# Analysis of Horizontal Gene Transfer and Clustering of Microbial ORFs by Use of a GRID Environment

**Hideaki Sugawara**[1] hsugawar@genes.nig.ac.jp
**Yoji Nakamura**[1] yojnakam@lab.nig.ac.jp
**Kazuho Ikeo**[1] kikeo@genes.nig.ac.jp
**Satoru Miyazaki**[1] smiyazak@genes.nig.ac.jp
**Takashi Gojobori**[1] tgojobor@genes.nig.ac.jp
**Kenji Satou**[2] ken@jaist.ac.jp
**Akihiko Konagaya**[3] konagaya@gsc.riken.go.jp

[1] Center for Information Biology and DDBJ, National Institute of Genetics (NIG), 1111 Yata, Mishima, Shizuoka 411-8540, Japan
[2] Graduate School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST), 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
[3] Bioinformatics Group, RIKEN GSC, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan

**Keywords:** GRID computing, VPN, comparative genomics, microbiology, horizontal gene transfer, clustering

## 1 Introduction

Tsunami of biological data and multiple views of the data analysis require an expandable and flexible information environment. GRID computing is expected to be the solution. We prepared a computational environment composed of 5 sites in OBIGrid [6] and succeeded in analyzing horizontal gene transfer and clusters of ORFs of more than 100 microbial genomes that were stored in the Genome Information Broker [1] as of May, 2003. This scheme is being applied to more than 300 thousands ORFs of genomic sequences of 124 microbial species.

## 2 Materials and Methods

Linux machines in NIG (64 CPUs), JAIST (68 CPUs), RIKEN GSC (10 CPUs), Japan Science and Technology Agency (JST) (66 CPUs) and Tokyo Medical and Dental University (TMD) (21 CPUs) are connected via LAN and VPN resulting in the computer resources of 229 CPUs. It is to be noted that the hardware specification and the version of Linux OS are diverse. In these machines, the Globus ToolKit version 2.4 [5] and OBIEnv [6] are installed. The database of ORFs is stored in the OBIEnv and applications are controlled by OBIEnv. In addition, a P2P server is set up in NIG to monitor the status of the OBIEnv machines. The screen dump of the P2P server is shown in Fig. 1. By use of the 229 CPUs in OBIEnv, we were able to analyze ORFs of more than 100 microbial genomes all together.

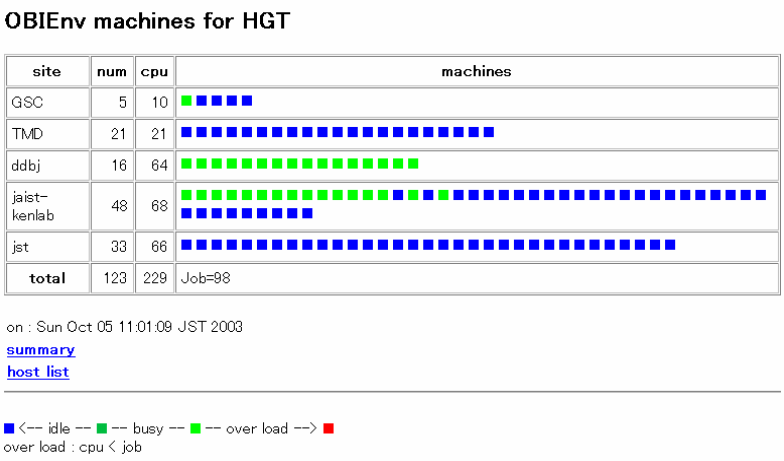


Figure 1: A screen dump of the P2P server that monitors the participating machines.

# 3 Results and Discussion

## 3.1 Horizontal Gene Transfer (HTG)

We performed all to all comparison of 124 genomic sequences from 112 microbes to analyze HTG based on a Markov model [3]. It will take about 60 months to accomplish the evaluation with a CPU of 2GHz. In OBIEnv, we assigned each one of 15,376 comparisons (i.e. 124 times 124 comparisons) to a CPU. The task was done in 18 days, although the network and some CPUs were down from time to time. We evaluated about 250,000 ORFs to find that 45,136 ORFs are probably transferred from alien species. We also identified the donor species of the 45,136 ORFs. The database of HTG will be published elsewhere.

## 3.2 Clustering of ORFs

We analyzed 354,606 ORFs of 119 microbial species by SODHO [2]. All to all comparison of the amino acids sequences will take 50 days with a CPU of 2 GHz. We segregated the large number of comparisons into 999 jobs and assigned a job to a CPU. Then the computation was done in parallel in 17 hours. We found 63,149 clusters that have more than 1 member and evaluated each cluster by use of InterPro [4]. The correspondence between the cluster and InterPro numbers was reasonable, e.g., as shown example in Table 1.

Table 1: A correspondence between a cluster defined by SODHO and InterPro numbers.

| Name of ORFs | IPR002528 | IPR001064 | IPR00687 |
|---|---|---|---|
| PFDSM[1850] \| Pfur_DSM3638:.faa_C10 | + | + | |
| AAVF5[101] \| Aaeo_VF5:.faa_C10 | + | | + |
| PAORS[366] \| Paby_ORSAY:.faa_C10 | + | + | |
| PHOT3[1861] \| Phor_OT3:.faa_C10 | + | + | |
| TTMB4[1686] \| Tten_MB4T:.faa_C10 | + | | |
| TTMB4[234] \| Tten_MB4T:.faa_C10 | + | | |

# Acknowledgments

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

[1] Fumoto, M., Miyazaki, S., and Sugawara, H., Genome Information Broker (GIB): Data retrieval and comparative analysis system for completed microbial genomes and more, *Nucleic Acids Res.*, 30(1):66–68, 2002.

[2] Naitou, K., Kawai, M., Kishino, A., Moriyama, E., Ikeo, K., Ina, Y., Ikezaka, M,. Satou, H., and Gojobori, T., The implementation of parallel processing for molecular evolutionary analysis using the highly parallel processor (in Japanese), *Joint Symposium on Parallel Processing'90*, 329–226, 1990.

[3] Nakamura, Y., Comparative genomics of prokaryotes with special reference to horizontal gene transfer, and its evolutionary implication, *Doctoral dissertation of the Graduate University for Advanced Studies*, 2003.

[4] `http://www.ebi.ac.uk/interpro/`

[5] `http://www.globus.org/`

[6] `http://www.obigrid.org/`