98–102 Nucleic Acids Research, 2002, Vol. 30, No. 1

RiceGAAS: an automated annotation system and database for rice genome sequence

Katsumi Sakata*, Yoshiaki Nagamura, Hisataka Numa, Baltazar A. Antonio[1], Hideki Nagasaki[1], Atsuko Idonuma[1], Wakako Watanabe[2], Yuji Shimizu[2], Ikuo Horiuchi[2], Takashi Matsumoto, Takuji Sasaki and Kenichi Higo

National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan, [1]Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, 446-1 Ippaizuka, Kamiyokoba, Tsukuba, Ibaraki 305-0854, Japan and [2]Mitsubishi Space Software Co. Ltd, 1-17-15 Sengen, Tsukuba, Ibaraki 305-0047, Japan

Received August 14, 2001; Revised and Accepted October 16, 2001

ABSTRACT

An extensive effort of the International Rice Genome Sequencing Project (IRGSP) has resulted in rapid accumulation of genome sequence, and >137 Mb has already been made available to the public domain as of August 2001. This requires a high-throughput annotation scheme to extract biologically useful and timely information from the sequence data on a regular basis. A new automated annotation system and database called Rice Genome Automated Annotation System (RiceGAAS) has been developed to execute a reliable and up-to-date analysis of the genome sequence as well as to store and retrieve the results of annotation. The system has the following functional features: (i) collection of rice genome sequences from GenBank; (ii) execution of gene prediction and homology search programs; (iii) integration of results from various analyses and automatic interpretation of coding regions; (iv) re-execution of analysis, integration and automatic interpretation with the latest entries in reference databases; (v) integrated visualization of the stored data using web-based graphical view. RiceGAAS also has a data submission mechanism that allows public users to perform fully automated annotation of their own sequences. The system can be accessed at http://RiceGAAS.dna.affrc.go.jp/.

INTRODUCTION

Rice is one of the major cereal crops and is the principal source of food for approximately half of the world's population. The genome organization of cereals exhibits a high degree of synteny (1) and rice, which has the smallest genome size (430 Mb) among the major cereal crops, provides the most suitable material for genome analysis.

The International Rice Genome Sequencing Project (IRGSP; see URL in Table 1) was established in 1998 with the goal of sequencing the entire rice genome (2). As of August 2001, 11 countries including Brazil, Canada, China, France, India, Japan, Korea, Taiwan, Thailand, UK and USA are actively involved in the project which is targeted to be completed by 2004. The Rice Genome Research Program (RGP; see URL in Table 1), based in Tsukuba, Japan, is one of the major sequencing sites for IRGSP. So far, >95 Mb of the rice genome has been sequenced by RGP and released to the public domain through the DNA Data Bank of Japan (DDBJ) and the integrated rice genome database INtegrated rice genome Explorer (INE) developed at RGP (3).

As in other large sequencing endeavors such as the human and the *Arabidopsis* genome projects, annotation is an essential part of the rice genome sequencing to enhance the value of the sequence data. With the acceleration of the sequencing efforts in IRGSP, a scheme to achieve a high-throughput and timely annotation and a reliable database to store and retrieve the analysis results have become indispensable. Thus an annotation system called Rice Genome Automated Annotation System (RiceGAAS) was developed to achieve a systematic and comprehensive annotation of accumulated sequence data. At the same time, a database containing the annotation of submitted rice genome sequences to GenBank has been made available to the public domain.

DATABASE CONTENTS

RiceGAAS stores the following data: (i) GenBank entry files of rice genome sequences; (ii) results of various analyses for the sequences including homology search against protein and rice expressed sequence tag (EST) databases, gene prediction using various programs as well as analysis of exons, splice sites, repeats and transfer RNA; (iii) predicted genes and long terminal repeats (LTRs) by RiceGAAS based on an algorithm which combines multiple gene prediction programs with homology search results; (iv) detailed information on predicted genes including plausible gene models. All of these data are completely produced automatically upon subjecting a genome sequence data for annotation.

