VIRTUAL LABORATORY MOLDYNGRID AS A PART OF SCIENTIFIC INFRASTRUCTURE FOR BIOMOLECULAR SIMULATIONS *

Andrii Salnikov [1], Ievgen Sliusar [1], Oleksandr Sudakov [1], Oleksandr Savytskyi [2], Alexander Kornelyuk [2]

[1] Information and Computer Center, National Taras Shevchenko University of Kyiv, Glushkova prosp., 2, Kyiv, Ukraine, cluster@cluster.kiev.ua, http://grid.org.ua
[2] Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine, Acad. Zabolotny Str., 150, Kyiv-03143, Ukraine, kornelyuk@imbg.org.ua, http://imbg.org.ua

Abstract: *Virtual laboratory MolDynGrid has been established for computational researches in structural biology and bioinformatics, especially for molecular dynamics (MD) simulations of biological macromolecules and their complexes. Biomolecular simulation usually requires a very high processing power and huge storage space for MD trajectories. In this study we describe the development of virtual laboratory infrastructure to run massive MD simulations on a Grid infrastructure in a distributed manner. The aim of MolDynGrid development was to provide an efficient infrastructure for automation of MD job processing in Grid. It will assist researchers to use grid technology without having them to be familiar with highly complicated internals of the Grid. Web-portal of MolDynGrid virtual laboratory has been developed (http://moldyngrid.org), which consists of computational, analytical and educational blocks and data base of MD simulations.*

Keywords: *grid, virtual laboratory, molecular dynamics, GROMACS, web-portal.*

1. INTRODUCTION

Molecular dynamics simulation is very important approach of structural bioinformatics because it makes possible modeling of the dynamical behaviour of proteins and their complexes in solution [1-3]. However, MD simulations are computationally very expensive and usually MD research requires a high processing power and huge storage space for trajectories. Grid computing is a very promising approach in order to encounter the limits of computational power for MD simulations [4]. Last time, biology and medicine are among the main directions, chosen for developing and implementing of grid technologies [5-7]. First of all, it concerns a development of novel drugs for medicine and creating databases of hereditary diseases. As an example, the BioinfoGRID project combines bioinformatics services and applications for molecular biology users with the Grid infrastructure of EGEE [8]. Grid-medicine is IT-infrastructure containing a specialized computer service, adapted for solving problems of processing biomedical data. Accordingly, Grid-medicine resources are computer resources, like specialized bases of medical data, medical devices and complexes [7].

The Ukrainian Academic Grid (UAG) has been implemented and functioning since 2005. The development of the first Grid segment in Ukraine, and in perspective a national-wide Grid network, give an opportunity of successful integration into the international scientific projects in Europe and over the world.

In the postgenomic era of modern biology, much attention is focused on the 3D structures and conformational flexibility of proteins which are novel targets for computational drug design [2,3]. Standard molecular dynamics simulations of biomolecular systems at full atomic resolution are typically limited to the nanosecond timescale and therefore need more powerful hardware and software systems. Common file sizes of MD simulation results may exceed 250 GBytes, computing time of single MD trajectory in nanosecond range usually exceed several month on one CPU.

Application of Grid technologies to MD simulations provides significant benefits of distributed computing and possibility to integrate

---

data storage and processing capabilities. However, the end users lack publicly available user-friendly tools for MD simulations that make application of grid middleware completely transparent. Web-interfaces are good candidates for such tools but available systems usually have limited features sets. For instance, BioSimGrid [9] is a powerful web-portal, but designed only as database of MD trajectories.

In this paper we describe the development of MolDynGrid virtual laboratory which was established for research in computational biology and bioinformatics that requires high processing power and huge storage space. Web-portal of MolDynGrid virtual laboratory has been developed (http://moldyngrid.org), which is composed of computational, analytical and educational blocks and data base of MD simulations. Web-portal contains a complete set of tools for MD simulation, storage and data analysis in grid environment. Virtual laboratory MolDynGrid is open for world-wide Internet access and efficiently used by scientists in Ukraine and other countries and will be integrated into the international research Grid-infrastructure.

