190–192 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg073

MOsDB: an integrated information resource for rice genomics

Wojciech M. Karlowski, Heiko Schoof, Vijayalakshmi Janakiraman, Volker Stuempflen and Klaus F. X. Mayer*

MIPS/Institute for Bioinformatics, GSF National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany

Received August 15, 2002; Revised and Accepted October 15, 2002

ABSTRACT

The MIPS Rice (*Oryza sativa*) database (MOsDB; http://mips.gsf.de/proj/rice) provides a comprehensive data collection dedicated to the genome information of rice. Rice (*O. sativa* L.) is one of the most important food crops for over half the world's population and serves as a major model system in cereal genome research. MOsDB integrates data from two publicly available rice genomic sequences, *O. sativa* L. ssp. *indica* and *O. sativa* L. ssp. *japonica.* Besides regularly updated rice genome sequence information, MOsDB provides an integrated resource for associated analysis data, e.g. internal and external annotation information as well as a complex characterization of all annotated rice genes. The MOsDB web interface supports various search options and allows browsing the database content. MOsDB is continuously expanding to include an increasing range of data type and the growing amount of information on the rice genome.

INTRODUCTION

Rice is one of the world's most important cereal crops and the second plant to have its genome sequenced. The first, *Arabidopsis thaliana*, is widely used as a reference model for different plant species, especially for dicotyledonous plants (1). Rice does not only serve as a model plant for monocotyledonous plants but also has an outstanding agronomic importance. Although the rice genome is 3.5 times larger than that of *Arabidopsis*, it has a small size (430 Mb) when compared to other grass species and exhibits a high degree of synteny and close evolutionary distance to other crop plants, such as wheat, maize and barley (2). Draft genome sequences have been reported in 2002 for two subspecies of rice: *japonica* (3) and *indica* (4). Both drafts were generated by whole-genome shotgun sequencing and are reported to cover more than 90% of the genomic DNA and 99% of all rice genes. In a complementary approach, The International Rice Genome Sequence Project follows a map based, clone oriented genome sequencing approach (5) and so far released more than half of the mapped genome of *Oryza sativa* L. ssp. *japonica.* Because of the outstanding importance of the rice genome, our aim is to develop a complete and comprehensive rice genome data resource. A special emphasis is directed towards the exhaustive analysis of the molecular inventory such as genes, transposable elements and in the future, regulatory elements. The MIPS *O. sativa* database (MOsDB; http://mips.gsf.de/proj/rice) provides researchers with a highly integrated information system of rice genome sequence and associated analysis data. It includes an up-to-date access to publicly available rice genomic sequences and provides various search options, data visualization, internal and external annotation information as well as a comprehensive PEDANT analysis of all annotated rice genes (6).

SYSTEM ARCHITECTURE

MOsDB was designed to easily accumulate the growing amount of data and integration of new data types. The database layout is implemented in MySQL and consists of three hierarchical levels: clones, contigs and genetic elements. Each of these layers contains tables to store information related to its type.

The clones module contains information about raw entries which were acquired from external sources. The contigs module contains final data, e.g. assembled sequences, which are in most cases already annotated. The genetic elements represent the main module in MOsDB and contain data describing annotations, clone coordinate information as well as all other accessory information, which have been acquired during the annotation and analysis process. Each element contains cross-reference information including methods to access corresponding external databases records as well as to scientific paper citations. It also contains fully searchable collection of alternate codes, assigned by different external sources. An additional feature of MOsDB are 'Virtual Contigs' (7), which allow assembling of raw contigs into bigger contiguous stretches of DNA and transferring annotation information to them. One of the major advantages of this approach is the possibility to annotate genes, which span two or more raw

*To whom correspondence should be addressed. Tel: +49 8931873584; Fax: +49 8931873585; Email: k.mayer@gsf.de

DNA clones. In MOsDB, chromosomes represent the highest level of virtual contigs. As soon as the sequencing of the rice genome will be completed and all chromosomes will be assembled, the annotation information (coordinates) will refer to this level.

Besides entry related components, like name, description, sequence etc., the MOsDB system provides researchers with additional information like confidence values, external comments and versioning of entries.

The MOsDB system allows assignment of confidence values to each annotated element in the database. This entry reflects the status and reliability of the particular annotation. The basic set of confidence codes was implemented using the Gene Ontology Evidence Codes (8). By default, the automatic retrieval system assigns the IEA value (Inferred from Electronic Annotation) to each annotation entry acquired from external sources. During the expert annotation and verification process, this entry is modified to achieve a higher confidence value. Our confidence system is evolving and new codes are added to enrich the original set of evidence levels.

