D834–D840 *Nucleic Acids Research, 2007, Vol. 35, Database issue*
doi:10.1093/nar/gkl945

MIPSPlantsDB—plant database resource for integrative and comparative plant genome research

Manuel Spannagl, Octave Noubibou, Dirk Haase, Li Yang, Heidrun Gundlach, Tobias Hindemitt, Kathrin Klee, Georg Haberer, Heiko Schoof and Klaus F. X. Mayer*

MIPS, Institute for Bioinformatics, GSF Research Center for Environment and Health, Ingolstädter Landstr. 1 85764 Neuherberg, Germany

Received August 15, 2006; Revised October 18, 2006; Accepted October 20, 2006

ABSTRACT

Genome-oriented plant research delivers rapidly increasing amount of plant genome data. Comprehensive and structured information resources are required to structure and communicate genome and associated analytical data for model organisms as well as for crops. The increase in available plant genomic data enables powerful comparative analysis and integrative approaches. PlantsDB aims to provide data and information resources for individual plant species and in addition to build a platform for integrative and comparative plant genome research. PlantsDB is constituted from genome databases for *Arabidopsis*, *Medicago*, *Lotus*, rice, maize and tomato. Complementary data resources for *cis* elements, repetive elements and extensive cross-species comparisons are implemented. The PlantsDB portal can be reached at http://mips.gsf.de/projects/plants.

INTRODUCTION

After the genome sequences of both *Arabidopsis* and rice are already available numerous plant genome sequencing are rapidly progressing and the near future a bouquet of plant genomes will be fully available. The ongoing projects circumvent both model genomes such as the model genomes for legumes *Medicago truncatula* and *Lotus japonicus* as well as crop genomes such as maize and tomato. Beside the need for comprehensive and structured information, genome and knowledge resources for the individual species and their genomes the availability of a range of plant genomes that represent a wide spectra and evolutionary range, bears the promise to undertake previously infeasible detailed and in-depth comparative analysis among different species. Without doubt, these analysis will give new insights into similarities and dissimilarities as well as specific characteristics of individual plant genomes. In addition, these information resources will help to elucidate genic elements that have not been discovered so far or have been difficult to detect. A prerequisite for detailed and in-depth cross-species comparisons and comparative phylogenetic analysis are consistent with detailed data resources. PlantsDB aims to address this task by applying a generic, highly flexible modular database infrastructure for a wide range of plant genomic data. The respective species databases are updated and new data are continuously integrated either through adjustment against external resources, or via the groups participation in a range of plant genome sequencing projects. The rapid cycle of data analysis and inclusion of analytical results into the respective databases thereby warrants rapid availability of the latest analytical results and data.

Although individual organism databases provide an important pillar of PlantsDB, the focus of PlantsDB is extending beyond individual genomes. PlantsDB also aims to make available resources that are species spanning and address, and support specific questions in comparative and integrative plant genomics. Topics and resources circumvent integrated resources for the detection and analysis of conserved ortholougous sequence markers (COS markers), repeat catalogs and classification systems for all plant species, comparative views and search opportunities and a *cis*-element database based on comparative sequence analysis. The PlantsDB resources are completed by the provided BioMOBY based web service opportunities that support seamless navigation and combination of services provided by PlantsDB and partner databases worldwide. PlantsDB can be accessed at http://mips.gsf.de/projects/plants.

PLANTSDB SYSTEM ARCHITECTURE

A modular rather than a data warehouse approach has been chosen for the MIPS plant genome resources. This enables flexible integration of new data sources and a high degree of flexibility for the development of individual modules.

*To whom correspondence should be addressed. Tel: +49 89 3187 3584; Fax: +49 89 3187 3585; Email: K.mayer@gsf.de
Present addresses:
Dirk Haase, Max Planck Institute for Plant Breeding Research, Carl-von-Linne Weg 10, 50829 Köln, Germany
Heiko Schoof, Max Planck Institute for Plant Breeding Research, Carl-von-Linne Weg 10, 50829 Köln, Germany

Core data, i.e. the genome sequence and annotation units, are implemented along the generic plant database system.

For each species a new instance of the data schema is initialized. The data schema consists of three hierarchical levels: *Clones*, *Contigs* and *Geneticelements*.