2. MD SIMULATIONS BACKGROUND

Simulation of molecular dynamics is performed by solution of classical or semi-empirical non-linear differential equations of motion for atoms:

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{f}_i\left(\vec{r}_1, ..., \vec{r}_i, ..., \vec{r}_N\right)$$

$$\vec{f}_i\left(\vec{r}_1, ..., \vec{r}_i, ..., \vec{r}_N\right) = -\frac{\partial U\left(\vec{r}_1, ..., \vec{r}_i, ..., \vec{r}_N\right)}{\partial \vec{r}_i},$$

where N – total number of atoms, $\vec{r}_i$ – coordinates of atoms, m_i – atom masses, $U\left(\vec{r}_1, ..., \vec{r}_i, ..., \vec{r}_N\right)$ – potential energy values, $\vec{f}_i$ – force acting at atom, t – time. These equations are solved numerically with step $\sim 10^{-18}$ s in time, and results are recorded with step $\sim 10^{-15}$ s. Numerical integration algorithm scales in time as $O(N_t)$, where N_t – total number of time steps. Potential energy $U\left(\vec{r}_1, ..., \vec{r}_i, ..., \vec{r}_N\right)$ depends on the distances $\left|\vec{r}_i - \vec{r}_j\right|$ between atoms and in the simplest case scales as $O\left(N^{1-2}\right)$. Total number of computing operations depends on the precision of potential approximation and scales as $O\left(N_t N^{1-2}\right)$. Taking into account the total number of atoms in macromolecules, it is easy to estimate the total number of computing operations for MD simulations – 10^{10-16} per 1 ns of dynamics. Total computing time may take several days or even months on one CPU. Data files grow to ~ 2.5 Gbyte per 1 ns of MD simulation.

Simulation data analysis is usually performed by filtering of trajectory data ($O(MN) - O(M \log N)$ operations), correlation analysis of trajectories ($O\left(M^2 N^2\right)$ operations), principal components analysis ($O\left(M^2 N^3\right)$ operations), building animations ($O(MN)$ operations), etc. These tasks requires from several minutes to several days of CPU time. Many application require simultaneous procession of several ($\sim 10^2$) trajectories to take into account influence of temperature, initial conditions, statistically significant amount of data.

Most of described tasks may be performed in parallel by distribution of data between processors. Computing of several trajectories and their analysis do not interfere with each other and thus can be performed independently on different grid clusters. Parallelization of MD trajectory computing may be performed in shared memory very efficiently. In distributed memory, the communication between processes ($O\left(p^{1-2}\right)$, where p – CPUs number) limits the parallelizing efficiency to about 2-4 nodes and significantly depends on latency of underline communication infrastructure.

Grid technologies provide cooperation of data storage and computing resources and proved to be rather efficient for MD simulations. Thus most stages of simulation may be performed in the background transparently to the end user. Personal human interference is usually required at stage of setup, monitoring and results examination. All interactive tasks may be preformed via web-interface.

3. VIRTUAL LABORATORY CONCEPTS

MolDynGrid virtual laboratory was established for interdisciplinary researches in computational structural biology and bioinformatics. In particular, this involves molecular dynamics simulations of biological macromolecules such as proteins, nucleic acids and their complexes. It is also important to compare MD simulations results with experimental NMR spectroscopy data on proteins dynamics. MolDynGrid virtual laboratory was established as a part of Ukrainian Academic Grid infrastructure (UAG).

One of the main aims of virtual laboratory is to provide an efficient utilization of Grid infrastructure for automation of MD job processing. It will assist biologists in using Grid for their needs without requiring them to be familiar with highly complicated IT internals.

Characteristics of resources needed for molecular dynamics calculations fit well with concepts of the

Grid. Historically, Grid has grown on top of the Internet, which is now available world-wide. The Internet itself began as information sharing tool for use by limited number of research organizations. Grid comes one step next in resource sharing – not only information can be shared, but also processing power, software, some devices and data storage capacities.

