for the analysis;
- The tools are disconnected from each other and this prevents users from building automatic workflows.
- The incompleteness of the data despite the attempts to collect them globally. (For several reasons, some organizations / institutes do not want to publish their data or at least there is a delay in the publication of data).

It is well known in bioinformatics that processing data in batch mode permits to save time and effort. Many phylogenetic pipelines have been developed: some are designed for general purpose such as Phylogenetic Diversity Analyzer (PDA) [10] or AIR-Appender, AIR-Identifier, and AIR-Remover (AIR) [11], while others are used for specific analyses, such as PhyloGena [12] which is a software for automated phylogenetic annotation of unknown sequences. Another example is a pipeline for processing and identification of fungal ITS sequences [13].

Phylogenetic analyses are very CPU consuming. That is the reason why most of common phylogenetic tools are executed with the aid of clusters. PhyML-MPI [14], for example is a parallel version of the well-known software PhyML [15] for the construction of phylogenetic trees. This is true also for genomics comparative analysis with the example of mpiBLAST [16], a parallelized version of BLAST [17]. BLAST was also ported on grid in the project EELA (E-science grid facility for Europe and Latin America) [18].

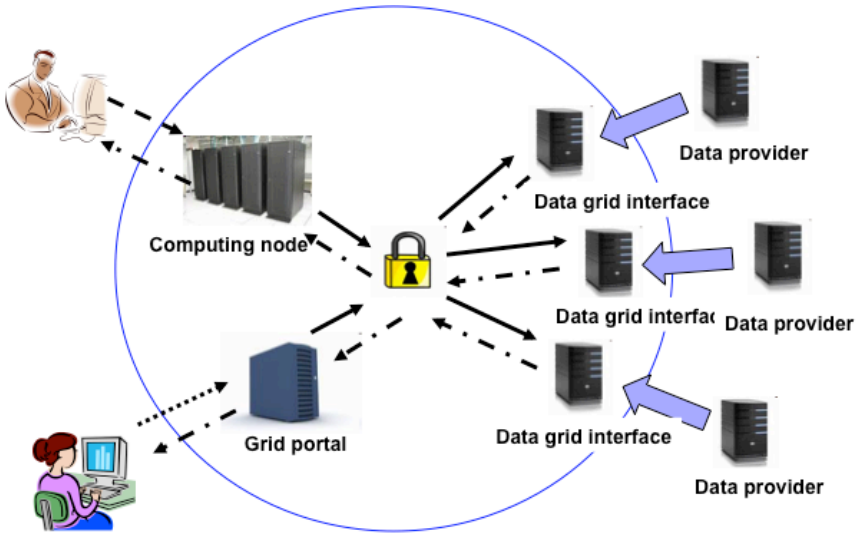

Figure 1. Architecture of g-INFO

The project g-INFO (Grid-based International Network for Flu Observation) is focused on influenza virus and is expected to address the bottlenecks mentioned above. The goal is to integrate influenza virus data sources into a federation of databases that can be queried on demand to process selected data through analysis pipelines (Figure 1). The CPU-intensive steps of the pipeline are deployed on grid resources to take advantages of its high security, heterogeneity and large-scale computation.

The architecture of g-INFO is shown in Figure 1. In this architecture, data providers are national repositories and international public resources on influenza. Each data provider has its own server(s) to store virus data. These servers export all - or only selected data - to a data grid interface server. A pipeline is deployed to be run daily on computing nodes and results are published on a web portal.

2. Implementation

The project 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*. [19]. 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). Initially designed to deploy docking jobs on the grid [20], 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 [21] (European Model for Bioinformatics Research and Community Education). In the following sub sections, the WPE will be introduced briefly as well as the integration of g-INFO within it.