The *Clones* module contains raw sequence entries. To assemble a representation of a particular genome sequence, these clone sequences are processed to remove overlaps and redundancy, ambiguous sequence or vector contamination. The *Contigs* module contains assembled sequences along with the information on how to assemble the contigs to longer sequences and pseudomolecules that represent complete chromosomes. The *Geneticelements* level is the main module within PlantsDB and contains data describing annotations, e.g. gene models, as well as all other accessory information acquired during the annotation and analysis process, e.g. versioning, cross-reference or evidence information.

PlantsDB has been implemented in a multi-tier-architecture using standard J2EE design patterns (http://java.sun.com/j2ee). This allows a component-oriented design. Middleware components communicate data in XML format. A business delegate layer provides a flexible interface to access data and methods, and is utilized by both the web presentation as well as to provide web service functionality (Figure 1).

This architecture eliminates the need to redesign retrieval and storage components for new database instances—for example new genomes—but makes existing patterns highly reusable and maintainable. On top of the middleware, a Java API provides easy programming access to all database content. This API is used for computational analysis as well as for large-scale data retrieval. The web presentation uses the Java Server Faces implementation of a Model View Controller pattern (http://java.sun.com/j2ee/javaserverfaces/) along with Cascading Style Sheets transforming the middleware-gained XML formatted database results.

For integration with remote databases, the web services-based interoperability solutions of the BioMOBY (http://www.biomoby.org) initiative are implemented (1).

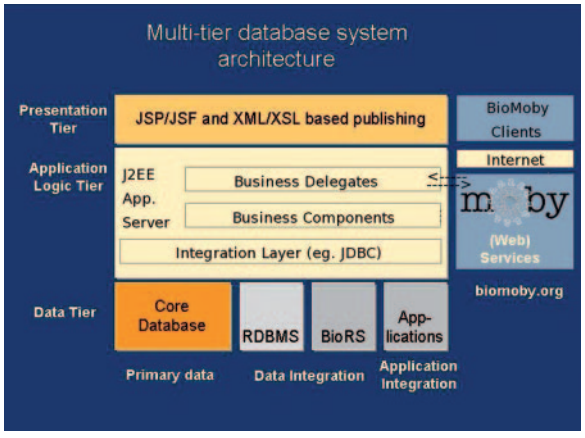

Figure 1. Database architecture of PlantsDB. PlantsDB is implemented as multi-tier architecture. Data tier, Application logic and presentation layer are seperated. BioMOBY driben web services are are driven via J2EE Business Delegates. The integration layer integrates data from core databases holding the primary data as well as from related relational database management systems (RDBMS), retrieval systems (BioRS) and applications.

PLANTSDB: DATA ACCESS AND RETRIEVAL

The MIPS plant genome resources provide access to all genomes included in common formats and similar interfaces. The main entry point for all databases included in PlantsDB can be found at http://mips.gsf.de/projects/plants. The web content is managed using the JBOSS 4.0.1 application server (http://www.jboss.org/) and Java Server Faces (http://java.sun.com/j2ee/javaserverfaces/). For every species and genome included all sequenced and annotated contigs can be retrieved. The contig names are linked to a report page that communicates detailed information on the respective entries and links to a list of the respective annotated genetic elements as well as to a graphical viewer. The genetic element name links to detailed report of analytical results for the respective protein sequence (Figure 2B). Specific genetic elements can be downloaded using the Genetic Element Retrieval System (GenERSys) tool (Figure 2C) as well as via web service. Cross-references in the report enable access to associated entries in external databases. Connection to the SIMAP database enables retrieval of sequence homologs from Viridiplantae, Fungi and Mammalia (2). For all genes, unspliced, spliced, coding DNA sequences as well as protein sequences are available and can be retrieved in HTML, XML or in FASTA format. In addition, for each species database a table containing complete lists of chromosomes, all genetic elements or all elements of a selected type can be browsed.

Together with other user selected tracks, genetic elements on a selected contig can be searched and viewed through Gbrowse (Figure 2D). Gbrowse is a Generic Genome Browser combining a relational database and interactive web interfaces for displaying and manipulating annotation on genomes (3).

Individual or cross databases queries can be undertaken using different search forms (Figure 2A). The query options include search by name (e.g. contig or genetic element name) as well as free-text searches. BLAST is used as homology search engine (4). Data sets from the plant projects at MIPS as well as Swissprot/SWALL and plant-specific EST and TC data sets are searchable.

