Munich Information Center for Protein Sequences Plant Genome Resources. A Framework for Integrative and Comparative Analyses[1[w]]

Heiko Schoof[2]*, Manuel Spannagl, Li Yang, Rebecca Ernst, Heidrun Gundlach, Dirk Haase, Georg Haberer, and Klaus F.X. Mayer

Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D–85354 Freising-Weihenstephan, Germany (H.S., M.S.); and Institute for Bioinformatics, GSF National Research Center for Environment and Health, D–85764 Neuherberg, Germany (H.S., M.S., L.Y., R.E., H.G., D.H., G.H., K.F.X.M.)

With several plant genomes sequenced, the power of comparative genome analysis can now be applied. However, genome-scale cross-species analyses are limited by the effort for data integration. To develop an integrated cross-species plant genome resource, we maintain comprehensive databases for model plant genomes, including Arabidopsis (*Arabidopsis thaliana*), maize (*Zea mays*), *Medicago truncatula*, and rice (*Oryza sativa*). Integration of data and resources is emphasized, both in house as well as with external partners and databases. Manual curation and state-of-the-art bioinformatic analysis are combined to achieve quality data. Easy access to the data is provided through Web interfaces and visualization tools, bulk downloads, and Web services for application-level access. This allows a consistent view of the model plant genomes for comparative and evolutionary studies, the transfer of knowledge between species, and the integration with functional genomics data.

The Munich Information Center for Protein Sequences (MIPS; http://mips.gsf.de) has been involved in maintaining plant genome databases since the Arabidopsis (*Arabidopsis thaliana*) genome project (Arabidopsis Genome Initiative, 2000). Genome databases and analysis resources have focused on individual genomes (Karlowski et al., 2003; Schoof et al., 2004) and aim to provide flexible and maintainable data sets for model plant genomes as a backbone against which experimental data, e.g. from high-throughput functional genomics, can be organized and evaluated. But model genomes also form a scaffold for comparative genomics, and much can be learned from genome-wide evolutionary studies (Goff et al., 2002; Bonnet et al., 2004).

However, genome-scale bioinformatics analyses are limited by the effort required for data integration (Wilkinson et al., 2005). Regarding plants only, there is already a plethora of data resources worldwide: Arabidopsis data is available from several genome databases, including The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org), The Institute for Genomic Research (TIGR; http://www.tigr.org), and the MIPS *Arabidopsis thaliana* Database (MAtDB; http://mips.gsf.de/proj/thal/db; Rhee et al., 2003; Wortman et al., 2003; Schoof et al., 2004). Several resources are emerging for *Medicago truncatula* (VandenBosch and Stacey, 2003; Cannon et al., 2005); rice (*Oryza sativa*) is available through TIGR, Oryzabase (Yazaki et al., 2004), or Gramene (Ware et al., 2002); and maize (*Zea mays*) is maintained at MIPS or MaizeGDB (Lawrence et al., 2004), to name just a few. Each resource generally has its own data models, formats, and standards. For cross-species analysis, data must be transformed into a representation that facilitates comparison. For updates, this must be repeated.

To overcome these limitations, an integrated cross-species plant genome resource is desirable. This would complement the efforts of species-specific databases that focus on providing the best possible annotation for that genome by providing a platform that facilitates comparative analyses. The challenges are as follows.

1. Maintaining current and comprehensive data on the most relevant plant model genomes and integrating this into the framework. Standardization of data representations and interoperability with primary databases can significantly ease the task of data integration and updates (Schoof, 2003). Several initiatives to provide the required technol-

[1] This work was supported by the Federal Ministry of Education and Research, Germany, through the Genomanalyse im biologischen System Pflanze (http://www.gabi.de) project. Work on the European Medicago and Legume Database is funded in the European Grain Legumes Integrated Project by the Sixth European Union Framework Programme of the European Commission (grant no. FP 6 FOOD–CT–2004–506223). Data and database integration for Arabidopsis is funded by the Fifth European Union Framework Programme PlaNet project (grant no. FP5 QLRI–CT–2001–00006).

[2] Present address: Max Planck Institute for Plant Breeding Research, D–50829 Cologne, Germany.

* Corresponding author; e-mail schoof@mpiz-koeln.mpg.de; fax 49–221–5062–413.

[w] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.059188.

ogy and to establish standards are well under way and are supported by numerous plant genome data resources (Wilkinson et al., 2005, and refs. therein) and thus will be discussed only superficially in this article.

