Nucleic Acids Research, 2005, Vol. 33, Database issue ***D651–D655***
doi:10.1093/nar/gki083

Rice Annotation Database (RAD): a contig-oriented database for map-based rice genomics

Yuichi Ito*, Kohji Arikawa[1], Baltazar A. Antonio[2], Isamu Ohta, Shinji Naito, Yoshiyuki Mukai, Atsuko Shimano, Masatoshi Masukawa, Michie Shibata, Mayu Yamamoto, Yukiyo Ito, Junri Yokoyama[3], Yasumichi Sakai[3], Katsumi Sakata[3], Yoshiaki Nagamura[2], Nobukazu Namiki, Takashi Matsumoto[2], Kenichi Higo[2] and Takuji Sasaki[2]

Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, 446-1 Ippaizuka, Kamiyokoba, Tsukuba, Ibaraki 305-0854, Japan, [1]Hitachi Science Systems Co. Ltd, 832-1 Horiguchi, Hitachinaka, Ibaraki 312-0034, Japan, [2]National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan and [3]Mitsubishi Space Software Co. Ltd, 1-6-1 Takezono, Tsukuba, Ibaraki 305-0032, Japan

Received June 24, 2004; Revised and Accepted October 12, 2004

ABSTRACT

A contig-oriented database for annotation of the rice genome has been constructed to facilitate map-based rice genomics. The Rice Annotation Database has the following functional features: (i) extensive effort of manual annotations of P1-derived artificial chromosome/bacterial artificial chromosome clones can be merged at chromosome and contig-level; (ii) concise visualization of the annotation information such as the predicted genes, results of various prediction programs (RiceHMM, Genscan, Genscan+, Fgenesh, GeneMark, etc.), homology to expressed sequence tag, full-length cDNA and protein; (iii) user-friendly clone / gene query system; (iv) download functions for nucleotide, amino acid and coding sequences; (v) analysis of various features of the genome (GC-content, average value, etc.); and (vi) genome-wide homology search (BLAST) of contig- and chromosome-level genome sequence to allow comparative analysis with the genome sequence of other organisms. As of October 2004, the database contains a total of 215 Mb sequence with relevant annotation results including 30 000 manually curated genes. The database can provide the latest information on manual annotation as well as a comprehensive structural analysis of various features of the rice genome. The database can be accessed at http://rad.dna.affrc.go.jp/.

INTRODUCTION

Rice is an important model plant because it is one of the fundamental foods for mankind and its genome has syntenic relationships with other major cereal crops (1). The Rice Genome Research Program (RGP) has been pursuing the sequencing of the rice genome since 1998 in collaboration with the International Rice Genome Sequencing Project (IRGSP). A clone-by-clone shotgun sequencing strategy has been adopted so that each sequenced clone could be associated with a specific position on the genetic map. The IRGSP also adheres to the immediate release of the genome sequence data to the public domain (2). Thus a high-throughput sequence production during the last six years has resulted in the accumulation of a huge amount of sequence data and annotation information (URL: http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl).

As we approach the post genome-sequencing era, it is indispensable that the accumulated annotation information should be integrated with other genomic data to facilitate map-based informatics. Towards this goal, the Japanese Ministry of Agriculture, Forestry and Fisheries (MAFF) initiated the Rice Genome Simulator Project (http://www.nias.affrc.go.jp/project/inegenome_e/simulator/simulator_outlook_e.htm) to coordinate the data produced from extensive genome analysis and other aspects of rice research with emerging technologies in computational biology. The main target of this research project is to establish simulation as a tool that would allow plant researchers and breeders to develop highly desirable varieties of rice. This will utilize a network that will link all fundamental rice databases and an informatics infrastructure

*To whom correspondence should be addressed. Tel: +81 298 38 2257; Fax: +81 298 38 2302; Email: yuito@staff.or.jp

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use permissions, please contact journals.permissions@oupjournals.org.

