A High-Throughput Annotation System of Rice Genome Sequence at RGP

Atsuko Idonuma¹ Nobukazu Namiki¹ Takashi Matsumoto¹
aido@staff.or.jp namiki@staff.or.jp mat@rias.affrc.go.jp

Satoshi Nobushima² Shigehiro Honda² Takuji Sasaki¹
nobu@mi.mss.co.jp honda@mi.mss.co.jp tsasaki@rias.affrc.go.jp

¹ Rice Genome Research Program, NIAR/STAFF, 446-1 Ippaizuka, Kaniyokoba, Tsukuba, Ibaraki 305-0854, Japan
² Mitsubishi Space Software Co., Ltd., 1-17-15 Sengen, Tsukuba, Ibaraki 305-0047, Japan

Keywords: annotation, rice, genome sequencing

1 Introduction
Rice is a major target plant for whole genome sequencing because it is a valuable source of food to an ever-expanding world population. Its genome size of nearly 430 Mb is the smallest among cereal crops and could therefore provide valuable information for understanding many agricultural crops. The Rice Genome Research Program (RGP) has already completed the high-quality sequencing and analysis of chromosome 1 which is the largest chromosome in the genome. Currently, sequencing of chromosomes 2, 6, 7, 8 and 9 is in progress as part of the International Rice Genome Sequencing Project (IRGSP). The entire genome is expected to be sequenced as at least phase 2 level by December 2002. Annotation of the sequence data produces the fundamental knowledge on the structure and function of all genes in an organism, and provides the essential tools for manipulation of biologically important genes.

We have constructed a high-throughput annotation pipeline system to facilitate efficient analysis, visualization and storage of the genome sequence data. Although we basically use the results of automated annotation, much of our effort is concentrated on editing the autopredicted genes to come up with the most accurate gene model based on existing evidence. In this report, we will describe our high-throughput annotation pipeline system with emphasis on the curation of the output of automated annotation using an editing tool called AnnotationPlot.

2 Annotation Pipeline
The high-throughput annotation pipeline system consists of automated annotation, curation of autopredicted genes and storage of all data in a relational database (Figure 1). Upon completion of sequencing to phase 3 level, the sequence data from each PAC/BAC clone is subjected to automated annotation using RiceGAAS (Rice Genome Automated annotation System, http://ricegaas.dna.affrc.go.jp/). The system automatically performs BLAST searches against rice EST and non-redundant protein database; prediction of coding regions using FGENESH GENSCAN, RiceHMM, MZEF, SplicePredictor; and structural analysis of the sequence using RepeatMasker, tRNAscan, MOTIF, SOSUI etc. The results of these analyses are integrated automatically to generate an autopredicted gene and transferred to the AnnotationPlot for manual editing. We also analyze the sequence using prediction programs which are trained for rice, namely, GeneMark.hmm (O. sativa, http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi) and GlimmerM (O. sativa, http://www.tigr.org/tdb/glimmerm/glmr_form.html) to improve the prediction accuracy. The results are combined with the output of RiceGAAS in the AnnotationPlot. For coding sequences with EST or protein matches, gene models are primarily constructed using these evidences. However for coding sequences without database matches, gene modeling is based only on the output of the
different prediction programs. In such cases, we are using a modified version of the TIGR Combiner (`ftp://ftp.tigr.org/pub/software/Combiner/`) to select a single best prediction for each gene model predicted by multiple gene prediction programs. Coding sequences predicted by a single prediction program are classified as singly predicted ORFs. The functions of the identified coding sequences are inferred from non-redundant protein database search with BLAST. Finally, the annotation for each clone is stored in the Rice Genome Annotation Database (RAD). All annotated sequence are released to the public domain through DDBJ/EMBL/GenBank database and accessible through the rice genome central database INtegrated rice genome Explorer (INE) on our website at http://rgp.dna.affrc.go.jp/.

3 Features of AnnotationPlot

The AnnotationPlot facilitates editing and refining of autopredicted genes (Figure 2). Among the major features of this tool are as follows: (1) provides an overview of the results of database searches and analysis with different prediction programs; (2) a scrollable frame allows viewing the entire sequence, (3) EST/protein searches can be altered by setting up threshold with BLAST score, E-value and identities; (4) EST/protein search alignments are presented, (5) automatic clustering and selection of the best EST/protein search results, (6) start/stop codons and splice sites are presented, (7) exon start/end positions are adjusted per one base, (8) translation of coding sequence to amino acid sequence and semi-automatic protein database search, (9) an annotation map is constructed using a png file output; and (10) all annotation results output are supported with csv format.

We have been using this AnnotationPlot for analysis of chromosome 7 sequence data. As of October 2002, we have annotated a total of 108 PAC/BAC clones covering about 12,271,623 bp of the chromosome. 2,733 predicted coding sequences including transposable elements and singly predicted ORFs were identified. This work is supported by a grant from the Ministry of Agriculture, Forestry and Fisheries (GS1201).