Grid Computing in Life Science

First International Workshop on Life Science Grid, LSGRID 2004
Kanazawa, Japan, May 31 – June 1, 2004
Revised Selected and Invited Papers
Preface

Researchers in the field of life sciences rely increasingly on information technology to extract and manage relevant knowledge. The complex computational and data management needs of life science research make Grid technologies an attractive support solution. However, many important issues must be addressed before the Life Science Grid becomes commonplace.

The 1st International Life Science Grid Workshop (LSGRID 2004) was held in Kanazawa Japan, May 31–June 1, 2004. This workshop focused on life science applications of grid systems especially for bionetwork research and systems biology which require heterogeneous data integration from genome to phenome, mathematical modeling and simulation from molecular to population levels, and high-performance computing including parallel processing, special hardware and grid computing.

Fruitful discussions took place through 18 oral presentations, including a keynote address and five invited talks, and 16 poster and demonstration presentations in the fields of grid infrastructure for life sciences, systems biology, massive data processing, databases and data grids, grid portals and pipelines for functional annotation, parallel and distributed applications, and life science grid projects. The workshop emphasized the practical aspects of grid technologies in terms of improving grid-enabled data/information/knowledge sharing, high-performance computing, and collaborative projects. There was agreement among the participants that the advancement of grid technologies for life science research requires further concerted actions and promotion of grid applications. We therefore concluded the workshop with the announcement of LSGRID 2005. More information about the workshop is available at: http://www.lsgrid.org/

This post proceedings contains the revised versions of the accepted papers of the LSGRID 2004 workshop. Ten regular papers were selected for inclusion in the postproceedings. The papers address the following issues:

- An Integrated System for Distributed Bioinformatics Environment on Grids
- Distributed Cell Biology Simulations with E-Cell System
- The Architectural Design of High-Throughput BLAST Services on OBIGrid
- Heterogeneous Database Federation Using Grid Technology for Drug Discovery Process
- Grid Portal Interface for Interactive Use and Monitoring of High-Throughput Proteome Annotation
- Grid Workflow Software for a High-Throughput Proteome Annotation Pipeline
- Genome-Wide Functional Annotation Environment for *Thermus thermophilus*
- Parallel Artificial Intelligence Hybrid Framework for Protein Classification
Parallelization of Phylogenetic Tree Inference Using Grid Technologies

Building a Biodiversity GRID

In addition to the regular papers, the postproceedings includes an invited keynote address by Hideaki Sugawara on:

Gene Trek in Procaryote Space Powered by a Grid Environment

and the following four papers presented in invited talks and posters in LSGRID 2004 (these papers were reviewed by the editors of the postproceedings)

EMASGRID: an NBBnet Grid Initiative for a Bioinformatics and Computational Biology Services Infrastructure in Malaysia

Development of a Grid Infrastructure for Functional Genomics

Mega Process Genetic Algorithm Using Grid MP

“Gridifying” an Evolutionary Algorithm for Inference of Genetic Networks Using the Improved GOGA Framework and Its Performance Evaluation on OBI Grid

We would like to acknowledge all the Program Committee members and all the additional referees for their work on reviewing the submitted papers. We also wish to thank all the authors and participants of the workshop for contributing to lively discussions and the exchange of knowledge and experiences on the Life Science Grid. It should also be mentioned that the workshop was independent but closely related to the Life Science Grid Research Group (LSG-RG) of the Global Grid Forum. More than half of the Program Committee members are also active members of the Life Science Grid Research Group. It would be difficult to organize such an international workshop without the continuous efforts of the LSG-RG. Finally, we wish to thank Fumikazu Konishi, Sonoko Endo, Maki Otani, Kyoko Hirukawa, Yuko Watada, Aki Hasegawa and Shigerv Takasaki for their help in organizing this workshop and editing the proceedings.

October 2004

Akihiko Konagaya
Kenji Satou
LSGRID 2004 was organized by the Special Interest Group on Molecular Bioinformatics (SIGMBI) of the Japanese Society for Artificial Intelligence, the Open Bioinformatics Grid (OBIGrid) Project, and the Japan BioGrid Project.