4. DISTRIBUTED FILE STORAGE SYSTEM

At the beginning of the development of Ukrainian Academic Grid infrastructure, only processing power and software sharing had been implemented. This was done by means of Nordugrid ARC Middleware. The next step was the implementation of data storage space sharing that was required for successful deployment of MD calculations in the Grid. So, we have implemented the storage infrastructure between clusters that support the MolDynGrid virtual organization [10].

Implementation was directed by users requirements. The users wanted to access storage resources from a single point and didn't care which particular storage element holds their data. An index is required to accomplish this task. When some data is going to be stored, a process of finding needed space should not require user intervention, still providing a single entry point to the user. File access policies should be the same no matter what particular storage element is used. User's personal certificate DN and VO membership are used to identify him on the Grid. So, access permissions must be set on the basis of this attributes. Of course, comfortable user interface should be available for operations with data independently of computational job management.

The first standard solution was the Replica Location Service provided by Globus Toolkit. It implements a data location index which acts as an entry point for automated storage operations. We had also chosen ARC's Smart Storage Element service instead of GridFTP as local storage agent. It supports setting access rights for each individual file and standard secure HTTP protocols. Comfortable user interface was implemented ourselves as a plug-in module for the Midnight Commander, a full-featured file manager for UNIX systems. Distributed storage operation scheme is shown at Fig. 1.

Development of user interface which provides access to distributed storage began with implementation of "stels" – a shell wrapper for ARC

file listing tool which transforms flat line file listing into a directory-tree like output. We simply parse all slashes and consider a part of name between slashes as a directory name, than we build file tree and show it the way like "ls" for local filesystem. But using command line still is not comfortable for users, especially for biology scientists that do not have experience working with Linux shell. So, we developed extended filesystem plugin for MC based on our "stels" wrapper.

There are several advantages of choosing MC extfs plugin as UI implementation:
- Easy to install – all that you need is to copy our script to plugins directory of MC and add plugin name to the list;
- After installation every user may use the plugin by creating an empty file with name in format <URL of RLS server> dot rls;
- Just pressing "Enter" on this file will provide an access to RLS, representing content as a directory tree ready to navigate in familiar way. You are free to use standard combinations like F5 to copy or F8 to remove files [11].

5. MOLDYNSUB_CLI

Several interfaces have been created with the aim of automated MD computation job management, from the command line tools to full-functioning web-portal. Our researches are based on MD simulations using GROMACS package [12]. Full cycle of computation consists of several steps: ion generation, energy minimization, position restraints and molecular dynamics. Each step require user to run a chain of programs provided by GROMACS package. We want to utilize a Grid computing power for these calculations, therefore a Grid job description is also required. To accomplish this, an automation script "moldynsub_CLI" providing step by step GROMACS cycle execution in Grid had been written and implemented here. All the parameters needed for different GROMACS applications and general Grid job description are gathered into one common file "params", which is used by the automation script. The script operates the following algorithm: it calls Grid tool for user authorization, if proxy certificate was not found or expired. The script generates default parameter description file, if it does not exist, and submits full-cycle GROMACS MD computation to Grid, if all required files exist in the current directory, or submits MD computation from checkpoint, if checkpoint file found.
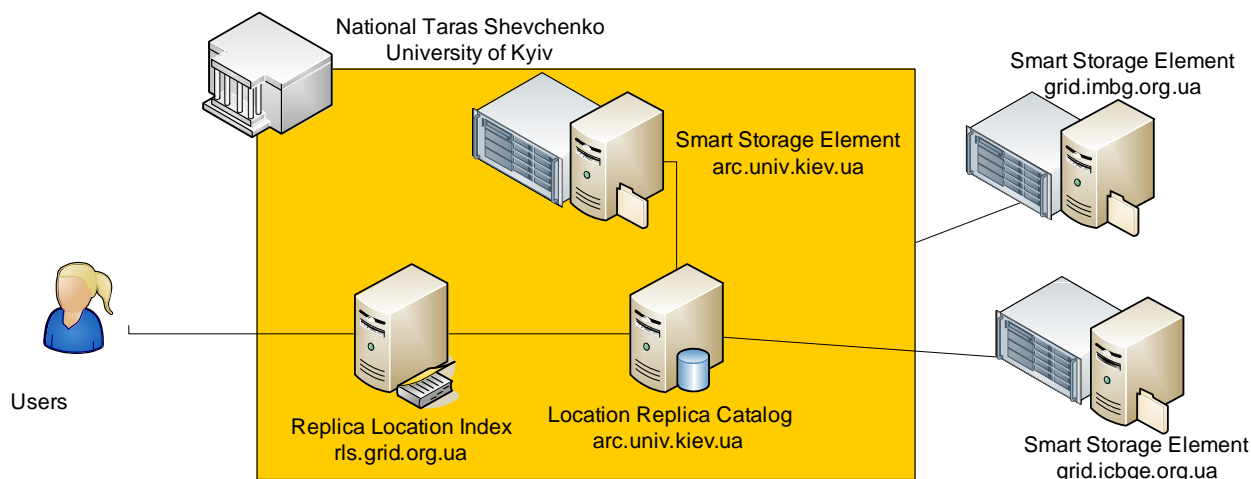
Fig. 1. – Distributed storage system of MolDynGrid virtual laboratory

