19

BUILDING A VIRTUAL LABORATORY FOR SCIENTIFIC EXPERIMENTATION IN MOLECULAR BIOLOGY

C. Garita[1], H. Afsarmanesh[2], O. Unal[2], L.O. Hertzberger[2]
[1]*Costa Rican Institute of Technology, COSTA RICA*
cesar@ic-itcr.ac.cr
[2]*University of Amsterdam, THE NETHERLANDS*
{hamideh, ozgul, bob}@science.uva.nl

This paper describes the application of the generic framework provided by the Grid-based Virtual Laboratory Amsterdam (VLAM-G), in the support of complex experimentation scenarios in the domain of molecular biology. The focus of the paper lies on both the analysis of some reference experimentation scenarios, and the on-going extension and tuning of Virtual Laboratory environment to better support advanced scientific experiments in this domain.

1. INTRODUCTION

The complexity associated with modern scientific experimentation in a large number of fields such as physics, astronomy, medicine, and biology is increasing at a very high pace. This is due to many factors, including the continuos growth in the amount and heterogeneity of data sets that need to be manipulated, the need to have proper access to shared resources among physically distributed centers with multi-disciplinary scientists or engineers, and the lack of widely-accepted ICT infrastructures supporting advanced experimentation activities in a comprehensive and flexible way. In the particular case of molecular biology, scientists of different organizations have long faced a strong need to collaborate in order to carry out their complex distributed experiments, involving for instance sharing and exchange of gene expression information, DNA sequencing data, and protein database entries. Proper integration of these kinds of information within a common experimentation framework becomes a crucial need for biologists all around the world in order to better study and understand specific biological processes.

In this context, this paper addresses the application of the generic framework provided by VLAM-G to complex experimentation scenarios and in specific, to the domain of molecular biology. The main focus of the paper therefore lies on one hand on the analyzing reference experimentation scenarios, and on the other hand on the on-going development activities related to the necessary extension and adjusting

L. M. Camarinha-Matos et al. (eds.), *Processes and Foundations for Virtual Organizations*

of the VLAM-G. This in-progress research activity has been carried out within the framework of the Flexwork[1] project.

Please notice that a comprehensive introduction to the molecular biology area is outside the scope of this document. Therefore, basic knowledge of this area is suggested for the reader of this paper. A comprehensive introduction to computational molecular biology can be found in (Setubal and Meidanis 1997).

The rest of this paper is organized as follows. Section 2 describes the general architecture and end-user interface of VLAM-G. Section 3 presents some of the general experimentation scenarios in molecular biology that have been identified and detailed within the Flexwork project, and that will be supported by VLAM-G infrastructure. Section 4 describes the current development activities regarding the VLAM-G support for the identified experimentation scenarios, as well as the potential future directions of the presented work. Finally, Section 5 summarizes the main conclusions of this paper.

2. THE VLAM-G ENVIRONMENT

Scientists from different areas such as physics, biology, and chemistry need to perform complex experiments that consist of many inter-related steps. However, typically, the tasks involved in these steps are carried out independently, and not through guided and specified steps within a single integrated framework. Thus, the idea of a Virtual Laboratory to better assist scientists with their experimentations comes into scene. In this context, the Grid-based Virtual Laboratory AMsterdam (VLAM-G) environment is a science portal for remote experiment control and collaborative distributed analysis for applied sciences, using cross-institutional integration of heterogeneous information and resources (Afsarmanesh, Belleman et al. 2002).

VLAM-G supports the fundamental information management requirements for a Virtual Laboratory (VL) framework (Afsarmanesh, Kaletas et al. 2001). Although there are many important requirements identified for VL in previous publications, in this paper we mainly will focus on the following:

- Management of large data sets produced by different devices. In order to extract valuable information from raw data sets produced by devices used within a given experiment (e.g. microarray and scanner devices), these sets need to be processed and integrated, which requires high-performance computing facilities.
- Provide a flexible and configurable collaboration environment. In order to allow definition and execution of experiments involving processes and devices offered by different organizations, the VL must provide an extremely flexible and configurable environment that scientist can easily use without being concerned about low-level details related to for instance, physical location, specific resource access operations, communication protocols, etc.
- Distributed resource management. Since the VL is used by different scientists to access resources located at different organizations, an underlying distributed

[1] This research was partially supported by the Dutch NWO Biomolecular Informatics Flexwork project (2002-2004) that involves the following partners: University of Amsterdam, Wageningen University, Catholic University of Nijmegen, and Leiden University Medical Center.