© 2005, the authors

including specialized softwares that will allow integration, visualization and simulation of genome information. We have developed a contig-oriented annotation database called the Rice Annotation Database (RAD) to integrate all relevant information on the structure and function of the rice genome with other databases on rice research. Similar rice genomics databases such as the Whole Rice Genome Automated Annotation Database of TIGR (3) and Rice Information System of Beijing Genomics Institute (4) which focus on an integrated facility for data-mining and comparative genomics are already available online. However, RAD provides a comprehensive database of manually curated annotation information as well as the results of analysis of various features of the annotated genes. In this paper, we present the concept, architecture and usability of RAD.

SYSTEM ARCHITECTURE

The primary concept of RAD is to provide a contig-oriented administration of the sequenced P1-derived artificial chromosome (PAC) and bacterial artificial chromosome (BAC) clones generated by the international sequencing collaboration. RAD is a relational database, which facilitates storage, query and visualization of annotation information such as sequence data, predicted genes and homology analysis of each PAC/BAC clone. These information are managed in contig or chromosome level to provide timely and general view of the data.

The overall architecture of RAD is shown in Figure 1. All the manual annotation information are sheared into genomic entities and converted into a contig every time a new dataset is inserted. Then, this information is managed as a contig in the relational database. The genes of the overlapping-region are consistently aligned with genes from the 'north' clone preceding genes from 'south' clone. This merge-system facilitates not only the efficient management of the annotation data but also provides updates on the gene structure as more annotation data are incorporated into the database. We used an ORACLE® database software for database administration server and the CGI-Perl package programs for database interface and application programs.

DATABASE CONTENTS

The genome sequencing of chromosomes 1 (5), 4 (6) and 10 (7) have been completed and the predicted genes have been manually curated. In addition to chromosome 1, the genome sequence of chromosomes 2, 6, 7, 8 and 9 assigned to RGP have almost been completed as well. As an initial analysis step, we utilize an automated annotation system called RiceGAAS (8), which combines the results of 14 kinds of gene prediction and analysis programs including the homology searches against rice full-length cDNA (9), rice expressed sequence tag (EST) and the NCBI non-redundant protein database. Subsequent manual curation (http://rgp.dna.affrc.go.jp/genomicdata/AnnSystem.html) facilitates accurate gene modeling of predicted genes based on available evidences.

The data for other chromosomes from the participating IRGSP members were obtained from the public database. We have integrated the flat files for the completed chromosomes into RAD thereby providing a central annotation database for IRGSP sequences. These operations successfully enhanced the structural features of the database. As of October 2004, the database contains a total of 215 Mb sequence with relevant annotation results including 30 000 manually curated genes.

