Gene Discovery and Presentation Systems for *Arabidopsis* Genome Sequencing Project at Kazusa DNA Research Institute

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

Keywords: gene discovery, gene presentation, *Arabidopsis* genome sequencing project

1 Introduction

Arabidopsis thaliana is a small weed in the mustard family *Brassicaceae* which has become the model organism for research in biology of higher plant. 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 [1]. 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 and a web-based data presentation system.

2 Description

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 [2].

Nucleotide sequences were translated in six frames using the universal codon table, and each frame was subjected to similarity search against the non-redundant protein database, nr, using the PSI-BLAST program [3]. Each local alignment, which showed E-value of 0.001 or less to known protein sequences, were extracted and stored. In order to predict exact donor/acceptor sites for splicing, alignments made by nap in AAT package [4] and Wise2 [5] were also examined. Potential exons were predicted by the computer programs Grail [6] and GENSCAN [7]. For localization of exon-intron boundaries, donor/acceptor sites for splicing were predicted by NetGene2 [8] and SplicePredictor [9]. To identify transcribed regions and structural RNA genes, nucleotide sequences were compared with the EST and RNA gene datasets extracted from GenBank [10] with the BLAST2 [3] program. For assignment of tRNA gene and structure of tRNA, prediction by the tRNA-scanSE [11] was carried out. Alignments made by gap in AAT [4] were also examined to fit EST sequences on genomic sequence. All the outputs were parsed and stored in the same format specified as GFF (Gene-Finding Format) [12]. 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 January, 1999, *Arabidopsis* Genome Displayer (http://www.kazusa.or.jp/arabi/displayer/), provides genomic information of 95,938,450 bases for 1,172 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] http://www.arabidopsis.org/AGI/AGI_data_release.html

[2] 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.

[3] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., Gapped BLAST and PSI-BLAST, *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[4] Huang, X., Adams, M.D., Zhou, H. and Kerlavage, A.R., A tool for analyzing and annotating genomic sequences, *Genomics*, 46:37–45, 1997.

[5] http://www.sanger.ac.uk/Software/Wise2/

[6] Uberbacher, E.C. and Mural, R.J., Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach, *Proc. Natl. Acad. Sci. USA*, 88(24):11261–11265, 1991.

[7] Burge, C.B. and Karlin, S., Finding the genes in genomic DNA, *Curr. Opin. Struct. Biol.*, 8(3):346–354, 1998.

[8] Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P. and Brunak, S., Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information, *Nucleic Acids Res.*, 24(17):3439–3452, 1996.

[9] Brendel, V. and Kleffe, J., Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA, *Nucl. Acids Res.* 26(20):4748–4757, 1998.

[10] Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J. and Ouellette, B. F., GenBank, *Nucl. Acids Res.* 26:1–7, 1998.

[11] Lowe, T.M. and Eddy, S. R., tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucl. Acids Res.* 25(5):955–964, 1997.

[12] http://www.sanger.ac.uk/Software/GFF/