6. WEB-PORTAL OF MOLDYNGRID VIRTUAL LABORATORY

Automation script and Midnight Commander plug-in were widely used in MolDynGrid virtual laboratory, while we came one step next in providing MD automation interfaces. The aim of web-portal development was the creating of web-based interfaces that eliminate usage of command line at cluster and provide access to virtual laboratory resources. In order to start working with MolDynGrid, the scientist just needs to open web-browser with personal certificate installed and have his credentials already delegated to MyProxy server. Next step is the navigation to the URL: http://moldyngrid.org. The web-page of MolDynGrid portal is available in Ukrainian now and its screenshot is depicted on Fig. 2.

The MolDynGrid web-portal is functioning on web-server with Apache backend running on Fedora Linux. All the modules use PHP CGI, without any Java applets. This is much faster solution and does not require installing Java JRE and application server like Tomcat on the server-side unlike existing web-portal solution based on GridSphere. Grid middleware UI required on web server.

MolDynGrid web-portal is divided into several blocks:

- Information block – contains documentation about portal and several virtual lab tools usage instructions;
- Computational block – provides interface to MDP parameters customization and job submission;
- MD simulations trajectories database – provides interface to previously calculated trajectories located on distributed storage.

Computational block is composed of two parts: parameter specification interface and job submission script. All users have an ability to save different profiles of parameters for job submission, modify their values and also specify job and trajectory names, mailing options and other grid parameters. Submission script is based on moldynsub_CLI tool which was developed earlier. To submit a job, a web-portal needs a user delegation, which is requested from MyProxy service.

Fig. 2. – Web-portal of MolDynGrid virtual laboratory

Information block contents are available to all guests directly from the home page. Access to other blocks providing interfaces to laboratory resources, is restricted. To gain access, it is nesessary to have a valid certificate signed by one of the Grid CA and to be recognized as a member of MolDynGrid VO [13]. Authorization procedure is controlled by "Authorization Gateway". Certificate validation implemented on Apache level, requesting client

certificate from browser. Certificate DN is used by CGI script to verify VO membership.

Already submitted and computed tasks control system implemented in management block are shown on Fig. 3. Output and error log listing is available for job owners. Also abilities to stop, cancel and rerun computational job are provided.