2. The flexibility to include additional data sets and novel data types. The number of at least partially sequenced plant genomes is increasing steadily. At the same time, high-throughput functional genomics experiments are yielding massive data sets that, on one hand, need to be interpreted in the context of the genome and, on the other hand, can enrich genome annotation. As new methodologies arise and our knowledge of how genomes work increases, new types of data, features, and novel ways to approach genome analysis need to be accommodated. For example, genome analysis previously focused on the protein coding genes, whereas recently, noncoding elements such as microRNAs have received intense attention (Reinhart et al., 2002).
3. The ability to work with heterogeneous data sets, i.e. completed genomes available on a clone-by-clone basis as well as partial genomes available through shotgun sampling of libraries prefiltered for coding regions. While the small plant genomes like Arabidopsis, Medicago, and rice have been/are being sequenced using a clone-by-clone approach, alternative approaches, including shotgun sequencing of methyl- or C_0t-filtrated libraries, are being tested for the large genomes like maize or wheat (*Triticum aestivum*; Palmer et al., 2003; Whitelaw et al., 2003). In other cases, large expressed sequence tag (EST) or bacterial artificial chromosome (BAC) end sequence (BES) collections may provide a first overview of genome content. While complete genomes can be represented by a single set of chromosome sequences (sometimes called pseudomolecules) that provide a coordinate reference for all annotation, the latter types of sequence collections are more dynamic: Many sequences may not be anchored to a specific chromosome, many are redundant, and the relations, e.g. overlaps between neighboring sequences, are often dynamic and subject to change. A comparative framework must be able to handle and compare these data and overcome the structural differences.
4. Easy access, both through clickable Web interfaces as well as through applications. Species-specific databases have seen several years of development and refinement of human interfaces that provide elegant ways to browse and access the data. Cross-species integration adds a new dimension: How to display data from several genomes at once? This problem is less critical for application access, where the prime requirement is the ability to access the data from different genomes through the same interfaces (i.e. using the same methods), enabling applications to work with any new genome without modification of the code.

The MIPS plant genome resources started out as a collection of species-specific databases but with the goal of merging these into an integrated, comparative framework (Schoof et al., 2004). Here, we present the design and implementation of the MIPS plant genome resources from the viewpoint of an integrative platform. The current status with respect to data content, analysis procedures, and access methods is described. Finally, we discuss the application of these resources for genome-scale cross-species analyses.

DESIGN

Instead of a single data warehouse, a modular approach was chosen for the MIPS plant genome resources. This allows for easy addition of new modules when new types of data need to be accommodated and for more flexibility in development of individual modules without being hampered by the complexity of the whole system. To be able to manage these data modules in a component-oriented manner, a multitier architecture following the Java 2 Platform, Enterprise Edition standard (http://java.sun.com/j2ee) was implemented (Mewes et al., 2004). For integration with remote databases, the Web services-based interoperability solutions of the BioMOBY (http://www.biomoby.org) initiative are implemented (Wilkinson and Links, 2002).

The core of the system, around which other data modules are organized, is a flexible data model for representing genome sequence and annotation (Karlowski et al., 2003). Three basic entities are defined: Clone, Contig, and GeneticElement. Clones store sequence and attached information that relates to a physical plasmid clone or similar, e.g. BAC sequences. To assemble a representation of a genome sequence, these clone sequences are processed to, for example, remove overlaps and redundancy, ambiguous sequence, or vector contamination, and then stored as Contigs. The Contig data module also stores information on how to assemble the contigs to longer sequences or eventually pseudomolecules representing whole chromosomes.

The third data module, GeneticElement, contains all features or elements that can be represented through coordinates on the genome sequence: protein coding genes, noncoding RNAs, repeats, sequenced markers, and transposons. This list is extensible whenever new features or elements are discovered and can utilize, for example, the Sequence Ontology (http://song.sourceforge.net) for semantic relationships, e.g. that a coding sequence is part of a transcript. GeneticElements can have subelements, e.g. exons or domains, that don't exist without the GeneticElement, e.g. exons exist only as parts of a transcript. To accommodate more abstract concepts, like "gene," GeneticElements can be grouped. In this way, all GeneticElements belonging to a gene (promoter, transcript, alternative transcripts, regulatory elements, cDNA matches, etc.) can be identified through a single group entry.

For every species, a separate physical instance of all three data modules is created to ensure scalability and separation of namespaces. The mapping of species to physical database instance is performed by the middleware. The database schema is available as supplemental data or on the MIPS Web site.

IMPLEMENTATION

Data Content and Sources

Arabidopsis

MAtDB (Schoof et al., 2004) contains both the original MIPS assembly of the Arabidopsis Genome Initiative (AGI) genome sequence (Arabidopsis Genome Initiative, 2000) as well as the TIGR version 5 assembly and annotation (Wortman et al., 2003). Manual curation of gene models and functional assignments, which partially continued at MIPS in parallel to the whole-genome reannotation performed by TIGR, is currently being mapped to TIGR version 5 to provide a merged annotation set. For now, both sets can be browsed independently. Through the European Union PlaNet project (Schoof et al., 2004), data from many European Arabidopsis resources are integrated through the use of BioMOBY Web services (Wilkinson and Links, 2002).

Medicago

The European Medicago and Legume Database (UrMeLDB; http://www.urmeldb.net) integrates data from both the international Medicago sequencing project (http://www.medicago.org/genome/) and the European Medicago project (http://www.eugrainlegumes.org; VandenBosch and Stacey, 2003). All publicly available Medicago genomic clone data, completed as well as unfinished, have been integrated into UrMeLDB. On a regular basis, newly available and updated clone sequences from the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL; http://www.ebi.ac.uk/embl) are retrieved and integrated into UrMeLDB. Currently (March 2005), 1,268 Medicago clone sequences are available, more than half of which can be mapped to a chromosome based on data from the University of California, Davis, Medicago physical map (http://mtgenome.ucdavis.edu).

Gene prediction and protein annotation of Medicago sequences are performed in collaboration with the International Medicago Genome Annotation Group (IMGAG; Cannon et al., 2005). MIPS provides genome sequences to Institut National de la Recherche Agronomique Toulouse (http://www.toulouse.inra.fr) for automated gene prediction using EuGène (Schiex et al., 2001). All bioinformatic analyses available in UrMeLDB rely on this gene call set. Updates of gene models will be recorded using the UrMeLDB versioning facilities. Thus, changes in annotation due to sequence reassembly or gene prediction improvement can be traced.