The download section of the web-interface provides ftp access to various data forms. This includes FASTA-formatted sequence files for all contigs and protein coding genes. Moreover this section contains functionality to create and download user defined Genome Annotation Markup Element files (GAMEXML) used by the Apollo Genome Browser (5). Apollo provides a detailed graphical viewer for genome data with more flexible interaction possibilities than a browser-based display. In addition, it allows interactive curation of the genome annotation and saves the results locally for future reference. Downloading a GAMEXML file of the region of interest enables the user to inspect (and modify) all gene annotation data, thus building the users own local, hand-curated annotation data set. This also provides an infrastructure for community-based distributed manual annotation. The edited GAMEXML files can be returned to us by email for inclusion in the database.

PLANTSDB ORGANISM COMPONENTS

Under the umbrella of PlantsDB several individual species databases are contained. Although the individual databases

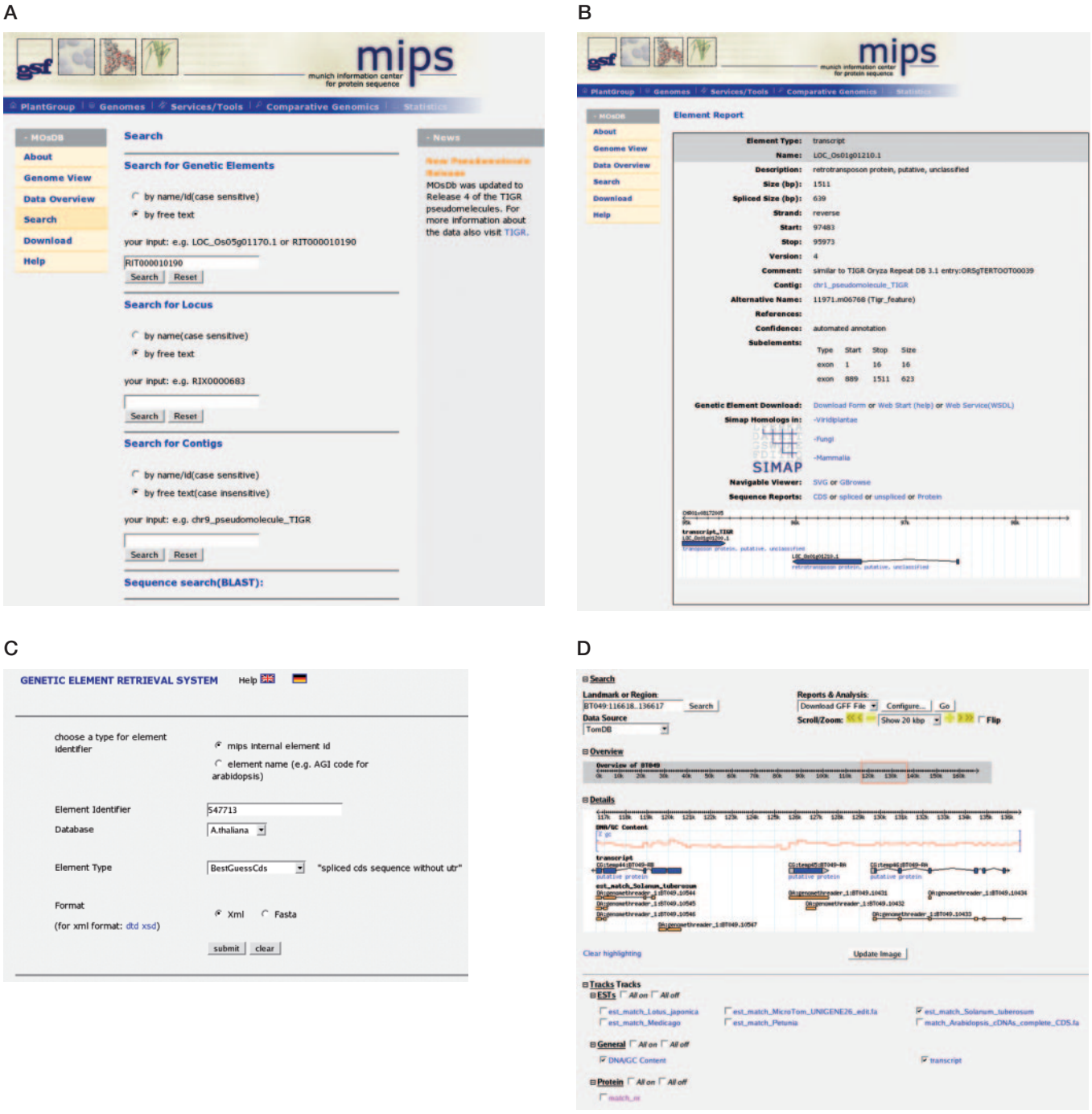

Figure 2. Web interfaces to the MIPS plant genome resources. (**A**) Users can search different databases fields by using the corresponding web form. (**B**) A genetic element report page with detailed information on a protein coding gene. Homologs in Viridiplantae, Fungi and Mammalia can be retrieved from the SIMAP database. (**C**) The Genetic Element Retrieval System (GenERSys) provides a tool to download specific genetic elements (e.g. promoter, UTRs, etc.). (**D**) A graphical view of transcripts on a contig section together with EST evidences as displayed by Gbrowse.

