Processing the Biomedical Data on the Grid Using the UNICORE Workflow System

Marcelina Borcz[1,2], Rafał Kluszczyński[2], Katarzyna Skonieczna[3,4], Tomasz Grzybowski[3], and Piotr Bała[1,2]

[1] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100 Toruń, Poland
[2] Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Pawińskiego 5a, 02-106 Warsaw, Poland
[3] Department of Molecular and Forensic Genetics, Institute of Forensic Medicine, Ludwik Rydygier Collegium Medicum, Nicolaus Copernicus University, Skłodowskiej-Curie 9, 85-094 Bydgoszcz, Poland
[4] The Postgraduate School of Molecular Medicine, Medical University of Warsaw, Żwirki i Wigury 61, 02-091 Warsaw, Poland

Abstract. The huge amount of the biological and biomedical data increases demand for significant disk space and computer power to store and process them. From its beginning the Grid has been considered as possibility to provide such resources for the life sciences community. In this paper authors focus on the UNICORE system which enables scientists to access Grid resources in a seamless and secure way. Authors have used the UNICORE middleware to automate experimental and computational procedure to determine the spectrum of mutations in mitochondrial genomes of normal and colorectal cancer cells. The computational and storage resources have been provided by Polish National Grid Infrastructure PL-Grid.

1 Introduction

The amount of data which biologists have to handle is usually too big to be stored and processed using desktop or even more powerful single computers in the laboratory. The scientists are forced to use clusters in supercomputer centers or even multiple resources combined into a single infrastructure. The problem is especially important for the processing biomolecular and genetic data necessary for diagnostics and research. A good example is analysis of the genomic data generated by high throughput experimental systems. With the increasing availability of such devices there is growing demand to process experimental data effectively and provide results in hours rather than days or weeks. Usually, processing of genetic data is performed on the dedicated hardware installed at the experimental facilities or hospital. With the introduction of new, low cost sequencing systems this model is no longer valid and there is strong demand to acquire external resources for storage and computations. The obvious choice are grids and clouds since they can provide required infrastructure. Because security of the data is here a concern, the grid solutions with the strong security model based on the X509 certificates is natural solution.

I. Caragiannis et al. (Eds.): Euro-Par 2012 Workshops, LNCS 7640, pp. 263–272, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Since there is number of the grid infrastructures around, the user can pick the most suitable according to the different metrics including cost. For the academic users, the natural choice is to use national grid infrastructures organized within the EGI framework. The PL-Grid - Polish Grid Infrastructure is a good example. It is available to the research and academic community free of charge and provides access through gLite and UNICORE middlewares.

UNICORE middleware [18] is one of the successful solutions used to build distributed infrastructure. Along with the UNICORE Rich Client (URC) it enables users to run programs, manage files and data transfers. It allows also to design and run workflows in a graphical environment which is easy to use. In this paper we present example solution which allows for the uniform access to the experimental data, their handling and analysis in the distributed environment which significantly reduces processing time.

The paper is organized as follows: the experiment is described in the first section of the paper. The second part focuses on the description of the UNICORE middleware and the UNICORE workflow system. It is followed by the specification of our solution and the experimental setup. The paper concludes with the summary and directions for further work.

2 Experiment

Recently an increasing number of studies have indicated that changes in mitochondrial genome sequence may contribute cancer phenotype [15,12]. Despite of that fact, the knowledge of mitochondrial DNA (mtDNA) variability in colorectal cancer is still limited, and until now the reliable spectrum of mtDNA somatic mutations has not been resolved [17]. Thus it is important to determine mtDNA variability in normal and tumor cells. Due to the usage of high-throughput GS FLX Instrument (Roche Diagnostics) up to 1 million reads (that correspond to the 1 million DNA fragments) of approximately 500 base pairs long could be generated independently in a single experiment [1]. The aforementioned ultra-deep DNA sequencing technology may aid the identification of low-level heteroplasmic, nucleotide variants. Therefore, we have used 454 sequencing technology to resolve mutational spectrum of the entire mitochondrial genome sequences of non-tumor and colorectal cancer cells. The particular aims of our research were:

- determination of the 18 complete mitochondrial genome sequences of tumor and matched non-tumor tissues obtained from 9 patients diagnosed with colorectal cancer,
- mtDNA sequences comparison with the reference sequence,
- mtDNA mutation identification,
- ultra high speed processing of mtDNA sequence data.