Ім'я задачі ▲ ▼	Ідентифікатор завдання Грід ▲ ▼	Дата розрахунку ▲ ▼	Автор ▲ ▼	Стан завдання ▲ ▼	Керування
htyrn_g41r	gsiftp://nordu.hpcc.ntu-kpi.kiev.ua:2811 /jobs/17537127135728387640 0793	2010-04-15 21:48:03	Oleksandr Savytskyi	DELETED	
htyrn_wild	gsiftp://nordu.hpcc.ntu-kpi.kiev.ua:2811 /jobs/12166127135727611612 38388	2010-04-15 21:47:56	Oleksandr Savytskyi	INLRMS: R	
htyrn_wild	gsiftp://arc.univ.kiev.ua:21 /job/29096126934913278824 4605	2010-04-15 21:47:49	Oleksandr Savytskyi	INLRMS: R	
htyrn_g41r	gsiftp://nordu.hpcc.ntu-kpi.kiev.ua:2811 /jobs/99021271357264186792 3962	2010-04-15 21:47:43	Oleksandr Savytskyi	INLRMS: R	
htyrn_g41r	gsiftp://nordu.hpcc.ntu-kpi.kiev.ua:2811 /jobs/11788127134853286291 3714	2010-04-15 19:22:10	Oleksandr Savytskyi	INLRMS: R	
htyrn_e196k	gsiftp://arc.univ.kiev.ua:21 /job/13469126816991011845 77548	2010-04-14 01:26:04	Andrii Salnikov	DELETED	

Fig. 3. – Tasks monitoring and management system

Trajectory database was created using MySQL backend. Interface to the database is available from restricted portal area. This interface provides information about available trajectories and properties (parameters of simulation) for each trajectory. When a new job is submitted from the Computational block, a new record in the database is created.

All this components, interoperating together, works as a single entity – MolDynGrid virtual laboratory portal, that make all MD simulations easy in the Grid.

7. FIRST APPLICATION RESULTS

Now MD simulations were performed for several proteins, such as HIV-1 protease, mammalian and eubacterial tyrosyl-tRNA synthetases, isolated functional N- and C-terminal domains of mammalian tyrosyl-tRNA synthetase, transfer RNAs and specific TyrRS–tRNA complexes [13].

Plumed plugin which was installed on GROMACS 4 was useful for the investigation results described above for free energy calculations with MD simulations [15].

All computed MD trajectories are located on the distributed file storage system in Grid and accessible for the future analysis. The resources of MolDynGrid are available free of charge for registered members upon observance of usage rules. Information about MolDynGrid is available from the web-site http://moldyngrid.org [16].

At the moment the Ukrainian Grid infrastructure (http://grid.org.ua/monitors) used for calculations consists of 26 clusters with about 2500 CPUs

available for Grid. MolDynGrid usually utilizes computing elements (CE) of 5 clusters (IMBG, KNU, ISMA, KPI, MAO) and storage elements (SE) of 2 clusters that correspond to 900 CPUs and 6 TBytes of disk space, respectively.
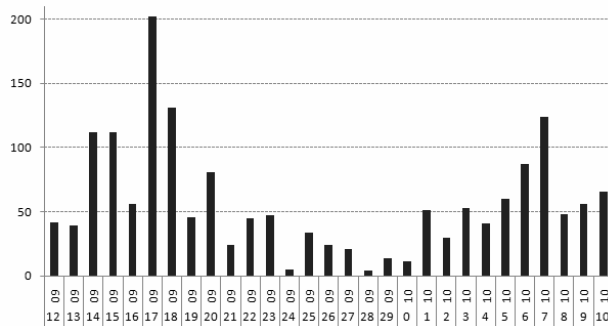

Fig. 4. – MolDynGrid computed tasks

Corresponding to visits statistics, the MolDynGrid web-portal has visitors from 9 different countries. The average number of daily visits is about 20 per day. Since 2009, about 2500 MD trajectories were successively computed using MolDynGrid resources. Figure 4 shows a MolDynGrid web-portal statistics and illustrates count of computed tasks per week for last year. The average utilization of computing resources by MD simulations is about 60 CPUs per day. Values are dependent on free resource in clusters, which available on that time and much of tasks can be in a queue status.
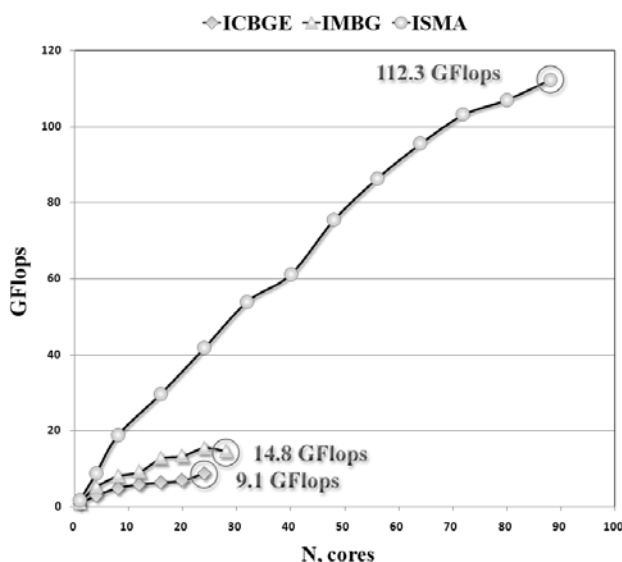