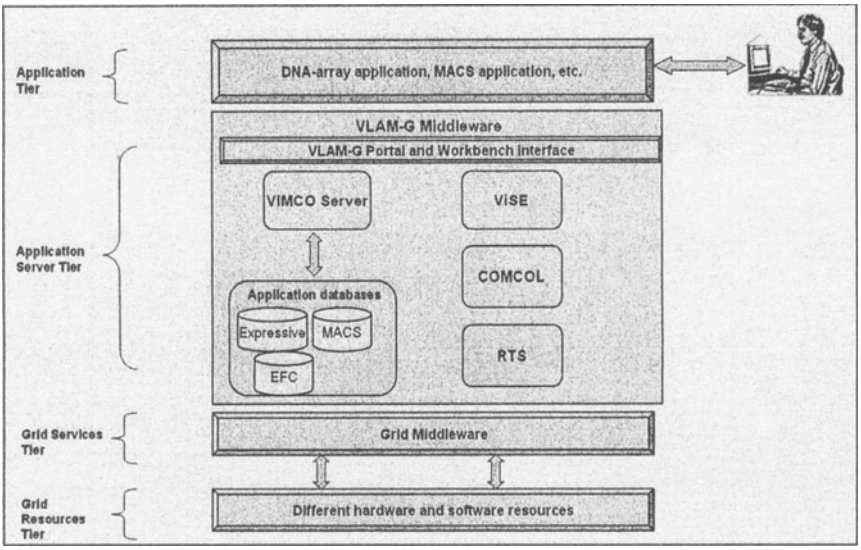

Figure 1. General VLAM-G architecture.

resource management platform is needed within the VL architecture to enable high-performance distributed computation required by many experiments.

In order to meet these requirements, the Grid technology has been adopted as the base in the design of the VLAM-G architecture. The Grid refers to an infrastructure that enables the integrated, collaborative use of high-end computers, networks, databases, and scientific instruments owned and managed by multiple organizations (Foster, Kesselman et al. 2002; Globus 2002). The VLAM-G platform takes advantage of the bag of services offered by Globus toolkit, which is the *de facto* standard in terms of Grid implementation options. Besides the Grid concept, other paradigms and approaches being applied within the VLAM-G architecture include the federated/distributed information management, and the Virtual Organization (VO) (see (Sheth and Larson 1990), (Afsarmanesh, Kaletas et al. 2001)). Proper integration of these technologies represents one of the main novel issues considered within the VLAM-G architecture.

An overview of VLAM-G multi-tier architecture is given in Figure 1. At the highest level, the application tier consists of different end-user application cases that are being developed as "VLAM-G proof of concepts", such as DNA array and Material Analysis of Complex Surfaces (MACS) application cases. At the application server tier, the VLAM-G *middleware* involves the following functional components: VIMCO (Virtual-Laboratory Information Management for Co-Operation), COMCOL (Communication & Collaboration), ViSE (Virtual Simulation and Exploration) and the Run Time System (RTS) (Afsarmanesh, Belleman et al. 2002). In brief, COMCOL provides facilities for communications and remote control of devices. VISE provides simulation facilities and a virtual reality environment for visualization. Furthermore, the RTS is a Grid-based run-time environment aiming at scheduling, dispatching and executing the tasks comprising an experiment (Belloum, Hendrikse et al. 2001). Finally, the VIMCO component supports the advanced federated information management requirements of the VLAM-G applications (Afsarmanesh, Kaletas et al. 2001). For instance, it provides