Executive Committee

Program Chair: Akihiko Konagaya (RIKEN GSC, Japan)
Organizing Chair: Kenji Satou (JAIST, Japan)
Exhibition: Rei Akimoto (Sun Microsystems, Inc., Japan)

Program Committee

Akiyama, Yutaka (AIST CBRC, Japan)
Ang, Larry (BII, Singapore)
Angulo, David (DePaul Univ., USA)
Arzberger, Peter (UCSD, USA)
Bala, Piotr N. (Copernicus Univ., Poland)
Goble, Carole (Univ. of Manchester, UK)
Farazdel, Abbas (IBM, USA)
Fukuda, Ken’ichiro (AIST CBRC, Japan)
Himeno, Ryutaro (RIKEN ACCC, Japan)
Hong, Gilnam (POSTECH, South Korea)
Kao, Cheng-Yao (NTU, Taiwan)
Konagaya, Akihiko (RIKEN GSC, Japan)
Konishi, Fumikazu (RIKEN GSC, Japan)
Lin, Fang-Pang (NCHPC, Taiwan)
Luo, Jingchu (CBI, Peking University, China)
Miyazaki, Satoru (NIG, Japan)
Nakamura, Haruki (Osaka Univ., Japan)
Matsuda, Hideo (Osaka Univ., Japan)
Matsuoka, Satoshi (TITECH, Japan)
Mohamed, Rahmah (UKM, Malaysia)
Napis, Suhaimi (UPM, Malaysia)
Ono, Isao (Tokushima Univ., Japan)
Palittapongarnpim, Prasit (BIOTEC, Thailand)
Rodrigo, Allen (Auckland, New Zealand)
Satou, Kenji (JAIST, Japan)
See, Simon (SUN, Singapore)
Sponsoring Institutions

We appreciate financial supports from the following institutes for LSGRID 2004.

- A Grant-in-Aid for Scientific Research on the Priority Area Genome Informatics from the Ministry of Education, Science, Sports and Culture of Japan
- Japanese Society for Artificial Intelligence
- RIKEN Genomic Sciences Center
- The Initiative for Parallel Bioinformatics, Japan
# Table of Contents

## Life Science Grid

**Gene Trek in Procaryote Space Powered by a GRID Environment**  
*Hideaki Sugawara* .............................................. 1

**An Integrated System for Distributed Bioinformatics Environment on Grids**  
*Kenji Satou, Yasuhiko Nakashima, Shin’ichi Tsuji, Xavier Defago,*  
*Akihiko Konagaya* ............................................ 8

**Distributed Cell Biology Simulations with E-Cell System**  
*Masahiro Sugimoto, Kouichi Takahashi, Tomoya Kitayama,*  
*Daiki Itou, Masaru Tomita* ............................... 20

**The Architectural Design of High-Throughput BLAST Services on OBIGrid**  
*Fumikazu Konishi, Akihiko Konagaya* ....................... 32

**Heterogeneous Database Federation Using Grid Technology for Drug Discovery Process**  
*Yukako Tohsato, Takahiro Kosaka, Susumu Date, Shinji Shimojo,*  
*Hideo Matsuda* ................................................ 43

**Grid Portal Interface for Interactive Use and Monitoring of High-Throughput Proteome Annotation**  
*Atif Shahab, Danny Chuon, Toyotaro Suzumua, Wilfred W. Li,*  
*Robert W. Byrnes, Kouji Tanaka, Larry Ang, Satoshi Matsuoka,*  
*Philip E. Bourne, Mark A. Miller, Peter W. Arzberger* .......... 53

**Grid Workflow Software for a High-Throughput Proteome Annotation Pipeline**  
*Adam Birnbaum, James Hayes, Wilfred W. Li, Mark A. Miller,*  
*Peter W. Arzberger, Philip E. Bourne, Henri Casanova* ........ 68

