MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome

Heiko Schoof, Paolo Zaccaria, Heidrun Gundlach, Kai Lemcke, Stephen Rudd, Grigory Kolesov, Roland Arnold, H. W. Mewes and Klaus F. X. Mayer*

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

Received September 17, 2001; Accepted September 18, 2001

ABSTRACT

***Arabidopsis thaliana* is the first plant for which the complete genome has been sequenced and published. Annotation of complex eukaryotic genomes requires more than the assignment of genetic elements to the sequence. Besides completing the list of genes, we need to discover their cellular roles, their regulation and their interactions in order to understand the workings of the whole plant. The MIPS *Arabidopsis thaliana* Database (MAtDB; http://mips.gsf.de/proj/thal/db) started out as a repository for genome sequence data in the European Scientists Sequencing *Arabidopsis* (ESSA) project and the *Arabidopsis* Genome Initiative. Our aim is to transform MAtDB into an integrated biological knowledge resource by integrating diverse data, tools, query and visualization capabilities and by creating a comprehensive resource for *Arabidopsis* as a reference model for other species, including crop plants.**

INTRODUCTION

The analysis of the *Arabidopsis* genome was published in December 2000 as a web resource and CD-ROM (1). MIPS contributed to data management, annotation, functional classification, whole genome analysis and intergenome comparison (2–4). The data generated by the *Arabidopsis* Genome Initiative have been compiled into the framework of MIPS *Arabidopsis thaliana* Database (MAtDB; http://mips.gsf.de/proj/thal/db). Additionally, the mitochondrial and chloroplast genome sequences were integrated. Efforts continue to improve genome sequence and annotation data. In collaboration with the The Institute for Genomic Research (TIGR; Rockville, MD; http://www.tigr.org) and The *Arabidopsis* Information Resource (TAIR; Stanford, CA; http://www.arabidopsis.org) teams, unambiguous locus codes for protein-coding, RNA and pseudogenes are consistently assigned across the databases.

In order to improve and update the annotation of predicted genes, an automated system has been developed that compares new sequence submissions in the EMBL database with MAtDB. If new expressed sequence tag (EST) or full-length cDNA sequences support new genes or different exon–intron structures for predicted genes, this information is used to update MAtDB and to link genes to cognate EST or cDNA sequences. At the same time, full-length cDNAs are used to annotate 5′- and 3′-untranslated regions. To incorporate individual knowledge or experimental evidence, an online submission form was created. It can be used to submit information related to the improvement of gene predictions, e.g. by pasting full-length cDNA or protein sequence, or for functional and background information. Comments are displayed directly in the gene reports including author information; however, corrections of gene predictions are curated by MIPS to ensure a high quality standard of gene prediction.

The challenge is to keep data up-to-date, i.e. to integrate the genome sequence with diverse biological data, as well as to provide useful tools to explore and evaluate the information. As a consequence of large amounts of published experimental data, MAtDB is rapidly evolving.

QUERY INTERFACE

A versatile query interface provides intuitive access to MAtDB and represents data compiled from multiple heterogenous sources. Three different views are provided: browse, search and special interest. To browse, a user can choose between graphical or list views and move along the genome, zooming in to a region of interest. Search provides ID, keyword and sequence queries besides a tool for formulating complex queries across multiple databases. Access by ‘special interest’ structures the data in tables and catalogs based on function, structure or inter-species comparison.

Using the MAtDB GenomeViewer, the user can choose a list or graphical display of a chromosome. The list view brings up a table of all the clones used for the genomic sequence and shows their length, orientation along the chromosome, overlap to neighboring clones, sequencing/annotation status as well as genetic or physical markers mapped to that clone. The graphical display shows an overview of the chromosome. A selected region can be enlarged to show the tiling path where individual clones can be selected for inspection. Physical and genetic

*To whom correspondence should be addressed. Tel: +49 89 3187 3584; Fax: +49 89 3187 3585; Email: kmayer@gsf.de markers are displayed, with genetic distances where available; clones are color-coded based on sequencing/annotation status.