Automated annotation of Medicago sequences, including gene prediction, is also performed at other sites (VandenBosch and Stacey, 2003; Cannon et al., 2005). Given the current accuracy of gene prediction, multiple alternative predictions can be valuable for individual genes, allowing possible alternative models to be considered (e.g. in planning experiments or designing primers). However, for genome-wide analysis, a common and comparable set of gene calls will be valuable as a reference. Efforts are under way with several Medicago genome annotation centers within the IMGAG (see http://www.medicago.org/genome/IMGAG.php) to develop a common reference set of gene model annotations for Medicago. This will be made available through the Web portals of the groups involved.

Grass Genomes: Rice and Maize

Both Arabidopsis and Medicago are dicotyledonous model plants. They are essential for scientific research but not of agronomic importance. This is an important difference from two other prominent plant genomes, rice and maize. These grass genomes have both a long tradition as model plants and at the same time are crop plants of major importance for the world food supply. Besides, they are representatives of the monocotyledonous plants, which diverged from the dicotyledonous plants approximately 150 to 250 million years ago. Thus, all different levels of comparative analysis between representatives of the monocotyledonous and the dicotyledonous plants are of special interest. A consistent representation of these genomes is indispensable to perform such analysis in an easy and straightforward manner. However, in some ways the data types and research focus differ between the species, e.g. the analysis of repeat elements and transposons plays only a minor role in Arabidopsis but is a major focus in maize (Messing et al., 2004). Thus, a generic data model applicable to all must be flexible enough to encompass species-specific data without affecting comparability.

For rice, the MIPS *Oryza sativa* database (Karlowski et al., 2003) includes the publicly available rice subsp. japonica cv Nipponbare assembly and annotation as provided by TIGR. When the International Rice Genome Sequencing Project genome sequence and annotation becomes available, that will be included in parallel.

The maize database includes 100 publicly available BAC sequences with manually curated gene predictions. Repeat detection and classification was also enhanced by manual efforts. These data provide an insight into the structure and composition of the maize genome and provide a basis for comparative and combinatorial analysis (Messing et al., 2004).

Access and Web Query Interface

The aim of the Web interface to the MIPS plant genome resources, available at http://mips.gsf.de/projects/plants/, is to provide access to all included genomes in a common format and tools for cross-species comparisons.

To browse data, the user can navigate in a genome-oriented way. Assuming one would, for example, start from the chromosome list, all contigs anchored to each chromosome can be retrieved. A contig report contains detailed information on the entry as well as links to sequence, EMBL database records, a list of annotated genetic elements, or a graphical viewer. The genetic element list links to reports on the protein genes or other features on display. Sequences can be viewed and downloaded as Hypertext Markup Language (HTML; http://www.w3.org/MarkUp/), Extensible Markup Language (XML; http://www.w3.org/XML/), or FASTA format. For protein coding genes, unspliced, spliced (transcript), and coding DNA sequences are available as well as protein sequences. Moreover, cross-references in the reports allow easy access to entries in external databases associated with the entry.

Alternatively, complete lists of all sequenced contigs, all genetic elements, or all elements of a selected type are available for browsing. The tables displayed on a given page can be sorted and filtered by clicking on a column heading or table cell content, respectively. The latter restricts the view to all rows that contain the value that was clicked. Somewhat separately, lists of clones can be browsed, by chromosome if linkage data is available.

To visualize and browse genetic elements on a specified contig, a graphical interface, DBBrowser, was developed (Fig. 1A). DBBrowser uses the scaleable vector graphics (SVG) graphics format and thus allows seamless zooming of the image as well as full editing of the downloaded vector graphics file. The controls for zooming and moving the image depend on the plugin used. For users that do not have an SVG plugin for their browser, an applet is provided that displays the SVG. The SVG images can be downloaded and extensively edited as vector graphics.

Search options include search by name, free text, or sequence. The free-text search option allows inspection of the content of all text fields, and it is available for individual genomes or across all databases. BLAST is used as a homology search engine (Altschul et al., 1997). The target databases for similarity searches include clones (completed and unfinished), contigs, and genetic elements (e.g. coding sequences). Besides data sets from the plant projects at MIPS (Arabidopsis, Medicago, maize, rice), Swissprot/SWALL and plant-specific data sets selected from the EMBL nucleotide database (e.g. all Arabidopsis ESTs) are searchable.