**Genome-Wide Functional Annotation Environment for Thermus thermophilus in OBIGrid**  
*Akinobu Fukuzaki, Takeshi Nagashima, Kaori Ide,*  
*Fumikazu Konishi, Mariko Hatakeyama, Shigeyuki Yokoyama,*  
*Seiki Kuramitsu, Akihiko Konagaya* ....................... 82
Parallel Artificial Intelligence Hybrid Framework for Protein Classification

*Martin Chew Wooi Keat, Rosni Abdullah, Rosalina Abdul Salam* ........................................... 92

Parallelization of Phylogenetic Tree Inference Using Grid Technologies

*Yo Yamamoto, Hidemoto Nakada, Hidetoshi Shimodaira, Satoshi Matsuoka* .............................. 103

EMASGRID: An NBNet Grid Initiative for a Bioinformatics and Computational Biology Services Infrastructure in Malaysia

*Mohd Firdaus Raih, Mohd Yunus Sharum, Raja Murzaferi Raja Moktar, Mohd Noor Mat Isa, Ng Lip Kian, Nor Muhammad Mahadi, Rahmah Mohamed* ............................................ 117

Development of a Grid Infrastructure for Functional Genomics

*Richard Sinnott, Micha Bayer, Derek Houghton, David Berry, Magnus Ferrier* ........................................... 125

Building a Biodiversity GRID

*Andrew C. Jones, Richard J. White, W. Alex Gray, Frank A. Bisby, Neil Caithness, Nick Pittas, Xuebiao Xu, Tim Sutton, Nick J. Fiddian, Alastair Culham, Malcolm Scoble, Paul Williams, Oliver Bromley, Peter Brewer, Chris Yesson, Shonil Bhagwat* ............................................ 140

Mega Process Genetic Algorithm Using Grid MP

*Yoshiko Hanada, Tomoyuki Hiroyasu, Mitsunori Miki, Yuko Okamoto* ........................................... 152

“Gridifying” an Evolutionary Algorithm for Inference of Genetic Networks Using the Improved GOGA Framework and Its Performance Evaluation on OBI Grid

*Hiroaki Imade, Naoaki Mizuguchi, Isao Ono, Norihiko Ono, Masahiro Okamoto* ........................................... 171

**Author Index** ................................................... 187
Gene Trek in Procaryote Space Powered by a GRID Environment

Hideaki Sugawara

Center for Information Biology and DNA Data Bank of Japan (DDBJ), National Institute of Genetics (NIG) 1111 Yata, Mishima, Shizuoka 411-8540, Japan
hsugawar@genes.nig.ac.jp

Abstract. More than 100 microbial genomes have been sequenced since 1995 and thousands of microbial genomes will be sequenced in a decade. It implies that millions of open reading frames (ORFs) will be predicted and should be evaluated. Therefore, we need a high throughput system to evaluate the predicted ORFs and understand functions of genes based on comparative genomics. We established and applied a protocol for the prediction and evaluation of ORFs to genome sequences of 124 microbial that were available from the International Nucleotide Sequence Database as of June, 2003. We could carry out the evaluation of about 300,000 predicted ORFs based on clustering and horizontal gene transfer analysis thanks to the GRID environment. This paper introduces mainly the scheme of the GRID environment applied to the comparative genomics.

1 Introduction

Genomes of procaryotes are doubtless small comparing to human genome. However, prokaryote genes are more diverse than human genes. The Genome Information Broker (GIB) [1] is a database of complete microbial genomes in public domain. It contained genome data of 124 microbes as of May, 2003. The number of open reading frames (ORFs) averages 3,000, namely, GIB is also a database of more than 300,000 genes in total. The space composed of procaryotes genes is vast and we need a space ship to trek there. Genome sequences are fundamental parts of the ship, an assembly of data mining tools is the engine and GRID computing is the booster. We successfully applied a GRID environment composed of 5 sites in OBIGrid (http://www.obigrid.org/) [2] to the comparative genomics of microbes registered in GIB.

