JARQ 43 (2), 103 – 113 (2009) http://www.jircas.affrc.go.jp

A Bioinformatics Resource for Crop Functional Genomics: GFSelector Module in Automated Annotation System, RiceGAAS

Katsumi SAKATA[1†], Hiroshi IKAWA[1,4†], Hiroyuki WATANABE[1,5], Ikuo A SHIKAWA[2], Yuji SHIMIZU[1], Ikuo HORIUCHI[1], Baltazar A. ANTONIO[3], Hisataka NUMA[3], Yoshiaki NAGAMURA[3] and Takashi MATSUMOTO[3]*

[1] Genome Informatics Department, Mitsubishi Space Software Co., Ltd. (Tsukuba, Ibaraki 305–0032, Japan)
[2] Research Team for Wheat and Barley Biotechnology, National Institute of Crop Science, National Agriculture and Food Research Organization (Tsukuba, Ibaraki 305–8518, Japan)
[3] Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences (Tsukuba, Ibaraki 305–8602, Japan)

Abstract

GFSelector (Gene Function Selector, http://alnilam.mi.mss.co.jp/rgadb/) has been developed to perform computational classification of gene models and assignment of unique biological function. It has been incorporated in RiceGAAS (http://ricegaas.dna.affrc.go.jp/usr/) which was designed to provide an analysis pipeline for user submitted genome sequences and comprehensive database for all rice gene models. The combined system facilitates accurate modelling of predicted rice genes, classification of gene structure, and assigning of function and GO (gene ontology) terms to the gene models. The reliability and accuracy are enhanced by integrating several reference databases into the system and generating multiple candidates for determining the function of the gene models. The pipeline is also fully automated thereby facilitating regularly updates of the rice gene models using the latest reference databases. Annotation of soybean, wheat and banana BAC (bacterial artificial chromosome) sequences was performed to test the applicability of the pipeline to other crops. As compared with the GenBank CDS (coding sequence) features, more than 83% of nucleotide-level sensitivity was obtained for the gene modelling by the pipeline. It was also confirmed that 95% of functional annotation by the pipeline was nearly equal or better than the corresponding GenBank CDS feature.

Discipline: Biotechnology
Additional key words: database, gene ontology, web-based system

Introduction

The International Rice Genome Sequencing Project (IRGSP) has completed a finished quality sequence of the rice genome (*Oryza sativa* L. ssp. *japonica* cv. Nipponbare). The genome size is estimated to be the smallest among the major crops belonging to the Gramineae family[10]. The genome organization of cereals exhibits a high degree of synteny[15], so that rice can serve as a principal model system for cereal genomics. With the availability of the rice genome sequence in the public domain, innovative researches in functional and applied genomics have been accelerated.

A standardized and reliable annotation is indispensable for efficient utilization of the genome sequence. Sequence information in databases continues to increase, for example, the number of sequences stored in GenBank

†These authors contributed equally to this work.
Present address:
[4] Research Division 1, Institute of the Society for Techno-Innovation of Agriculture, Forestry, and Fisheries (Tsukuba, Ibaraki 305–0854, Japan)
[5] Department of Applied Biological Chemistry, Tamagawa University (Machida, Tokyo 194–8610, Japan)
*Corresponding author: e-mail mat@nias.affrc.go.jp
Received 21 March 2008; accepted 3 September 2008.

has increased at a rate of 1.4 times per year (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). At the same time, revisions of previously submitted entries render a lot of obsolete information. Thus, a regularly updated annotation using the latest reference databases is highly desirable.

Several annotation databases have been developed for the rice genome sequence and are currently available in the public domain. The Osa1 database stores rice gene models and functional characterization based on manual annotation[25] and features a rice repeat database and a Distributed Annotation System (DAS)[5]. The Rice Annotation Database (RAD) has been developed to provide a contig-oriented annotation[11]. It facilitates structural analysis for rice genome such as codon usage and splice site statistics. The Rice Information System of Beijing Genomics Institute (BGI-RIS) is based on the draft genome sequence of the *indica* rice variety *93–11*[28]. In addition to the genome annotation, the cDNA sequences of *japonica* rice[14] have been mapped on the *indica* genome. Also, a jamboree style annotation has been implemented for the IRGSP genome sequence and rice full-length cDNA sequences[14] and a database called RAP-DB has been developed to provide access to the manually curated annotation[18]. As for the cereal genomes, Gramene has been developed for comparative genomics with emphasis on rice[13]. It stores various types of data such as genes, genomes, proteins, markers, quantitative trait loci, comparative maps and literature citations using the rice genome sequence as an anchor.

All the databases mentioned above are based on manual annotation and lack a scheme to update the stored data. The RiceGAAS (Rice Genome Automated Annotation System, http://ricegaas.dna.affrc.go.jp/) is a web-based automated system for prediction of coding regions in the rice genome with automated update scheme[20]. However, it could not facilitate functional annotation of the gene models. We have developed the GFSelector and incorporated it into the RiceGAAS to perform a computational classification of gene models and assignment of unique biological functions. In the present paper, we will discuss the functional annotation by the GFSelector, the fully automated pipeline by the combined system of GFSelector and RiceGAAS, and its application to cereal crop analyses.

Materials and methods

Annotation of rice genomic sequences was performed using all rice BAC and PAC (P1-derived artificial chromosome) sequences (phase 2 and phase 3) in the HTG (high-throughput genomic) division of GenBank which are automatically checked daily by a data collecting system. Nearly 4,400 rice BAC/PACs have been analyzed so far. Automated annotation for genomic sequences from other crops was evaluated using 4 wheat BACs (EF426565, DQ537335, 537336, 537337), 2 banana BACs (AY484588, 484589) and 4 soybean BACs (EF533695, 533698, 533700, 533702).