The stored data are visualized using a web-based graphical view. To display the annotation for the collected rice genome sequences, RiceGAAS has three types of web pages: (i) a

*To whom correspondence should be addressed. Tel: +81 298 38 2199; Fax: +81 298 38 2302; Email: ksakata@nias.affrc.go.jp

Table 1. Useful URLs for RiceGAAS and related sites

RiceGAAS	http://RiceGAAS.dna.affrc.go.jp/
IRGSP	http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/seqcollab.pl
RGP home page	http://rgp.dna.affrc.go.jp/
INE	http://rgp.dna.affrc.go.jp/giot/INE.html
Rice Genome Annotation	http://RiceGAAS.dna.affrc.go.jp/rgadb/
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
GENSCAN	http://genes.mit.edu/GENSCAN.html
RiceHMM	http://rgp.dna.affrc.go.jp/RiceHMM/index.html
MZEF	http://argon.cshl.org/genefinder/
SplicePredictor	http://bioinformatics.iastate.edu/cgi-bin/sp.cgi
printrepeats	http://www.sanger.ac.uk/Software/sequencing/docs/printrepeats/
RepeatMasker	http://ftp.genome.washington.edu/
tRNAscan	http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
HMMER	http://pfam.wustl.edu/hmmsearch.shtml
ProfileScan	http://www.isrec.isb-sib.ch/software/PFSCAN_form.html
MOTIF	http://www.motif.genome.ad.jp/MOTIF3.html
PSORT	http://psort.ims.u-tokyo.ac.jp/
SOSUI	http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html
PLACE-SignalScan	http://www.dna.affrc.go.jp/htdocs/PLACE/signalscan.html
MAFF DNA Bank	http://www.dna.affrc.go.jp/
Comparison table of gene prediction	http://RiceGAAS.dna.affrc.go.jp/rga-bin/col_accur.pl

tabulated list of collected bacterial artificial chromosome (BAC) and P1-derived artificial chromosome (PAC) clones for each chromosome; (ii) an annotation map for each clone; (iii) information for a predicted gene. The tabulated list of clones includes the clone name, accession number, chromosome location, clone size and date when the sequence was collected and analyzed. An analysis result for each clone can also be searched by using the chromosome number and clone name from the Rice Genome Annotation Database page (see URL in Table 1). An annotation map is linked from the clone name on the tabulated list. One of the distinctive features of RiceGAAS is that all analysis results for a sequence are integrated in an annotation map, as shown in Figure 1. An information window for a predicted gene is linked from the predicted gene on the annotation map. An example of the information window for a predicted gene is shown in Figure 2.

Furthermore, text files of some analysis results including BLAST, HMMER and RepeatMasker stored in the system can be directly retrieved using a keyword search mechanism from the Rice Genome Annotation Database page of the system. This search function is useful, for example, when a user tries to find predicted genes with significant homology to proteins in a particular functional category of interest.

INTEGRATED PROGRAMS AND SYSTEM FLOWCHART

RiceGAAS integrates several programs for prediction and analysis of protein-coding gene structure. It is designed specifically to provide an efficient system and comprehensive database for annotation of rice genome sequences. All rice genome sequences in GenBank are automatically collected by the system. A keyword search is executed on a daily basis against the GenBank database to collect the sequences in high-throughput genomic (HTG) sequence division of *Oryza sativa*. As of August 2001, a total of 1011 sequence data have been collected. An automated annotation is performed for the collected HTG phase 2 or 3 sequences and then made available to the public via the Internet.

A total of 14 analysis programs are incorporated into the system. These include BLAST for homology search against protein database and rice EST database; GENSCAN and RiceHMM for gene domain prediction; MZEF for exon prediction; SplicePredictor for splice site prediction; Printrepeats and RepeatMasker for repetitive sequence detection; tRNAscan for transfer RNA prediction; HMMER, ProfileScan and MOTIF for homology search against amino acid sequence motif database; PSORT for protein localization site prediction; SOSUI for classification and secondary structure prediction of membrane protein; and PLACE-SignalScan for *cis*-element detection (see URLs in Table 1). RiceHMM refers to a gene domain prediction program developed at RGP and targeted exclusively for rice. It is based on the probability scheme of the hidden Markov model and redefined by using a catalog of rice ESTs composed of nearly 15 000 cDNAs (4).