Fig. 5. – GROMACS 4 benchmarks on clusters with different communication subsystems

It is known that the different clusters of Ukrainian Grid have different communication subsystems [17]. Figure 5 illustrates the much higher efficiency of MD calculations when the ISMA

cluster with Infiniband communication was used in comparison with the IMBG and ICBGE clusters with 1 Gb Ethernet interconnection. Heterogeneity of parallel computing system significantly reduces parallelizing efficiency [18]. Nevertheless in grid-computing parallelizing is performed inside computing clusters and communication between different jobs is absent, so the total performance of grid is equal to the sum of clusters performances.

Grid provides more benefits for MD besides performance increase. This involves more accurate models, longer trajectories, analysis of statistically meaningful ensemble of trajectories ($\sim 10^{2-3}$), searching the most stable molecules conformations and "scientifically meaningful" trajectories that demonstrate special molecules features. Such simulations require several months on single cluster and may be performed in Grid for several days. Some results of IMBG cluster and Ukrainian Grid application for MD simulations is described in Table 1.

Table 1. MD simulations characteristics

Characteristic	Grid	IMBG cluster
CPUs	900	32
Storage, TBytes	6	0.5
MD trajectory duration, ns	100	10
Total number of trajectories since 2008	2500	63
"scientifically meaningful" trajectories	74	6

8. CONCLUSIONS

MolDynGrid virtual laboratory was successfully established and developed for molecular dynamics simulations in Grid and now provides the powerful facilities for solving of resource consuming problems of biomolecular modeling. The efficiency of MolDynGrid was proved by application to different problems, such as MD calculations of some proteins and nucleic acids. It is important to note that the resources of our virtual laboratory may be used also by biologists that are not familiar with low and middle level of Grid functionality.

A future development of MolDynGrid virtual laboratory which includes the implementation of analytical block with the interfaces to the most important algorithms, and integration with gLite middleware [14, 17] is now in progress.

9. ACKNOWLEDGMENT

10. REFERENCES

[1] J. Norberg, L. Nilsson. Advances in biomolecular simulations: methodology and recent applications. *Quot. Rev. Biophys.* (36) (2003). pp. 257-306.

[2] M. Karplus, J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology* (9) (2002). pp. 646-652.

[3] D. Kovalskyy, V. Dubyna, A. E. Mark, A. Kornelyuk. A molecular dynamics study of the structural stability of HIV-1 protease under physiological conditions: The role of Na+ ions in stabilizing the active site. *Proteins* (58) (2005). pp. 450-458.

[4] F. Gagliardi, B. Jones, F. Grey, M. E. Bégin, M. Heikkurinen. Building an infrastructure for scientific Grid computing: status and goals of the EGEE project. *Philos Transact A Math Phys. Eng. Sci.* 363 (1833) (2005). pp. 1729-1742.

[5] L. Milanesi, I. Merelli. High performance Grid based implementation for genomics and protein analysis. *Stud Health Technol Inform.* (120) (2006). pp. 374-380.

[6] B. G. Kim, N. D. Nhan, S. Lee, S. Hwang, B. Breton. DrugScreener-G: towards an integrated environment for Grid-enabled large-scale virtual screening and drug discovery. *Fourth IEEE International Conference on eScience (2008).* pp. 666-671.

[7] A. I. Kornelyuk, O. P. Mintser. Modern computational Grid-technologies and their applications in medical researches. *Medical Informatics and Engineering* 1 (1) (2008). pp. 23-29.