Several analysis programs have been incorporated in the analysis pipeline. A similarity search based on the BLASTP[2] program against a non-redundant protein database available from the ftp site of NCBI BLAST as 'nr' is mainly used to assign a function of gene model. The 'nr' database has entries from GenPept, Swissprot, PIR, PRF, PDB, and NCBI RefSeq (http://www.ncbi.nlm.nih.gov/blast/blast_databases.html). A similarity search based on the BLASTN program against 'dna_all' database is used to select the significant similarity to the nucleic acid sequences of rice and *Arabidopsis thaliana* cDNAs. The 'dna_all' is a nucleotide sequence database with entries from DDBJ, EMBL and GenBank (ftp://ftp.dna.affrc.go.jp/pub/dna_all/). The gene prediction programs assembled in the pipeline are GENSCAN[4] for *Arabidopsis thaliana*, GENSCAN for maize, RiceHMM (http://rgp.dna.affrc.go.jp/RiceHMM/) and FGENESH[21]. An exon prediction program, MZEF[26] is also incorporated.

Results

1. Algorithm for assigning function and classification of gene model

We have developed an algorithm for assigning function and classification of the gene models (Fig. 1) using the manual annotation system of the Rice Genome Research Program (http://rgp.dna.affrc.go.jp/genomicdata/AnnSystem.html) as a model. The GFSelector program integrates the similarity search results which are interpreted using keyword search techniques. It is designed to find a significant similarity to a known protein hidden in miscellaneous hits of similarity search.

The BLASTP report is examined based on two criteria. First, if a full definition line in 'Alignment' contains predicted information such as 'putative', 'unknown', 'hypothetical', and 'probable', the corresponding BLASTP hit is eliminated. In this process, misspelled descriptions such as 'putatative' are considered and the hit which contains such descriptions is also eliminated. Second, the full definition line in the BLASTP report is analyzed to determine the relationship to biological function. In the 'nr' database, we found some entries with definitions not related to biological functions such as 'awaiting functional assignment'. We constructed a database including such descriptions. If a full definition line in the BLASTP re-

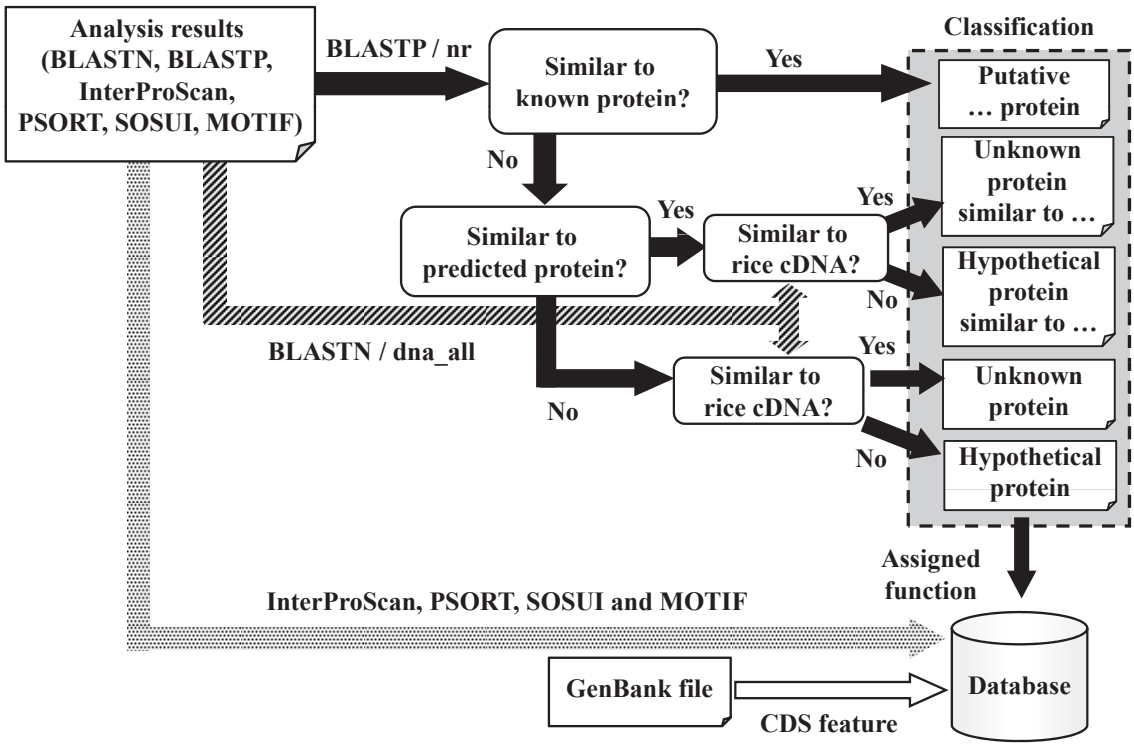

Fig. 1. Flowchart for assigning gene function and classification

port contains a description in the database, the corresponding hit is eliminated.

Even after eliminating BLASTP hits which contain predicted information and/or information not related to biological functions, a BLASTP report sometimes includes redundant hits which contain same words based on different but similar sequences. Thus, in order to provide a concise evidence for computational analysis, the system was also designed to keep the candidates for the assigned function after eliminating the redundancy.