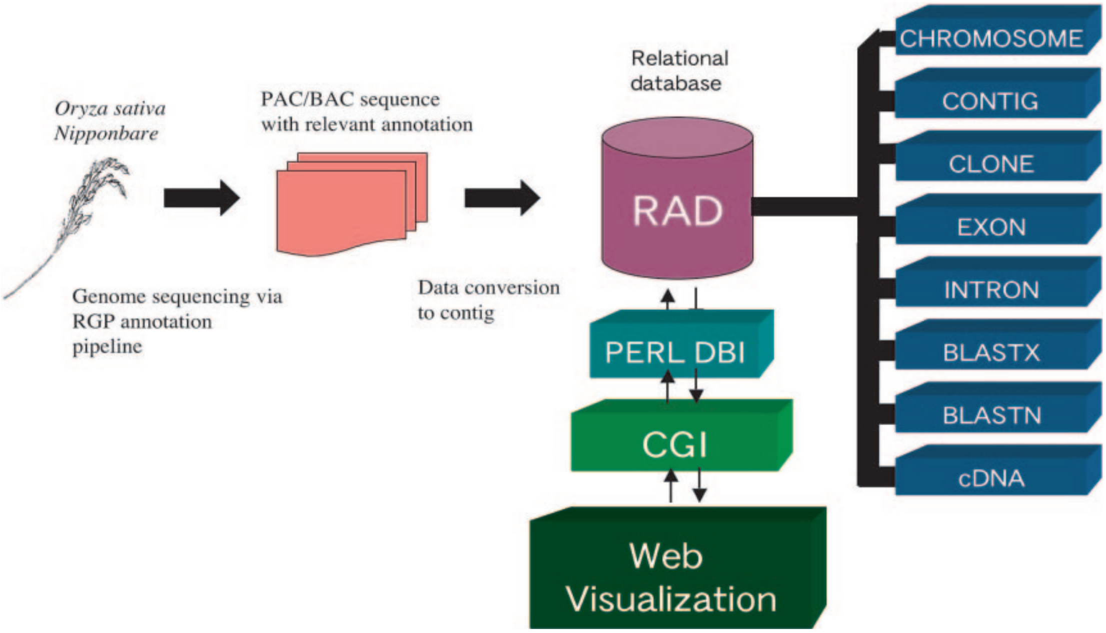


Figure 1. System architecture of the RAD. All the manual annotation information are sheared into genomic entities and converted into a contig every time a new dataset is inserted. Subsequently, these information are managed as a contig in the relational database to facilitate concise view of the merged annotations, user-friendly query and statistical analysis.

DATA VIEW

The annotation information with sequence data can be viewed in a user-friendly numerical and graphical web presentation form. This facilitates map-based informatics particularly in associating a genome sequence with a specific position in a genetic map. On the top page of RAD, users can initially view the status of annotation for each chromosome. By clicking a contig on the chromosome, users can access the merged contig-level annotation page as shown in Figure 2. The merged contigs are visualized at the center frame and the viewing region is highlighted as yellow on chromosome image at the upper frame. The contig-level annotation page also contains information of the predicted genes, pseudogenes, ESTs, LTRs, and the results of various prediction programs such as RiceHMM, Genescan, Genscan+, Fgenesh and GeneMark. The position of genetic marker is indicated on the map that can be adjusted to various window sizes (2M, 1M, 500K, 200K, 100K and 50K). Pointing on a gene shows the gene attributes such as gene ID, strand, position, identification, etc. on a table in the lower frame. These genomic, amino acid and DNA sequences displayed on the screen are downloadable.

By clicking a clone on a particular contig, users can view the clone-level annotation page. The page contains a graphical display (middle frame) and a tabular form (lower frame) of the annotation data as well as the PAC/BAC sequence and its quality information (upper frame). The annotation data contain the predicted genes, protein homology, EST and the rice full-length cDNA hits with links to respective public databases. The quality file describes the sequence data with a Phrap score of <30 and other specific features such as repeats and LTR. The PAC/BAC sequence data are also downloadable.

By subsequently clicking on a gene on the tabular form links the gene annotation, where users can view further information of gene attributes such as the position of exons, number of exons, gene identification, EST hit, rice full-length cDNA hit, mRNA, GC, gene length, etc. The results of homology searches such as BLASTP and BLASTX, as well as full-length cDNA hits are tabulated with corresponding links to the entries in public databases.

QUERY

A text-based keyword search function allows overall survey of specific genes and clones throughout the genome on the top page. Gene search can be carried out using gene accession, gene name, EST hit as well as full-length cDNA hit. Clone