2.1. WISDOM Production Environment

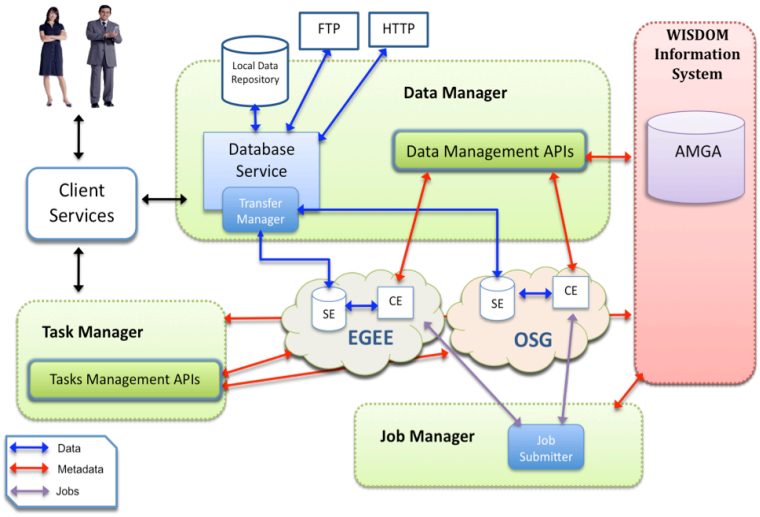

Figure 2. Architecture of the WPE

The WPE can be seen as a kind of meta-middleware as it is 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 follow different technology standards. Based on this meta-middleware, it is possible to simply build web-services that interact with the services of the system. The meta-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 2). 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.

As presented in the Figure 2, 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 [22] (ARDA Metadata Grid Application) to store all meta-data needed by the *Data Manager* and the *Job Manager*.

The project g-INFO does not use WPE *Data Manager*. Instead, it uses AMGA directly to store influenza virus data and meta-data. Because of the small size of 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 data on the grid and manage their replica. Indeed, it is easier and faster to access directly an AMGA server rather that accessing a *Storage Element* (SE) on the Grid for such small files. The next section presents how the influenza data are stored on the AMGA server.

2.2. Data preparation

NCBI has been selected as the first data provider for the g-INFO prototype. The influenza data are made available at NCBI on their FTP server: ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/. Influenza sequences for nucleotide, protein and coding regions are provided in FASTA format. Every day, an update occurs at 9 a.m. GMT when a script is triggered to download all new FASTA files to a local directory on the server. Another script extracts these files to get data (i.e. the sequences of the new viruses) and meta-data (i.e. the information about the new viruses). Data and meta-data extracted are then stored in the AMGA repository with the following structure (or schema in AMGA's terminology):

- /Phy_Workflow/NCBI_Sequences/updates
- /Phy_Workflow/NCBI_Sequences/metadata
- /Phy_Workflow/NCBI_Sequences/coding
- /Phy_Workflow/NCBI_Sequences/nucleotide
- /Phy_Workflow/NCBI_Sequences/protein

The "*updates"* collection is used to log the changes in the database on a daily basis. The "*coding", "nucleotide"* and "*protein"* collections include sequences of influenza virus. The "*metadata"* collection includes additional information about the virus: virus host, country where the sequenced strain was isolated, year when the strain was sequenced, virus subtype, number of segments sequenced, sequence length, strain name, including additional information such as the city where the strain was isolated and the precise date the sample was collected, association number of the corresponding protein sequence, association number of the corresponding coding sequence, age of the patient affected by the virus, sex of the patient affected by the virus and additional information.

The whole process of downloading, extracting and updating is logged for the purpose of monitoring and properly handling any issue or inconsistency such as network connection or data redundancy. The schema defined within AMGA is dynamic

since it can be modified by the client, even at runtime. This feature of AMGA provides us with the possibility of easily importing data from other data sources besides NCBI.

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 server. From a specific AMGA server, AMGA APIs enable the access to the data stored in other servers [23].

2.3. Current phylogenetic pipeline

The first workflow implemented for g-INFO is a well-known and widely used phylogenetic pipeline, which builds a phylogenetic tree from a curated alignment of sequences. This workflow, depicted on Figure 3, is the chained execution of three bioinformatic algorithms:

- A selected pool of sequences (composed of old and brand new sequences) is aligned by MUSCLE (Version 3.7) [24];
- The alignment in FASTA format generated during the first step is curated with Gblocks (Version 0.91b) [25] to remove low quality sequences;
- The cleaned alignment is converted into PHYLIP format to be injected in PhyML software (Version 3.0 aLRT). The output of PhyML is a phylogenetic tree in standard Newick format.