Selecting a clone returns a list of annotated features on this clone. This can be sorted based on coordinates, feature type, (functional) classification or detected protein motifs. A search box provides a free text search across all annotation on this clone. The Jaba AnnotationViewer tool displays annotated genes along with the output of several gene prediction tools. It permits the user to reconsider the judgments of the annotator who modeled a particular gene. Features can be selected and a separate window shows information like code, name, coordinates and sequence. Three-frame translation or highlighting of sequence features like start/stop codons or splice sites can be activated. In addition, an alternative, java-free graphical representation displays the annotated features color-coded by type.

Selecting a feature displays a detailed report compiled from various data sources like the core MAtDB annotation database or the PEDANT database of precomputed analyses (5) containing functional or structural attributes related to the protein sequence, such as PROSITE motifs, PFAM or SCOP classifications. Links are provided to retrieve sequences (protein, spliced/unspliced DNA), to analyses (like precalculated FASTA or BLAST searches, multiple alignments) or to external databases (e.g. EMBL, TIGR *A.thaliana* database, PlantsP, LGICdb).

SEARCHING MAtDB

The search section of MAtDB provides direct access by locus, clone or INTERPRO ID. To search by sequence, BLAST, FASTA and pattern search are available. FASTA and pattern search run against all protein sequences in MAtDB, whereas the user can select different protein and DNA databases for BLAST. In order to provide a fast, sophisticated, customizable query tool across all annotation, integration of the underlying data sources is performed by the BioRS indexing and data integration software. The user interface is built from applications relying on the BioRS retrieval system and its communication layer rather than using individual software modules, and is accessible via the 'advanced search' link. The output format can be customized to display selected fields, or downloaded in various formats, e.g. FASTA for sequences or XML representation.

TABLES AND CATALOGS

Several catalogs have proven to be especially useful for the classification of genes. During annotation, genes were classified as known proteins, hypothetical proteins or as having some degree of similarity to other proteins or EST sequences. For functional classification, the EC number catalog and the MIPS functional catalog (5) are both available. For some protein families, e.g. resistance genes and receptor-like kinases, curated tables are available. Functional and structural classification by automated methods to generate catalogs is performed by PEDANT, and lists of all genes sharing a common functional or structural domain can be retrieved. Intergenome comparison has its own section where, for example, the functional composition of the complete genomes of *Arabidopsis*, yeast, *Caenorhabditis elegans* and *Drosophila* can be compared based on INTERPRO domain assignments.

REDUNDANCY VIEWER

One notable result from the analysis of the complete *Arabidopsis* genome sequence was the identification of large duplicated segments with interesting biological features. This suggests a polyploidization event in the evolution of the *Arabidopsis* genome, and subsequent fusion and reorganization of chromosomes (1). A java-based tool to visualize and inspect these duplications was developed and integrated into the GenomeViewer section of the MAtDB web interface. The data of the original duplication analysis can be inspected in an interactive, graphical environment. The user can select any duplicated segment from an overview of the five *Arabidopsis* chromosomes and explore the fine structure and retrieve genes in the duplications.

DATA INTEGRATION

Sequence-related information of an organism cannot be adequately described by listing its genes. The essence of function is interaction in the context of the cellular localization of proteins. For this reason, genome databases must advance from a collection of genes and proteins to the integration of other types of data. For instance, the yeast database at MIPS (CYGD; 6) contains protein interaction and gene expression data, and the methods are reusable for the *Arabidopsis* genome. They include an interactive pathway prediction tool, Amphora, which can be used to cluster and analyze expression profiles.

Bioinformatics methods often fall short in the detailed assignment of functional properties and their organism specific description. Thus, manual annotation by experts for certain protein families should be available in a whole-genome database. In our collaboration with several partners, we have established methods that integrate selected data into the gene reports in MAtDB either locally or by linking to relevant information in external databases.