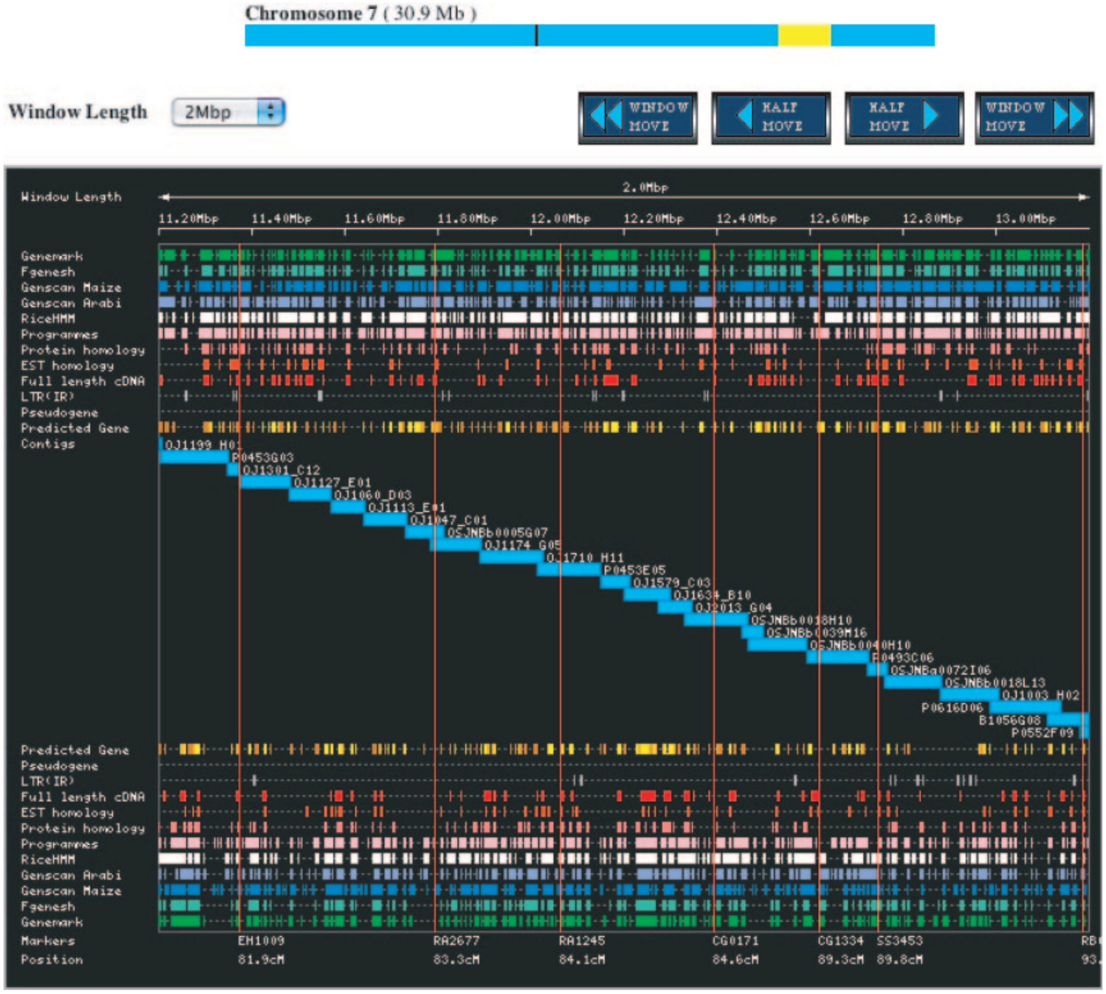
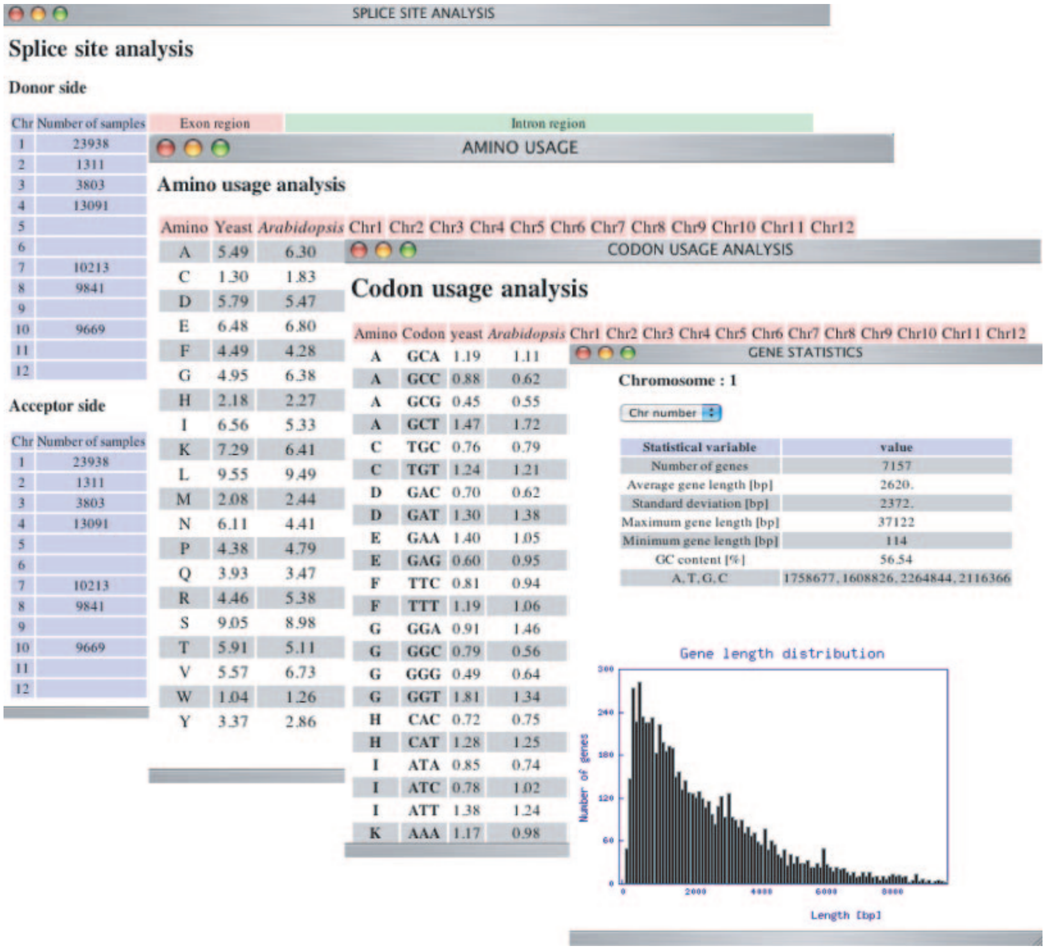