[8] L. Milanesi, I. Merelli. High performance GRID based implementation for genomics and protein analysis. *Stud. Health Technol. Inform.* (120) (2006). pp. 374-380.

[9] B. Wu. A Web. Grid portal implementation of BioSimGrid: a biomolecular simulation database. B. Wu, D. Matthew, et al. http://eprints.soton.ac.uk/22809/01/Wu_04.pdf: *University of Southampton and University of Oxford*, 2004.-March 31.

[10] M. Zynovyev, S. Svistunov, O. Sudakov, Yu. Boyko. Ukrainian Grid infrastructure:

practical experience. *Proceedings of 4-th IEEE Workshop IDAACS 2007*, Dortmund, Germany 6-8 September 2007, pp. 165-169.

[11] A. O. Salnikov, I. A. Sliusar, O. O. Sudakov, Yu. V. Boyko, A. I. Kornelyuk. Implementing the file storage system in the ukrainian academic Grid infrastructure. *Abstracts of 21th International CODATA Conference "Scientific Information for Society-from Today to the Future,* Kyiv, Ukraine 5-8 October 2008, p. 31.

[12] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. Berendsen. GROMACS: fast, flexible and free. *J. Comput. Chem.* (26) (2005). pp. 1701-1718.

[13] A. O. Salnikov, I. A. Sliusar, O. O. Sudakov, Yu. V. Boyko. Ukrainian GRID-segment authorization system. *Proceedings, III International Conference "Electronics and Applied Physics",* Kyiv, Ukraine 25-27 October 2007, p. 163.

[14] gLite 3.1 home page, 2008. http://glite.org

[15] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia and M. Parrinello. PLUMED: a portable plugin for free energy calculations with molecular dynamics. *Comp. Phys. Comm.* (180) (2009). p.1961.

[16] A. O. Salnikov, I. A. Sliusar, O. O. Sudakov, O. V. Savytskyi, A. I. Kornelyuk. MolDynGrid virtual laboratory as a part of Ukrainian academic Grid infrastructure. *Proceedings of the 5-th IEEE Workshop IDAACS 2009*, Rende (Cosenza), Italy 21-23 September 2009, pp. 237-240.

[17] Ukrainian academic Grid community, 2010. http://grid.bitp.kiev.ua

[18] O. O. Sudakov, Yu. V. Boyko, T. V. Necheporuk, T. P. Korotkova. Performance optimization of computing cluster with weakly interconnected components. *Mathematical Machines and Systems.* – 2004. – N 4. – P. 57 – 65.

Andrii O. Salnikov, *master of science in applied physics, Ph.D. student of Radiophysics Faculty, engineer of Parallel Computing Laboratory at Information and Computer Center of National Taras Shevchenko University of Kyiv. Scientific interests: grid and high performance computing.*

Ievgen A. Sliusar, *master of science in applied physics, Ph.D. student of Radiophysics Faculty, engineer of Parallel Computing Laboratory at Information and Computer Center of National Taras Shevchenko University of Kyiv.*

Scientific interests: grid and high performance computing.

Oleksandr O. Sudakov, *candidate of physical and mathematical sciencies (Ph.D.). Associate Professor of Medical Radiophysics Department, Head of Parallel Computing Laboratory at Information and computer Center National Taras Shevchenko University of Kyiv. Graduated from Radiophysics Faculty Taras Shevchenko Kyiv University in 1996. Ph.D. thesis in Procession of magnetic resonance tomography signals in 2002. Scientific interests: high performance computing, physical processes in biological systems.*

Oleksandr V. Savytskyi, *master of science in biology, Ph.D. student and engineer of Protein Engineering and Bioinformatics Department at the Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine.*

Scientific interests: computational biology, grid and high performance computing.

Alexander I. Kornelyuk, *doctor of sciences, professor. Head of Department of Protein Engineering and Bioinformatics at the Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine. Scientific interests: protein engineering, computational biology, bioinformatics, protein dynamics, grid in biology and medicine.*