A system flowchart of RiceGAAS is shown in Figure 3. The computer tasks are programmed to finish the entire process from the beginning of analyses to visualization of results within 48 h.

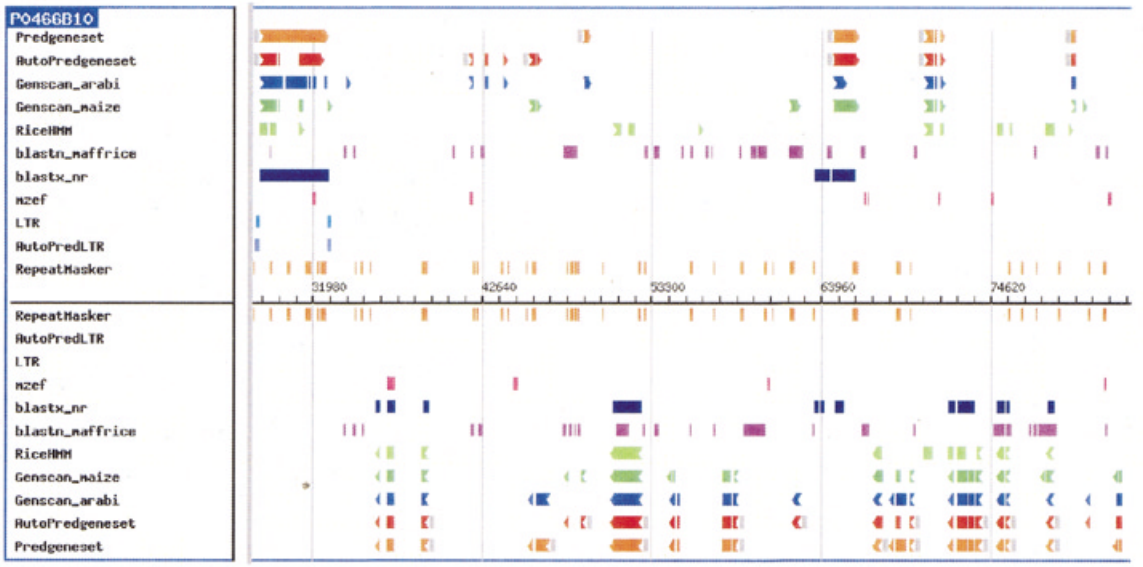

Figure 1. An annotation map for P0466B10 clone. The upper and lower parts of the map show the analysis results for the normal and complementary strands, respectively. The horizontal line on the middle shows the position in nucleotide base pairs (bp). Predgeneset represents the manually predicted gene from submitted sequences in the collected GenBank entry file. AutoPredgeneset represents the coding region of automatically predicted gene by the system. The translation initiation and termination domains for each predicted gene are indicated by an arrow tail and head, respectively. A predicted gene in Predgeneset or AutoPredgeneset is linked to a corresponding predicted gene page. The gray colored box located in the upstream region of the predicted gene represents *cis*-elements in the region detected using plant *cis*-acting regulatory DNA elements (PLACE) as a reference database. Clicking the box shows a tabulated list of these *cis*-elements. The regions detected by the prediction programs such as Genscan_arabi, Genscan_maize and RiceHMM as well as the results of homology search against rice EST and protein databases are also represented. LTR shows the predicted LTR in the GenBank entry file, whereas AutoPredLTR shows the automatically predicted LTR by the system. RepeatMasker shows the detected repetitive sequence by the RepeatMasker program. All objects are clickable and linked to the corresponding page of analysis results.

RiceGAAS uses several reference databases to facilitate a comprehensive analysis of the genome sequence. These include the National Center for Biotechnology Information (NCBI) non-redundant protein database (nr) for homology search by BLASTX and BLASTP; the MAFF RICE database of the Ministry of Agriculture, Forestry and Fisheries (MAFF) DNA Bank (see URL in Table 1) for homology search against rice EST database by BLASTN, PLACE database for *cis*-element detection (5), Pfam and PROSITE databases for homology search against amino acid sequence motif database using HMMER and ProfileScan, respectively. These reference databases are uploaded from the original sites and stored in the system and regularly updated at the original sites. Therefore, users can avail of the latest entries in reference databases to analyze their sequences. This is achieved with an automatic re-analysis function so that if a reference database is updated, RiceGAAS immediately uploads the new one, which is used for analysis. Furthermore, the rice genome sequences collected from GenBank are periodically re-analyzed. The re-analysis frequency is 150 days for BLASTX and 20 days for other programs. Every time the re-analysis is done, the gene prediction is automatically re-executed.