Mitochondrial DNA sequences were determined with 454 sequencing technology. In brief, mtDNA molecules were amplified in two overlapping fragments of about 8500 base pairs in length [10]. Next, the amplicones were nebulized into 400–500 base pairs fragments long, clonally amplified in emPCR and sequenced with GS FLX Instrument

(Roche Diagnostics). During the sequencing reaction digital images were subsequently captured by the CCD camera. The raw image data from a single experiment (approximately 30GB in size) were next converted with the use of GS Run Processor (Roche Diagnostics) into base-called results. Reports of the base-calling analysis were generated with GS Reporter (Roche Diagnostics). Thereafter obtained reads were aligned to a revised Cambridge Reference Sequence (rCRS) [3] with the use of a GS Reference Mapper (Roche Diagnostics), which enables mtDNA mutations to be detected. Roche provides software for the RedHat Linux free of charge to the equipment. All components use popular and standardized file formats (eg. fasta, ece) which makes it easy to interoperate.

3 Data Storage and Data Processing Infrastructure

The processing of the sequenced data requires significant computational resources not available as part of the typical experimental setup. We have decided to integrate GS FLX Instrument with the PL-Grid distributed infrastructure [13] as the data storage and processing system.

3.1 PL-Grid Infrastructure

The Polish Grid Infrastructure provides the Polish scientific community with an IT platform based on computer clusters, enabling research in various domains of e-Science. PL-Grid infrastructure enables scientists carrying out scientific research based on the simulations and large-scale calculations using the computing clusters as well as provides convenient access to distributed computing resources. The infrastructure supports scientific investigations by integrating experimental data and results of advanced computer simulations carried out by geographically distributed research teams. Polish Grid Infrastructure is a part of a pan-European infrastructure built in the framework of the EGI (European Grid Initiative), which aims to integrate the national Grid infrastructures into a single, sustainable, production infrastructure. PL-Grid infrastructure is both compatible and interoperable with existing European and worldwide Grid frameworks.

UNICORE is one of the middlewares offered to the users which enables access in a seamless and secure way. The UNICORE PL-Grid infrastructure is based on the resources provided by the main computational centers in Poland. It is built of individual sites which are described by the list of provided hardware, services and installed applications (see. Figure 1).

The PL-Grid portal provides users with the forms to perform registration and application for resources. It also delivers interface for the users to obtain self-generated certificates necessary to access the grid. The process itself uses an additional module registering users in the virtual organization. As a part of the user's registration there is performed data replication to VOMS and UVOS server depends on the middleware the user want to use. The portal allows to register any certificate issued by CA approved by EUGridPMA organisation [8] or additional Simple CA certificates used only internally in PL-Grid.

This solution allows to hide from the user less intuitive parts of the certificate issuing process and facilitates placement in the UVOS additional information in the form of groups or attributes. These informations are necessary to grant users with the proper privileges.

3.2 UNICORE

UNICORE is a system which provides an easy and secure access to the Grid resources. The system has been developed since 1997 [18] and has been successfully used in many European scientific projects contributing significantly to the increase of the popularity and applications of distributed computing. UNICORE is part of the European Middleware Initiative [9] where, together with gLite and ARC middlewares, developers continue their contribution to the Grid standards by increasing interoperability, manageability, usability and efficiency of the Grid services. The UNICORE system is also used by the National Grid Initiative infrastructures such as PL-Grid which provides resources to the scientific communities.

One of the main advantages of the UNICORE system is set of the UNICORE clients. There are three types of them: UNICORE Rich Client (URC), UNICORE Command-line Client (UCC) and High Level API (HiLA). UNICORE Rich Client enables users to design and run workflows in a graphical way which does not require knowledge of particular language used to define and store workflows. This capability is important for the end users since most scientific experiments have the workflow structure: the output of one step of the experiment is the input for another one. The URC allows for easy, graphical creation of the workflows and for the management of their execution. In the Grid context, workflow is the automation of the processes which involves the orchestration of a set of Grid services or agents in order to solve a problem or to define a new service [20]. UNICORE workflow system implements loops and if-else statements so the user can design sophisticated scientific experiments and run them just by pressing the run button. Once designed, the workflow can be saved and used many times with different data and parameters. All files created in the workflow can be managed in a graphical way in the client. Two-layered structure of the UNICORE workflow system allows to plug-in domain-specific workflow languages, workflow engines and various brokering strategies.