are physically separate, database structures are identical and user interfaces and services provided are similar. This is a prerequisite for easy and intuitive navigation as well as for comparative studies. At the time of writing the PlantsDB system comprises genome databases for *Arabidopsis*, *Medicago*, *Lotus*, Maize, rice and tomato. However, due to the generic and modular architecture new and upcoming plant genome databases can be rapidly installed.

Arabidopsis thaliana

MAtDB2 [MIPS *A.thaliana* database, (6)] contains both the original MIPS assembly of the Arabidopsis Genome Initiative (AGI) genome sequence (7) as well as the TAIR version 6 containing refined assembly and updated annotation. To provide a merged annotation set, gene models and functional assignments that have been manually curated at MIPS were mapped to the TAIR version 6 assembly.

Legumes: *M.truncatula* and *L.japonicus*

The genome of *M.truncatula* is currently being sequenced. The European Medicago and Legume Database (UrMeLDB) integrates data from both the international *Medicago* sequencing project as well as the European GLIP project. UrMeLDB aims to provide comprehensive, state-of-the-art analysis of the contained sequences, using tools selected and customized to perform well on *Medicago*. All publicly available *Medicago* genomic clone data are continuously integrated into UrMeLDB. Updates are undertaken on a regular basis and newly available as well updated clone sequences are integrated into UrMeLDB. As of July 2006, 2001 *Medicago* clone sequences are available.

Gene prediction and protein annotation of *Medicago* sequences are performed in collaboration within the frame of the International Medicago Genome Annotation Group [IMGAG, (8)]. All bioinformatic analyses available in UrMeLDB build on this gene call set. Updates of gene models are recorded by using the UrMeLDB versioning facilities. Thus, changes in annotation due to sequence reassembly as well as improvement of gene prediction can be traced. Besides gene prediction, annotation of the genomic sequences involves detection of noncoding features like transposons, repeats and RNA genes. Many of these tasks are performed through automated bioinformatic analysis, and the results are integrated into the genome databases.

In addition to *M.truncatula*, *L.japonicus* is the second legume species being sequenced. The MIPS Lotus genome database provides access to both BAC sequences derived from the Lotus sequencing project (9) and annotation performed using the same standards as for *Medicago*. At the time of writing (July 2006), 1384 BAC sequences were available.

Grass genomes: rice and maize

Both rice and maize have a long tradition as model plants. At the same time, they are crop plants that are of major importance for the worlds food supply. Both are representatives of the monocotyledonous plants, which diverged from the dicotyledonous plants ~140–180 million years ago (10). Consequently all different levels of comparative analysis between representatives of the monocotyledonous and the dicotyledonous plants are of special interest.

For rice the MIPS data resource MIPS *Oryza sativa* database (MOsDB) (11) includes the publicly available *Oryza sativa* ssp. *japonica* cv. Nipponbare assembly and annotation as provided by TIGR (12). Beside that, the International Rice Genome Sequencing Project (IRGSP) genome sequence and annotation has been integrated as well. Having annotation from two different sources in place enables users to compare and evaluate gene models of interest.

The maize database includes 100 publicly available BAC sequences and two megatile sequences from chromosome 1 and 9 (>14 Mbps) with manually curated gene predictions (13). With the ongoing sequencing efforts for the maize genome, we will continue to integrate sequence and annotation data into the MIPS PlantsDB section as they become publicly available.

Tomato

Tomato (*Solanum lycopersicum*, formerly *Lycopersicon esculentum*) belongs to the family of Solanaceae (nightshade) that circumvent many agriculturally important crops such as potato, tobacco, pepper, egg-plant as well as ornamental plants and medical plants.