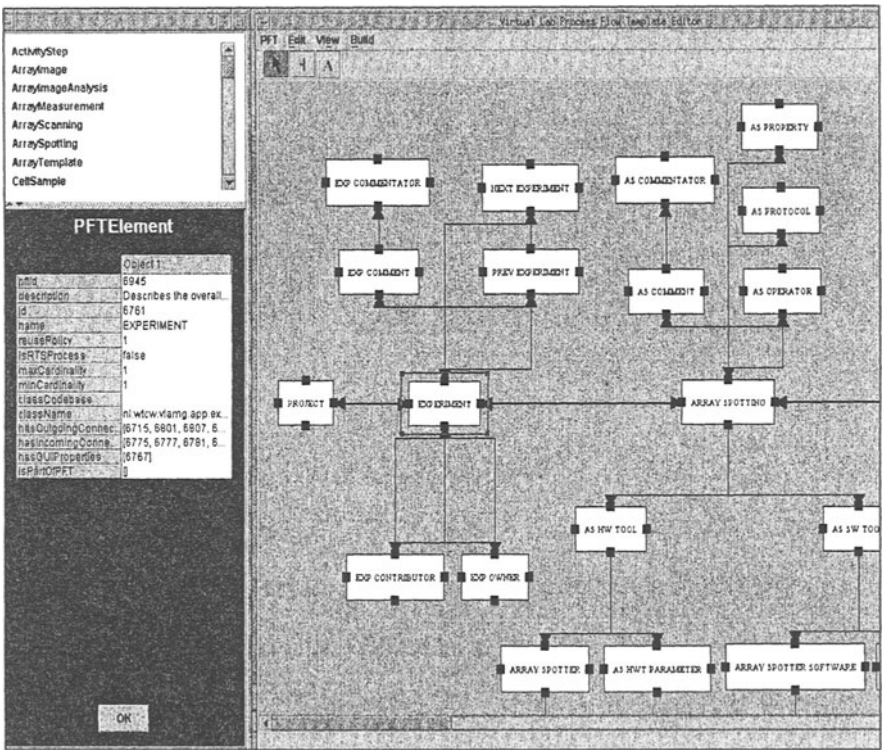

Figure 2. GUI for VLAM-G PFT editor.

access to databases of different application domains: MACS from Physics domain, and EXPRESSIVE for gene expressions. These databases store both the steps involved in experiments and the results of the experiments (Kaletas, Afsarmanesh et al. 2001).

Please also notice that since most scientific experiments contain a large number of steps, graphical user interfaces and tools need to be developed as part of the VLAM-G environment in order to assists scientists from different disciplines with the definition and execution of their experiments. One of these tools developed for VLAM-G is the Process Flow Template (PFT) Editor (Kaletas, Afsarmanesh et al. 2002), which is used by the administrator or an experienced scientist user from an application domain in order to define steps involved in a typical experiment. Figure 2 shows the layout of the PFT Editor prototype. For each VLAM-G application (e.g. MACS), a corresponding PFT is defined through this editor, and then it is stored in VIMCO to guide the experimenter throughout the experimental procedure (Kaletas, Afsarmanesh et al. 2002). Using this graphical interface, the user can also instantiate an experiment template by going step by step through the defined template, and providing the required information for each step into the forms that will be displayed.

During the execution of an experiment, the scientist may need to analyze the resulting data sets. This process may also include some steps requiring high computational power, distributed computing, a variety of analysis tools, etc. For this purpose, the user can again benefit from the utilities provided by Grid-based

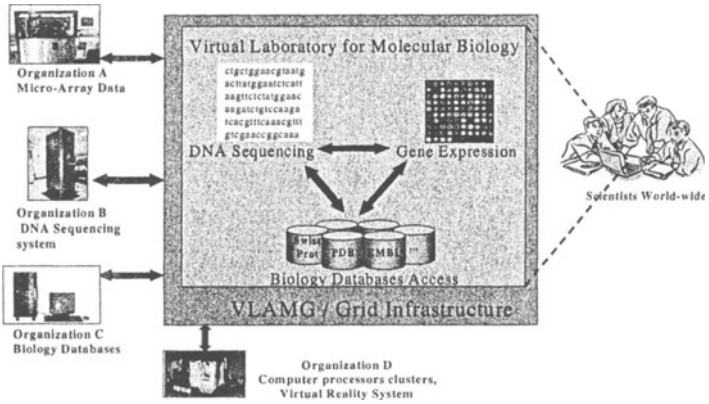

Figure 3. Analysis of Reference Experimentation Scenarios.