The UNICORE server side components fully adopt grid services paradigm. The computational systems are called Target Systems, and are built of two main UNICORE services: UNICORE/X and TSI which enable access to computing and data resources. All communication between the user and grid services is streamed through the UNICORE Gateway service to ensure secure access. In the Polish National Grid Infrastructure main UNICORE services are located at the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) at the University of Warsaw. ICM provides central services such as the registry of all sites' endpoints and Unicore Virtual Organization Service acting as a users management service. The UVOS is populated with the user data received from the PL-Grid portal which contains information about users, their privileges and certificates. This data is synchronized with the UVOS and is used to authorize and authenticate users.

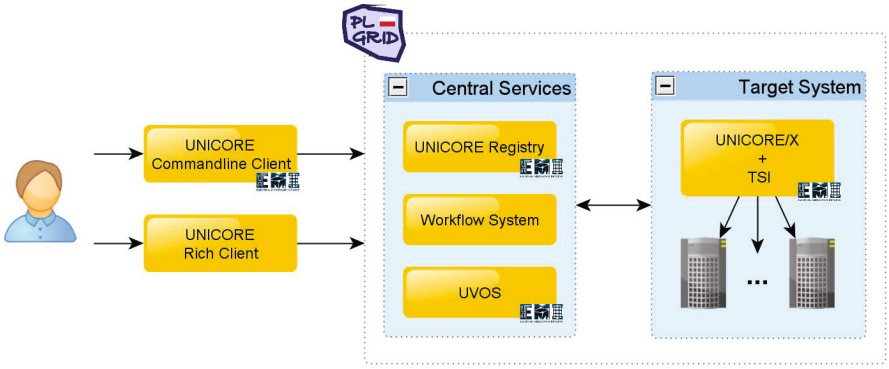

Fig. 1. Diagram of UNICORE infrastructure setup deployed in PL-Grid

UNICORE infrastructure in PL-Grid is built based on the on European Middleware Initiative (EMI) releases, which are distributed from EMI repository as RPM packages [7] which allows for easy installation and configuration with only few additional dependencies.

3.3 Storage

PL-Grid infrastructure allows users to organize themselves in a scientific groups (called teams), which can apply for common allocations, share computational resources and data. In the case of data, members of a group have access to special storage at every site enabling them to share data. UNICORE provides access to such storage by a simple configuration of Target System. A default access privileges (unix `umask`) can be set to ensure that newly created folders and files are accessible by default to all of the group members.

A typical PL-Grid target system running UNICORE middleware offers three types of storages:

- **user space** (also known as file-space), which contains files created for running tasks, it is usually fast filesystem (like Lustre [11]);
- **global storage**, which contains files not related with particular task, it is usually slower, but more reliable then user space;
- **group storage**, which contains files that can be accessible by all of the group members, it can be slow but is also reliable.

The PL-Grid infrastructure offers access to the group storage also from other grid middlewares enabling users to use different tools for data access.

3.4 UNICORE FTP

The key issue for integration of the experimental setup with the storage infrastructure is efficient data transfer. UNICORE/X offers three built-in protocols: BFT, RBYTEIO

and SBYTEIO. In the case of data upload, all of them transfer data through the UNICORE Gateway which introduces communication overhead resulting in low transfer speed. The solution is a new protocol called UNICORE File Transfer Protocol (UFTP) which has been recently implemented [16]. In order to transfer data with UFTP an additional UFTPD service has to be installed next to the UNICORE/X. The data transfer is therefore performed directly between UFTPD and UNICORE client. Data transfer uses dynamic firewall port opening mechanism by using passive FTP connections which allows to bypass the UNICORE Gateway. This solution is considered to be more secure then statically opened port ranges usually used for example by the GridFTP server [2].