The overall process for assigning function is as follows: (i) The BLASTP hits containing predicted information and/or not related to biological functions are deleted. (ii) The remaining hits are divided into two classes based on the database from which each sequence was obtained, namely, similarities against protein databases such as Swissprot and similarities against nucleotide databases such as GenBank. (iii) For each class, BLASTP hits are clustered based on a word contained in the hit. The clustering configures a maximum of three groups. If one of the first informative two words in a hit is the same as one of the first informative two words in another hit, these hits are clustered into the same group. A hit with the highest probability score in each group is selected to represent the group. In this step, the informative words are extracted from a full definition line in BLASTP report by eliminating a 'gi' identifier and a concatenated identifier for the database from which the corresponding sequence was obtained. (iv) The selected hits (maximum of three) become the 'Similar Protein Data from Protein Databases' and 'Similar Protein Data from Nucleotide Databases'. (v) The group with the highest probability score in 'Similar Protein Data from Protein Databases' is selected and the words contained in the corresponding hit are assigned as a predicted function. In the absence of 'Similar Protein Data from Protein Databases', the group with the highest probability score in 'Similar Protein Data from Nucleotide Databases' is selected. (vi) The BLASTN result against 'dna_all' database is then examined to select the significant similarity to the nucleic acid sequences of rice and *Arabidopsis thaliana* cDNAs. The result is searched for similarity if it contains the corresponding name of the species and other qualifiers such as 'cDNA' or 'mRNA'.

The gene model is named using three kinds of qualifiers, namely, 'putative', 'unknown', and 'hypothetical' based on the IRGSP annotation standard (http://rgp.dna.affrc.go.jp/genomicdata/AnnSystem.html). A 'putative protein' reflects the most probable gene function output from GFSelector. It indicates a significant similarity between the amino acid sequence of the gene model and the sequence of the known protein contained in NCBI 'nr'.

An 'unknown' qualifier is used for a gene model that has significant similarity to the nucleic acid sequences of rice cDNAs but no significant similarity to proteins in 'nr' or similar proteins with no adequate functional annotations. If GFSelector finds some functional information regardless of adequacy in the similar protein data, ancillary information is attached to the protein name such as 'unknown protein similar to X'. The gene model categorized into 'hypothetical protein' has no significant similarity to any database. Ancillary information is also attached to indicate similarity to a database entry if some functional information regardless of adequacy is detected.

The main view with the summary of the output for functional annotation of a gene model is shown in Fig. 2. The function assigned by GFSelector and the results of similarity search are listed. In addition, the GFSelector output also contains analyses for amino acid sequences. The analyses include Hmmer (http://www.biology.wustl.edu/gcg/hmmerpfam.html) against Pfam database[3], ProfileScan[23] against PROSITE database[9], MOTIF (http://motif.genome.jp/) against PROSITE database[9], PSORT[17], and SOSUI[8]. GO terms[22] based on functional domains identified by InterProScan[19] are also incorporated. The alignment against amino acid sequence motifs can be confirmed by a link to the corresponding graphical view of the gene model.

2. Analysis pipeline for user submitted genome sequences

The analysis pipeline has been significantly improved from the former-version of RiceGAAS[20] by adding the functional assignment and classification of gene models. Furthermore, the gene modelling algorithm was modified to improve the accuracy and the analysis pipeline was accelerated so that the whole process could be accomplished within 24 hours. The gene modelling in RiceGAAS was basically designed to select the most plausible gene model from the gene prediction programs,