A data submission mechanism is also prepared for public use to enable users to annotate their own sequences. User submitted sequences are annotated in the same way as collected sequences from GenBank. For this purpose, a user needs to register and enter a password which serves as security control to prevent disclosure of the submitted sequence and analysis results to another user.

AUTOMATIC INTERPRETATION OF CODING REGIONS

RiceGAAS automatically integrates multiple analysis results and interprets coding regions which configure the open reading frames (ORFs) correctly. The algorithm for gene domain prediction in RiceGAAS was designed based on the concept of combining multiple gene prediction programs with homology search results. This reduces significant false positive rate and minimizes missing many correct exons, often encountered using a single prediction program (6). The standard procedure used by annotators at RGP serves as the fundamental structure in designing the algorithm. Furthermore, it was designed to predict as many genes as possible in order to increase the chance of finding out the correct genes among the candidate genes.

The algorithm consists of the following four stages:

1. Score the exons given by prediction programs (GENSCAN for *Arabidopsis*, GENSCAN for maize and RiceHMM) based on the probability that a particular sequence segment is an exon (7). If the predicted exon overlaps a region detected by another prediction program or homology search against rice EST or protein database, the score for the exon is increased.
2. Re-assemble predicted genes which overlap with predicted LTRs by omitting the exon which overlaps an LTR and reconfiguring exons between LTRs.
3. Select a gene model from the results of the prediction programs by choosing predicted genes based on the average exon scores. The algorithm is considered not to predict multiple genes at each domain on the sequence.