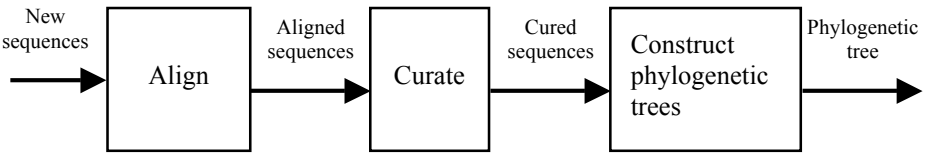

Figure 3. Current phylogenetic pipeline

In addition to all algorithm results, an annotation file is generated from all metadata stored in our AMGA database. At the end of g-INFO workflow execution (all instances together), when all results have been retrieved from grid storage elements (SE) and archived on a server, a global archive is created, transferred and deployed on a portal hosted by the HealthGrid association.

The current pipeline is composed of one main program and three consecutive blocks:

- Workflow launcher: the main program launches and monitors the multiple instances of the pipeline which can run simultaneously;
- MUSCLE block to align sequences;
- Gblocks block to curate aligned sequences;
- PhyML block to construct phylogenetic trees.

To run a bioinformatics algorithm on the grid using the Wisdom Production Environment (WPE), it must be encapsulated into a service package that contains:

- A main program written in shell language (bash) which recovers input files and software installation package, installs the software or compiles the source code, executes the bioinformatics tool and stores results and log files on the grid;
- A shell script that contains a list of functions (Function pack)

- A configuration file that specifies a series of environment variables and current algorithm parameters;
- Possible secondary programs (in Perl or Java for example) to perform various tasks like file format conversion for example;

When all the WPE elements are installed on the local server and all the relevant bioinformatic algorithms are correctly packaged into WPE services, the g-INFO Workflow launcher handles the creation of tasks for these services in the WPE Task Manager. In addition to the set of parameters, the Workflow launcher needs an input file containing a list of AMGA requests. The line syntax for this file is:

```
NAME_OF_THE_WORKFLOW_INSTANCE; AMGA query to select sequences
```

Below is an example of the input file to the Workflow launcher

```
Segment_4_Spain_2009;selectattr
/Phy_Workflow/NCBI_Sequences/nucleotide:sequence
'/Phy_Workflow/NCBI_Sequences/nucleotide:FILE=/Phy_Workflow/NCBI_
Sequences/metadata:FILE and
/Phy_Workflow/NCBI_Sequences/metadata:host="Human" and
/Phy_Workflow/NCBI_Sequences/metadata:segnum="4" and
/Phy_Workflow/NCBI_Sequences/metadata:year="2009" and
/Phy_Workflow/NCBI_Sequences/metadata:country="Spain" and
/Phy_Workflow/NCBI_Sequences/metadata:subtype="H1N1"'
```

The input file is scanned line after line by the Workflow launcher in order to define the request to be executed on the AMGA server to get a particular pool of sequences (in this example, all H1N1 segment 4 sequences isolated from Spanish patients in 2009) on which the analysis workflow will be launched.

Execution of a workflow instance corresponds to the creation of a task for the first g-INFO service (Muscle) using the WPE Task Manager. Each workflow is a chain of services which create tasks for the next service. Taking example of the workflow depicted on Figure 3, the Muscle service launches a Gblocks task, creating itself a PhyML task at the end of its execution.

After all the workflow instances are launched, the g-INFO main script monitors the task execution, retrieves the task results and log files, creates a global archive with all the results of the workflow instance, creates a backup of the global result archive on a local server and transfers results to the HealthGrid web server for future display after launching a cleaning script to the grid file catalogue (LFC) used.

The numerous tasks created for all the workflow instances are executed by special jobs called “Wisdom Agents” which are “pilot jobs” created and renewed by the WPE Job Manager. Each pilot job is like an empty shell that regularly requests the Task Manager for a task to execute. A given pilot job doesn’t end after a simple task execution but can carry out a very large number of tasks until the user proxy expiration. As many workflow instances as needed for our molecular epidemiology analyses, they can be run simultaneously using Wisdom Agents on the computing resources available to the EGEE Biomed Virtual Organizational and on the regional grid AuverGrid in Auvergne.