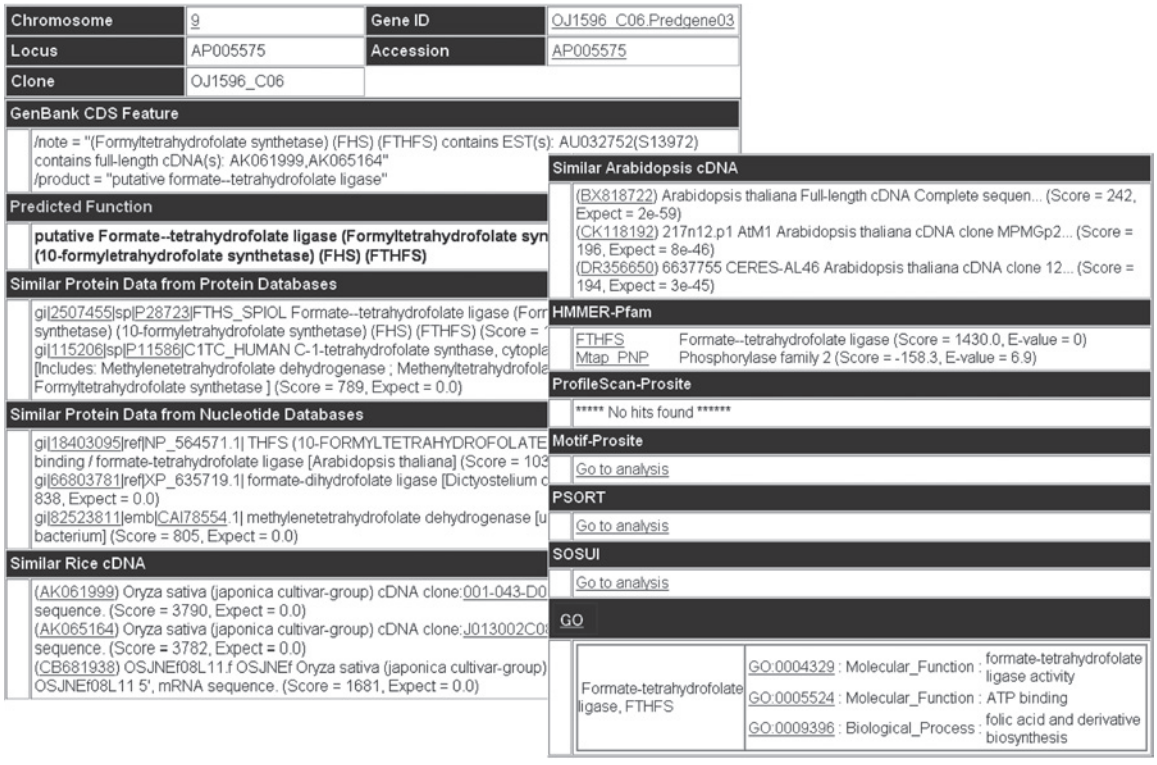

Fig. 2. An example of web view which shows the functional annotation of gene model 'OJ1596_C06.Predgene03'
The functional assignment described in the corresponding GenBank entry is indicated in the 'GenBank CDS Feature'. The functional assignment by GFSelector is indicated in the 'Predicted Function'. The alignment against amino acid sequence motifs can be viewed through a link on 'Gene ID'. Significant similarities are tabulated in 'Similar Protein Data from Protein Databases', 'Similar Protein Data from Nucleotide Databases', 'Similar Rice cDNA', and 'Similar Arabidopsis cDNA'. The tabulated similarity has a link to the corresponding entry of the reference database and KOME database developed by large-scale rice cDNA project[14]. Links to some analyses of amino acid sequence of the gene model are also provided. The term from gene ontology based on functional domains identified by InterProScan[19] is indicated at the bottom.

namely, GENSCAN for *Arabidopsis thaliana*, GENSCAN for maize and RiceHMM, and to insert internal exons from an exon prediction program, MZEF. Selection is based on a score for each gene model which represents the average exon score from each gene prediction program. The original exon score is automatically evaluated in the system. If a predicted exon overlaps with another exon predicted by a different gene prediction program or a genome region similar to rice EST (expressed sequence tag) or protein, the predicted exon is emphasized by adding a value to the original exon score.

Two major modifications were incorporated in the process mentioned above. First, the monocot matrix of FGENESH, which has been evaluated as the most accurate gene prediction program for rice[24] was added. However, the exon score scale in FGENESH is different from GENSCAN and RiceHMM and should be normalized. We investigated the correlation among the scores and induced the coefficients to normalize the scores from the gene prediction programs to enable to compare the results from different prediction programs. Second, the condition was revised to detect an exon that overlaps a genomic region similar to a protein. In the former-version pipeline, only overlaps were checked. In the modified pipeline, the coding frame of predicted exon is compared with the results of BLASTX to check if the amino acid sequence from the exon has significant similarity with the protein mapped on the genome.

Fig. 3 shows a flowchart and an example for gene modelling by the modified pipeline. In the example, GENSCAN for Arabidopsis predicted a gene model of eight exons. GENSCAN for maize predicted two gene models of two and three exons respectively. FGENESH and RiceHMM showed the same prediction result and each program predicted a gene model of five exons. A BLASTX hit overlapped with an internal exon predicted by the GENSCAN for Arabidopsis and terminal exons predicted by the GENSCAN for maize, FGENESH and RiceHMM. The coding frame of the internal exon and the terminal exons was the same as that of the protein mapped on the genome. The number with an underline for each predicted exon shows the overlap score (O-score) for the exon appearing in the second step in the flowchart (Fig. 3 (A)). For example, the start/end position and the type (initial, internal or terminal) of the initial exon by the GENSCAN for Arabidopsis were the same as the GENSCAN for maize, FGENESH and RiceHMM. Thus, the O-score for the initial exon by the GENSCAN for Arabidopsis was 1.0. On the other hand, the start/end position and the type (initial, internal or terminal) of the terminal exon by the RiceHMM were the same as the GENSCAN for maize and FGENESH, and the terminal exon had an overlap with a BLAST result (BLASTX) with a coding frame the same as the protein mapped on the genome. Thus, the O-score for the terminal exon by the RiceHMM was 1.6 (1.0 by an overlap with another prediction program and 0.6 by an overlap with a similarity search result). The number in parenthesis shows the average of O-score per exon by the predicted gene model. The average exon score for each predicted gene model appearing in the fourth step in the flowchart (Fig. 3 (A)) was calculated from the O-score and B-score of predicted exon as the formula 3) in Fig. 3 (A). In this case, the average exon score by the FGENESH and RiceHMM was greater than the GENSCAN for Arabidopsis and the GENSCAN for maize, and the gene model by the FGENESH and RiceHMM was selected and showed in the 'Autopredgeneset' row.

3. Stored data and update mechanism

As of February 2008, nearly 70,000 GenBank rice CDSs and 145,000 automatically predicted gene models by RiceGAAS are stored in the database (http://ricegaas.dna.affrc.go.jp/rgadb/). The difference between the number of gene models analyzed by the system and the estimated number of genes for the rice genome[12] was due to the inclusion of 567 Nipponbare clones overlapping the minimum tilling path of the 12 rice pseudomolecules (http://rgp.dna.affrc.go.jp/E/IRGSP/download.html), 456 non-Nipponbare clones such as *indica* BAC clones, and redundant genes in the overlapping regions of adjacent BAC/PAC clones. Furthermore, the automated system has been designed to generate all possible candidates for gene models including predicted transposons which could also be useful in understanding the genome structure. A guide for filtering the gene models is available through http://ricegaas.dna.affrc.go.jp/filtering_guide.html.

To update the rice gene models stored in the database, the processes described in Fig. 1 are daily executed. All rice genome sequences in GenBank are daily checked by a data collecting system. A keyword search is automatically executed against GenBank database to collect the sequences in HTG sequence division of *Oryza sativa*. Automated analyses of genes are executed for the updated HTG phase 2 and 3 rice sequences in GenBank. Therefore, the analyses by GFSelector become available in a day after release of the sequence. The re-analysis for the non-updated sequence is automatically executed every 180 days. Distributions of CpG island are also calculated for all collected rice BAC/PAC sequences in a 100 bp window at 10 bp intervals (http://ricegaas.dna.affrc.go.jp/CpG/).

