Real-Time Mapping System of cDNAs and Genomes Using Grid Computing

Yusuke Saito[1]
yuusuke.saito@hp.com

Manabu Gomi[1]
manabu.gomi@hp.com

Hideo Matsuda[2]
matsuda@ist.osaka-u.ac.jp

Naohisa Goto[3]
ngoto@gen-info.osaka-u.ac.jp

Ken Kurokawa[3]
ken@gen-info.osaka-u.ac.jp

Teruo Yasunaga[3]
yasunaga@gen-info.osaka-u.ac.jp

[1] Hewlett-Packard Japan, Ltd., Tennoz Central Tower, 2-2-24 Higashishinagawa, Shinagawa-ku, Tokyo 140-8641, Japan
[2] Graduate School of Information Science and Technology, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
[3] Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan

Keywords: cDNA, genome, mapping, grid computing, Globus Toolkit

1 Introduction

DataGrid group of BioGrid project [1], in which we are joining, is aiming to establish a methodology which enables to choose target proteins easily through the process of drug discovery. For that purpose, the group has been developing intelligent database services which federate various biological databases seamlessly. As one component of the services, we have been developing a real-time mapping system of cDNAs and genomes.

In choosing a target protein for a drug discovery, the exon-intron structure of the corresponding gene and the information about the gene expression may also be required. The system, which we have been developing, is planned to enable users to get such the sequence information immediately.

A lot of mapping methods are already published and there are excellent gene databases like Ensembl [2]. One of the distinguished features from these existing ones is that the user can complete the process of genome research such as homology search, mapping calculation and acquisition of detailed data only by inputting query sequence into this system.

2 Methods

2.1 Architecture of Mapping System

We have been developing a web-based system which contains both the pre-calculated mapping database of cDNAs and genomes for known cDNAs and the mapping engine for a new cDNA(Figure 1). The query sequence can be a part of cDNA or mRNA and genome sequence. If the query sequence has high homology with a cDNA sequence in the mapping database, the user can get the information immediately from the result of database search. On the other hand, if it does not have homology with any cDNA sequence in the database, mapping of the query sequence and the genome is calculated using the high performance mapping program, Spidey [5]. Because this process is time consuming, we apply grid computing technology to improve the process, as described in detail later. In addition, since the database contains ortholog information about various species such as human and mouse, it can also be used for a comparative genomics tool.

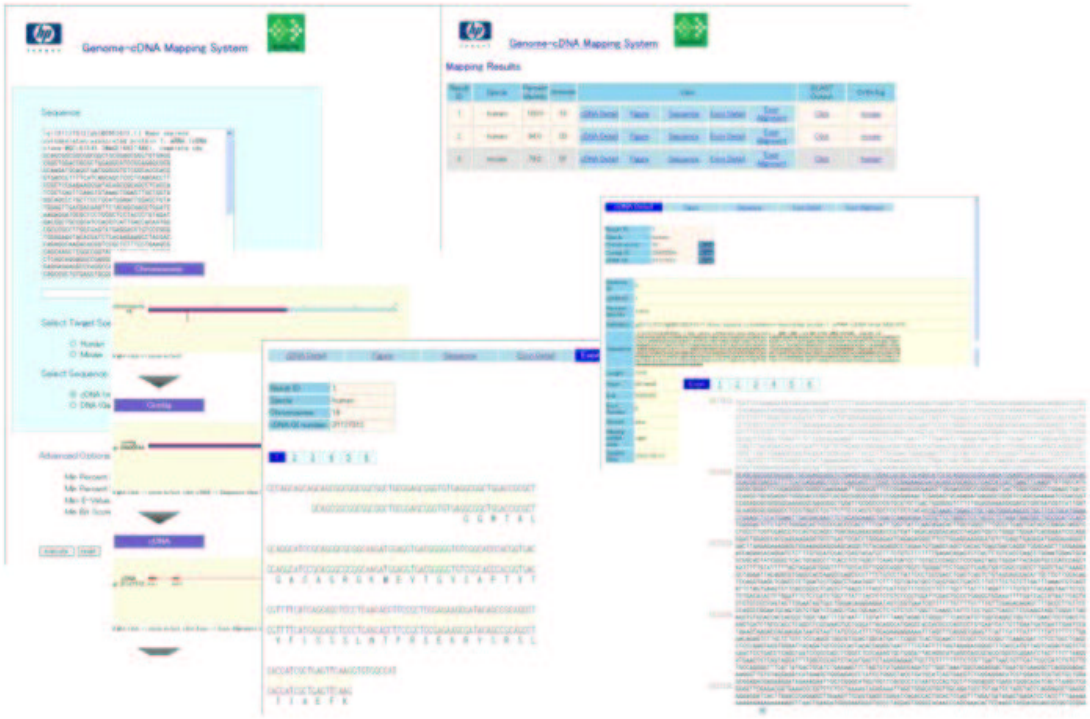

Figure 1: Web interface of mapping system.

2.2 Improvement Using Grid Computing Technology

Since mapping calculation requires long time, it is difficult to get the result immediately. To improve this situation, we adopted grid computing technique to accelerate the calculation. We divided the cDNAs and genomes data and distributed the computation of them for many computers in network via Globus Toolkit [3] which is one of grid middleware. Moreover we have to compute again whenever the sequence data is updated frequently. We have installed software for updating genome data automatically, and have connected the mapping calculation with the download schedule of the data. This system is implemented on OGSA-DAI [4] architecture of Globus Toolkit 3.0 and prepares APIs to be used from remote programs via SOAP messages.

3 Discussion

In this research we have been developing a system which can be used as a useful method for genome research. To make it more practicable, additional information such as SNPs, transposon and alternative splicing may be required. In addition to that, we would like to improve this system according to the remarkable progress of grid technology.

4 Acknowledgments

This study was performed through IT-program of Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] BioGrid Project: `http://www.biogrid.jp/`

[2] Ensembl: `http://www.ensembl.org/`

[3] Globus Toolkit: `http://www.globus.org/`

[4] OGSA-DAI: `http://www.ogsadai.org.uk/`

[5] Spidey: `http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/`