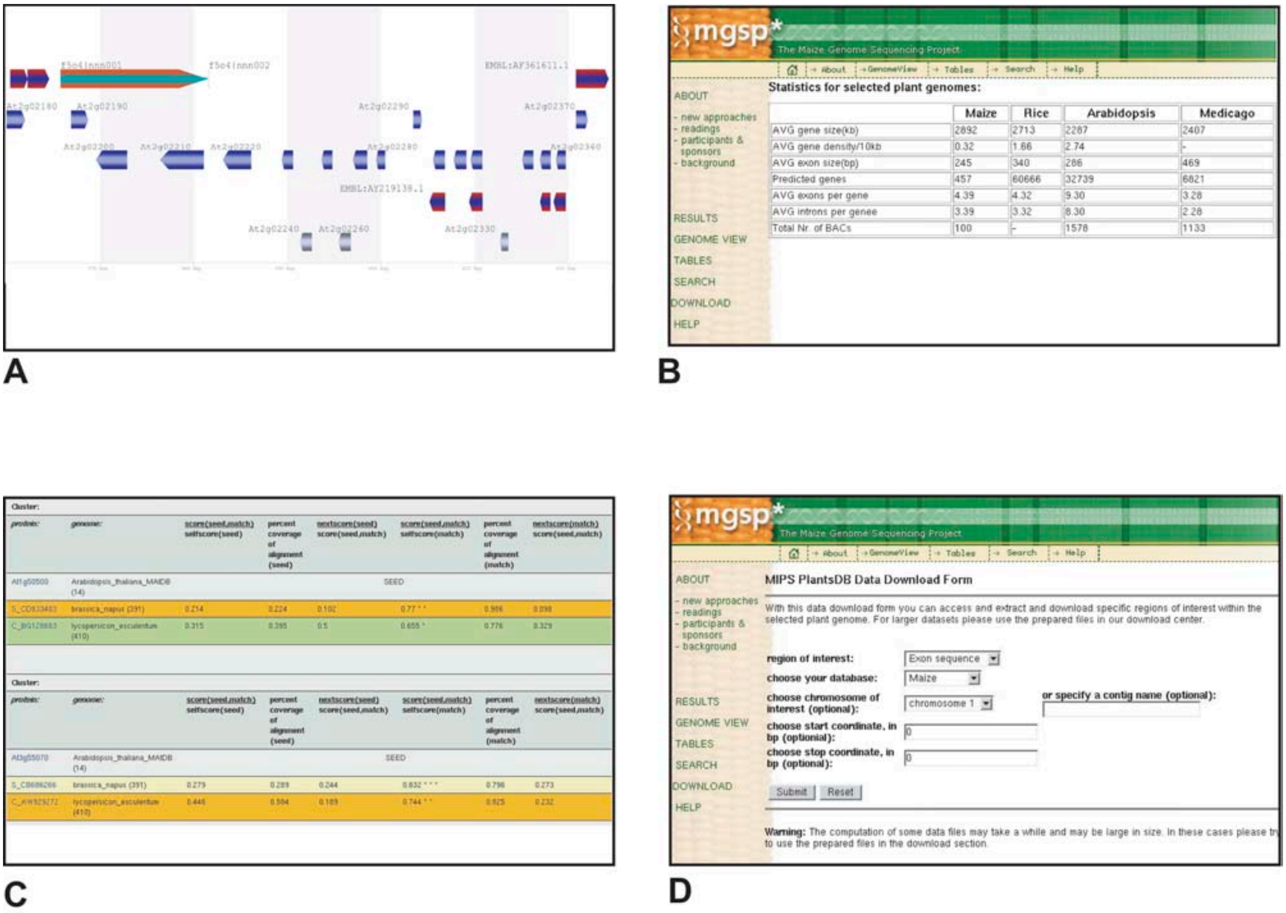
Finally, the download section provides ftp access to various data downloads. This includes FASTA-formatted sequence files for all clones/contigs and protein coding genes. Besides this, the download section contains functionality to create and download a Genome Annotation Markup Elements file (GAMEXML; http://xml.coverpages.org/game.html) for a specified contig and coordinate range. The GAMEXML format is used by the Apollo Genome Browser (Lewis et al., 2002). 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. Thus, downloading a GAMEXML file of the region of interest enables the user to inspect (and modify) all gene annotation data, thus building the user's 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 e-mail for inclusion in the database.

Web Services

Bioinformatics tools and databases are most commonly accessible through Web interfaces that allow navigation using a standard browser. While it is relatively easy to provide a human-readable presentation of data in this way, the data is not easily accessed from applications in order to be integrated with remote data sets. Solutions to this problem have been screen scraping, where the HTML code of Web pages was parsed in order to grab the information therein, or the import of bulk data dumps into local data warehouses to make the data available for integration. Maintaining current data is hard, as the HTML representation or data dump format may change, requiring changes to the parser, and as data needs to be reimported into the local warehouse for every update.

Recently, Web services have been suggested as a solution to the problem of data and analysis resource interoperability and data integration for biology and bioinformatics (Schoof et al., 2004; Wilkinson et al., 2005). We have decided to support BioMOBY, an emerging standard that builds on Web services and extends these with ontologies that describe data structure, semantics, and service types (Wilkinson and Links, 2002; Wilkinson et al., 2005). Through BioMOBY Web services, data from the plant genome resources at MIPS can be directly accessed from applications, allowing analyses to be performed remotely or included in the Web presentation 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.

Currently, 35 BioMOBY-compliant Web services provide retrieval (22) and analysis (13) functionality (see http://www.eu-plant-genome.net [select "Tools"]). These allow retrieval of protein and DNA sequences for keywords, EMBL accessions, or AGI locus codes (Schoof et al., 2002). The retrieved sequences can be



mgsp* The Maize Genome Sequencing Project

Statistics for selected plant genomes:

	Maize	Rice	Arabidopsis	Medicago
AVG gene size(kb)	2892	2713	2287	2407
AVG gene density/10kb	0.32	1.66	2.74	-
AVG exon size(bp)	245	340	286	469
Predicted genes	457	60666	32739	6821
AVG exons per gene	4.39	4.32	9.30	3.28
AVG introns per genee	3.39	3.32	8.30	2.28
Total Nr. of BACs	100	-	1578	1133