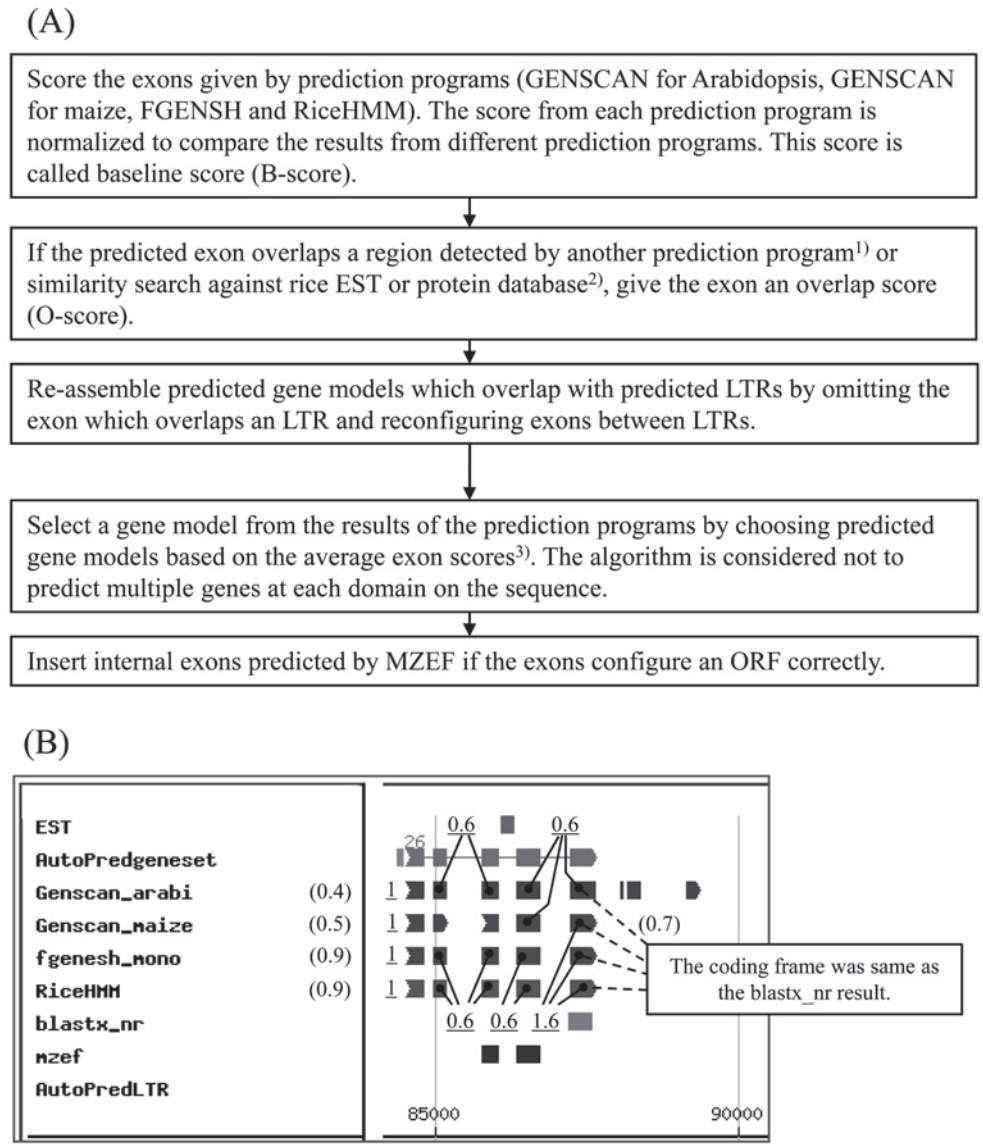

Fig. 3. Gene modelling by the analysis pipeline

(A): Flowchart. 1): Overlap with another prediction program: if the exon is internal exon, the O-score is 0.6; else if the exon is initial or terminal exon, the O-score is 1.0. 2): Overlap with a similarity search result: the O-score is 0.6. 3): {Sum of (B-score)+(O-score) for all exons in a gene model} / {No. of exons in a gene model}.
(B): Example.

Discussion

1. Evaluation of automated functional annotation and update mechanism

The GFSelector annotation was compared with the RGP manually curated annotation to evaluate the software implementation. The functions of 3,406 gene models on rice chromosome 1 as predicted by GFSelector were compared with the corresponding descriptions in GenBank entries which had been manually annotated by RGP. Among 761 RGP annotated putative proteins, 95% (723) were also categorized as putative proteins by GFSelector. On the other hand, among 765 putative proteins categorized by GFSelector, 95% (723) were also categorized as putative proteins by RGP. In terms of functional assignment, 92% of 765 putative proteins predicted by GFSelector were assigned with similar function as the RGP annotation. About 83% (399/482) of unknown proteins and 95% (2,048/2,159) of hypothetical proteins categorized by GFSelector were the same as RGP. The slightly low rate for unknown proteins (83%) may be caused by the accumulation of the rice cDNA sequences between the dates when manual annotation and computational annotation were executed. These results suggest

that GFSelector showed relatively reliable functional assignment of predicted rice genes.