2 Materials and Methods

The machines were connected by VPN in OBIGrid for our study as shown in Fig. 1. They were Linux machines in National Institute of Genetics (NIG) (64CPUs), Japan Advanced Institute of Science and Technology (JAIST) (68-CPUs), RIKEN Genomic Sciences Center (GSC) (10CPUs), Japan Science and
Fig. 1. The GRID environment used for the analysis of microbial genomes. Linux machines in the participating nodes were connected by VPN Technology Agency (JST) (66CPUs), and Tokyo Medical and Dental University (TMD) (21CPUs). Thus the total number of CPUs is 229. In these machines, the Globus ToolKit version 2.4 [3] and OBIEnv [2] were installed. To monitor the status of the OBIEnv machines, a P2P server is set up in NIG. The snapshot of the monitoring is introduced in Fig. 2 and demonstrates that the work load was actually distributed to CPUs in the OBIEnv.

<table>
<thead>
<tr>
<th>site</th>
<th>num</th>
<th>cpu</th>
<th>machines</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSC</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>TMD</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>DDBJ</td>
<td>16</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>aist-kanlab</td>
<td>40</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>jet</td>
<td>30</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>118</td>
<td>222</td>
<td></td>
</tr>
</tbody>
</table>

on: Fri Dec 06 09:12:29 JST 2003
summary
heat list

Fig. 2. Monitoring of the GRID environment. Small boxes in the table represent each CPU in nodes that participated in the project. Color codes are explained in the table.
In DDBJ [7], we had identified open reading frames (ORFs) of the microbial genomes by our own protocol and stored them in a database. The protocol and details of the identification will be published elsewhere. We analyzed these ORFs by use of a GRID environment in two ways. In the case of microbial strains, genes are transferred among species, namely, horizontal gene transfer (HGT) occurs after species are established during the evolution [4]. Since we have genome sequences, we will be able to carry out a comprehensive analysis of HGT. In addition to the HGT analysis, it is another important issue to infer functions of ORFs from sequence data. The basic method of the inference is clustering ORFs expecting that ORFs in a same cluster will share the same function.

3 Results and Discussions

3.1 Horizontal Gene Transfer (HTG)

We constructed a gene model for each microbial genome to evaluate if the candidate ORFs is intrinsic or introduced by HTG [4]. If a candidate ORF is largely deviate from the model, the ORF may be from other species. Therefore, we had to construct 124 models and then compare a number of ORFs in 124 genomes with all the models. In this way, we are able to identify donor of ORFs of HTG as well.

We estimated that the computation of the HTG analysis took about 60 months with a CPU of 2GHz. In OBIEEnv, the task was divided into 17,689 jobs as

Fig. 3. Distribution of the HTG jobs in OBIEEnv. The ORF sequences are distributed among machines and programs of Simulation.pl and Predseqs.c identify candidates of horizontally transferred genes. The total number of the JOBs was 17,689
shown in Fig. 3 to be completed in 18 days, although the network and some CPUs were down from time to time. Therefore, it will be quite feasible to repeat the analysis of HTG, whenever the data is update and a new genome is determined.

Results of the HTG analysis are stored in a database and a sample view of the database is introduced in Fig. 4. The top block in the figure displays

**Fig. 4.** Some results of horizontal transfer of genes (HTG). Results of the HTG analysis are stored in a database. A view of the database is a matrix of species vs genes in a species. This figure is a matrix of *E. coli* K12 (MG1655) genes and species whose genome sequences are publicly available. Bluish color of the cell in the matrix represents the probability of horizontal transfer of *E. coli* genes. The darker the color is, the higher the probability is. The warm color shows the probability of the donor of the horizontal transfer genes of the species to *E. coli*
ORFs of *E. coli*. Identifiers of ORFs in our system are written vertically there. Colors of cells under each ORF tell the probability of horizontally transferred gene. A blank cell means that the ORF is intrinsic to *E. coli*. On the other hand, a dark blue cell notifies that the corresponding ORF is probably transferred from other species, namely, horizontally transferred from a donor species of the ORF. The left tree in Fig. 4 displays a phylogenetic tree of microbes whose genome sequences are publicly available, e.g., from GIB. The donor species of the ORF notified by the dark blue cell is identifiable in Fig. 4, if you look down the corresponding column until you find a warm color cell. The red cell strongly suggests that the ORF of *E. coli* was transferred from the species in the same line as the red cell is. Thus the database of HTG provides information about not only genes horizontally transferred but also donor species.