Figure 2. Web presentations of genome information at contig-level. A graphical view of the gene prediction programs, homology searches, genetic markers in alignment with the merged sequence can be viewed.



Splice site analysis

Donor side

Chr	Number of samples	Exon region	Intron region
1	23938		
2	1311		
3	3803		
4	13091		
5			
6			
7	10213		
8	9841		
9			
10	9669		
11			
12			

Acceptor side

Chr	Number of samples
1	23938
2	1311
3	3803
4	13091
5	
6	
7	10213
8	9841
9	
10	9669
11	
12	

Amino usage analysis

Amino	Yeast	*Arabidopsis*	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Chr12
A	5.49	6.30												
C	1.30	1.83												
D	5.79	5.47												
E	6.48	6.80												
F	4.49	4.28												
G	4.95	6.38												
H	2.18	2.27												
I	6.56	5.33												
K	7.29	6.41												
L	9.55	9.49												
M	2.08	2.44												
N	6.11	4.41												
P	4.38	4.79												
Q	3.93	3.47												
R	4.46	5.38												
S	9.05	8.98												
T	5.91	5.11												
V	5.57	6.73												
W	1.04	1.26												
Y	3.37	2.86												

Codon usage analysis

Amino	Codon	yeast	*Arabidopsis*	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Chr12
A	GCA	1.19	1.11												
A	GCC	0.88	0.62												
A	GCG	0.45	0.55												
A	GCT	1.47	1.72												
C	TGC	0.76	0.79												
C	TGT	1.24	1.21												
D	GAC	0.70	0.62												
D	GAT	1.30	1.38												
E	GAA	1.40	1.05												
E	GAG	0.60	0.95												
F	TTC	0.81	0.94												
F	TTT	1.19	1.06												
G	GGA	0.91	1.46												
G	GGC	0.79	0.56												
G	GGG	0.49	0.64												
G	GGT	1.81	1.34												
H	CAC	0.72	0.75												
H	CAT	1.28	1.25												
I	ATA	0.85	0.74												
I	ATC	0.78	1.02												
I	ATT	1.38	1.24												
K	AAA	1.17	0.98												