The combined system of GFSelector and RiceGAAS automatically presents a regularly updated annotation for all rice gene models using the latest reference databases. Fig. 4 shows a representative case of the automated update (for rice PAC sequence, P0022F10). Two genomic regions with significant similarity to recently registered protein in the Swissprot database were detected in the annotation map on December 2007 and a different gene modelling was performed for the gene model 22 (see 'Autopredgeneset' row) based on the similarity search re-

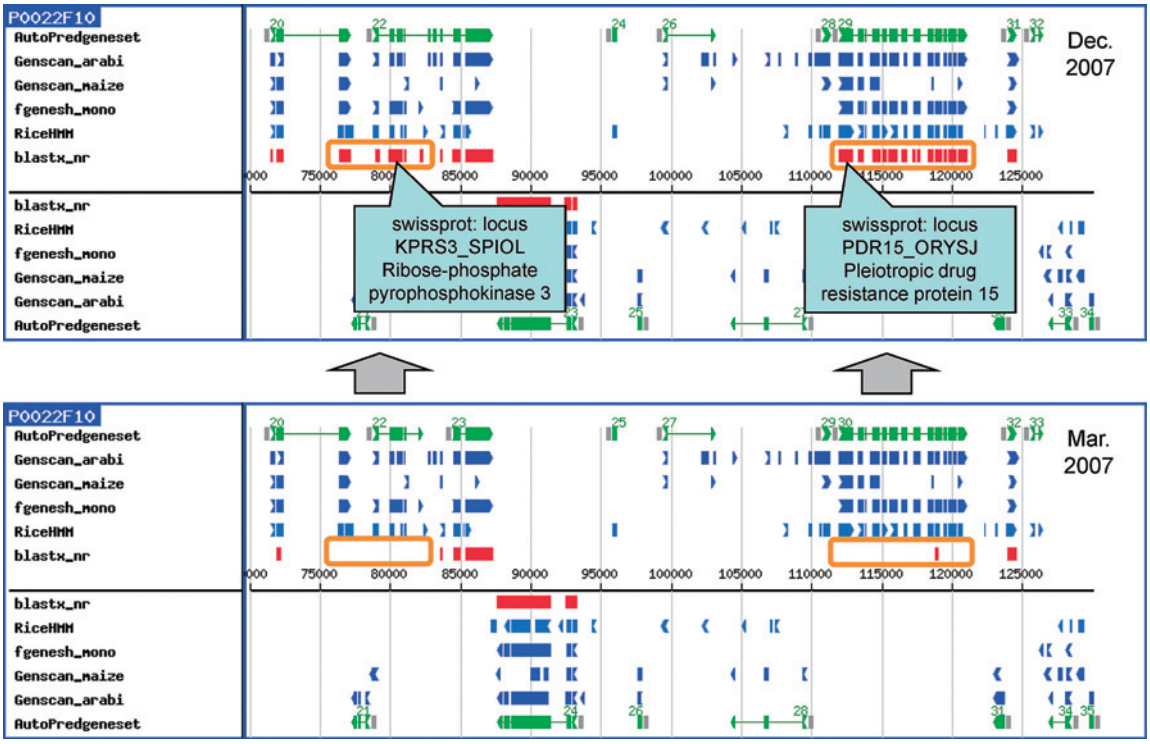

Fig. 4. Automated update of annotation based on the latest available information

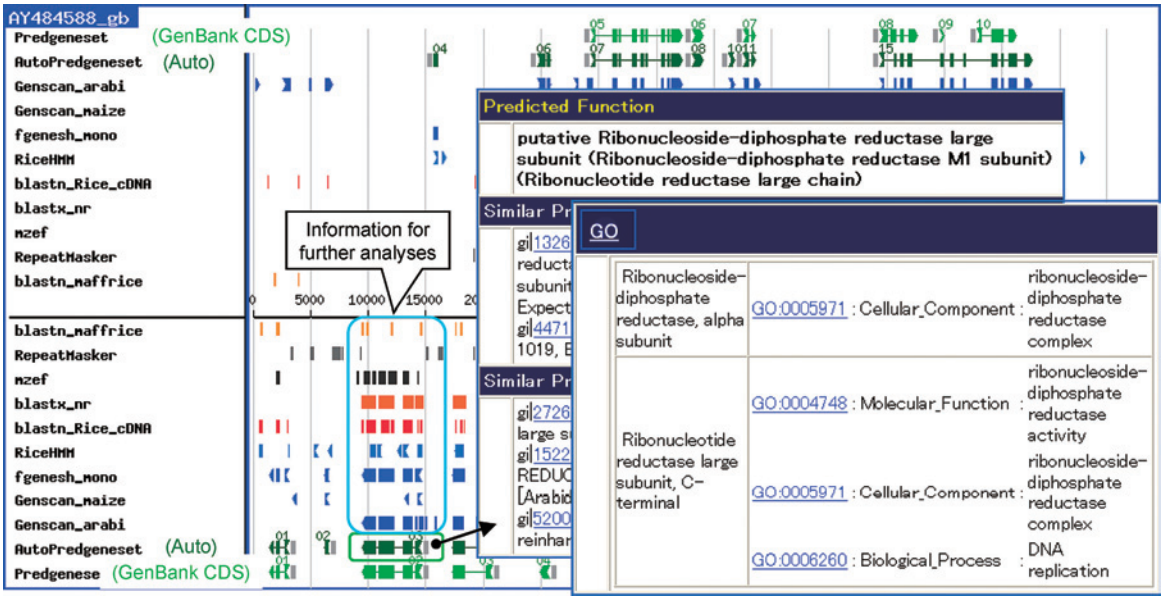

Fig. 5. Annotation map and functional assignment for banana BAC sequence (AY484588)

sults.

2. Utility for crop functional genomics

The combined system of GFSelector and RiceGAAS can be utilized in various ways for crop functional genomics. Using the analysis pipeline (http://ricegaas.dna.affrc.go.jp/usr/), user submitted genome sequences can be analyzed. We evaluated automated annotation by the pipeline based on the BAC sequences of wheat (total length: 808 kbp), banana (total length: 156 kbp) and soybean (total length: 593 kbp). These BAC sequences had been submitted to GenBank by the other research groups and the CDS features were assigned independently from our research.