Another feature of the MOsDB system is the ability to store and display externally supplied and expert comments. In general, the expert comments are displayed on the detailed element entry view and for further improvement and adjustment of annotation data the supplied information is being integrated into MOsDB in an expert supervised mode.

Due to the draft nature of rice genome, data are changing rapidly. Therefore, we have introduced a versioning system into MOsDB, which allows us to store several versions of each sequence entity and annotation data. The highest version number represents the newest record and is displayed on the web. In the near future, it will be possible to retrieve the data for different versions from the web interface.

DATA CONTENT AND SOURCING

MOsDB integrates data from two publicly available rice genomic sequences, *O. sativa* L. ssp. *indica* and *O. sativa* L. ssp. *japonica.*

The genome sequence of *O. sativa* ssp. *indica* was obtained as a result of a whole genome shotgun project and the DNA sequence is provided in a form of contigs and scaffolds assemblies. Both sets of sequence data have been integrated into MOsDB. As soon as the genome annotation process will be completed (scheduled for 2002/2003, see later), this data will be also available in MOsDB.

The genome sequence of *O. sativa* ssp. *japonica* is a result of a traditional chromosome by chromosome and clone by clone sequencing strategy. The International Rice Genome Sequencing Project (IRGSP) integrates 13 collaborating centers, which are contributing DNA sequences as well as annotation data to the public databases. We source the information about available contig updates directly from sequencing laboratories, retrieve the data from the GenBank database and integrate it into MOsDB data structure. The updates of MOsDB are run on weekly basis and include new or updated sequence data for completed as well as for unfinished (not yet assembled) contigs.

At the time of submission of this paper, the MOsDB provides annotation data only for *O. sativa* ssp. *japonica.* This information is expanded and verified by bioinformatic approaches and additional available data resources. The putative protein encoding sequences are compared against the MIPS Sputnik EST cluster collection (9), assembled specific for rice as well as other grass transcriptomes (e.g. maize, barley and sorghum). The high scoring sequences are analyzed for the intron–exon junction and information about the UTRs is extracted. Moreover, each element stored in MOsDB is cross-linked to the PEDANT system (10), which provides a complex characterization of gene products using a suite of bioinformatics tools. Currently, it enables to retrieve information about protein sequence statistics, protein localization prediction and several protein structural analyses, e.g. detection of low complexity regions, non-globular regions, coiled coil regions and transmembrane region prediction. It also provides results of HMM search against the Pfam database and a set of results of PSI-BLAST similarity searches against various databases, e.g. non-redundant protein database, database of protein sequences with known 3D structure, databases of proteins with manually assigned functional categories, SCOP database of protein domains, the COG database, among others.

ACCESS AND WEB QUERY INTERFACE

The MOsDB interface and query system is part of a new integration framework for genomic data: the Genome Annotation Management System (GAMS; http://mips.gsf.de/proj/gams). GAMS is composed out of a multi-tier architecture and in case of MOsDB it contains three layers. The first component is the database layer, as described above. In case of MOsDB, it provides access to the MySQL database. The second one is a business logic layer where the information from the underlying databases are unified via XML and where the actual logic for the web-client requests reside. These logical components consist out of Enterprise Java Beans (EJBs). Finally, a presentation layer, which uses an advanced Model View Controller (MVC) concept, provides a web interface. User interfaces were designed as a combination of HTML code and Java Server Pages (JSPs).

The MOsDB web interface is divided into two views: browse and search. To browse the user can navigate in a genome-oriented way, e.g. starting from a list of individual chromosomes (12 + 1). As a next step, information about contig constituents anchored to the respective chromosome can be retrieved. By entering a detailed contig view, a list of annotated genetic elements can be accessed. The element names represent links, which can be used to display information about the DNA sequences (spliced and unspliced) and protein sequence (in case of protein encoding genes). Moreover, on this view users have the option to display the results of PEDANT analysis corresponding to the particular entry. For all genetic elements, a list of alternative names with links to the appropriate databases is provided. At the time of writing, the annotation data is available only for *O. sativa* ssp. *japonica.*

The second method of accessing the MOsDB database is the search mode. Currently, MOsDB supports two ways of

searching of the rice genomic data: by sequence homology or by keyword search. MOsDB provides BLAST and FASTA as a homology search engine allowing input of custom sequences. The target databases for DNA searches include clones (completed and unfinished contigs), contigs (completed and mostly annotated contigs) and annotated genetic elements sequences (spliced coding sequences). In addition, MOsDB provides a search against a database of annotated proteins.