### 3.2 Clustering of ORFs

We clustered 354,606 ORFs in total by SODHO [5] to retrieve information on functions of proteins from amino acid sequences of ORFs. With a CPU of 2 GHz, it would have taken 50 days to complete the clustering. In OBIEnv, we separated the all to all comparison of ORFs into 999 jobs to assign a job to a CPU for a parallel computing. In this way, it took only 17 hours to complete the computation. We evaluated the quality of the cluster by mapping to the motif in InterPro [6]. An example of the mapping is available in Table 1. We may

<table>
<thead>
<tr>
<th>Name of ORFs</th>
<th>IPR002528</th>
<th>IPR001064</th>
<th>IPR00678</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFDSM [1850] Pfur.DSM3638:.faa.C10</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>AAVF5 [101] Aaeo_VF5:.faa.C10</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>PAORS [366] PabyORSAY:.faa.C10</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>PHOT3 [1861] Phor_OT3:.faa.C10</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>TTMB4 [1686] Tten_MB4T:.faa.C10</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTMB4 [234] Tten_MB4T:.faa.C10</td>
<td>+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. A result of analysis of ORFs by use of InterPro. All the ORFs are compared with the database of InterPro. This table shows InterPro IDs of 6 ORFs that are member of a cluster found by SODOH

expect that member ORFs of a cluster share the same motif among them, if the clustering is biologically meaningful. The six member ORFs share the same InterPro ID of IPR002528, although some of the ORFs hit to other InterPro IDs as well. Therefore, the correspondence the cluster in Table 1 is comparatively well defined. The precise and global evaluation of the clustering by SODHO is still under way.
3.3 Need of Computer Resources for Comparative Genomics

We analyzed 124 microbial genomes that had been disclosed by May, 2003. You will find genomes of 185 microbial strains as of August, 2004, i.e., 4 genomes a month were recently sequenced on average. We are able to expect that a number of microbial genomes will be continuously sequenced. Some microbes are closely related and even multiple strains belong to the same species. Other microbes are distant in the phylogenetic tree with each other. Therefore, the comparative genomics will be a powerful tool to understand especially the dynamics of genomic evolution and gene functions of microbes. Based on this observation, we are afraid that we need keep expanding the computer resources to compare a number of objects. Otherwise, we will not be able to complete a set of analysis before a new genomic sequence is available.

The computation will be left behind a tidal wave of genomic data, if an expandable and flexible large scale computing facility. The International Sequence Database (INSD) exceeded 30 millions entries and 30 giga nucleotides in the year of 2003 and will keep expanding every year as much as 1.5 times. Protein sequence database such as InterPro will expand too. In this study, we did not use a GRID environment to matching every candidate ORFs to InterPro. We need certainly a GRID environment or a large scale PC cluster to apply the InterPro analysis to 185 microbes as of August 2004 and afterwards.

The INSD already includes hundreds of genomic sequences of such wide variety of species as viruses, microbes, plants, animals and human. The INSD will capture thousands of genomes in the near future. Therefore, it is obvious that GRID environment will be the infrastructure of comparative genomics that traverse all the species to understand the universe of life phenoma.