3. Results and discussions

3.1. Current status of g-INFO

The g-INFO prototype is currently in the last phase of internal testing and will be switched into production mode before the end of March 2010. First of all, flu virus sequences from at least one data provider are daily collected to update several AMGA tables through a fully automated process. Then, the workflow launcher is able to launch and monitor several workflow instances simultaneously. It can be launched using two different modes:

- **Local mode**: this is the test mode that permits to test the execution of the different services on a local server (same configuration and operating system than a grid node). The scripts are executed sequentially on the User Interface. File transfer and storage are accomplished using standard grid mechanism. The LFC and grid storage elements are used to store data produced.
- **Grid mode**: this is the normal mode of the pipeline. The pipeline is executed on the EGEE and AuverGrid e-infrastructures using the resources accessible to the Biomed VO or to the Auvergrid VO.

Figure 4. Results web page of the g-INFO website

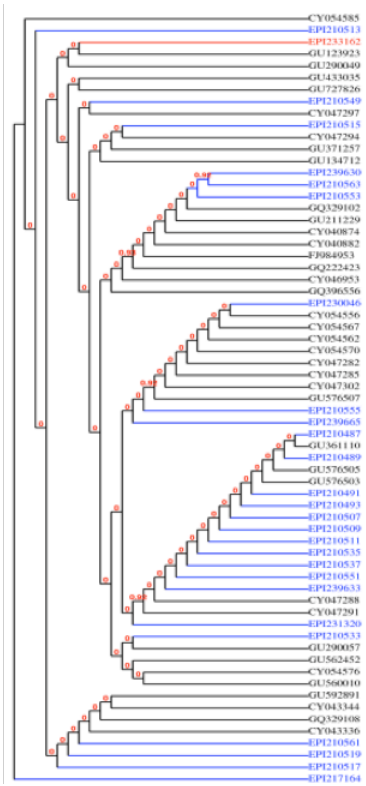

Figure 5. Phylogenetic tree of 65 sequences to monitor resistance to oseltamivir

The workflow results are automatically transferred to a specific web server on which a website has been developed to make all results available to the scientific

community on a daily basis. In addition to the information about the flu virus and the description of the g-INFO project, a result web page has been setted up (Figure 4) to allow biologists to download data of each workflow instance and each service such as the alignment generated by MUSCLE or the raw phylogenetic tree created by PhyML.

However, due to the lack of tree visualization tool fully compatible with grid systems, the current version of the g-INFO workflow doesn't contain yet a service to convert raw phylogenetic trees to a more readable representation. On the other hand, a list of graphic tools for tree analysis is suggested to the end user. As an illustration, figure 5 presents the example of a phylogenetic tree generated using the tree view tool TreeDyn [26] available at http://www.phylogeny.fr taking as input 65 sequences (H1N1, 2009, nucleotide, segment 6, human host, Europe) among which, 38 sequences are from NCBI and 27 sequences are from GISAID *(*Global Initiative on Sharing Avian Influenza Data*)* with specific information about resistance (branches in red color) or sensitivity (branches in blue color) to Oseltamivir. This example illustrates how the g-INFO project can be used to monitor specific features of the virus strains.

3.2. Discussions

The current implementation of the g-INFO portal has only one workflow aiming at building automatically phylogenetic trees from new sequences. Future developments will bring additional features such as search or customized workflow execution. The grid specific features available in the current version are:

- The data extracted from NCBI is stored on a grid database using AMGA. As a consequence, additional databases can be easily interfaced to g-INFO. There are two ways to do it: either by downloading the database on the AMGA server where NCBI data are already stored or parsing it to add the new sequences, or by installing an AMGA server at the database location and parsing data of specific interest into the AMGA database there. The second approach has the interest of leaving the data where they are produced and to give its owner a complete control of their usage. This is the approach used to share medical data in a Cancer Surveillance Network under deployment in Auvergne [27].
- The workflow is deployed on grid computing and storage elements through the Wisdom Production Environment. This makes it very easy to change the chain of bioinformatics algorithms but also to mobilize as many computing resources as needed. In the present prototype version, the power of the grid is not yet fully exploited because parallelism is currently only implemented on the input data: only one processor is used for each workflow instance, because we use PhyML.