In order to make UFTP protocol more secure, all file transfers are always initiated by the UNICORE/X server. To start file transfer client contacts UNICORE/X, determines the parameters of the transfer and then connects to the UFTPD to perform the actual transfer. The UNICORE/X passes client IP and transfer parameters to the UFTPD service which waits for connections for defined period of time. Communication between UNCIORE/X and UFTPD is done through a secure "command port" accessible only by defined UNICORE/X servers which prevents non authorized usage. Another advantage of the UFTP protocol is ability to establish multiple parallel TCP connections per data transfer which significantly speeds up transfer.

4 Automation of the Experiment

The applications used as parts of the workflow developed for the processing of the genetics data work under the Roche License. Key components are parallelized using Message Passing Interface. The programs from the FLX suite are compiled for the Red Hat Linux and are installed by the site administrators on the PL-Grid target systems. Therefore it is important to make sure that only privileged users have permission to run it. This can be done using the group management available in the PL-Grid. Groups are mapped into the operating system's groups. In this way the access to the installed applications is granted only for the proper team members for whom the license has been issued.

The developed workflow consists of three programs from the FLX program set. First, the GS Run Processor processes raw images generated by the FLX Instrument (Roche diagnostics). This data, which can exceed tens of gigabytes, is put into the UNICORE storage in an automatic way. After a new file or directory with data is created as a result of experiment, the UNICORE Commandline Client program is used to put automatically data to the PL-Grid UNICORE target system storage (see Figure 2). Because of the size of the data a UFTP protocol is used for transfers. Appropriate access rights to the group storage ensure privacy.

The GS Run Processor has several components. We run the `runAnalysisPipe` script for full processing of the acquired data and use the GS_LAUNCH_MODE environmental variable to set MPI mode enabling us to use multiple worker nodes on the target system. The next part of the workflow can be run simultaneously on the available resources. The GS Reporter generates all reports files from CWF files created in the

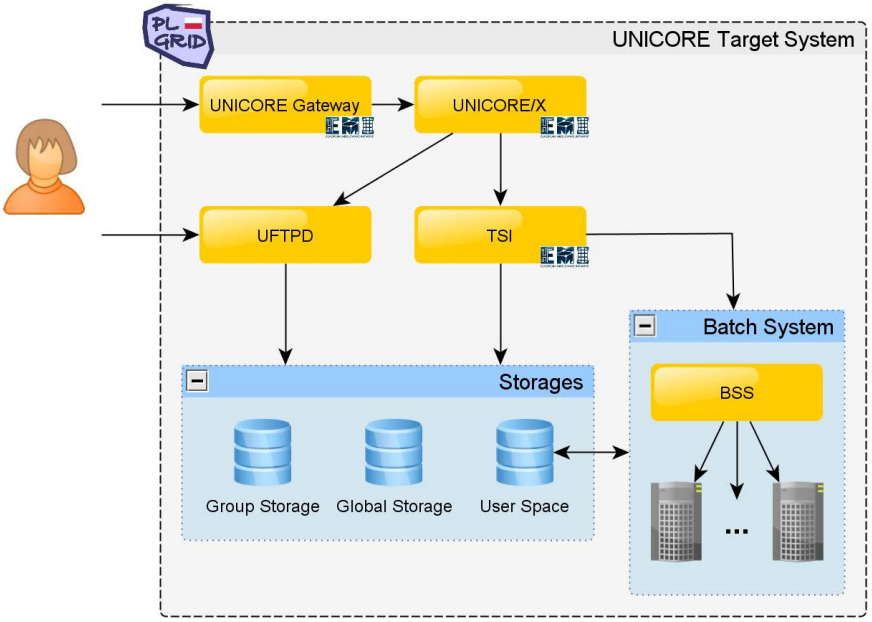

Fig. 2. The detailed view of the UNICORE target system with the different types of storage

previous job. The GS Reference Mapper consists of two steps which create the mapping project and align the reads against a reference sequence. The example workflow edited in the UNICORE Rich Client is shown in the Figure 3.

The execution of the workflow ends when all workflow components are finished. The workflow branches are run in parallel on different target systems to speed up execution. Each workflow component represent application which can be run in parallel using MPI on the target system. The workflow can be stored on the disk and then loaded to the URC. The UNICORE uses its own format for internal representation of the workflows which is translated to the single work assignments. The significant advantage of the UNICORE Rich Client is automatic generation of the necessary files based on the graphical representation of the workflow. Therefore the knowledge of particular workflow definition language is not required and workflow creation and modifications are performed using intuitive GUI.

