Nucleic Acids Research, 2004, Vol. 32, Database issue ***D373–D376***
DOI: 10.1093/nar/gkh068

MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics

Heiko Schoof[1,2,*], Rebecca Ernst[1], Vladimir Nazarov[1], Lukas Pfeifer[1], Hans-Werner Mewes[1,2] and Klaus F. X. Mayer[1]

[1]Institute for Bioinformatics (MIPS), GSF National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany and [2]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Received September 15, 2003; Revised and Accepted September 29, 2003

ABSTRACT

Arabidopsis thaliana **is the most widely studied model plant. Functional genomics is intensively underway in many laboratories worldwide. Beyond the basic annotation of the primary sequence data, the annotated genetic elements of *Arabidopsis* must be linked to diverse biological data and higher order information such as metabolic or regulatory pathways. The MIPS *Arabidopsis thaliana* database MAtDB aims to provide a comprehensive resource for *Arabidopsis* as a genome model that serves as a primary reference for research in plants and is suitable for transfer of knowledge to other plants, especially crops. The genome sequence as a common backbone serves as a scaffold for the integration of data, while, in a complementary effort, these data are enhanced through the application of state-of-the-art bioinformatics tools. This information is visualized on a genome-wide and a gene-by-gene basis with access both for web users and applications. This report updates the information given in a previous report and provides an outlook on further developments. The MAtDB web interface can be accessed at http://mips.gsf.de/proj/thal/db.**

INTRODUCTION

Since the publication of its sequence (1), the *Arabidopsis* genome has established itself as a model data set not only for plant biologists. Its availability has sparked numerous functional genomics projects, and genome-scale data are being generated by high-throughput approaches. On one hand, these data need to be evaluated in the genome context, e.g. by mapping large-scale expression data onto metabolic pathways or functional categories. On the other hand, the experimental output will again be the basis for further analyses, e.g. detection of regulatory elements in the promoters of co-expressed genes. Genome databases are therefore faced by the necessity of integrating this kind of experimental data. How can they form a resource for analysis and the transfer of knowledge? Our approach is to apply state-of-the art bioinformatics methods on a whole-genome scale. The results are integrated into the database and correlated with experimental data, where necessary by manual curation. This allows them to again be utilized as a data resource for computations. This process can be iterated to continuously enhance the information content of the data set.

The following requirements have to be met: (i) update of information, (ii) design and implementation of flexible data models that can evolve with new data, (iii) integration of heterogeneous data, e.g. through the use of standard ontologies, (iv) comprehensive views and visualization of complex information for interactive access, (v) simple interfaces for direct access through applications to render data computable and to allow for transfer across species (2). While automation is the only way to achieve a high data throughput, manual curation is necessary to ensure quality (3). Intuitive tools to explore and evaluate the information enable users to inspect the data on which annotations are based and thus assess the confidence of the assignments.

Here we report our efforts to meet these challenges within the MIPS *Arabidopsis thaliana* database [MAtDB, http://mips.gsf.de/proj/thal/db (4)].

DATA INTEGRATION AND UPDATE

All data from the *Arabidopsis* Genome Initiative (AGI) have been integrated into MAtDB (serving as primary data repository for sequences and annotation produced by the European effort within the AGI) (1). In addition to the nuclear genome, the mitochondrial and chloroplast genome sequences were integrated.

A major advance has been the integration of full-length cDNA data produced by various projects (5). In the original data set published by the AGI in 2000, only 9% of all genes had been characterized experimentally. The remaining gene models were based on intrinsic information only, similarity to known proteins or partial coverage by EST sequences (1). Now, over 25 000 full-length cDNA sequences have been aligned to their cognate genes, providing experimental

*To whom correspondence should be addressed. Tel: +49 89 3187 3586; Fax: +49 89 3187 3585; Email: h.schoof@wzw.tum.de

evidence for the gene models of ~50% of all protein-coding genes. This allows a large-scale validation of the prediction procedures adopted by the AGI. A study by Haas *et al.* (6) confirms the value of extensive manual curation of gene predictions: 62% of AGI gene models were confirmed by the cDNA data, while fully automated algorithms are expected to perform at ~45% (7). Additionally, the full-length cDNA data allows us to assess features like alternative splicing, UTRs, transcription start sites and micro-exons (6).