VLAM-G environment, since it is possible to define experiments including different analysis and visualization modules as part of the experiment itself.

Currently, a first prototype implementation of these user interface tools is available within VLAM-G and is being extended to be fully functional.

3. REFERENCE EXPERIMENTATION SCENARIOS

In this section, some preliminary reference scenarios are described supporting potential collaborative experiments among a group of scientists in molecular biology. Namely, the main goal here is to describe experimentation scenarios in the area of molecular biology specifically related to DNA sequence analysis, gene expression analysis, and/or biology database access. Please notice that these scenarios correspond to experiments that would be ultimately carried out by scientists from different organizations using the VLAM-G architecture and tools described in the previous section. Through VLAM-G, seamless and protected access to heterogeneous/distributed resources available from different locations can be offered to users that have specific needs (and restrictions on access rights) in order to perform a given experiment (see also Figure 3). Depending on the experimentation scenarios, shared resources may include data sources, I/O devices, computational infrastructure components, etc. In the case of the Flexwork project, shared resources are mostly related to molecular biology data, specialized bio-informatics services (e.g. blasting), and high-performance computational services.

In collaboration with partners within the Flexwork consortium, some real-case examples of potential experimentation scenarios have been defined. In order to better understand these scenarios, let us consider four major autonomous organizations (e.g. universities, research centers, private enterprises) that wish to share resources and collaborate through the virtual laboratory to perform advanced scientific experimentation.

The main role of each organization (or site) is described next (see also Figure 3):

- Organization A: generation, preparation and analysis of DNA-microarray data related to genome expression.
- Organization B: generation, preparation and analysis of biological data regarding DNA sequences and annotations.

- Organization C: provision of the proper data access interfaces to biology databases (e.g. genome, sequences, and protein databases) that are locally updated and maintained.
- Organization D: provision of database management for the integration of information from different sources, high-performance computing infrastructure, and advanced experiment simulation and visualization facilities.

Being a part of the virtual laboratory, scientists from these organizations, as well as other authorized scientists, can perform different experiments on the available resources in their environment. Please notice that the interrelations between the first three partner organizations form a triangle involving experiments encompassing the areas of DNA sequencing, gene expression and biology databases. Several individual experiments may address specific problems in each corner (area) of this triangle (e.g. gene expression alone), but it is also possible and natural to identify more complex situations in which scientists conduct experiments involving activities and resources from all three corners of the triangle. In this direction, several specific experimentation scenarios are defined for Flexwork and described below (for a more detailed description of a specific micro-array experiment model using the VLAM-G approach, see (Kaletas, Afsarmanesh et al. 2001)).

Scenario Case 1: Sequencing experiment in Organization B

A research project in Organization B with the aim to identify certain genes in plant species X has revealed a number of candidate genes. These candidate genes have been isolated and have been sequenced. Researchers now have the task to further characterize the genes and would make use of services available in sites A, C and D. As a first step, researchers try to find genes similar to the ones under study in species X and other species in order to deduce their function. Blast searches are carried out in Site C against public nucleotide and protein sequence databases (e.g. EMBL and SWISS-PROT). As next step, information about expression patterns is gathered, either in Site C or perhaps in A. Microarray databases containing results from gene expression studies are searched to check whether expression information is available for one or more of the genes. With this information, the researchers go back to Site C to search cellular pathway databases to identify functions for the gene products. With the acquired information, experiments are designed to test whether the gene functions identified *in silico* reflect the true biological functions.

In summary, the main experiment steps are represented in Figure 4. Please notice that DNA sequencing block is in fact composed of internal subblocks that are connected forming a subexperiment on their own. This kind of "experiment composition" is also possible following the VLAM-G experiment modeling and execution approach.