It is up to the user to decide where the output of each workflow step is stored. Files can be directed to the global or group storage or simply stored in the directory of the task. In the context of workflows the second option is better since it usually takes less time to copy the data between jobs.

It should be mention that a user has access to all files generated by each job despite their physical location. Data can be downloaded to the local computer or permanently stored in the distributed storage. The workflow can be extended to include additional programs. For example Blast [4], Clustal [6] or R [14] applications can be used for further processing.

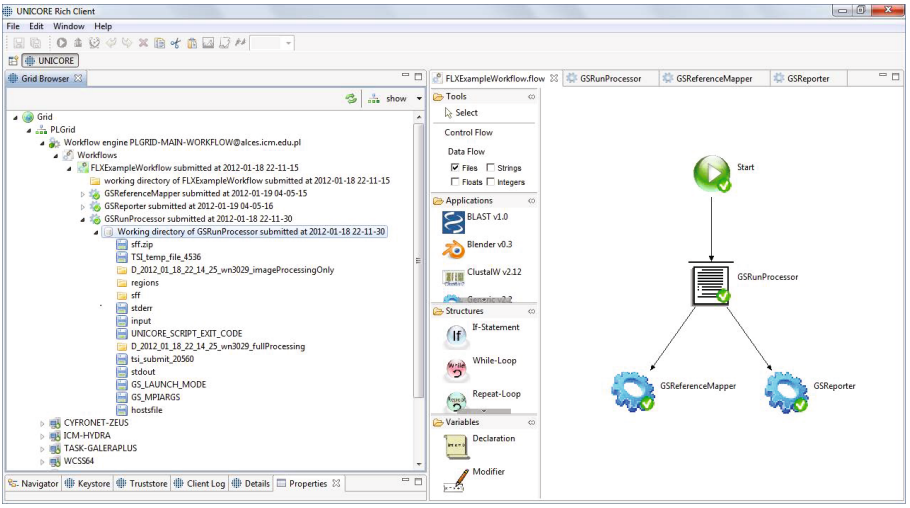

Fig. 3. The UNICORE Rich Client view on the designed workflow to process acquired biomedical data

5 Results

The described setup has been used to store and to process experimental data on the PL-Grid infrastructure. Single sequencing experiment generated ca. 30 GB of data which is organized in the 834 files of the size of about 33 MB each. The transfer of the individual file from the workstation attached to the sequencer takes few second, the overall time to transfer experiment's results was 3-4 hours. One should note that this process has been fully automated and did not require user assistance. Once data is stored on the Grid it can be easily used as the input for analysis.

Table 1. Performance results of sequence analysis for various hardware. The different hardware systems are available to the user through UNICORE target systems provided by the PL-Grid.

Processor type	Cache size	Interconnect	Number of cores	Time (hours)
Intel Xeon CPU @ 3.60GHz	1024 kB	none	1	70.0
AMD Interlagos Opteron Processor 6272 @ 2.10GHz	2048 kB	Gigabit	64 (1 x 64 cores)	6.5
AMD Opteron Processor 6174 @ 2.20GHz	512 kB	Gigabit	64 (8 x 8 cores) 96 (8 x 12 cores)	4.5 4.5
Intel Xeon CPU, X5660 @ 2.80GHz	12288 kB	Infiniband	64 (8 x 8 cores) 96 (8 x 12 cores)	2.5 2.5

The important benefit of the UNICORE workflow system is that workflow once developed can be run on the different target systems. Therefore we were able to run the analysis workflow on the available hardware using parallel execution. The performance of the GS Run Processor, the most time consuming step, presented in the Table 1 shows significant speedup for 64 cores used compared to the single core run. Further increase of the number of cores has no effect on execution time. One should note significant difference in the execution time between different processors. This is caused by the different size of the internal cache which is the largest for the Xeon processor. UNICORE allows user for easy pick up optimal one.