4. Conclusion and perspectives

4.1. Conclusion

The g-INFO project is a success in terms of international collaboration. Indeed this project is lead by Vietnamese and French researchers having a common goal: showing the relevance of the grid computing to address and impact emerging health threats, particularly the flu pandemics. The use of grid for such purposes has been permitted

through the use of the WISDOM Production Environment that can be seen as a meta-middleware providing generic services that just abstract the specific resources and provide a generic management of data and jobs so the application services can use any of the underlying systems in a very transparent way. The WPE brings the possibility to take advantage of the heterogeneity and dynamism of the grid technology.

The g-INFO grid-based pipeline deployed daily provides an updated picture of H1N1 situation in terms of molecular indicators relevant to the virus evolution. g-INFO is expected to spare from epidemiologists the heavy tasks of collecting all data available and processing them on his/her own machine. It does not aim at replacing the existing services made available on databases such as NCBI where experienced users can design their own workflow. It is aimed at providing a complementary service to the public health research community by producing common interest epidemiologic indicators on all available data.

The current g-INFO prototype paves the way for the adoption of grids for pandemics monitoring and represents a step forward responding to the needs of the research community concerning the federation of all the influenza data sources to avoid incompleteness and provision of tools not limited in terms of sequences' length, number of sequences, or workflow possibilities.

4.2. Perspectives

The current phylogenetic pipeline is just a starting point. The work in perspective includes the access to other influenza databases in addition to NCBI. As the possibility to access non-public databases is also considered, a security framework must be developed to allow the data owner to keep privileges on his own data.

A close dialogue with biologists and epidemiologists has been started in order to develop workflows with specific interest. In collaboration with the Institute of Biotechnology of the Vietnamese Academy of Sciences, a dedicated implementation of g-INFO is under study to monitor avian flu strains. Indeed, while H1N1 has received most of the attention in the recent months, the H5N1 virus keeps evolving.

Virologists have identified regions on the virus genome (protease cleavage sites, glycosylation sites, epitopes and binding site), which are particularly relevant for its transmission and pathogenicity. We are currently investigating the specific monitoring of these sites of interest and the customization of the workflow for this purpose.

A request expressed by experts is the possibility for g-INFO users to create their own pipeline and store their analysis results. This takes us away from the initial vision for g-INFO about providing daily results without any effort for the users. On the other hand, it is also the natural following step as it is expected that the daily information provided by g-INFO would trigger questions and therefore tools to further dig into the data should be made available together with the daily indicators. In consequence, the g-INFO portal should provide additionnal features for users such as search engine for virus data and possibility to create,and execute customized workflows.

5. Availability and requirements

The project is freely accessible, using a web browser at http://g-info.healthgrid.org/

Project name: g-INFO

Home page: http://g-info.healthgrid.org/

Operating systems: The pipeline requires Linux with gLite middleware

Programming languages: Java, C, Bash script

License: GNU-GPL

Acknowledgements

The work described in this article was partly supported by grants from the European Commission (EGEE, EMBRACE, EUAsiaGrid), the French Ministry of Research (GWENDIA) and the regional authorities (Conseil Régional d'Auvergne, Conseil Général du Puy-de-Dôme, Conseil Général de l'Allier).

The Enabling Grids for E-sciencE (EGEE) project is co-funded by the European Commission under contract INFSO-RI-031688. The EMBRACE project is co-funded by the European Commission under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092.

Auvergrid is a project co-funded by CNRS, Conseil Regional d'Auvergne and the European Commission. The GWENDIA project is supported by the French ministry of Research.

EUAsiaGrid ("Towards a common e-Science infrastructure for the European and Asian Grids") is a project co-funded by the European Commission as a Coordinated and Support Action within the 7th Framework Programme (FP7-INFRA-223791).