Within MAtDB, new cDNA data is automatically gathered from the EMBL nucleotide database using a BioRS™ keyword query (BioRS™, a biological retrieval system, is a software product of Biomax Informatics AG, http://www.biomax.de). The retrieved sequences are aligned against the *Arabidopsis* genome using SplicePredictor (8). However, as a significant portion of full-length cDNA sequences represent splicing anomalies or artifacts, only alignments consistent with annotated gene models are integrated automatically. Remaining discrepancies are checked manually and commented whenever appropriate. A graphical view of the cDNA alignments allows MAtDB users to check these and form their own conclusions. In a few cases, single nucleotide mismatches or indels between cDNAs and genomic sequences lead to incompatible gene models. If the open reading frame suggested by the cDNA is interrupted by a stop or frameshift in the corresponding genomic sequence, sequencing errors in the genomic sequence are probable.

EST sequences, though partial, can in many cases be used in a similar way once they have been assembled to tentative consensus clusters (9). This is performed by the Sputnik system (10) and integrated with MAtDB. Over 60% of *Arabidopsis* protein-coding genes are matched by at least one EST.

Beside the protein-coding genes, non-coding RNA (ncRNA) genes are receiving increasing interest (11). Known RNA genes as well as predicted tRNA genes are included in MAtDB from the start. Recently, in an effort to identify new candidate ncRNA genes, 509 EST sequences not mapping to any annotated gene were analysed with the INFERNAL package (12). In all, 302 matched at least one of the models derived from the Rfam database (12) and could thus be classified, e.g. as nucleolar RNAs (snoRNAs).

Flanking sequence tags (FSTs) derived from insertion mutagenesis experiments are being produced in high numbers (13). Currently, the data generated by the GABI-Kat (http://www.mpiz-koeln.mpg.de/~GABI-Kat/GABI-Kat_homepage.html) and SIGNAL (http://signal.salk.edu) projects are gathered automatically by a BioRS query from the EMBL nucleotide database and mapped onto the genomic sequence. A graphical display shows the location of the matches with respect to the structure of the closest gene. Over 90 000 FST sequences have been unambiguously anchored to the *Arabidopsis* genome. This procedure results in over 50% of protein-coding genes having a FST match within the coding region, and 90% within a 2 kb region surrounding the gene (see Fig. 1).

Expert knowledge, e.g. on specific protein families, must find its way into the genome databases. For this purpose, MAtDB contains an external annotation section. Data are stored in XML format, allowing rapid integration of any tabular data. A generic display based on XML stylesheet

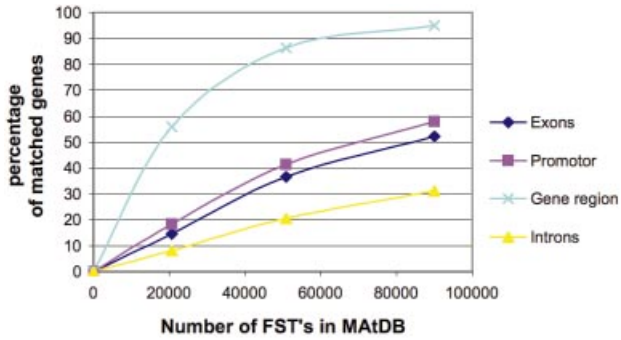

Figure 1. Percentage of protein-coding genes matched by a FST (see text). With increasing numbers, insertion mutagenesis experiments are expected to achieve saturation of all genes. At three timepoints, the amount of FST matches then within MAtDB was plotted against the percentage of matched genes. Gene region: FSTs matching within 2 kb upstream of the start or downstream of the stop codon. Promoter: FSTs matching within 1 kb upstream of the start codon. Exons: FSTs overlapping a coding exon. Introns: FSTs overlapping an intron in the coding sequence.

transformations allows sorting and subselection of the tables by web users. More than 20 data sets representing diverse protein families have been contributed so far (http://mips.gsf.de/cgi–bin/proj/thal/framesetter?about&externalanno.html). In addition, a comment form allows web users to add their knowledge to any gene. Submitted comments are displayed directly in the gene reports including author information, whereas corrections to gene predictions are curated at MIPS to ensure a high quality standard for gene models.