Acknowledgements

The GRID computer environment in NIG was supported by Life Science System Division of Fujitsu Limited. This work has been partly supported by BIRD of Japan Science and Technology Agency (JST) and also partly by the Grant-in-Aid for Scientific Research on Priority Area “Genome Information Science”, Ministry of Education, Sports, and Science (MEXT), Japan

References


6. http://www.ebi.ac.uk/interpro/

7. DDBJ (http://www.ddbj.nig.ac.jp/), EMBL (http://www.ebi.ac.uk/) and GenBank (http://www.ncbi.nlm.nih.gov/)
An Integrated System for Distributed Bioinformatics Environment on Grids

Kenji Satou\textsuperscript{1,2}, Yasuhiko Nakashima\textsuperscript{3}, Shin’ichi Tsuji\textsuperscript{3}, Xavier Defago\textsuperscript{4,5}, and Akihiko Konagaya\textsuperscript{6}

\textsuperscript{1} School of Knowledge Science, Japan Advanced Institute of Science and Technology \texttt{ken@jaist.ac.jp}
\textsuperscript{2} BIRD, Japan Science and Technology Agency (JST)
\texttt{y-nakashima@pu.jp.nec.com, s-tsuji@pj.jp.nec.com}
\textsuperscript{3} NEC Software Hokuriku, Ltd.
\texttt{defago@jaist.ac.jp}
\textsuperscript{4} School of Information Science, Japan Advanced Institute of Science and Technology
\texttt{konagaya@gsc.riken.jp}

Abstract. In this paper, an integrated system called OBIEnv, which has been developed on OBIGrid, is described. In addition to automatic database transfer and deployment, it provides various functionalities for transparent and fault-tolerant processing of bioinformatics tasks on Grid. A feasibility study on the analysis of horizontal gene transfer was done using 119 heterogeneous Linux nodes in 5 different sites, and OBIEnv proved its applicability to practical problems in bioinformatics.

1 Introduction

Based on the explosive yield of experimental data on biomolecules, a research area called bioinformatics has grown remarkably in the past decade. This growth has been caused by several factors, such as, success of genome projects on human and model organisms, developments of new protocols and equipments for high-throughput experiments in molecular biology, and rapid growth of the Internet and WWW. Furthermore, traditional and advanced technologies in computer science have been applied to accelerate all the aspect of life sciences. Scientists nowadays are routinely using computers and networks in experiments, search, analysis, and visualization.

While the amount of experimental data is increasing exponentially, focus of genome analysis is shifting from identification of facts (e.g. the sequences and structures of genes and proteins) to conjecture of relationships, like gene network\cite{1} and interaction pathways\cite{2} in a cell. Since the latter requires combinatorial computation, and the search space itself is exponentially extending\cite{3},
researchers in bioinformatics are suffering from chronic shortage of computing power. Therefore, utilization of Grid computing technology is strongly needed also in this research area.

Starting from metacomputing in late 80’s, the concept of Grid computing has evolved giving rise to a variety of research projects[4, 5, 6], which can be classified according to different aspects: Compute Grids, Data Grids, Access Grids, Campus Grids, etc.[7] Among them, Open Bioinformatics Grid (OBIGrid) project is one of the comprehensive Grid project in Japan. Aiming at the construction of a practical infrastructure for bioinformatics, OBIGrid[8, 9] has adopted a virtual private network (VPN) in order to guarantee transparency and security issues, and several research and development projects are running based on it.

In this paper, an integrated system called OBIEnv, which has been developed on OBIGrid, is described. Though it adopts the Globus Toolkit as its basic middleware for remote command execution, it provides various and original functionalities for distributed, high-throughput, adaptive, and transparent computing, as well as database management in bioinformatics.

2 Design Issues

In the earliest stages of OBIEnv project, we assumed the following things as a basis for the design of our system.