Scenario Case 2: Gene Expression Analysis in Organization B

In this case, let us suppose that microarrays at Site B have been produced from cDNA libraries of species X, and experiments have revealed a number of interesting genes that are uniquely expressed under a certain condition. Attempts are made to identify the functions of these genes in a similar fashion described under Case 1. Furthermore, the biologist needs to obtain detailed information about proteins related to the specific gene, for which he/she uses a "block" offered by Organization C. Finally, he/she also needs to analyze and visualize the DNA sequence of a

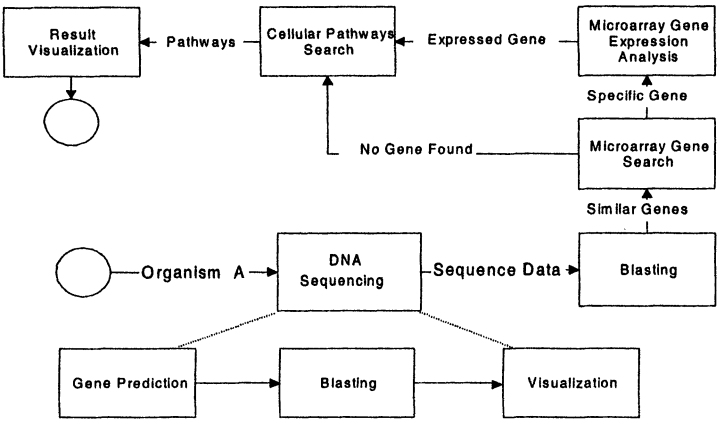

Figure 4. Extended DNA sequence analysis scenario.

specific gene (accessing DNA sequence information from Site A). The main steps involved in this scenario are depicted in Figure 5.

Scenario Case 3: Access to specialized databases in Organization C

Biological databases often contain information derived from computer analysis and that has not been tested experimentally. Thus, in this case specialized databases are set up, which link data available in Site C to a number of other databases available not only at the same location (e.g. nucleotide sequence, protein sequence, and cellular pathway databases), but also with databases containing experimental databases, such as microarray databases in Site A or databases containing experimental annotations in Site B. Access to these "extended" databases can be modeled as a value-added data provision service as depicted in Figure 6. In this way, researchers accessing databases in Site C would have transparent access to databases from Sites A or B making data acquisition extremely convenient. Please notice that the objective of each of the scenario diagrams described in this section is to ultimately model them using the PFT tool described previously, and then to execute them on the Grid-based VLAM-G platform. As such, these scenario diagrams provide a good reference for the coming definition of the final experiments.

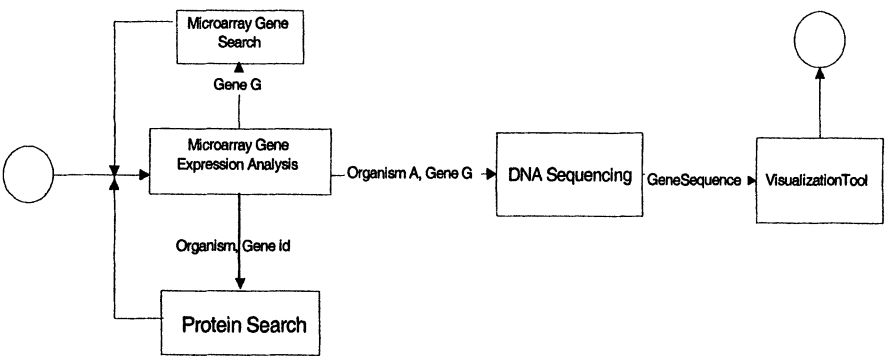

Figure 5. Gene expression analysis scenario at Organization B.