The keyword option allows searching of MOsDB using names of contigs or genes. This module allows searching using current codes as well as alternate codes. The free text search option allows inspection of the content of all text fields. Two of the MOsDB modules (clones and elements entries) are available for these options.

FUTURE DEVELOPMENTS

Rice is being used as the major model and reference system for monocotyledonous plants. Therefore, MOsDB is actively addressing the need for integrated, publicly available resources for rice genomics by providing a database system of rice DNA sequences and the exhaustive analysis of the individual genetic elements, an approach which is complementary to work carried out at various other sites (11,12). Our database is continuously expanding to include an increasing range of data type as well as to build comparative genome sequence views. As in the MIPS *A. thaliana* database (MAtDB) (13), we strongly encourage external input from our database users and shortly we will provide the opportunity to submit comments and corrections via our web interface. Finally, improvements of the web query interface, which will include a graphical chromosome browser and gene annotation information viewer, among others, are already underway.

Our work on establishing a pipeline to annotate the non-completed parts of the rice ssp. *japonica* and *indica* genomes is currently in progress. The annotation process includes mapping of rice Sputnik EST clusters data onto genomic DNA and *ab initio* prediction of the genes using FGeneSH software package (14). In addition, we are implementing alternative annotation methods for gene structure prediction, like a cross-species sequence comparison (15,16). The last step in our annotation pipeline is the curation of acquired data by gene family experts. To provide a high quality curated data we plan to implement a system, which will allow an automatic integration of external expert and community annotation.

DATA DOWNLOAD

The MOsDB FTP (ftp://ftpmips.gsf.de/rice) site provides a variety of data downloads, including whole DNA sequence collections for contigs and scaffolds (*O. sativa* ssp. *japonica* and *indica*) and protein sequences for *O. sativa* ssp. *japonica* genome.

ACKNOWLEDGEMENTS

Work on MOsDB database system is funded in the GABI project by the German Federal Ministry of Education and Research (BMBF) (0312270/4).

REFERENCES

1. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
2. Moore,G., Devos,K.M., Wang,Z. and Gale,M.D. (1995) Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.
3. Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma,H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.
4. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
5. Sasaki,T., Matsumoto,T., Baba,T., Yamamoto,D., Wu,J., Katayose,Y. and Sakata,K. (2001) The International Rice Genome Sequencing Project: progress and prospects. In Khush,G., Brar,D.S. and Hardy,B. (eds), *Rice Genetics IV (Proceedings of the Fourth International Rice Genetics Symposium 2000)*, 189–196.
6. Frishman,D., Mokrejs,M., Kosykh,K., Kasternmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K., Volz,A., Wagner,C., Fellenberg,M., Heumann,K. and Mewes,H.-W. (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 201–211.
7. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
8. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
9. Rudd,S., Mewes,H.-W. and Mayer,K.F.X. (2003) Sputnik: a database for comparative plant genomics. *Nucleic Acids Res.*, **31**, 128–132.
10. Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.
11. Ware,D., Jaiswal,P., Ni,J., Pan,X., Chang,K., Clark,K., Teytelman,L., Schmidt,S., Zhao,W., Cartinhour,S., McCouch,S. and Stein,L. (2002) Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.*, **30**, 103–105.
12. Yuan,Q., Quackenbush,J., Sultana,R., Pertea,M., Salzberg,S.L. and Buell,C.R. (2001) Rice bioinformatics: analysis of rice sequence data and leveraging the data to other plant species. *Plant Physiol.*, **125**, 1166–1174.
13. Schoof,H., Zaccaria,P., Gundlach,H., Lemcke,K., Rudd,S., Kolesov,G., Arnold,R., Mewes,H.W. and Mayer,K.F. (2002) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.*, **30**, 91–93.
14. Salamov,A.A. and Solovyev,V.V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.*, **10**, 516–522.
15. Morgenstern,B., Rinner,O., Abdeddaim,S., Haase,D., Mayer,K.F., Dress,A.W. and Mewes,H.W. (2002) Exon discovery by genomic sequence alignment. *Bioinformatics*, **18**, 777–787.
16. Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.