Management System for Sequencing Data of Human Genome

— As a Part of *ALIS* —

Mika HIRAKAWA [1]
mika@tokyo.jst-c.go.jp

Kensaku IMAI [1]
imai@tokyo.jst-c.go.jp

Akira OHYAMA [2]
akr@troy.hydra.mki.co.jp

Fumihiko KIKUCHI [1]
fukiku@tokyo.jst-c.go.jp

[1] Bioinformatics division, Department of Advanced Databases
Japan Science and Technology Corporation (JST)

[2] Bioscience Systems Department,
MITSUI Knowledge Industry Co., LTD.(MKI)

Abstract

ALIS (Advanced Life science Information Systems) is dedicated to supporting and encouraging large scale human genome research by creating and distributing databases and providing the computing environment. We report on the primary status of ALIS project and our WWW service site (`http://www-alis.tokyo.jst-c.go.jp` *). The primary stage of the project has three aspects: large-scale human genome sequencing, construction of an integrated human genome database and development of supporting function for the database.*

1 Introduction

We developed The Sequencing Data Management System to store and maintain the data. The aim of the system is entry of the sequencing data from four different teams into the database, evaluation of the data and making up automatically the WWW pages to make the data available to public. The four research institutes are organizing the JST sequencing teams to supply the human genome sequencing data for JST. We fund them. It is a preliminary status effort for the large-scale human genome sequencing in Japan.

2 Data Entry Subsystem

The Data Entry Subsystem pre-processes the data before their registration into database. In this process, heterogeneity of the data as the teams submitted is absorbed. The data dealt in this system are trace data of sequencer, report files from assembler, edited fragments and the consensus nucleotide sequences. Data transfer from Macintosh (the collection device for the ABI sequencer) to UNIX is accomplished by FTP using a Macintosh program, Fetch. After that system identifies the types of the data classified by directories, then registers them into the database (using SYBASE as DBMS). This system has WWW interfaces to input the basic information on the data directly by the sequencing teams.

3 Data Evaluation Subsystem

The Data Evaluation Subsystem consists of tools on the X-window to browse assembly and trace to prove the consensus sequence data. It is our policy to guarantee the quality of sequence by raw data. This system executes homology search to annotate the sequence and check the accuracy for use. The main function of the Data Evaluation Subsystem is to give automatically annotated data to the database. Our advisory committee suggests that the sequence data from large-scale sequencing must have automatic annotations as follows: homologous regions for known genes, ESTs, STS-makers and repeat sequences, restriction map and candidate regions for genes or protein coding regions. We have integrated the Bioccelerator from Compugen Ltd. as a search engine into the subsystem. The Bioccelerator accelerates dynamic programming algorithms such as Frame Search, Smith-Waterman and Profile Search. The subsystem will search the latest releases of public databases and store the results in the database.

4 Data Presentation Subsystem

The goal of the preliminary status of the project is to publish the sequencing data on our WWW site. This system generates Web pages from the sequencing data in the database. The sequencing data are shown on the page of hierarchy. The WWW pages with some JAVA applets show relationship among chromosome, target region of sequencing, determined region and clones contigs. The nucleotide sequences with source and strategy information can be seen by clicking characters of the clone name. The teams' information pages are constructed with collaboration from the teams.

5 The ALIS Project

The aim of ALIS Project is construction of an informative resource for multidisciplinary research after human genome sequencing is completed. At the second status we try to rearrange various public human genome related information sources. These sources will be merged with our mega-sequencing data. We also will prepare for genome analysis tools to utilize the genome sequence data effectively.