Mutant phenotype information can be valuable for discovering the function of genes. MAtDB includes a phenotype database that allows searching for mutants on the basis of phenotypic traits. The hierarchical structure of the phenotype catalog was designed to aid researchers observing a phenotype to query the database and identify characterized mutants with a similar phenotype. The root categories 'What', 'Where' and 'When' were chosen to make the description of a phenotype as intuitive as possible. The 'What' category ranges from overall plant appearance to specific cell types, the 'Where' category allows restriction to specific affected organs and the 'When' category is useful for annotating developmental stages. An extra branch, named 'Response', is used for conditional phenotypes or response phenotypes, like hormone insensitivity. Many categories include a free text field, e.g. for annotation of precise conditions for conditional phenotypes.

FUTURE DIRECTIONS

MAtDB will incorporate comprehensive information on the *Arabidopsis* genome while developing data representations, visualization and query tools for new data types and integrating data from our collaborators. MAtDB is part of the European PlaNet (http://mips.gsf.de/proj/planet) project that aims to interconnect several European plant databases, e.g. *Arabidopsis* insertion mutants will be linked to MAtDB. In an effort to facilitate data exchange and comparison, a mapping of the MIPS functional catalog to Gene Ontology (GO) categories (7) has been created in collaboration with M. Ashburner and, together with Biomax Informatics AG, the MIPS functional catalog is being expanded and streamlined for greater collinearity with GO.

On the other hand, the *Arabidopsis* genome will be used as a backbone for annotation and representation of data from other species. Aligning annotation from *Arabidopsis* to another plant species allows transfer of knowledge and enhances the value of data, as has been tested with tomato and rice sequences (8,9).

DATA DOWNLOAD AND STABLE LINKS

Complete sets of *Arabidopsis* sequences and annotation can be downloaded from ftp://ftpmips.gsf.de/cress. Please contact the MAtDB administrator if you require large-scale data that is not available at this location or retrievable via BioRS. If you wish to link to the gene reports from your own site, please only use the URL http://mips.gsf.de/cgi-bin/proj/thal/search_gene?code=At1g10000 and replace the locus code.

ACKNOWLEDGEMENTS

We wish to thank our collaborators who have contributed their data or agreed to link their database to MAtDB: B. Haas and O. White (TIGR), A. Kozik and R. Michelmore (UC Davis), S. Shiu and A. Bleecker (University of Wisconsin), F. Fana and M. Gribskov (San Diego Supercomputer Center), Benoit Lacombe (Institut Pasteur), B. Weisshaar (MPIZ Cologne) and all who submitted comments, e.g. to the online form. Dirk Haase (MIPS) and Martin Ruopp (Biomax Informatics AG) created the Redundancy Viewer, and Matthias Fellenberg (Biomax Informatics AG) created the Amphora expression profile analysis tool. BioRS is developed by Biomax Informatics AG. MAtDB is funded in the GABI project by the 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. The European Union Arabidopsis Genome Sequencing Consortium (1999) Sequence analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, **402**, 769–777.
3. European Union Chromosome 3 Arabidopsis Genome Sequencing Consortium, The Institute for Genomic Research and Kazusa DNA Research Institute (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 820–823.
4. Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University Sequencing Consortium, The European Union Arabidopsis Genome Sequencing Consortium and Institute of Plant Genetics and Crop Plant Research (IPK) (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 823–826.
5. 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.
6. Mewes,H.W., Frishman,D., Gruber,C., Geier,B., Haase,D., Kaps,A., Lemcke,K., Mannhaupt,G., Pfeiffer,F., Schueller,C., Stocker,S. and Weil,B. (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 37–40. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 31–34.
7. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature*, **25**, 25–29.
8. Ku,H.M., Vision,T., Liu,J. and Tanksley,S.D. (2000) Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl Acad. Sci. USA*, **97**, 9121–9126.
9. Mayer,K., Murphy,G., Tarchini,R., Wambutt,R., Volckaert,G., Pohl,T., Dusterhoft,A., Stiekema,W., Entian,K.D., Terryn,N. *et al.* (2001) Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. *Genome Res.*, **11**, 1167–1174.