Tomato is currently being sequenced within the frame of an international consortium (http://www.sgn.cornell.edu/). Within the frame of the international and EU sequencing efforts, the MIPS tomato database aims to develop toward a main information gate for tomato and Solanaceae genomics. Currently, the MIPS tomato database contains 50 finished BAC sequences including annotation from different sources, and will constantly be updated and integrate sequences that become available, and are released from the international tomato project.

COMPARATIVE AND INTEGRATIVE TOOLS AND RESOURCES

The MIPS Repeat Element database (mips-REdat) and the MIPS Repeat Element catalog (mips-REcat)

Plants genomes are notorious for containing large amounts of repetitive elements. The content ranges from 20 to 30% in Arabidopsis (125 Mb) and rice (430 Mb) to over 90% in huge genomes like wheat (16 000 Mb). Mobile elements, the main group of repetitive elements, have first been discovered in maize in the 1950s by Barbara McClintock (14). Their role has been redefined several times: first as controlling elements, then as parasitic and selfish junk DNA with only deleterious effects and more recently as a necessary evolutionary task force. More and more evidence is accumulating, that mobile elements contribute to genetic diversity by shuffling functional blocks (15–17). Regardless of their role, repeat elements can be used as fossil traces from the past to study genome evolution. Prerequisite for such comparative analyses are consistent cross-species annotation and classification of repeat elements.

The umbrella term repeat element covers a large and heterogeneous group of genetic elements, which to make matters worse, are often degenerated and fragmented by insertions into each other. We have developed an Automated Nested Genetic Element Annotation pipeline (ANGELA) to detect such complex fragmented and nested structures on long, even chromosome sized sequences. Repeat identification, element defragmentation and data extraction of ANGELA are based on two main external components: an extendable database of plant repeats elements (mips-REdat) and a generic repeat classification schema (mips-REcat).

MIPS Repeat Element database

mips-REdat represents an exhaustive collection of plant repetitive elements. The content has been collected and compiled from several public sources, like Repbase (18), TIGR repeats (19), TREP (http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml) or PlantSat (20) as well as additional private collections. The database is continuously extended by reparsing updated repeat collections and by newly detected repeat elements from our pipeline. To remove identical repeats the sequences are clustered using Vmatch (http://www.vmatch.de) with a 98% identity limit, taking the longest sequence as a representative for the non-redundant set. This procedure concurrently removes incomplete sequences, which are part

of a longer template. The current version 4.3 of mips-REdat contains 6489 non-redundant sequences that add up to 23 Mb. mips-REdat does not only store sequences. Additional information with respect to source (institution, Genbank ID), description, keywords, literature (Pubmed ID), organism (NCBI taxonomy-ID), sequence completeness (full-length or partial element) and most importantly, classification keys that link the repeat sequences to mips-REcat.

MIPS Repeat Element catalog

mips-REcat has been designed for automated repeat element annotation and flexible data retrieval. It integrates and extends existing repeat classifications into a systematic hierarchical tree structure. The machine-readable key (e.g. 02.01.05) facilitates data extraction at different levels of detail. Repetitive elements are divided into three main groups:

(i) Simple Sequence Repeat (e.g. micro-, minisatellite and satellite),
(ii) Mobile Element (Retrotroelement, DNA transposon and Helitron),
(iii) High Copy Number Gene (e.g. RNA gene, histones).

An additional category for additional attributes enables the assignment of general features, like replication type or chromosome location and the annotation of sequence attributes, like partial/complete sequence or nesting level. The categories in mips-REcat are characterized by the following data fields: key, description, definition, regular expression for assignment by keyword, pubmed IDs and mappings to the TIGR and RepeatMasker classifications.

All repeat sequences in mips-REdat have been classified by several complementary approaches: regular expression search for REcat keywords of the description and keyword fields, source specific classifications (TIGR, Repbase, RepeatMasker) and an hmm search (21) for transposon specific protein signatures (transposase, GAG, PR, RT, INT). LTR-retrotransposons have been checked for completeness by inspecting the presence of left and right solo LTRs.

mips-REdat and mips-REcat web-interface

The web-interface offers three different entry points to browse and retrieve designated sequences from mips-REdat: A toggle menu of the mips-REcat tree, a toggle menu of the taxonomy tree and input fields for a combined classification and taxonomy search.

For each category the two tree menus show the sum of sequences available from mips-REdat including all corresponding lower categories. The numbers are linked to a list of the associated repeats containing additional information regarding source, classification, literature, genbank IDs and the opportunity for sequence download.

Depending on the respective focus of interest, it is possible to retrieve a customized subset of mips-REdat (e.g. all LTR-retrotransposons from grasses or all repeat sequences from maize) as RepeatMasker or FASTA format.

