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

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

1 Introduction

To understand the entire genetic system in higher plants, we initiated large-scale sequencing 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]. We selected the clones containing DNA markers on each chromosome from Mitsui P1 and TAC libraries. The authenticity of the selected clones was examined by anchoring both ends of the clones onto YAC tiling paths of the chromosomes by PCR. The nucleotide sequence of each clone was determined according to the shotgun based strategy with an approximate redundancy of 10.

The finished sequences were subjected to similarity search and computer-aided analysis for prediction of protein and RNA coding regions in an semi-automatic system. The nucleotide sequences are available on the public DNA database and on our web database.

As of October, 1998, the nucleotide sequences of 174 P1 and TAC clones (12,242,188 bp) have been released. Sequence information on 110 clones (8,138,210 bp) with annotation has been detailed on our web site. A total of 1,927 potential protein-coding genes and/or gene segments were identified in annotated regions.

2 Improved Annotating Procedure

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 [2]. 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 [3] and Wise2 [4] were also examined.

Potential exons were predicted by the computer programs Grail [5] and GENSCAN [6]. For localization of exon-intron boundaries, donor/acceptor sites for splicing were predicted by NetGene2 [7] and SplicePredictor [8].

To identify transcribed regions and structural RNA genes, nucleotide sequences were compared with the EST and RNA gene datasets extracted from GenBank [9] with the BLAST2 [2] program. For assignment of tRNA gene and structure of tRNA, prediction by the tRNA-scanSE [10] was carried out. Alignments made by gap in AAT [3] 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) [11]. Then the contents were combined to create an HTML-based form by annotation composing system for gene-modeling process.

3 Data Presentation Site

A web site KAOS (Kazusa *Arabidopsis* data Opening Site) was originally designed to present information on the *Arabidopsis* genome sequencing project at Kazusa DNA Research Institute, which includes the nucleotide sequences with annotation, physical maps of the chromosomes, status of the project etc.

An *Arabidopsis* Genome Displayer is a new feature of KAOS. The purpose of this service is to enable users to browse the annotated sequence data deposited from all the sequencing teams of AGI through an user-friendly graphic display system and search engines. Gene structures proposed in the annotated sequences as well as those predicted by computer programs (detailed above) are presented and each graphic item is a hyperlink to detailed information of the corresponding area. This site provides standard view of the *Arabidopsis* genome information for research community. The genome displayer will be available through the world wide web site KAOS at *http://www.kazusa.or.jp/arabi/.*

Acknowledgments

This work was supported by the Kazusa DNA Research Institute Foundation.

References

[1] http://genome-www3.stanford.edu/cgi-bin/Webdriver?MIval=atdb_registry_info.html#agi-info

[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucl. Acids Res.*, 25:3389–3402, 1997.

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

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

[5] 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*, 88:11261–11265, 1991.

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

[7] Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P. and Brunak, S., Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information, *Nucl. Acids Res.*, 24:3439–3452, 1996.

[8] 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.*, 26:4748–4757, 1998.

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

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

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