Cluster:

protein:	genome:	score(seed,match)
		selfscore(seed)
			percent coverage of alignment (seed)	nextscore(seed)
				score(seed,match)
					score(seed,match)
					selfscore(match)
						percent coverage of alignment (match)	nextscore(match)
							score(seed,match)

At1g50500	Arabidopsis_thaliana_MAtDB (14)				SEED		
S_CD833483	brassica_napus (391)	0.214	0.224	0.102	0.77 * *	0.986	0.098
C_BG128683	lycopersicon_esculentum (410)	0.315	0.395	0.5	0.655 *	0.776	0.329

Cluster:

protein:	genome:	score(seed,match)
		selfscore(seed)
			percent coverage of alignment (seed)	nextscore(seed)
				score(seed,match)
					score(seed,match)
					selfscore(match)
						percent coverage of alignment (match)	nextscore(match)
							score(seed,match)

At3g55070	Arabidopsis_thaliana_MAtDB (14)				SEED		
S_CB686266	brassica_napus (391)	0.279	0.289	0.244	0.832 * * *	0.796	0.273
C_AW929272	lycopersicon_esculentum (410)	0.446	0.584	0.189	0.744 * *	0.825	0.232

mgsp* The Maize Genome Sequencing Project

MIPS PlantsDB Data Download Form

With this data download form you can access and extract and download specific regions of interest within the selected plant genome. For larger datasets please use the prepared files in our download center.

region of interest:	Exon sequence ▾
choose your database:	Maize ▾
choose chromosome of interest (optional):	chromosome 1 ▾
or specify a contig name (optional):	
choose start coordinate, in bp (optionial):	0
choose stop coordinate, in bp (optional):	0

Warning: The computation of some data files may take a while and may be large in size. In these cases please try to use the prepared files in the download section.

Figure 1. Web interfaces to the MIPS plant genome resources. A, Graphical view of an Arabidopsis contig generated by DBBrowser, showing protein-coding genes in blue, cDNA-matches in blue and red, pseudogenes in grey, and transposons in green and red. B, Statistics overview of the four currently available plant genomes, including average gene size, gene density per 10 kb, exon size in base pairs, exons and introns per gene, as well as the current total of predicted genes and BAC contigs. C, A section from an ortholog list for Arabidopsis, *Brassica napus*, and tomato (*Lycopersicon esculentum*), showing two groups of tentative orthologs detected by the best bidirectional hit method (Rudd et al., 2005). It is available at http://www.eu-plant-genome.net under Search. The scores are, from left to right: the ratio score/selfscore(seed), the alignment score of the pair divided by the score of the alignment of the Arabidopsis sequence against itself. For a perfect match between two identical proteins, this is 1. It gives a measure of the conservedness of the sequences, i.e. high values reflect sequences that are more similar and more probably true orthologs. Ratios above 0.6 receive one asterisk, above 0.7 two, above 0.8 three, and very closely related sequences with a score/selfscore ratio above 0.9 four asterisks. The ratio nextscore(seed)/score is the score of the second best hit in Arabidopsis divided by the score of the best match pair. If there is a second match with the same alignment score, this ratio is 1. If there is no other match, it is 0. This reflects paralogs or closely related sequences, which can hinder the use as COS markers as they lead to ambiguous ortholog relationships or mapping results. The next column gives the ratio score/selfscore, using the selfscore of the alignment of the Brassica or tomato protein against itself. Especially for collections of partial protein sequences (in this case the Brassica and tomato proteins, which are derived from ESTs), the selfscore of the partial protein will be significantly lower than a full-length match. The last column gives the ratio nextscore(match)/score, using the second best hit in Brassica or tomato. The color reflects a relative quality measure; COS marker suitability increases from green to light yellow and is highest for sequence pairs marked in burned orange. D, This form allows the user to select customized sequence sets for download. The user selects a sequence region of interest, e.g. all exons, all introns, all untranslated regions, all noncoding RNAs, or some amount of sequence upstream of coding genes. Then, the genome of interest is selected. The retrieval can be restricted to a single chromosome or contig, and a range of coordinates thereon. On submission, a multiple FASTA file is generated and can be downloaded.

passed into BLAST services for comparison against selected sequence data sets from the MIPS plant databases. Annotation can be retrieved, e.g. all GeneticElement entries (see above) on a given Contig, or the coordinates of a given GeneticElement in General Feature Format (http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml). These services can be used either through a Web interface at http://www.eu-plant-genome.net (select "Search"), using stand-alone tools such as Taverna (Oinn et al., 2004), or by using the BioMOBY libraries to build custom applications.

These services implement a public application interface that provides consistent access to all plant genome resources over the Internet. On the one hand, these allow a programmer to build applications that retrieve data as required, e.g. retrieving genomic sequences, then iterating over all protein-coding genes and extracting the upstream sequences based on their coordinates. On the other hand, BioMOBY also provides tools with more user-friendly interfaces that allow point-and-click discovery of data (Wilkinson et al., 2005). As the services provided by the MIPS plant genome resources 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.

Analyses: Tools and Results

The first overview of a genome can be achieved through some basic statistics, like average gene length, exon number, gene density, or GC content. These are regularly calculated from the plant genome databases and made available in the Web interface (Fig. 1B).