First, the automated gene modelling was evaluated. We investigated the nucleotide-level sensitivity and specificity for the CDSs contained in the GenBank entries (the number of CDSs contained in the GenBank entries was 20 CDSs in 4 wheat BACs, 23 CDSs in 2 banana BACs and 6 CDSs in 4 soybean BACs). The sensitivity (Sn) was determined as follows:

$$Sn = \frac{\text{No. of nucleotide predicted positives (both the automated system and GenBank CDS features)}}{\text{No. of nucleotide predicted positives (GenBank CDS features)}}$$

The calculated sensitivity was 97% for wheat, 83% for banana and 93% for soybean. On the other hand, a sensitivity of 85% was calculated for the rice genome sequence (http://ricegaas.dna.affrc.go.jp/rga-bin/col_accur.pl). Although the automated system had been developed based on the rice genome sequence, the results suggest that the sensitivity is not correlated to the type of organism closely or distantly related to rice.

The specificity (Sp) was determined as follows:

$$Sp = \frac{\text{No. of nucleotide predicted positives (both the automated system and GenBank CDS features)}}{\text{No. of nucleotide predicted positives (the automated system)}}$$

The calculated specificity was 14% for wheat, 74% for banana and 11% for soybean. The reasons for the low specificity for the wheat and soybean BACs were: (i) The automatically predicted gene models included DNA transposons and retrotransposons (54 out of 214 gene models for wheat and 6 out of 96 gene models for soybean were transposons and retrotransposons), but the GenBank CDSs did not include DNA transposons and retrotransposons for the wheat and soybean BACs. (ii) The GenBank entries of wheat and soybean BACs contained few CDSs (the corresponding gene density was 40 kb/gene for wheat and 99 kb/gene for soybean). It was suggested from the related articles[1,6,7,27] that the GenBank entries of wheat and soybean BACs limited the CDSs to research targets of the authors. (iii) The automated system has been designed to generate all possible candidates for gene models.

For the rice genome sequence, a specificity of 59% was calculated (http://ricegaas.dna.affrc.go.jp/rga-bin/col_accur.pl). The results suggest that the specificity is also not correlated to the type of organism closely or distantly related to rice.

It was also suggested that the specificity was improved by integrating the GFSelector into the pipeline: (i) The GFSelector assigned transposons or retrotransposons to 54 automatically predicted gene models among 214 gene models in the four wheat BACs. The specificity for the wheat BACs will be improved from 14% to 28% by eliminating the gene models assigned transposons or retrotransposons by the GFSelector. (ii) The GFSelector presented the predicted gene models three kinds of qualifiers. The qualifier which means most probable is 'putative', and it indicates a significant similarity between the gene model and a known protein. The specificity for the four wheat BACs will be further improved from 28% to 35% by choosing the 85 gene models qualified as 'putative'.

Second, the functional annotation by the system was evaluated. We investigated 20 CDSs with completely the same nucleotide sequence between the GenBank CDS feature and automatically predicted gene model (Table 1). Seventeen of 20 CDSs (85%) were almost equivalently annotated based on the GenBank CDS feature and automated annotation, or equivalent candidates were presented by both systems. Two CDSs (10%) had an appropriate function assigned automatically based on the reference database updated later than the corresponding GenBank entry. Thus, 19 of 20 CDSs (95%) had an automatic annotation nearly equal or better than the corresponding GenBank CDS feature. The annotation map with functional assignment by the system for banana BAC sequence (AY484588) is shown in Fig. 5. Three GO terms were automatically assigned for a gene model. The results of similarity search and prediction programs will be useful for further analysis of the genome sequence such as splice variants etc.

The combined RiceGAAS and GFSelector has other potential applications for crop functional genomics. Using the similarity search through RiceBLAST[16] (http://riceblast.dna.affrc.go.jp/), a query sequence such as crop

Table 1. Comparison between the GenBank CDS feature and automated annotation by the system

CDS	GenBank feature	Automated annotation	Comment
Almost equivalent annotation was performed between the GenBank CDS feature and automated system (14 CDSs).			
EF426565.30	unknown protein	unknown protein similar to katanin p80 subunit-like protein	
EF426565.37	gamma gliadin 2	putative Gamma-gliadin precursor	
EF426565.39	gamma gliadin 3	putative Gamma-gliadin precursor	
DQ537335_2.19	X-type HMW glutenin	putative Glutenin, high molecular weight subunit DX5 precursor	
DQ537335_2.20	protein kinase	putative protein kinase	
DQ537336.10	Y-type HMW glutenin	putative Glutenin, high molecular weight subunit 12 precursor	
DQ537336.52	X-type HMW glutenin	putative Glutenin, high molecular weight subunit DX5 precursor	
DQ537337.05	receptor kinase 2	putative receptor kinase 2	
DQ537337.09	Y-type HMW glutenin	putative Glutenin, high molecular weight subunit DY10 precursor	
DQ537337.27	X-type HMW glutenin	putative Glutenin, high molecular weight subunit DX5 precursor	
DQ537337.28	protein kinase	putative protein kinase	
AY484588.16	retrotransposon-like protein	putative gag-proteinase polyprotein	
AY484588.09	hypothetical protein	hypothetical protein similar to Os04g0495900	
EF533702.02	LYK8	putative LYK8	
Another candidate provided by the automated system was equivalent to the GenBank CDS feature (3 CDSs).			
DQ537335_1.23	globulin 1	putative Glutenin, high molecular weight subunit 12 precursor	The candidate was: gi|110341790|gb|ABG68030.1| globulin 1 [*Triticum aestivum*].
DQ537336.08	globulin 1	putative Glutenin, high molecular weight subunit 12 precursor	The candidate was: gi|110341795|gb|ABG68034.1| globulin 1 [*Triticum aestivum*].
DQ537337.07	globulin 1	putative Glutenin, high molecular weight subunit 12 precursor	The candidate was: gi|110341801|gb|ABG68039.1| globulin 1 [*Triticum aestivum*].
Appropriate function was automatically assigned based on the reference database updated later than the corresponding GenBank entry (2 CDSs).			
AY484589.01	hypothetical protein	putative Zinc finger A20 and AN1 domain-containing stress-associated protein 12 (OsSAP12)	The assignment was based on: gi|75133829|sp|Q6Z541|SAP12_ORYSJ Zinc finger A20 and AN1 domain-containing stress-associated protein 12 (OsSAP12).
AY484589.05	crinkly4-like protein	putative Nodulation receptor kinase precursor (Does not make infections protein 2) (Symbiosis receptor-like kinase) (MtSYMRK)	The assignment was based on: gi|71152016|sp|Q8L4H4|NORK_MEDTR Nodulation receptor kinase precursor (Does not make infections protein 2) (Symbiosis receptor-like kinase) (MtSYMRK).
A function was automatically assigned based on a low BLASTP score (1 CDS).			
AY484588.11	hypothetical protein	putative Solution Structure Of Rrm Domain In Protein Bab28521	The assignment was based on: gi|157880646|pdb|1WEX|A Chain A, Solution Structure Of Rrm Domain In Protein Bab28521 (Score = 33.5, Expect = 7.6).