– Nodes are heterogeneous in its hardware and software. While there might be Linux boxes and clusters, big SMP machines with Unix (e.g. Sun Fire 15k) can coexist on a Grid.
– Unlike cluster computing, in case of distributed computing with nodes scattered over the Internet, instability caused by node and network failures is frequent.
– Bandwidth of wide area network is drastically narrower than that of local area network.
– While there exist many types of computation in bioinformatics, one of the most frequently executed computation is processing a great number of similar tasks independent to each other. For example, all-to-all homology search with a sequence database is a typical one. It requires a job including many BLAST executions with the same search space (database) and different query sequences. This type of computation is classified into embarrassingly parallel processing, and it is well-suitable for Internet-wide distributed computing.
– Though we assume the above type of computation, it is impossible to estimate how long time is needed to process each task. In general, it depends on the application, its input, node characteristics (e.g., CPU type, memory size, I/O speed), and so on.

Next, we specified the requirements to develop an easy-to-use bioinformatics environment on Grid.
Deployment and update of common databases and applications is essential. It should be performed automatically as much as possible. Of course, network bandwidth and computing power should be used effectively for fast transfer.

Information about the existence of databases and applications in a computation node has to be reported, gathered, and accessible for transparent computing. Furthermore, this information should be managed transparently to the user.

Database transfer and job processing must be adaptive to the dynamically changing situation of nodes and networks, including slowdown and failure.

Finally, we designed OBIEnv as shown in Fig. 1. It consists of the following three components:

- **P2P server** for gathering node information about what kind of hardware, software, and database are deployed in each node. Update of the information is done in bottom-up manner (reporting from nodes to P2P server).
- **Job dispatcher** for easy and transparent distributed processing. Given a set of similar tasks, it dispatches them to the nodes with required environment. While it is running with the specified constraints, it occasionally searches a set of appropriate nodes by making a query to the P2P server, then adapts to dynamically changing situation of nodes (e.g. addition and retire of nodes).
– **Standard Software Environment (SSE)** to deal with platform heterogeneity. By assuming that all the nodes have essentially the same set of software including GNU tools, programming languages, and bioinformatics applications, a user can safely scatter a set of tasks to heterogeneous Unix/Linux systems. Besides the above free software, a series of programs is included for job dispatching, database update, and node information reporting to P2P server. Wrapper programs for bioinformatics applications are also included for easy execution in the nodes dispersed over OBIGrid.

Note that in this model, a representative user account is adopted. Globus Toolkit requires user accounts and home directories in all remote nodes for job processing. However, we think this requirement is too strict for our purpose. As shown in Fig. 1, local and site-level authentication is sufficient, and given a rich SSE, the need for user’s permanent program files on remote machines could be reduced.

### 3 Implementation Issues

On the P2P server, a PostgreSQL daemon is running for search and update of information. Fig. 2 illustrates what kind of information is stored in it.

To avoid too frequent access to the P2P server, most of the hardware, benchmark, and OS information are not frequently updated. Database information is typically updated daily. Information about job dispatching and database transfer by rsync is updated on demand. In order to prevent malicious updates from inside the OBIGrid, access to the port of PostgreSQL is controlled by using iptables on the P2P server.

The job dispatcher is implemented in Perl and Java. Given a set of tasks and constraints, the dispatcher checks user authentication. Then, based on the

<table>
<thead>
<tr>
<th>Constraint variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOST</td>
<td>host names</td>
</tr>
<tr>
<td>SITE</td>
<td>site names</td>
</tr>
<tr>
<td>SMP</td>
<td>the number of CPUs</td>
</tr>
<tr>
<td>MEM</td>
<td>memory size</td>
</tr>
<tr>
<td>HDD</td>
<td>disk size</td>
</tr>
<tr>
<td>CPU_MODEL</td>
<td>model name of CPU</td>
</tr>
<tr>
<td>CLOCK</td>
<td>CPU clock</td>
</tr>
<tr>
<td>INDEX</td>
<td>memory index by nbench-byte</td>
</tr>
<tr>
<td>IINDEX</td>
<td>integer index by nbench-byte</td>
</tr>
<tr>
<td>FINDEX</td>
<td>float index by nbench-byte</td>
</tr>
<tr>
<td>DB</td>
<td>database names</td>
</tr>
<tr>
<td>DBV</td>
<td>database names and versions</td>
</tr>
</tbody>
</table>