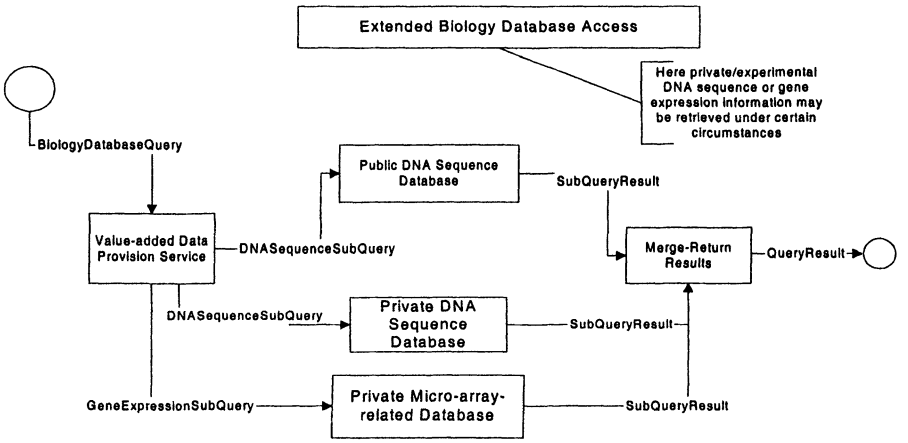

Figure 6. Extended biology database access scenario.

4. CURRENT DEVELOPMENTS AND FUTURE DIRECTIONS

In this section, the current status of development activities regarding the application of VLAM-G within the Flexwork molecular biology domain is described, as well as some possibilities for future development directions.

Since VLAM-G provides seamless high-performance computing facilities, distributed data integration and collaboration environment, it has been chosen as the base infrastructure for the Flexwork environment. Thus, all modules to be developed for Flexwork will be integrated into VLAM-G environment. Also since re-usability and modularity are among the key properties of VLAM-G, this integration can be easily achieved. Currently, the analysis, design and development efforts regarding the application of the existing VLAM-G prototype within the specific Flexwork domain have focused on:

- Information management requirements analysis. A requirement analysis and preliminary information modeling phase has been carried out for the distributed information management functionalities needed to support the experimentation scenarios described in Section 4. In brief, the identified requirements include the definition of an integrated database schema providing seamless access to distributed data sources, data modeling for the micro-array experiment (see (Kaletas, Afsarmanesh et al. 2001)), data modeling for DNA sequence analysis, and data modeling for integration of biology databases. Due to space limitations, the detailed description of the specific data models are outside the scope of this paper and constitute the subject of future publications.
- Adaptation, design and development of Distributed Annotation System (DAS) (DAS 2002). In order to establish a flexible and interoperable data integration framework, one of the main issues being considered in Flexwork is the use of available standards in molecular biology domain. For this reason, the Distributed Annotation System (DAS) is applied for modeling DNA sequence

and annotation data. A prototype of a DAS server and web-based clients have already been implemented for Flexwork.

- EXPRESSIVE database. In the same way that DAS is used for sequencing information, the EXPRESSIVE data model has been developed and supported by VLAM-G in order to represent and manage the gene expression information related to micro-array experiments (Kaletas, Afsarmanesh et al. 2001).
- Provision of a high-performance BLAST service. As mentioned in a previous section, biologists need to use modules or tools to find similar matches for protein or nucleotide sequences using techniques such as BLAST. BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs that compare a protein or nucleotide sequence query within a database of known sequences. In this respect, a version of a BLAST service has been set up in Amsterdam, on a powerful machine with Grid installed on it to demonstrate the feasibility of this kind of experiments and how it can be accessed from the virtual laboratory by external users. When compared to other existing BLAST implementations, this version runs much faster using the high performance computing facilities provided by Grid.

Among the plans for future VLAM-G related developments within Flexwork we can mention: the incorporation of virtual collaboration among different organizations, federated database framework to also support the proprietary data and information access rights, full integration of DAS system and BLAST services into VLAM-G, and provision of integrated access to biology databases. These points are briefly addressed in the paragraphs below.

Nowadays, within the molecular biology domain there is an increasing demand among organizations to collaborate and share information and resources with other partners. In order to achieve this global collaboration in a proper way, all the rules for each kind of partner relationship must be defined, and contracts must be prepared among participating organizations. Furthermore, access rights and visibility levels on local resources need to be clearly defined based on the established contracts. Being an emerging paradigm, the concept of Virtual Organizations best suits these requirements (Garita 2001), (Afsarmanesh, Kaletas et al. 2001). Therefore, it is important to explicitly support VO life-cycle functionalities within VLAM-G and this will be one of the research directions to follow in the future. The necessary components for the support of VO in VLAM-G include for instance a catalogue to store information related to VOs and VO partners, and a federated information management framework (Kaletas, Afsarmanesh et al. 2002).