Cross-referencing of gene reports in MAtDB to other databases, e.g. TIGR (http://www.tigr.org) and TAIR [http://www.arabidopsis.org (14)], is based on AGI locus codes. These provide unambiguous identifiers for all *Arabidopsis* protein-coding, RNA and pseudogene genetic elements and are consistent across databases. They have the format At1g00010, where At stands for *Arabidopsis thaliana*, followed by the chromosome number, followed by g for 'gene', followed by a five-digit number with which all loci on a chromosome are numbered, starting at the top. Generally, AGI locus codes are incremented by 10 for the next locus, in order to leave room for as yet undiscovered genes. MAtDB maintains a list of current and obsolete locus codes (http://mips.gsf.de/cgi-bin/proj/thal/display_codes.pl), while TAIR has taken over the responsibility of code administration and maintains a code history (http://www.arabidopsis.org/tools/bulk/locushistory/index.html).

It is worthwhile to mention that these codes are carefully maintained to be consistent even while reannotation efforts are ongoing at several sites (15). However, a frequent misunderstanding is that these codes will reference the same protein in all databases, in particular the nucleic acid sequence archives. The locus code always references a defined genome location, and not a specific protein transcribed from that location. There may be several gene models for a single AGI locus code, in the case of alternative transcription or splicing even within one database.

WEB ACCESS

The MAtDB web interface allows access of the data through graphical or list browsing, searching by keywords, names or

sequences, and through precompiled tables that summarize general, functional, structural or comparative features (4). New features include precomputed FASTA homology scores provided by the SIMAP database [see Mewes *et al.*, this issue (16)] or protein domain detection using SESAM (17).

To be more flexible with respect to extensions of the data model, a new graphical interface, dbBrowser, was developed. For more demanding graphical exploration and interactive curation of the genome annotation, a standalone solution is less restrictive than a web interface. For this purpose, the already widely used Apollo Genome Browser developed by the Berkeley Drosophila Genome Project and Ensembl was adopted (18).

Keyword queries across user-selected fields in multiple databases are available via the query interface of BioRS. BioRS also allows for downloading batch results such as FASTA sequence files or annotation data in XML format.

ACCESSING MAtDB VIA WEB SERVICES: TOWARDS INTEGRATION OF DISTRIBUTED BIOLOGICAL DATA SOURCES

Genome-related data sources become increasingly complex and heterogeneous. Thus, the interpretation of the data by interactive browsing becomes limited and interfaces to allow for computational access become a necessity. Solutions for interoperability between databases are needed (2). This problem is approached by the PlaNet project of European plant databases (http://www.eu-plant-genome.net). It will interconnect the partner databases and allow centralized access while the data remain distributed. Database interoperability eliminates the need to locally warehouse all data. Thereby, the most current data are always at hand.

The BioMOBY project [http://www.biomoby.org (19)] was designed to require minimum standardization, retaining flexibility, while achieving a maximum of integration. To overcome restrictions of heterogeneous data models and formats, BioMOBY offers a technology that is based on a central registry of available services. These are implemented as web services that retrieve or operate on data. BioMOBY defines data objects in XML, which are kept lightweight to reduce the effort required for standardization. The central registry acts as a broker, removing the need for clients to know the data sources.

A first step towards interoperability was achieved by the implementation of BioMOBY web services operating on MAtDB data. The services currently implemented retrieve an AGI locus code, an *Arabidopsis* protein sequence or an EMBL entry by a given keyword or AGI locus code. Nevertheless the power of such a service is by no means restricted to the retrieval of such a particular piece of information. Applications may be pipelined, i.e. the output of one application or data resource can serve as the input of another service. In this way it is possible to generate workflows: e.g. a keyword can be used to retrieve a list of AGI locus codes annotated with 'disease resistance', which can then be used to retrieve NASC codes from the BioMOBY service implemented by the Nottingham *Arabidopsis* Stock Center (NASC, http://www.arabidopsis.info), leading on to the corresponding phenotypes through another service from the NASC.

FUTURE DIRECTIONS

The concept and content described here have established MAtDB as a valuable resource for biologists and bioinformaticians worldwide. The focus of MAtDB development will be towards providing biological organization of genome-related data. The main requirements are continuous data collection, timely integration of new data and analysis methods, and interoperability with distributed resources.

Our database integration efforts include both local and global plant databases. For global interoperability, access to MAtDB data and functionality through BioMOBY-compliant services will be provided within the PlaNet project. Locally, we are developing MAtDB technologically and integrating it with the other databases at MIPS in cooperation with the Genome Research Environment project (GENRE, http://mips.gsf.de/projects/gams). We will not restrict development to the *Arabidopsis*-only view but aim for a comprehensive and integrated view of available plant genomes. To facilitate this, the MIPS *Oryza sativa* database [MOsDB (20)] and the only recently emerging maize database share a common database design, application logic and presentation with MAtDB.