MotifDB *cis*-element database and CREDO (*cis* regulatory element detection online)

One of our research goals are the comprehensive discovery of candidate transcription factor binding sites (TFBs) in plant genomes to enable and stimulate the study of regulatory networks in plants. Based on a combination of phylogenetic footprinting and motif discovery within promoters of co-expressed genes among *Arabidopsis* and *Brassica* promoters we determined candidate *cis*-regulatory elements in *A.thaliana* (22). Statistical significance of detected motifs has been confirmed and enrichment for specific functional categories like GO annotations and KEGG pathways both for co-expressed gene sets as well as for detected motifs were determined.

Results of this study have been integrated within the comparative section of PlantsDB as a browsable database (http://mips.gsf.de/proj/plant/webapp/expressionDB/index.jsp). PlantsDB supports queries for their individual genes of interest as well as for co-expressed genes and for detected candidate TFBs. In addition, users interested in specific biological processes can look up candidate motifs that are significantly over represented in categories of interest as well as genes within these categories that contain an instance of the respective motif candidates. In summary, the MotifDB *cis*-element database content supports researchers to identify new candidate genes and functional links. Future directions aim to integrate comparative *cis* element directed analysis between more species as well as the inclusion of expression data derived from species other than *Arabidopsis*.

Complementary to the combined usage of expression data and phylogenetic data for the detection of conserved *cis*-regulatory elements for species, which so far lack exhaustive expression data, we developed an integrated analysis pipeline to compare and integrate results derived from different methods and using complementary tools (23). CREDO (*Cis*-Regulatory Element Detection Online) integrates, combines and visualizes the analysis of AlignACE, DIALIGN, FootPrinter, MEME and MotifSampler, and therefore facilitates the comparison of their results. It enables to run each of the algorithms simultaneously on a given dataset and summarizes the outputs of all programs graphically, in tables and within a XML file. Almost all parameters of the algorithms applied can be adapted. The CREDO web form provides a structured parameter selection form by default advanced parameters are hidden and preset values are being used. For expert users the opportunity to change these parameters and refine the analysis is provided. CREDO provides three different and widely applied presettings. Presetting 1 has been designed for users, who aim to carry out phylogenetic footprinting with closely related species. The second setting has been designed for phylogenetic footprinting with more distantly related species, and finally a presetting for users who set out to search for conserved sequence motifs in co-expressed genes is provided. Analysis results are depicted in a graphical overview. For each input sequence the motifs detected by the individual algorithms along with a summary view are displayed. This view summarizes the motifs found by all programs.

The graphical representations of motif occurrences are linked to underlying analytical data. The result pages include links to three pop-up windows that contain the table of found motif data, input sequences and chosen parameters, respectively. The motif table provides all important motif data and sequence logos for each motif detected.

Conserved orthologous sequence (COS) markers

Orthology is one of the most important concepts in bioinformatics. If two genes are orthologs, they are usually believed to fulfill the same or very similar functions in their respective organism. For comparative mapping experiments, it is crucial to know sets of orthologous genes that cover all target species. On the other hand, it is important that each of the genes occurs as single-copy in its respective genome. We developed an application called 'Conserved Ortholog Sequence (COS) Markers' that satisfies these constraints. Putative ortholog pairs are determined based on bidirectional best hits (bbh). Second best hits are used to delete clusters that contain paralogous genes.

The application is accessible via a web-based user interface (http://mips.gsf.de/cgi-bin/proj/planet/cos/seedSelect.pl). In the first step, the user selects the so-called seed species. In a second form, the other (non-seed) organism(s) are specified. Additional parameters for adjusting the default values for orthology and paralogy filters can be defined as well. Distinction of seed and non-seed species is caused by the strategy for building the individual clusters. An initial set of clusters is built from all putative orthologs between the seed and the first non-seed species. For additional non-seed species clusters are extended if a putative ortholog between seed and the non-seed species is found. Similarity between two non-seed genes is not considered. The processing results in a 'star-like' topology of the clusters, where each non-seed gene has a bbh relation to the seed species. Result sets are displayed along with the most important similarity parameters and the match quality is depicted by a color code. Detailed information on each cluster can be retrieved and a multiple alignment of the cluster can be performed.