An important tool for comparative genomics is the prediction of orthologs between genomes. We use a tool initially designed for the detection of putative conserved orthologous set (COS) markers (Fulton et al., 2002; Rudd et al., 2005) to extract possible orthologs from the Similarity Matrix of Proteins (SIMAP; http://mips.gsf.de/proj/simap/) database of all-against-all protein similarity searches (Mewes et al., 2004; Güldener et al., 2005). Basically, best bidirectional FASTA matches are selected as tentative orthologs, with a filter to select against proteins that have a close paralog in any species. The resulting lists of predicted orthologs can be browsed, or only orthologous sets with a representative in each of a selected set of species selected (Fig. 1C). These ortholog tables will form the basis for displaying information transferred from putative orthologs in more fully annotated species, e.g. the Medicago protein reports could include functional annotation from predicted orthologous proteins in Arabidopsis.

The consistent application interface to all plant databases at MIPS facilitates the implementation of sophisticated query tools. We have set up a sequence export tool that allows users to download specific sequence data sets such as all first introns of all protein-coding genes on a selected contig, or a selected number of base pairs upstream of all start codons in a genome (Fig. 1D). This will soon also be available as a Web service.

USAGE EXAMPLES

The Web service interface to the MIPS plant genome resources provides a versatile and powerful, yet easy to use, access for remote users, enabling them to create their own analyses. An example workflow that can be realized in this way would retrieve putative Gene-Ontology (GO; Ashburner et al., 2000) terms for all the genes on a Medicago contig (Fig. 2). It can be executed using either a step-by-step procedure with the Gbrowse form-based BioMOBY client available at http://www.eu-plant-genome.net, or Taverna for one-click execution (see supplemental data for a Taverna XML workflow definition file). It starts by retrieving a Medicago contig using a free-text search with a BAC clone name. Then the MIPS plant databases are queried for all GeneticElements on this contig, and the protein sequence for each is retrieved. The latter is used to execute a BioMOBY BLAST search against the Arabidopsis proteome. Another BioMOBY service extracts the AGI LocusCodes (Schoof et al., 2002) of matching proteins, which serve as inputs to the AGI-LocusCode-to-GO BioMOBY service provided by the *Arabidopsis thaliana* Insertion Database (Pan et al., 2003) to retrieve putative GO terms.

This can be extended to a whole-genome comparative analysis by retrieving GO term annotation for all Arabidopsis genes with a putative ortholog in both Medicago and rice, compared to all Arabidopsis genes with no detected ortholog in Medicago or rice. This workflow is available as supplemental data or from the MIPS Web site. It first retrieves all identifiers for all protein-coding genetic elements in the MIPS Arabidopsis database. Then, it queries a service that returns putative orthologs between Arabidopsis, Medicago, and rice (based on best bidirectional hits extracted from the MIPS SIMAP database; Güldener et al., 2005). Extraction of the Arabidopsis identifiers only from the output yields Arabidopsis proteins with orthologs in both Medicago and rice. A difference set operator available as a local processor within Taverna is used to compare the list of all Arabidopsis identifiers to this list, returning Arabidopsis proteins with no ortholog detected in both Medicago and rice. Both lists of AGI locus codes are used to independently query the *Arabidopsis thaliana* Insertion Database GO BioMOBY service, returning all GO terms annotated to these proteins. Up to this point, the workflow can be built by point-and-click using existing processors and services, and the result could be saved as an Excel file for further analysis. In the example provided, some lines of script code were added (Taverna provides BeanShell processors for this task) that count, for each GO term in the list, how often this term occurs in the input data. Thus, the number of proteins annotated with each GO term is returned, both in Arabidopsis proteins with putative orthologs and, for comparison, in Arabidopsis proteins with no ortholog detected in both Medicago and rice.

The result is provided as an Excel file in the supplemental data; the most frequent GO terms are, unsurprisingly, GO:0000004 biological_process unknown and GO:0008372 cellular_component unknown. However, both these terms are underrepresented in the set of Arabidopsis proteins with putative orthologs, suggesting that conserved proteins are more likely to

