A DNA Sequencing Project Monitoring and Sequence Data Quality Control System

Akira Ohyama [1]
akr@mitsui-knowledge.co.jp

Masaki Takazawa [2]
takazawa@kazusa.or.jp

Atsuko Kubota [2]
kubota@kazusa.or.jp

Nobuyuki Miyajima [2]
miyajima@kazusa.or.jp

[1] Mitsui Knowledge Industry Co. Ltd.
3-7-4 Kojimachi, Chiyada-ku, Tokyo 102, Japan

[2] Kazusa DNA Research Institute
1532-3 Uchino, Yana, Kisarazu, Chiba 292, Japan

Abstract

A solution system for the bottlenecks of a large-scale DNA sequencing project is now under development in the Kazusa DNA Research Institute. The system provides an integrated sequencing project management capabilities. Its functions are ranged from monitoring of running equipments, viewing of every data flow of DNA clones, contigs and raw data from automatic sequencers, commands for setting instructions to specified clones, a multiple alignment editor, to sequencing data quality control functions.

1 Introduction

A large scale DNA sequencing project generates a huge amount of data flow of either genomic or complementary DNA clones, contigs and randomly selected fragments, which are mutually related in fairly complicated manners. Although high-speed automatic DNA sequencers are introduced to these projects, there still remains critical bottlenecks in the course of a project. One bottleneck is located in the reconstruction of consesus sequences from many random DNA fragments. Other difficulties are also encountered when trying to overview the progress of indivisual sequencing job. A solution system for these sequencing bottlenecks is now under development in the Institute. In addition to the software solution to these bottlenecks, this system will provide various monitoring functions of the project resources.

[1]大山 彰、 三井情報開発株式会社、〒 102 東京都千代田区麹町 3-7-4
[2]高沢 昌樹、 かずさ DNA 研究所、 〒 292 千葉県木更津市矢那内野 1532-3
[2]久保田 敦子、 かずさ DNA 研究所、 〒 292 千葉県木更津市矢那内野 1532-3
[2]宮嶋 伸行、かずさ DNA 研究所、〒 292 千葉県木更津市矢那内野 1532-3

2 System

The current version of the software runs on the Sun OS 4.1.3, Solaris 2.3 and OSF/1. The programs are written in C on X-Window System and OSF Motif version 1.2. This system can be directly conneted to DNA fragments assemblers like INHERIT.

1. Project Progress and Equipment Status Monitoring
   In any large-scale DNA sequencing project, clones are concurrently assigned to different stages of process. Thus, the Project Progress Monitor is to be incorporated to overview each project progress or the specified project. The program also provides some graphical overviews of every clone currently running, as well as finished clones. Any automated analytical equipment would occasionally fall into various trouble, eventhough a high MTBF is accomplished for these equipment. The Equipment Monitor shows the current availability of each equipment in the system.

2. DNA Data Management
   In the system, an identifier is uniquely provided for each cDNA or DNA clone. Using this ID, the DNA Viewer enable users to overview the attributes of any clones, contigs and DNA fragment, as well as image data from Northern Blotting, electrophoretic radiogram from automatic DNA sequencers.

3. Sequence Quality Control
   The quality of the consensus sequences which are reconstructed from many overlapped fragments of randomly selected ones from the original clone. The redundancy of participant fragments to each consensus nucleotide determines the accuracy of the consensus sequences. The Sequence Quality Control adopts-two stage scoring. One is for the quality of the sequenced fragments, others is for the contig and consensus sequences.

4. Multiple Alignment Editing
   The big bottleneck in a sequencing project exists in the edit work of consensus sequences which are deribed from a DNA sequence assembler. Almost all the results from a DNA fragment assembler require a validation of human eyes, this could be completed after time-comsuming monotonous checking works of every sequence and electrophoretic data.The Multiple Alignment Editor provides total multiple alignment editing environments.

3 Conclusions

In a large-scale DNA sequencing project, existing critical bottlenecks could be conquered by an integrated software and hardware connected to network which enables effective use of availalbe equipments and human resources and realize high accuracy of the DNA data just sequenced.