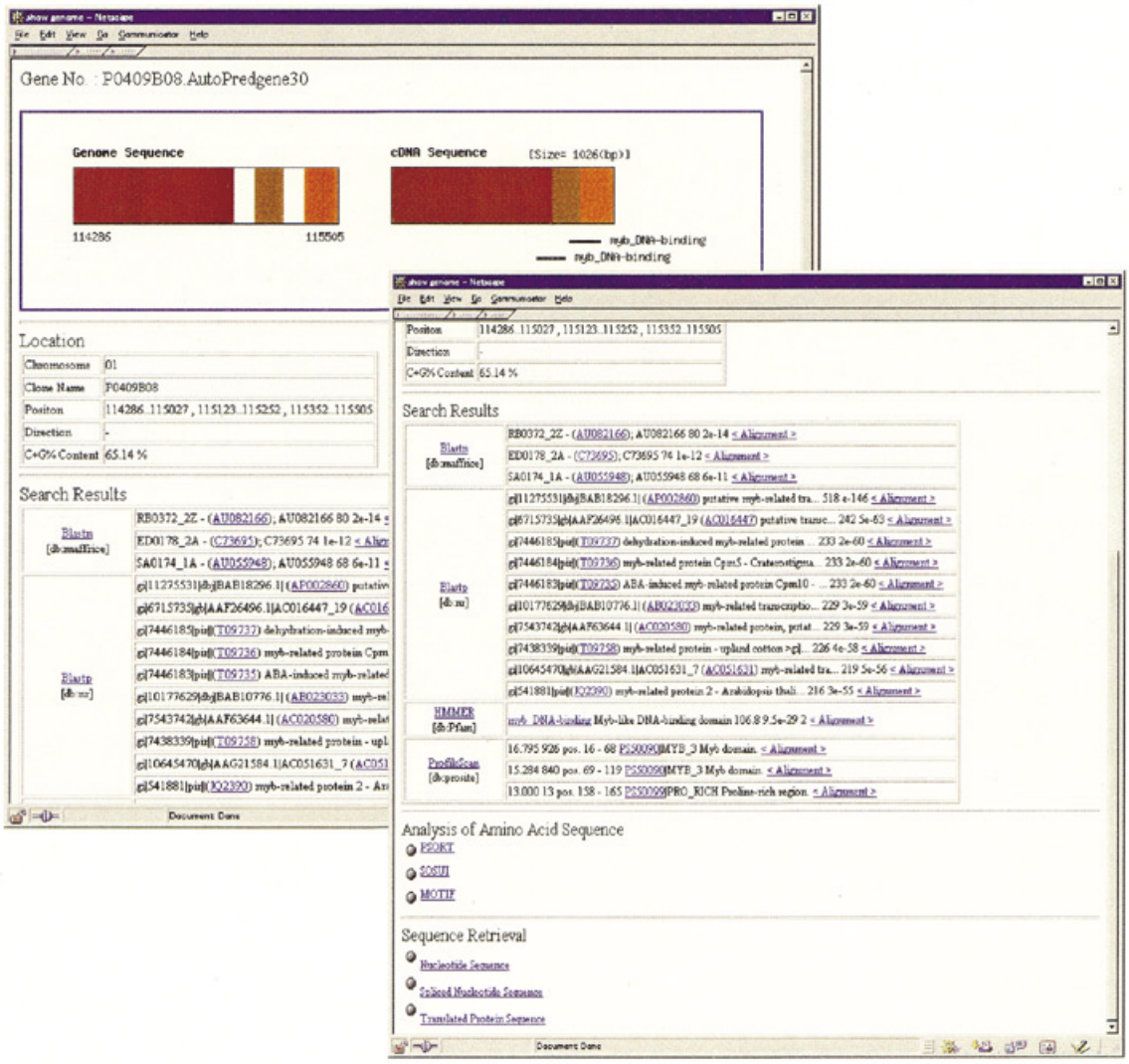

Figure 2. An information window for a predicted gene (P0409B08.AutoPredgene30). RiceGAAS provides detailed information for each predicted gene such as: (i) structure of predicted gene; (ii) position on the clone; (iii) homology search results against protein, rice EST and amino acid sequence motif databases; (iv) links to some analyses of amino acid sequence of the predicted gene such as protein localization site prediction; and (v) links to the pages containing nucleotide and amino acid sequence of the predicted gene.

4. Insert internal exons predicted by MZEF if the exons configure an ORF correctly.

We evaluated the algorithm by comparing genes automatically predicted by RiceGAAS with manually predicted genes in which the results of gene prediction programs and homology search are edited to come up with the most plausible gene model. The number of nucleotides included in the automatically predicted genes and manually predicted genes, and the number of nucleotides included in both the automatically and manually predicted genes, were counted. Then, the sensitivity, which refers to the number of nucleotides in manually predicted genes covered in automatically predicted genes, and specificity, which refers to the number of nucleotides in automatically predicted genes covered in manually predicted genes, were calculated. The results are shown in the table of gene prediction page (see URL in Table 1). As of October 2001, the sensitivity was calculated to be 0.77 whereas the specificity was calculated to be 0.71. This indicates that ~74% of the predicted genes are the same at nucleotide level between automatically and manually predicted genes. This accuracy of the predicted gene was determined by using 6728 manually predicted genes.

RiceGAAS also provides automatic prediction of the following LTRs:

1. Detect tandem or inverted repeats. At this stage, the input sequence is divided into 4 kb fragments and the homology between each fragment and the input sequence is examined.
2. Select repeats with regions showing homology to transposable elements between them.

FUTURE PLANS

Following the progress of rice genome sequencing, accumulated data will require more powerful automated annotation. Accelerating automated annotation by upgrading the BLAST server will be achieved by the end of 2001. This improvement will be the final step required to accomplish the analysis process overnight