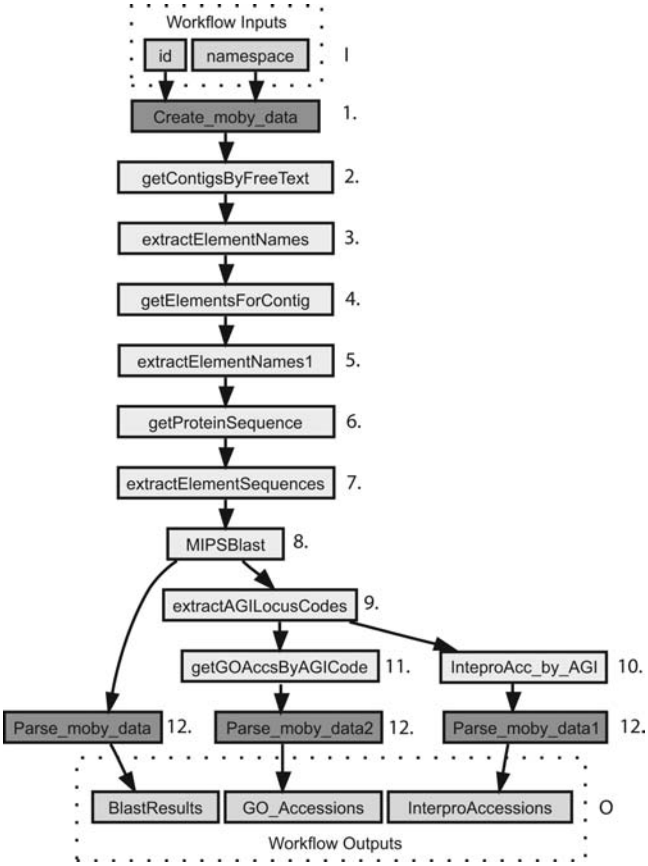


Figure 2. Annotation by similarity to Arabidopsis. This example workflow generated in Taverna retrieves putative GO terms and Interpro accessions for all genes on a Medicago contig by finding similar Arabidopsis genes using BLAST and retrieving the GO and Interpro accessions associated with these. To try this workflow, download Taverna at http://taverna.sourceforge.net and load the XML workflow description provided as supplemental data or on the MIPS Web site. The workflow can also be executed step by step using the Gbrowse Web client at http://mips.gsf.de/cgi-bin/proj/planet/gbrowse/gbrowse_moby and selecting the services as listed in the figure. Example input values are namespace = "MIPS_Contig_Medicago" and id = "mth2-4o4." Services and input/output objects used in this Workflow: I Inputs: namespace, id. The namespace defines which database is searched in, e.g. MIPS_Contig_Medicago for the Medicago contig sequence database. The id is the name of the contig to search for. 1, Create_moby_data creates the MOBY object required as input to BioMOBY services from the input strings. 2, getContigsByFreeText does a free-text lookup, in case the contig name was not entered precisely, e.g. with wrong case. 3, extractElementNames returns a BioMOBY object with only the name. 4, getElementsForContig retrieves all GeneticElements (e.g. protein-coding genes) located on the given contig. 5, extract ElementNames (see 3). 6, getProteinSequences. 7, extractElementSequences transforms the XML returned by getProteinSequences into BioMOBY GenericSequence objects as consumed by the MIPSBlast service. 8, MIPSBlast executes a BLAST search against the MAtDB Arabidopsis protein database. 9, extractAGILocusCodes parses only the AGI locus codes of matches in the BLAST output. 10, InterproAcc_by_AGI retrieves the Interpro accessions assigned to Arabidopsis proteins. 11, getGOAccsByAGICode retrieves the GO terms assigned to Arabidopsis proteins. 12, Parse_moby_data makes the MOBY output human readable. O, Outputs: InterproAccessions, GO_Accessions, BlastResults. The services used in this workflow are made available by Nottingham Arabidopsis Stock Centre (http://arabidopsis.info), John Innes Centre (http://www.jic.ac.uk), and MIPS (mips.gsf.de). See http://www.eu-plant-genome.net for details.

be annotated with functional or localization information.

Executing this workflow requires several hours and 1 GB of RAM on a standard workstation, as tens of thousands of Web service calls are executed, returning 89,444 GO term-protein associations. This may seem tedious to anyone accustomed to answering such questions using, for example, SQL queries on a data warehouse. However, Taverna and the Web service architecture are fully capable of this kind of analysis and allow enormous flexibility in the generation of queries, without the need for warehoused data or knowledge of SQL. Instead, distributed data sources can be combined on the fly. Instead of using precalculated orthologs, a similar workflow could incorporate a BLAST service to calculate similarities on the fly, thus allowing a user to start with his or her own set of proteins and building a workflow to discover which of these have homologs in the MIPS plant resources.

DISCUSSION

Plant genome and associated data have revolutionized plant research during the last years. The individual model plant research communities have benefited greatly from appropriate storage, communication, and display of genome data, which are a prerequisite for the sustainable maintenance of genome annotation and an indispensable informational infrastructure. And the results from genome research and genome-scale analyses have had significant impact. Even from partial genomes such as collections of BESs, detailed insights into the composition and characteristics of particular genomes can be gained (Palmer et al, 2003; Whitelaw et al., 2003; Messing et al., 2004). In addition, the potential of combinatorial analysis of heterogeneous data sets such as fingerprinted contig data and BES resulting in an ordered genomic alignment of sequence and gene tags has been demonstrated recently (Messing et al., 2004). Highly automated, multilayered analytical pipelines for detailed analysis and annotation have become accepted necessities for plant genome analysis. Continuous further development of such procedures and the appropriate feed-in/feed-out into data management systems already have a long history at MIPS (Mayer et al., 1999; Arabidopsis Genome Initiative, 2000; Schoof et al., 2004) and are indispensable for plant genome analysis (Haberer et al., 2004; G. Haberer, S. Young, A. Bharti, H. Gundlach, C. Raymond, G. Fuks, E. Butler, R.A. Wing, S. Rounsley, C. Nusbaum, B. Birren, K.F.X. Mayer, and J. Messing, unpublished data). While these analyses have so far required local access to the database or the creation of a data warehouse for the specific task, more and more these will be possible remotely. This is an essential step, as maintaining and integrating databases for multiple species locally becomes increasingly tedious (Wilkinson et al., 2005). Remote access through Web services can ensure availability of current data while providing the flexibility for novel queries and combinatorial analyses. Besides, with a generic framework in place, any analysis can be easily rerun for every new species made available within the framework.