EST will be located on a rice BAC/PAC. The RiceBLAST result has a link to the annotation map of the corresponding rice BAC/PAC, and each gene model on the map is linked to the functional annotation by GFSelector. Thus a user can observe similar and/or adjacent rice gene models for the query sequence with structural and functional information of the rice gene models. Furthermore, using the keyword search mechanism (http://ricegaas.dna.affrc.go.jp/rgadb/index.html#KeyGFS), rice gene models containing a keyword such as user's target function can be easily retrieved.

Conclusion

GFSelector performs a computational classification of gene models and assignment of unique biological functions. The combined system of GFSelector and RiceGAAS provides a fully automated annotation for user submitted genome sequences using the latest reference databases and covers a whole process from gene modelling to functional annotation of the gene models. Comparison of the annotation with the GenBank CDS feature for other cereal crops as well as dicot plants suggests that a relatively reliable annotation can be obtained from automated annotation. The system also provides analyses based on regularly updated rice gene models. As a tool for analysis of genomic sequences, it can be used even by small groups with limited infrastructure and bioinformatics staff. Therefore the system could serve as an important bioinformatics resource for crop functional genomics.

The automated annotation provided by the system is mainly for protein coding sequences. The main process of the system is a prediction of protein coding gene and an assignment of a function for the predicted gene. As a future work, an automated annotation module for non-coding RNA will be developed.

Acknowledgments

The authors would like to thank Dr. Kenichi Higo for suggestions and discussions, Dr. Takuji Sasaki for giving us the opportunity to develop the system, and Dr. Shoshi Kikuchi for providing the link to KOME database (http://cdna01.dna.affrc.go.jp/cDNA/). This research was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Project GS-1302, SY-1101, SY-1103).

References

1. Aert, R., Sagi, L. & Volckaert, G. (2004) Gene content and density in banana (*Musa acuminata*) as revealed by genomic sequencing of BAC clones. *Theor. Appl. Genet.*, **109**, 129–139.
2. Altschul, S. F. et al. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
3. Bateman, A. et al. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
4. Burge, C. & Karlin, S. (1997) Prediction of complete gene structure in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
5. Dowell, R. D. et al. (2001) The distributed annotation system. *BMC Bioinform.*, **2**, 7.
6. Gao, S. et al. (2007) Rapid evolution and complex structural organization in genomic regions harboring multiple prolamin genes in the polyploid wheat genome. *Plant Mol. Biol.*, **65**, 189–203.
7. Gu, Y. Q. et al. (2006) Types and rates of sequence evolution at the high–molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics*, **174**, 1493–1504.
8. Hirokawa, T. et al. (1998) SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
9. Hulo, N. et al. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
10. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
11. Ito, Y. et al. (2005) Rice Annotation Database (RAD): A contig-oriented database for map-based rice genomics. *Nucleic Acids Res.*, **33**, D651–D655.
12. Itoh, T. et al. (2007) Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. *Genome Res.*, **17**, 175–183.
13. Jaiswal, P. et al. (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.*, **34**, D717–D723.
14. Kikuchi, S. et al. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science*, **301**, 376–379.
15. Moore, G. et al. (1995) Grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.
16. Nagamura, Y. et al. (2003) RiceBLAST: A comprehensive homology search for rice specific sequences. *Genome Inform.*, **14**, 533–534.
17. Nakai, K. & Horton, P. (1999) PSORT: A program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–35.
18. Ohyanagi, H. et al. (2006) The Rice Annotation Project database (RAP-DB): Hub for *Oryza sativa* ssp. *japonica* genome information. *Nucleic Acids Res.*, **34**, D741–D744.
19. Quevillon, E. et al. (2005) InterProScan: Protein domains identifier. *Nucleic Acids Res.*, **33**, W116–120.
20. Sakata, K. et al. (2002) RiceGAAS: An automated annotation system and database for rice genome sequence. *Nucleic Acids Res.*, **30**, 98–102.
21. Salamov, A. A. & Solovyev, V. V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.*, **10**, 516–522.
22. The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

23. Thompson, J. D. et al. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Bioinformatics*, **10**, 19–29.
24. Yu, J. et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
25. Yuan, Q. et al. (2005) The Institute for Genomic Research Osa1 rice genome annotation database. *Plant Physiol*., **138**, 18–26.
26. Zhang, M. Q. (1998) Identification of protein-coding regions in *Arabidopsis thaliana* genome based on quadratic discriminant analysis. *Plant Mol. Biol*., **37**, 803–806.
27. Zhang, X. C. et al. (2007) Molecular evolution of lysin motif-type receptor-like kinases in plants. *Plant Physiol*., **144**, 623–636.
28. Zhao, W. et al. (2004) BGI-RIS: An integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res*., **32**, D377–D382.