Chromosome : 1

Statistical variable	value
Number of genes	7157
Average gene length [bp]	2620.
Standard deviation [bp]	2372.
Maximum gene length [bp]	37122
Minimum gene length [bp]	114
GC content [%]	56.54
A, T, G, C	1758677, 1608826, 2264844, 2116366

Figure 3. Examples of analysis page. Users can view the general characteristics at the statistical analysis page.

search can be performed based on clone accession and clone name. Moreover, advanced search functions allow gene search based on homology level classifications such as a same protein, putative protein, similar protein, unknown or hypothetical protein, as well as clone search based on marker positions and other features.

ANALYSIS

The statistical analyses of various features of the sequence such as the GC-content, the length of exon and intron, splice site sequence, base/codon usage to the predicted genes, etc. are provided in RAD. These analyses facilitate a general characterization of the predicted gene models and the entire genome. Moreover, the Gene Ontology (GO) annotations (10) obtained using Interpro (11) can be viewed as a pie chart. The predicted genes are classified based on the GO. Similarly, the classification of the gene function based on a criterion of MIPS (http://mips.gsf.de/about/) can be viewed. Genome-wide similarity search (BLAST) is also implemented. These analyses give an overview of the various features of the genome sequence and allow correlation of the sequence information with gene expression, GO pathways, as well as genome-wide comparison with other organisms.

Table 1. List of various analyses incorporated in RAD[a]

Statistical analysis
Number of genes
Average gene length
GC-content
Splice site analysis
Amino usage analysis
Codon usage analysis
Functional classification based on GO
Functional classification based on MIPS
Genome-wide homology search
Query (Gene/Clone)
Data export function

[a] The implemented analysis tools are modified and updated on a regular basis to provide a comprehensive analysis of the rice genome.

The analysis contents are shown in Figure 3 and summarized in Table 1.

FUTURE DIRECTIONS

RAD has been developed by keeping pace with the progress of the rice genome sequencing project. The IRGSP will complete the rice genome sequencing by the end of year 2004. The manually curated annotation dataset of all chromosomes will be incorporated as well to provide an overview of all the genes that comprise the rice genome. Also, RAD will be improved to facilitate data mining of elucidated genes and provide a robust tool for understanding gene networks in the post-genome era.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of all members of the Rice Genome Research Program especially Dr Kyozo Eguchi for advice and encouragement. We also appreciate the technical support of Mitsubishi Space Software Co., Ltd, and the NIAS DNA Bank for providing the server for RAD. This work was funded by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Simulator Project SY-1102).

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. Yuan,Q., Ouyang,S., Liu,J., Suh,B., Cheung,F., Sultana,R., Lee,D., Quackenbush,J. and Buell,C.R. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.*, **31**, 229–233.
4. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y., Ren,X. *et al.* (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.*, **32**, D377–D382.
5. Sasaki,T., Matsumoto,T., Yamamoto,K., Sakata,K., Baba,T., Katayose,Y., Wu,J., Niimurra,Y., Cheng,Z., Nagamura,Y. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.
6. Feng,Q., Zhang,Y., Hao,P., Wang,S., Fu,G., Huang,Y., Li,Y., Zhu,J., Liu,Y., Hu,X. *et al.* (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.
7. The Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. *Science*, **300**, 1566–1569.
8. Sakata,K., Nagamura,Y., Numa,H., Antonio,B.A., Nagasaki,H., Idonuma,A., Watanabe,W., Shimizu,Y., Horiuchi,I., Matsumoto,T. *et al.* (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res.*, **30**, 98–102.
9. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al.* (2003) Collection, mapping, and annotation of over 28000 cDNA clones from *japonica* rice. *Science*, **301**, 376–379.
10. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, 258–261.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.