We will explore methods to transfer the richness and depth of data available for *Arabidopsis* to other plant genomes. In this view, MAtDB intends to be a fundamental data source within a platform for comparative genomics.

DATA DOWNLOAD AND STABLE LINKS

Complete sets of *Arabidopsis* sequences and annotation can be downloaded from ftp://ftpmips.gsf.de/cress. This includes lists of EST matches or functional classification, based partially on automatic, similarity-based assignments. Specialized dumps can be generated and downloaded using BioRS (http://biors.gsf.de:8111/searchtool/searchtool.cgi). 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 with an AGI locus code.

ACKNOWLEDGEMENTS

We wish to thank our collaborators who have contributed their data or agreed to link their database to MAtDB and all who submitted comments, e.g. to the online form. We thank all our colleagues at MIPS for support and contributions. BioRS™ is developed by Biomax Informatics AG (http://www.biomax.de). MAtDB is funded by the GABI (http://www.gabi.de, BMBF FKZ 0312270/4) and PlaNet (http://www.eu-plant-genome.net, EU Framework V; QLRI-CT-2001-00006) projects.

REFERENCES

1. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature*, **408**, 796–815.
2. Schoof,H. (2003) Towards interoperability in genome databases: the MAtDB (MIPS *Arabidopsis thaliana* database) experience. *Comp. Funct. Genomics*, **4**, 255–258.
3. Schoof,H. and Karlowski,W.M. (2003) Comparison of rice and *Arabidopsis* annotation. *Curr. Opin. Plant Biol.*, **6**, 106–112.

4. Schoof,H., Zaccaria,P., Gundlach,H., Lemcke,K., Rudd,S., Kolesov,G., Arnold,R., Mewes,H.W. and Mayer,K.F.X. (2002) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.*, **30**, 91–93.
5. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.
6. Haas,B.J., Volfovsky,N., Town,C.D., Troukhan,M., Alexandrov,N., Feldmann,K.A., Flavell,R.B., White,O. and Salzberg,S.L. (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.*, **3**, RESEARCH0029.
7. Pertea,M. and Salzberg,S.L. (2002) Computational gene finding in plants. *Plant Mol. Biol.*, **48**, 39–48.
8. Usuka,J., Zhu,W. and Brendel,V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211
9. Zhu,W., Schlueter,S.D. and Brendel,V. (2003) Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. *Plant Physiol.*, **132**, 469–484.
10. Rudd,S., Mewes,H.W. and Mayer,K.F.X. (2003) Sputnik: a database platform for comparative plant genomics. *Nucleic Acids Res.*, **31**, 128–132.
11. Reinhart,B.J., Weinstein,E.G., Rhoades,M.W., Bartel,B. and Bartel,D.P. (2002) MicroRNAs in plants. *Genes Dev.*, **16**, 1616–1626.
12. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
13. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.
14. Huala,E., Dickerman,A.W., Garcia-Hernandez,M., Weems,D., Reiser,L., LaFond,F., Hanley,D., Kiphart,D., Zhuang,M., Huang,W. *et al.* (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.
15. Wortman,J.R., Haas,B.J., Hannick,L.I., Smith,R.K.,Jr, Maiti,R., Ronning,C.M., Chan,A.P., Yu,C., Ayele,M., Whitelaw,C.A. *et al.* (2003) Annotation of the *Arabidopsis* genome. *Plant Physiol.*, **132**, 469–484.
16. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Güldener,U., Mannhaupt,G., Münsterkötter,M., Pagel,P., Strack,N., Stümpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
17. Strack,N. and Mewes,H.W. (1999) SESAM: Seed Extraction Sequence Analysis Method. *Proceedings of the German Conference on Bioinformatics GCB '99*. pp. 59–65.
18. Lewis,S.E., Searle,S.M.J., Harris,N., Gibson,M., Iyer,V., Richter,J., Wiel,C., Bayraktaroglu,L., Birney,E., Crosby,M.A. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.
19. Wilkinson,M.D. and Links,M. (2002) BioMOBY: An open source biological web services proposal. *Brief. Bioinform.*, **3**, 331–341.
20. Karlowski,W.M., Schoof,H., Janakiraman,V., Stuempflen,V. and Mayer,K.F.X. (2003) MOsDB: an integrated information resource for rice genomics. *Nucleic Acids Res.*, **31**, 190–192.