A High-throughput Gene Discovery System for *Arabidopsis* Genome Sequencing Project

Yasukazu Nakamura **Satoshi Tabata**
ynakamu@kazusa.or.jp tabata@kazusa.or.jp
Laboratory of Gene Structure 2, Kazusa DNA Research Institute
1532-3 Yana, Kisarazu, Chiba 292-0812, Japan.

1 Introduction

Arabidopsis thaliana is a small weed in the mustard family *Brassicaceae* which has become the model organism for research in plant biology. The 130-megabase genome of the plant is organized into five chromosomes and contains an estimated 20,000 genes. To understand the entire genetic system in this plants, we initiated large-scale sequencing project of the *Arabidopsis thaliana* genome. We are taking part in sequencing of the entire bottom arm and portions of the top arm of chromosome 5, and also the top arm of chromosome 3 along the line of the international agreement of the Arabidopsis Genome Initiative (AGI). The entire genome is scheduled to be sequenced by July of the year 2000. During the process of annotating genomic sequence of clones on chromosome 3 and 5, we have constructed a system for high-throughput gene modeling process. Examples of our gene finding process and features of deduced genes will be presented.

2 Materials and Methods

We selected the clones containing DNA markers on each chromosome from P1, TAC and BAC libraries. The nucleotide sequence of each clone was determined according to the shotgun based strategy as described in previous paper [1].

Nucleotide sequences were subjected to similarity search against non-redundant databases. Potential exons, gene models and intron-exon boundaries were predicted by computer programs. Then all outputs were parsed and stored in a database. The libraries and computer programs used for automated analyses were the same as those described in the previous report [2].

3 Description

We developed a protocol to automate the execution of similarity search and gene prediction programs. The results are parsed and loaded into a web based display system named *Arabidopsis* Genome Displayer. Displayer shows relationship of the features in the database along a genomic sequence. Simultaneously, annotation-making interface allows manual editing of gene model showing tentative sequences of nucleotide and protein and images of exon-intron organization. An annotator perform database searches on each working model during gene-modeling process. After careful editing process, the most reasonable gene model on a region is saved into in-house database as an deduced gene. High-throughput analysis of *Arabidopsis thaliana* genomic sequences have been carried out with the assistance of the system.

As of October 1, 1999, Kazusa DNA Research Institute has released the nucleotide sequences of 383 P1, TAC and BAC clones (in total, 23,097,056 base pairs). Analyzed information will be detailed on our web site (http://www.kazusa.or.jp/arabi/) and DDBJ/EMBL/GenBank

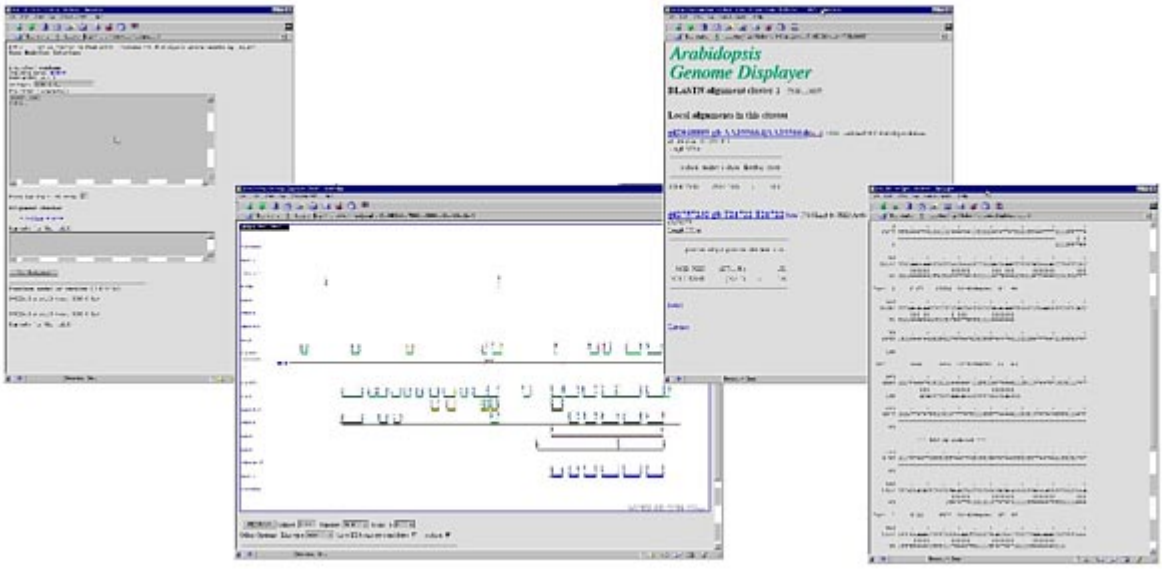

Figure 1: Examples of gene-modeling process on the system.

international DNA Databases. Also, *Arabidopsis* Genome Displayer is a public database (http://www.kazusa.or.jp/arabi/displayer/), which provides genomic information of 85,127,092 bases for 1,050 clones as AGI total. This service enable users to browse original annotation and re-computational information for all sequences nucleated by AGI participants.

Acknowledgments

We thank Takaharu Kimura, Mitsuyo Kohara, Atsuko Kubota, Shinobu Nakayama and Sayaka Shinpo for their excellent technical assistance. This work was supported by the Kazusa DNA Research Institute Foundation.

References

[1] Sato, S., Kotani, H., Nakamura, Y., Kaneko, T., Asamizu, E., Fukami, M., Miyajima, N., and Tabata, S., Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones, *DNA Research,* 4:215–230, 1997.

[2] Nakamura, Y., Sato, S., and Tabata, S., Annotation and presentation systems for *Arabidopsis* genome sequencing project at Kazusa DNA Research Institute, *Genome Informatics 1998*, Universal Academy Press, 361–362, 1998.