The similarity matrix of proteins (SIMAP)

An important tool for comparative genomics is the prediction of homologs and orthologs between genomes. Within protein coding gene reports, possible orthologs in other plant species (as well as in fungi, mammals, etc.) can be extracted from the SIMAP database, a matrix for precomputed homologies of protein sequences (2,24).

SIMAP covers the similarity space formed by more than 4 million amino acid sequences from public databases (including UNIPROT Swissprot, all databases hosted at MIPS) and completely sequenced genomes. Sequence similarity searches are performed using FASTA (25) on low complexity masked regions, storing hits with a Smith-Waterman Score $\geqslant 80$ in the matrix. SIMAP links are provided both from the 'comparatice genomics' section of PlantsDB as well as from all individual gene reports. SIMAP data are retrieved from the SIMAP database protein reports using Enterprise Java Beans (EJB). The user can limit the search space to specific taxonomic groups (e.g. searching hits only in plants). The result report provides an overview about all relevant hits with the option to inspect the alignment in detail along with relevant scores and other features.

Interfacing the SIMAP database bears the advantage that precomputed FASTA homology scores and Smith–Waterman alignments can be retrieved on the fly. Compute intensive *de novo* homology searches are circumvented. Moreover, time and compute costs for orthology and paralogy assignment via suitable methods such as INPARANOID (26) are massively reduced by the replacement of the necessary BLAST alignments required as INPARANOID inputs by the precomputed SIMAP FASTA scores (24).

Links to external databases using web services

One of the most important goals of plant databases at MIPS is to make data accessible in a user-friendly fashion. Many of the newer features of PlantsDB reported in this article are aiming to further improve usability for the research community. However, in recent years, a growing demand for interfaces that allow for direct access by scripts became obvious (27). Although classic browser-based applications are in principle also accessible by programs (so-called screen-scraping), this approach is very tedious as every web-interface uses individual sets of parameters and different result formats. Service oriented architectures offer a solution by using standardized interfaces and protocols. Especially, web services are gaining wide acceptance for providing automated access to data and computational resources. In the context of bioinformatics, BioMOBY has emerged as a special flavor of web services, extending these by the addition of ontologies for the description of data structures, semantics and service types (28). BioMOBY was chosen as middleware within the frame of the PlaNet consortium and proved to be of special value for the task of data and resource interoperability and data integration between remote sites (1).

The BioMOBY webservices implemented at MIPS allow for direct access on the data from our plant genome resources by applications. They enable analyses to be performed remotely and provide a way to include MIPS data in web presentations without the need for a local copy. As the data format and semantics are defined in the BioMOBY ontologies, this provides solutions to the issues of synchronizing data and format updates.

To date 41 BioMOBY compatible web services are implemented and provide retrieval and analysis functionality. Analysis services provide access to BLAST search against organism specific sequence databases, but also on the COS marker application (see above). Another type of services are result parsers which serve for interlinking services to build complete pipelines, so-called workflows (see below). As the services implement full middleware access to the underlying databases, these tools can be used to remotely perform large-scale and complex comparative analyses across species that were previously possible only with local access to the data. This showcases the usage of web services for data integration tasks and serves as an example for the extension of locally stored data by information from remote sites. In a similar fashion, further data types like literature references or known phenotypes will be added to the gene report in the near future.

Many of the web services only implement a small piece of functionality. This makes them easily combinable to more complex tasks in a Lego-like manner. A setup of many web services with the output of one service serving as input for the next is called a workflow. Several example workflows are stored on our web site (http://mips.gsf.de/projects/plants/PlaNetPortal/taverna_workflows.html) for download. They are stored in the file format used by Taverna (29).

Thus, users who have installed Taverna can download and execute them from their computer.

DIRECTIONS

PlantsDB embraces a wide spectrum of data resources. Beside classical genome centric resources focused toward the individual species and their genomes resources communicating detailed analysis data and spanning a wider range of genomes. PlantsDB supports the challenges and potential of comparative plant genomics for academic and applied research. Modular and generic design of the individual organism genome components ensures comparability of data. Important steps toward providing species spanning resources and integrative comparative analysis and analytical views are already provided. With the huge amount of plant genomic data from different species and wide spectra of clades and families, future focus will be directed toward on one hand integrating the new genomes, and subject them to comprehensive and detailed bioinformatic and analytical analysis procedures within the framework of PlantsDB. A second important direction of resource and infrastructure development will be to support the increasing comparative analytical efforts by providing structured data resources as well as integrative and comparative analytical frameworks and views.

