ALIS: Data Management Systems for Human Genome Sequencing

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
Katsuji Matsumura katsuji@tokyo.jst.go.jp
Takehiko Itoh takehiko@tokyo.jst.go.jp
Masako Kuroda kuroda@jst.go.jp

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

1 Introduction

The ALIS (the Advanced Lifescience Information Systems) Project has commissioned four laboratories to execute sequencing of human genome regions to demonstrate the feasibility of large-scale human genome sequencing in Japan. The sequence data of genome project should be release as soon as possible with mapping information on an international consensus, so the mission of JST is acquisition all of data from the ALIS sequencing project and publishes data by our web. Our systems are able to process sequencing information for web publication with automatically drawn maps. Large-scale human genome sequencing will be more common strategy to search genes by improvement of technology for sequencing. ALIS sequencing management systems will apply to acquire mega-scaled human genome sequences for biological researches in general laboratories.

2 Data acquisition

Since 1995, the JST ALIS (Advanced Life Science Information Systems) Project has acquired contiguous mega-scale sequences of human genome. JST gave four laboratories, which have a principal investigator with the motivation to sequence the human genome through more than 1 mega-bases, commissions to execute large-scale sequencing of the human. This first attempt of generation data is going well and about 15 million bases are sequenced by the March 1999. Sequencing is conducted at individual laboratories, with strategies determined by each site. There are many varieties even in same kinds of data, because of differences of sequencers, software and their conditions. We have been working on the acquisition of data from the laboratories and designing a database to maintain it. At first we design a master database to treat the data from each sequencing team equally. The database stores experimental sequencing data and information to identify the target regions, sources and investigators in the team. Sequencing data from each laboratory consist of consensus sequences of each clone and continuous regions, results of assembly of fragment data and sequence traces by MOs and tapes. The data collection system checks all data by check algorithm to evaluate their file formats and data types to determine whether or not they can be registered into the database. The system uses SYBASE as DBMS for the master database. The system also post-processes among registered data to give relational information. The data except from storage media is entered through Web interfaces. Pre-viewer on Web is available for maintenance of map information. DNA markers and found genes are annotated on sequences by using marker search programs.

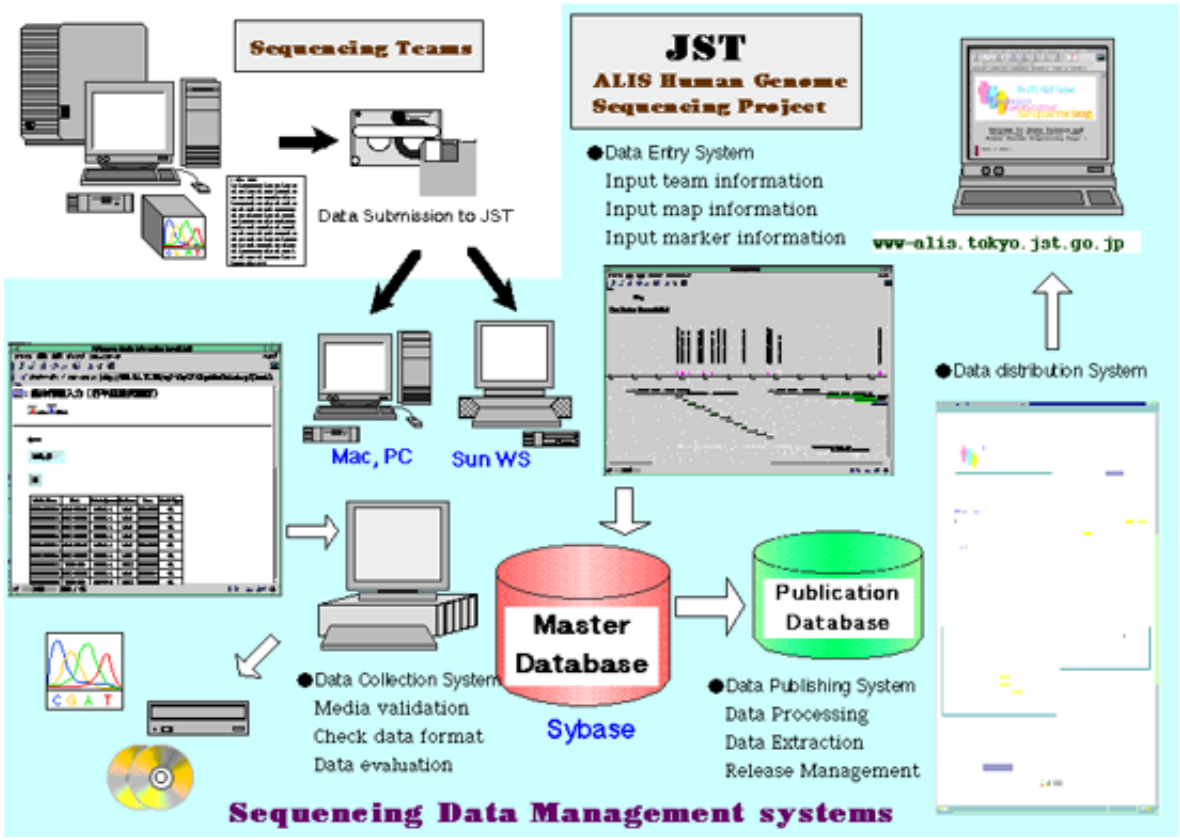

Figure 1: Image of ALIS Sequencing Management Systems

3 Data publication

Data publication system has another database to manage data used for WWW pages. All the web pages are generated by perl scripts automatically with the data from the database. The database manages and maintains data by the region and release date. ALIS Human Genome Sequencing Pages provide three features of maps for target regions finished nucleic acid sequences and profile of sequencing sites. All the mapped objects have links to reach detail information. There are three ways to see maps and sequences from top-page, selecting a link of chromosome, band of target regions and sequencing team. The top-page also has direct link to download sequence data.

4 Prospect

Using ALIS data management systems will be able to acquire any sequence data in the files suited to our directory structure. Improvement of sequencing technology makes easier to get mega-scaled sequence data, so it will be power-full strategy to search genes in near future. On the other side mega-scaled sequencing generates huge data, so it is not easy to manage and release data in general biological laboratories. Our system will be able to support such a mega-scaled biological sequencing projects.