6 Conclusions

The usage of the storage and processing power of the distributed infrastructures allowed for significant reduction of the analysis time of the results of mitochondrial genome sequence runs which was the bottleneck in the process. The time required for the data processing needed for the final analysis has been reduced to hours instead of several days. Available distributed storage allowed for tracking multiple data which opens up field for detailed statistical analysis. The data processing has been automated and simplified based on the UNICORE workflows. In practice all stages of the data processing are run automatically reducing manual work and unnecessary delays. Practically unlimited storage and fast and reliable data transfer mechanisms allow to offer users flexible and easy to use storage accessible simply on the Internet. This functionality is especially attractive to biological and medical users who can focus on their research instead of mastering computer science details necessary to process data.

7 Future Work

Current set up for high-throughput GS FLX Instrument is a starting point for building dedicated infrastructure to process biomedical data. We plan to significantly increase data processing functionality by adding new applications and by creation more complicated workflows which will automate analysis. We also plan to provide more flexibility to biologists and medicians using the Gridbeans for `gsMapper` and other applications used. A GridBean is a plugin for UNICORE Rich Client which provides graphical interface to an application and allows for easy change of the parameters used for execution. Thanks to this users will be able to modify workflow execution without assistance of the IT staff and thus will receive further flexibility in running complicated data analysis.

Acknowledgments. This research was supported in part by the PL-Grid Infrastructure and by the EMI project (UE RI-261611). The study was supported by the Ministry of Science and Higher Education (grant no.: NN 301 075 839).

References

1. 454 sequencnig technology website, http://www.454.com
2. Allcock, W.: GridFTP: Protocol Extensions to FTP for the Grid. Global Grid ForumGFD-R-P.020 (2003)
3. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N.: Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23, 147 (1999)
4. Borc, M., Kluszczyński, R., Bała, P.: BLAST Application on the GPE/UnicoreGS Grid. In: Lehner, W., Meyer, N., Streit, A., Stewart, C. (eds.) Euro-Par 2006 Workshops. LNCS, vol. 4375, pp. 245–253. Springer, Heidelberg (2007)
5. Benedyczak, K., Stolarek, M., Rowicki, R., Kluszczyński, R., Borcz, M., Marczak, G., Filocha, M., Bała, P.: Seamless Access to the PL-Grid e-Infrastructure Using UNICORE Middleware. In: Bubak, M., Szepieniec, T., Wiatr, K. (eds.) PL-Grid 2011. LNCS, vol. 7136, pp. 56–72. Springer, Heidelberg (2012)
6. Clustal software website, http://www.clustal.org
7. EMI Software Repository, http://emisoft.web.cern.ch/emisoft
8. EUGridPMA organisation website, http://www.eugridpma.org
9. European Middleware Initiative (EMI) website, http://www.eu-emi.eu
10. Fendt, L., Zimmermann, B., Daniaux, M., Parson, W.: Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. BMC Genomics 10, 139 (2009)
11. Lustre product website, http://www.lustre.org
12. Petros, J.A., Baumann, A.K., Ruiz-Pesini, E., et al.: mtDNA mutations increase tumorigenicity in prostate cancer. Proc. Natl. Acad. Sci. U.S.A. 102, 719–724 (2005)
13. PL-Grid project website, http://www.plgrid.pl
14. R environment website, http://www.r-project.org
15. Sánchez-Aragó, M., Chamorro, M., Cuezva, J.M.: Selection of cancer cells with repressed mitochondria triggers colon cancer progression. Carcinogenesis 31, 567–576 (2010)
16. Schuller, B., Pohlmann, T.: UFTP: High-Performance Data Transfer for UNICORE. In: Romberg, M., Bala, P., Müller-Pfefferkorn, R., Mallmann, D. (eds.) Proceedings of UNICORE Summit 2011, Forschungszentrums Jülich. IAS Series, vol. 9, pp. 135–142 (2011) ISBN 978-3-89336-750-4
17. Skonieczna, K., Malyarchuk, B.A., Grzybowski, T.: The landscape of mitochondrial DNA variation in human colorectal cancer on the background of phylogenetic knowledge. Biochim. Biophys. Acta 1825, 153–159 (2011)
18. UNICORE middleware website, http://unicore.eu
19. UVOS project website, http://uvos.chemomentum.org
20. Yu, J., Buyya, R.: A taxonomy of scientific workflow systems for grid computing. SIGMOD Record 34(3), 44–49 (2005)