OVERCOMING BIOINFORMATICS CHALLENGES WITH SUPERComputING

Satoru Miyano, Dr.Sc.
Human Genome Center, Institute of Medical Science, The University of Tokyo

The Human Genome Center of University of Tokyo was founded in 1991 as the central research center for the Japanese Human Genome Project by the Ministry of Education, Science, Sports and Culture. It started as a single laboratory called Laboratory of Genome Database. Minoru Kanehisa was the head of this first laboratory and ran the "Genome Informatics Project" (1991 – 1995) which is the informatics part of the Japanese Human Genome Project. Prior to the project, he initiated the Genome Informatics Workshop (GIW) in 1990 that is currently called the International Conference on Genome Informatics. The acronym GIW has been used for this bioinformatics conference from the start. From 2007, GIW will be held in various countries in the Asia Pacific region and GIW 2007 will be held at Biopolis in Singapore on 3 – 5 December 2007.

Since 1996, the Human Genome Center (HGC) is directed by Yusuke Nakamura and its final aim is to make personalized medicines by identifying the novel molecular markers of the diseases and develop genetic diagnosis and molecular targeted therapy. Table 1 summarizes the activities of the laboratories for achieving this aim. HGM has a supercomputer system which is leased for six years, with the total budget at 600M JPY. The current system was installed in January 2004 and will be replaced with a new system in January 2009 which will have around one peta flops computing ability in total. Supercomputing is considered as the critical technology to advance the activities of HGM. On the other hand, recently, RIKEN has started a new big project “Development and Application of NEXT-Generation Supercomputer” located in Kobe with a budget of 1150M JPY (spread over seven years) which will develop a supercomputer system with 10 peta flops computing ability and will challenge the development of life science applications to extend the frontiers of the life sciences. HGM is leading the subproject “Data Analysis Fusion” in the project “Research and Development of Integrated Life Simulation Software” by RIKEN which currently comprises four peta-scale computing subprojects: molecular scale simulation, cell scale simulation, systemic organ scale, and data analysis fusion. The goal of the data analysis project is to create a peta-scale information technology for drug target discovery and personalized medicine which will fuse “individual” and “general” models with various heterogeneous and high dimensional data. More precisely, it will develop a system for large-scale SNP and phenotype association analysis for “individual models”; create computational and statistical methods for inferring biological networks enhanced with the cutting-edge technology for simulation called data assimilation for “general models”; and establish a peta-scale computational strategy by fusing them. For example, the current supercomputer can compute gene networks of 1000 nodes from several hundreds siRNA knockdown microarray data but it becomes feasible to enlarge this size to 30 000 nodes by 2010 (see Fig. 1). This project is a collaboration of researchers from RIKEN SNP Research Center, Tokyo Institute of Technology and Institute of Statistical Mathematics.

The NIH roadmap for biomedical research describes “… All of these techniques generate large amounts of data, and biology is changing fast into a science of information management.” (http://nihroadmap.nih.gov/bioinformatics/). This is also the case for the HGC that is going into a new era based on supercomputing.
## Table 1. Laboratories of Human Genome Center

<table>
<thead>
<tr>
<th>Laboratories</th>
<th>Faculty Members &amp; Research Topics</th>
</tr>
</thead>
</table>
| Molecular Medicine and Genome Technology | Yusuke Nakamura (Professor, Head, Director)  
Toyomasa Katagiri (Associate P), Yataro Daigo (Project Associate P)  
Hidewaki Nakagawa (Assistant P), Koichi Matsuda (Assistant P)  
Hitoshi Zembutsu (Assistant P), Ryuji Hamamoto (Assitant P)  
- Genes playing significant roles in human cancer  
- Genes responsible for hereditary diseases  
- Human genome analysis  
- Fukuyama-type congenital muscular dystrophy  
- Studies for genetic background of the bone and joint diseases  
- Biobank Japan (300 000 patient cases)                                                                                   |
| Genome Database & Sequence Analysis      | Minoru Kanehisa (Head, Adjunct Professor)  
Tetsuo Shibuya (Lecturer), Shuichi Kawashima (Asstiant P)  
Toshiaki Katayama (Assitant P), Michihiro Araki (Asstiant P)  
- Systematic analysis of the gene universe for classification of gene functions  
- Community databases for experts' knowledge acquisition  
- Disease pathways and disease ontology  
- New interfaces to KEGG  
- Algorithms and software for analyzing large-scale data in genomics, proteomics, metabolomics, and glycomics.  
- Computational technologies for chemical genomics                                                                 |
| DNA Information Analysis                 | Satoru Miyano (Head, Professor)  
Seiya Imoto (Associate P), Noriko Goto (Project Associate P)  
Rui Yamaguchi (Project Lecturer), Masao Nagasaki (Assistant P)  
- Computational systems biology  
- Gene network analysis and its application to drug target gene discovery  
- Modeling and simulation of biological pathways  
- Software tool Cell Illustrator development  
- Toxicobioinformatics  
- Data assimilation technology for biological systems  
- Systems biomedical technology for cancer biomarker discovery  
- Peta-flops scale supercomputing technology for data analysis fusion in the integrated life simulation software project                                                                 |
| Functional Analysis in silico            | Kenta Nakai (Head, Professor)  
Kengo Kinoshiba (Associate P), Riu Yamashita (Project Assistant P)  
- DBTBS: Database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics  
- DBTSS (DataBase of Transcriptional Start Sites)  
- Analysis of gene regulatory elements  
- Analysis of CpG islands around promoters and their tissue-specificity  
- Motif discovery program development  
- Analysis of alternatively-spliced protein isoforms  
- Protein 3D structure analysis and prediction                                                                 |
| Statistical Genetics                     | Mark G. Lathrop (Head, Adjunct Professor)  
Ryo Yamada (Associate P)  
- Statistical genetics  
- Genetic analysis of autoimmune and other diseases  
- SNP analysis                                                                                                              |
| Public Policy                            | Kaori Muto (Associate P)  
- ELSI                                                                                                                        |
Fig. 1. Fenofibrate (a drug for hyperlipidemia) response HUVEC gene network of the size 1049 computed from 270 siRNA knockdown microarray data by using Bayesian network and nonparametric regression. It took several weeks to obtain high quality networks on the HGC supercomputer system. However, it deals with only a few percent of gene products. Peta-scale computing enables us to compute networks of more than 30,000 genes.