Furthermore, the development of the DAS system described previously is complete, but it runs as a standalone server. Since the aim is to use VLAM-G as the base infrastructure, the DAS system needs to be fully integrated into the VLAM-G environment. It is also necessary to inter-link sequence information with the EXPRESSIVE gene expression information, in order to enable scientists carrying out their experiments as defined in the example scenarios. Similarly, the BLAST module that has been developed also needs to be integrated into VLAM-G.

Finally, as mentioned before, scientists need to reach biology databases (e.g. PDB, EMBL) at some point within their experiments. In fact, the BLAST module also runs on copies of these large molecular databases. In order to access these databases, there must be a single interface provided to the scientist so that the

complexity of accessing different databases remains hidden from him. Thus, another next step in Flexwork is to provide an integrated access to these databases through VLAM-G environment.

5. CONCLUSIONS

This paper described how the virtual laboratory framework offered by the VLAM-G infrastructure can be effectively applied to support advanced scientific experimentations in the specific field of molecular biology. Namely, through the VLAM-G infrastructure, it is possible for different organizations and scientists world-wide to collaborate and share resources to carry out complex experiments in this application domain. The presented experimentation scenarios have been defined within a multi-disciplinary team involving both biologists and computer scientists. These scenarios are crucial as reference points for these scientists in order to have a common understanding of the tasks that are required to be supported within the VLAM-G framework. Finally, the paper also presented a brief summary of the main VLAM-G-related development activities that have been carried out within Flexwork, and provided several possible directions in terms of extensions and future work.

6. REFERENCES

1. Afsarmanesh, H., R. Belleman, et al. (2002). "VLAM-G: A Grid Based Virtual Laboratory." Scientific Programming Journal Special Issue on Grid Computing **10**(2): 173-181.
2. Afsarmanesh, H., E. Kaletas, et al. (2001). "A Reference Architecture for Scientific Virtual Laboratories." Future Generation Computer Systems **17**(8): 999-1008.
3. Afsarmanesh, H., E. C. Kaletas, et al. (2001). The Potential of Grid, Virtual Laboratories and Virtual Organizations for Bio-sciences. 28th Conference on Current Trends in Theory and Practice of Informatics, SOFSEM 2001, Piestany, Slovakia.
4. Belloum, A. S. Z., Z. W. Hendrikse, et al. (2001). The VL Abstract Machine: A Data and Process Handling System on the Grid. High Performance Computing and Networking (HPCN) Europe, Amsterdam, The Netherlands.
5. DAS (2002). A Distributed Annotation System (DAS), http://www.biodas.org.
6. Foster, I., C. Kesselman, et al. (2002). The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration.
7. Garita, C. (2001). Federated Information Management for Virtual Enterprises. Informatics Institute. Amsterdam, The Netherlands, University of Amsterdam, Doctoral Thesis.
8. Globus (2002). The Globus Project, http://www.globus.org/
9. Kaletas, E. C., H. Afsarmanesh, et al. (2001). EXPRESSIVE - A database for micro-array gene expression studies. Amsterdam, University of Amsterdam, Informatics Intitute.
10. Kaletas, E. C., H. Afsarmanesh, et al. (2002). Virtual Laboratories and Virtual Organizations Supporting Biosciences. 3rd IFIP Working Conference on Infrastructures for Virtual Enterprises ProVE'02, Algarve, Portugal.
11. Kaletas, E. C., H. Afsarmanesh, et al. (2001). Virtual Laboratory Experimentation Environment Data Model. Amsterdam, University of Amsterdam, Informatics Intitute.
12. Setubal, J. and J. Meidanis (1997). Introduction to computational molecular biology, Brooks/cole Publishing Company.
13. Sheth, A. and J. Larson (1990). "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases." ACM Computing Surveys **22**(3): 183-236.