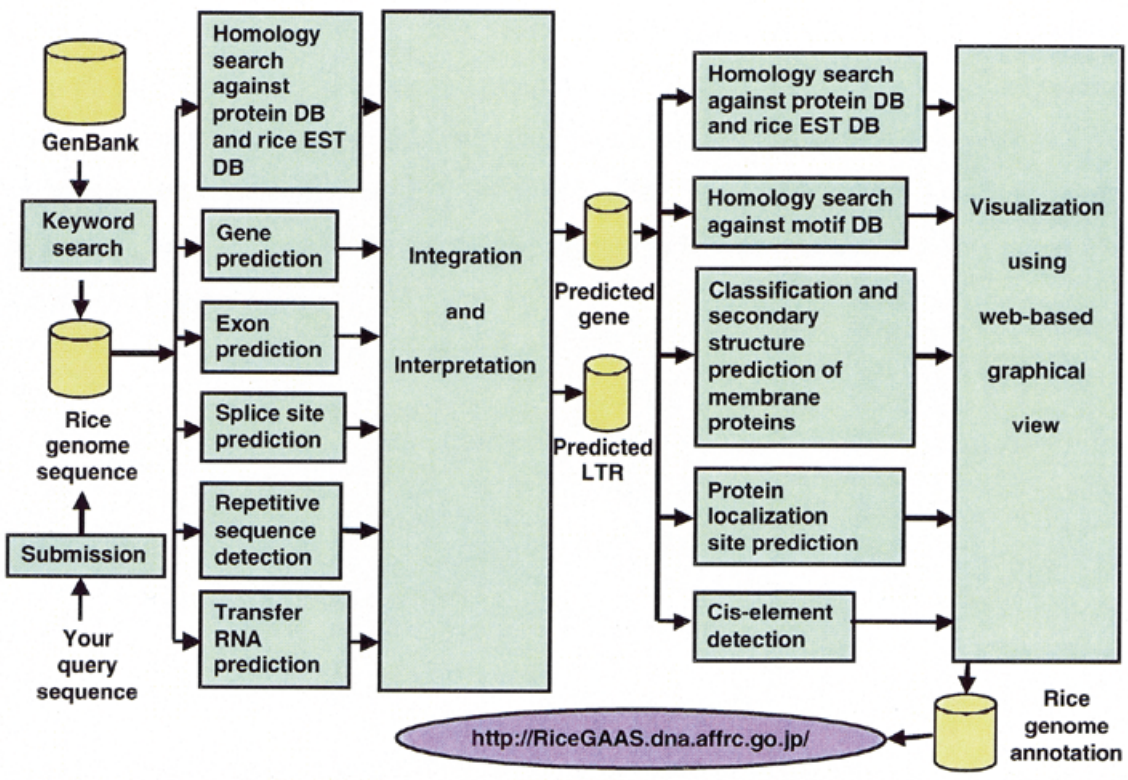

Figure 3. System flowchart of RiceGAAS. Sequences collected from GenBank and submitted by users are stored in the genome sequence database. The sequence is analyzed for homology against protein and rice EST databases, gene prediction using various programs as well as analysis of exons, splice sites, repeats and transfer RNA. The results are integrated and automatically interpreted and the predicted genes and LTRs are stored in respective databases. Predicted genes are further analyzed to come up with the most plausible gene model. All analysis results are visualized using web-based graphical view and are made accessible via the Internet.

(12 h). The algorithm for integration and interpretation of analysis results will be improved to make differences between the manually and automatically predicted genes insignificant. Also, some functions will be added for the user's convenience including: (i) filtering BLAST search results based on the expect value; (ii) highlighting keywords in the retrieved files by the keyword search mechanism; (iii) examining homology of predicted genes against public databases containing nucleotide sequences other than rice ESTs; and (iv) estimating the unique function of predicted gene based on a homology search.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Project GS-1302).

REFERENCES

1. Moore,G., Devos,K.M., Wang,Z. and Gale,M.D. (1995) Grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.
2. Sasaki,T. and Burr,B. (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.*, **3**, 138–141.
3. Sakata,K., Antonio,B.A., Mukai,Y., Nagasaki,H., Sakai,Y., Makino,K. and Sasaki,T. (2000) INE: a rice genome database with an integrated map view. *Nucleic Acids Res.*, **28**, 97–102.
4. Sakata,K., Nagasaki,H., Idonuma,A., Watanabe,W., Kise,M. and Sasaki,T. (2000) RiceHMM: gene domain prediction program for rice genome sequence. *Abstracts of 4th Annual Conference on Computational Genomics*. p. 31.
5. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga T. (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.*, **27**, 297–300.
6. Fortna,A. and Gardiner,K. (2001) Genomic sequence analysis tools: a user's guide. *Trends Genet.*, **17**, 158–164.
7. Burge,C. and Karlin,S. (1997) Prediction of complete gene structure in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.