The authors acknowledge fruitful discussions with John Fitzpatrick and Muzna Mirna from CDC Atlanta.

The authors acknowledge the support of the FKPPL and FVPPL International Associated Laboratories as well as EUAsiaGrid (FP7) project.

References

[1] WHO website, *Pandemic (H1N1) 2009 - update 87*, [Online], Available: http://www.who.int/csr/don/2010_02_12/en/index.html

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

[3] Robert G. Webster, *Predictions for Future Human Influenza Pandemilcs*, The Journal of Infectious Diseases 1997; 176(Suppl 1):S14 – 19.

[4] Muzna Mirza, *Global Public Health Grid - WHO-CDC Public Health Informatics Initiative: Value Proposition and Pilot Projects*, Public Health Information Network Conference, 2009.

[5] 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.

[6] Squires et al. BioHealthBase: *Informatics support in the elucidation of influenza virus host pathogen interactions and virulence*, Nucleic Acids Research (2008) vol. 36 (Database issue) pp. D497.

[7] Influenza virus database, [Online]. Available: http://influenza.big.ac.cn/about/About.jsp.

[8] Influenza sequence database, [Online], Available: http://flu.lanl.gov/

[9] 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.

[10] Casillas, S., and A. Barbadilla, 2004. PDA: a pipeline to explore and estimate polymorphism in large DNA databases. Nucleic Acids Res. 32**:** W166–W169
[11] S. Kumar, A. Skjaeveland, R. Orr, P. Enger, T. Ruden, B. H. Mevik, F. Burki, A. Botnen, and K. S. Tabrizi, *Air: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses*, BMC Bioinformatics, vol. 10, no. 1, pp. 357+, October 2009.
[12] K. Hanekamp, U. Bohnebeck, B. Beszteri, and K. Valentin, *Phylogena-a user-friendly system for automated phylogenetic annotation of unknown sequences*, Bioinformatics (Oxford, England), vol. 23, no. 7, pp. 793-801, April 2007.
[13] H. R. Nilsson, G. Bok, M. Ryberg, E. Kristiansson, and N. Hallenberg, *A software pipeline for processing and identification of fungal its sequences*, Source code for biology and medicine, vol. 4, no. 1, January 2009.
[14] Phylm-mpi, [Online], Available: http://atgc.lirmm.fr/phyml/versions.html.
[15] Guindon S, Gascuel O., *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*, Systematic Biology. 2003 52(5):696-704.
[16] A. Darling, L. Carey, and W. Feng, *The Design, Implementation, and Evaluation of mpiBLAST*, 4th International Conference on Linux Clusters: The HPC Revolution 2003 in conjunction with ClusterWorld Conference & Expo, June 2003.
[17] 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.
[18] M Cárdenas, V Hernández, R Mayo, I Blanquer, J Pérez-Griffo, R Isea, L Nuñez, HR Mora, M Fernández, *Biomedical Aplications in EELA*, Studies in Health Technology and Informatics 2006;120;397-400.
[19] gLite middleware, [Online], Available: http://glite.web.cern.ch/glite/default.asp
[20] 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.
[21] EMBRACE Grid, [Online], Available: http://www.embracegrid.info/page.php.
[22] 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.
[23] Ahn, S., Kim, N., Lee, S., Nam, D., Hwang, S., Koblitz, B., Breton, V., and Han, S. 2009. Performance analysis and optimization of AMGA for the large-scale virtual screening. *Softw. Pract. Exper.* 39, 12 (Aug. 2009), 1055-1072.
[24] Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.
[25] Castresana, J. *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis*, Molecular Biology and Evolution 17 (2000), 540-552.
[26] Chevenet F., Brun C., Banuls AL., Jacq B., Chisten R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics. 2006, Oct 10;7:439.
[27] Catherine Quantin, Gouenou Coatrieux, François André Allaert, Maniane Fassa, Karima Bourquard, Jean-Yves Boire, Paul de Vlieger, Lydia Maigne and Vincent Breton, New Advanced Technologies to Provide Decentralised and Secure Access to Medical Records: Case Studies in Oncology, Cancer Informatics 2009/7, 217-229.