ALIS Sequencing Database
for large scale human genome project

Mika HIRAKAWA
mika@tokyo.jst.go.jp
Kensaku IMAI
imai@tokyo.jst.go.jp
Hiroko YAMAGUCHI
yamako@tokyo.jst.go.jp
Junko SHIMADA
sjunko@tokyo.jst.go.jp
Kazuo TAKEHANA
take3@tokyo.jst.go.jp
Itaru SUZUKI
itarun@tokyo.jst.go.jp
Masahiko SUZUKI
masa@tokyo.jst.go.jp
Fumihiko KIKUCHI
fukiku@tokyo.jst.go.jp

Bioinformatics division, Advanced Databases Department,
Japan Science and Technology Corporation (JST)

Abstract

The goal of the Advanced Life Science Information Systems (ALIS) project is construction of an entire human genome database that will provide an efficient source of information for researchers after the human genome has been sequenced. We have initiated this project to encourage large scale human genome sequencing and to develop systems for genome data management and data publishing by World Wide Web. It has been 2 years since the project began and our first attempt at human genome sequencing is going well and more than 4M bases of well-edited human genome sequences have been acquired. The human genome project is progressing and international consensus releasing data generated from the project has been defined. We have been improved on our sequencing database to adapt the situation. Recently we organized collection and publication system for the genome sequencing data.

1 Introduction

We have been developing computer systems to manage data of large scale human genome sequencing by JST sequencing teams. The mission of JST is acquisition of all data from the sequencing project and publication of these data through our web site. We have established a consistent procedure from raw data to published data. ALIS Sequencing Database was actualized by systematization of these of data processing steps.

2 Collection of Team Data

Sequencing data from each laboratories, e.g. consensus sequences of clones and regions, assembly layout files and sequencer traces, by MOs and tapes. All data and information are stored in the master database through some checking programs. File formats and a variety of data is checked, then data is submitted into the master database design on SYBASE. The files not to put into database by registration error are possible to recover manually. The data complementation is confirmed by offset search. Sequences of assembled unit are arranged to consensus sequences by pairwise alignment and located at their original positions. The database can maintain the up-dated latest consensus sequence data after a chain of data collection processes. We can prepare the information added to sequencing data in the master database, such as materials, experimental method, sources and maps. The locations of each objects on the map of sequence ready source is calculated by map editing tool and the data stored in the master database.

3 Release Sequence Data

The data for public release is extracted from the master database. The conditions for retrieving these data are also stored in the database, and they are available at any time to get complete data for publication. The web pages and the objects on the pages are generated dynamically by cgi programs. The maps of the regions to plan to sequence and progress of sequencing are also made by cgi. The maps show sequencing source clones and standard markers on the region. Our new sequencing pages have chromosome map for planed sequencing regions, framework maps with the Genethon markers, sequence ready clone maps with standard markers, experimental information and consensus sequences. The consensus sequence data can be selected from fasta format and annotated EMBL like format. All the data are available from our WWW site(http://www-alis.tokyo.jst.go.jp).