ACKNOWLEDGEMENTS

The work has been supported within the GABI project of the German Ministry of Science and Education as well as by funding by the European Commission (framework 6) within the Grain Legumes Integrative Project (GLIP). Funding to pay the Open Access publication charges for this article was provided by the GSF Research Center for Environment and Health.

Conflict of interest statement. None declared.

REFERENCES

1. Wilkinson,M., Schoof,H., Ernst,R. and Haase,D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. *Plant Physiol*., **138**, 5–17.
2. Rattei,T., Arnold,R., Tischler,P., Lindner,D., Stumpflen,V. and Mewes,H.W. (2006) SIMAP: the similarity matrix of proteins. *Nucleic Acids Res*., **34**, D252–D256.
3. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
4. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
5. Lewis,S.E., Searle,S.M., Harris,N., Gibson,M., Lyer,V., Richter,J., Wiel,C., Bayraktaroglir,L., Birney,E., Crosby,M.A. *et al*. (2002) Apollo: a sequence annotation editor. *Genome Biol*., **3,** RESEARCH0082.
6. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K.F. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res*., **32**, D373–D376.
7. Initiative,T.A.G. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
8. Cannon,S.B., Crow,J.A., Heuer,M.L., Wang,X., Cannon,E.K., Dwan,C., Lamblin,A.F., Vasdewani,J., Mudge,J., Cook,A. *et al*. (2005) Databases and information integration for the *Medicago truncatula* genome and transcriptome. *Plant Physiol*., **138**, 38–46.
9. Young,N.D., Cannon,S.B., Sato,S., Kim,D., Cook,D.R., Town,C.D., Roe,B.A. and Tabata,S. (2005) Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. *Plant Physiol*., **137**, 1174–1181.
10. Bell,C.D., Soltis,D.E. and Soltis,P.S. (2005) The age of the angiosperms: a molecular timescale without a clock. *Evolution Int. J. Org. Evolution*, **59**, 1245–1258.
11. Karlowski,W.M., Schoof,H., Janakiraman,V., Stuempflen,V. and Mayer,K.F. (2003) MOsDB: an integrated information resource for rice genomics. *Nucleic Acids Res*., **31**, 190–192.
12. Yuan,Q., Ouyang,S., Wang,A., Zhu,W., Maiti,R., Lin,H., Hamilton,J., Haas,B., Sultana,R., Cheung,F. *et al*. (2005) The institute for genomic research Osa1 rice genome annotation database. *Plant Physiol*., **138**, 18–26.
13. Bruggmann,R., Bharti,A.K., Gundlach,H., Lai,J., Young,S., Pontaroli,A.C., Wei,F., Haberer,G., Fuks,G., Du,C. *et al*. (2006) Uneven chromosome contraction and expansion in the maize genome. *Genome Res*., **16**, 1241–1251.
14. McClintock,B. (1956) Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol*., **21**, 197–216.
15. Orgel,L.E. and Crick,F.H. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604–607.
16. Kazazian,H.H. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.
17. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nature Rev. Microbiol*., **3**, 722–732.
18. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res*., **110**, 462–467.
19. Ouyang,S. and Buell,C.R. (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res*., **32**, D360–D363.
20. Macas,J., Meszaros,T. and Nouzova,M. (2002) PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*, **18**, 28–35.
21. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
22. Haberer,G., Mader,M.T., Kosarev,P., Spannagl,M., Yang,L. and Mayer,K.F.X. (2006) Large-scale Cis-element detection by analysis of correlated expression and sequence conservation between *Arabidopsis* and *Brassica oleracea*. *Plant Physiology*, **142**, in press.
23. Hindemitt,T. and Mayer,K.F. (2005) CREDO: a web-based tool for computational detection of conserved sequence motifs in noncoding sequences. *Bioinformatics*, **21**, 4304–4306.
24. Arnold,R., Rattei,T., Tischler,P., Truong,M.D., Stumpflen,V. and Mewes,W. (2005) SIMAP—the similarity matrix of proteins. *Bioinformatics*, **21**, ii42–ii46.
25. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
26. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*., **314**, 1041–1052.
27. Buetow,K.H. (2005) Cyberinfrastructure: empowering a 'third way' in biomedical research. *Science*, **308**, 821–824.
28. Wilkinson,M.D. and Links,M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform*., **3**, 331–341.
29. Oinn,T., Addis,M., Ferris,J., Marvin,D., Senger,M., Greenwood,M., Carver,T., Glover,K., Pocock,M.R., Wipat,A. *et al*. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.