This comparative aspect is essential to unleash the full potential of mining sequence data. Comparative analyses have been demonstrated to be extremely powerful in yeasts and vertebrates (Mouse Genome Sequencing Consortium, 2002; Kellis et al., 2003; Bejerano et al., 2004; International Chicken Genome Sequencing Consortium, 2004; Rat Genome Sequencing Project Consortium, 2004) and genome-scale comparative analyses are emerging in plants as well (Goff et al., 2002; Bonnet et al., 2004). While many of these analyses will remain too computationally demanding to be executed remotely, a generic framework for management of data from multiple species is essential. In our experience, this framework again simplifies the task of making data available through Web services, thus bringing the power of comparative analyses to remote users, even if with limited performance.

CONCLUSION

We will continue to work toward a consistent view of the model plant genomes, which entails intense collaboration with international partners to synchronize our efforts with other plant genome initiatives, databases, and analysis resources. To this end, BioMOBY-compliant Web services are increasingly helpful. At the same time, they enable a completely new user experience, providing a remote application interface that can be harnessed by tools such as Taverna to bring comparative analyses to remote users without the need for programming or warehousing. While current services focus on sequence, gene structure, and function annotation, integration of modules to handle expression data from microarray experiments is expected to add significantly to the value of the resource. Further plans are to enable browsing syntenic regions and viewing predicted orthologs or homologs in any gene report. To this end, the integration with existing MIPS systems like SIMAP (Mewes et al., 2004) and Pedant (Riley et al., 2005) will be extended.

Plant genome databases will have to evolve with our increasing knowledge of how genomes work. But at the same time, well-structured integrated knowledge resources and new query and access interfaces can facilitate research and uncover some of their secrets.

ACKNOWLEDGMENTS

Several students contributed code and programming effort. We thank Arthur Zimek for work on DBBrowser, Mirjam Maier and Peter Kral for work on the sequence export, Elisabeth Wischnitzki and Patrick Tischler for work on the COS markers, and Daniela Foelsl for work on the BLAST BioMOBY services. Thomas Rattei has helped with SIMAP access. Tobias Hindemitt assisted with graphics and Web design. We thank all members of the BioMOBY community, the IMGAG, and the PlaNet project for cooperation and helpful discussions on data exchange and standards.

Received January 5, 2005; revised April 6, 2005; accepted May 3, 2005; published July 11, 2005.

LITERATURE CITED

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796–815

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:** 25–29

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science **304:** 1321–1325

Bonnet E, Wuyts J, Rouze P, Van de Peer Y (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. Proc Natl Acad Sci USA **101:** 11511–11516

Cannon SB, Crow JA, Heuer ML, Wang X, Cannon EKS, Dwan C, Lamblin AF, 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

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell **14:** 1457–1467

Goff SA, Ricke D, Lan TH, 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

Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martínez J, Pérez-Ortín JE, et al (2005) CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res (Database issue) **33:** D364–D368

Haberer G, Hindemitt T, Meyers BC, Mayer KF (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. Plant Physiol **136:** 3009–3022

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432:** 695–716

Karlowski WM, Schoof H, Janakiraman V, Stuempflen V, Mayer KFX (2003) MOsDB: an integrated information resource for rice genomics. Nucleic Acids Res **31:** 190–192

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:** 241–254

Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V (2004) MaizeGDB, the community database for maize genetics and genomics. Nucleic Acids Res **32:** 393–397

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Ricter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, et al (2002) Apollo: a sequence annotation editor. Genome Biology **3:** RESEARCH0082

Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N, et al (1999) Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature **402:** 769–777

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, et al (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA **101:** 14349–14354

Mewes HW, 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:** 41–44

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520–562

Oinn T, Addis M, Ferris J, Marvin D, Greenwood M, Carver T, Pocock MR, Wipat A, Li P (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics **20:** 3045–3054

Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. Science **302:** 2115–2117

Pan X, Liu H, Clarke J, Jones J, Bevan M, Stein L (2003) ATIDB: Arabidopsis thaliana insertion database. Nucleic Acids Res **31:** 1245–1251

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428:** 493–521

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev **16:** 1616–1626

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, et al (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res **31:** 224–228

Riley ML, Schmidt T, Wagner C, Mewes HW, Frishman D (2005) The PEDANT genome database in 2005. Nucleic Acids Res **33:** D308–D310

Rudd S, Schoof H, Mayer KFX (2005) PlantMarkers: a database of predicted molecular markers from plants. Nucleic Acids Res **33:** D628–D632

Schiex T, Moisan A, Rouzé P (2001) EuGène: an eukaryotic gene finder that combines several sources of evidences. *In* O Gascuel, M-F Sagot, eds, Computational Biology. LNCS 2066. Springer-Verlag, Heidelberg, pp 111–125

Schoof H (2003) Towards interoperability in genome databases: the MAtDB (MIPS Arabidopsis thaliana database) experience. Comp Funct Genomics **4:** 255–258

Schoof H, Ernst R, Mayer KFX (2004) The PlaNet consortium: a network of European plant databases connecting plant genome data in an integrated biological knowledge resource. Comp Funct Genomics **5:** 184–189

Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KFX (2002) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res **30:** 91–93

VandenBosch KA, Stacey G (2003) Summaries of legume genomics projects from around the globe: community resources for crops and models. Plant Physiol **131:** 840–865

Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, et al (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Res **30:** 103–105

Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, et al (2003) Enrichment of gene-coding sequences in maize by genome filtration. Science **302:** 2118–2120

Wilkinson MD, Schoof H, Ernst R, Haase D (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case. Plant Physiol **138:** 5–17

Wilkinson MD, Links M (2002) BioMOBY: an open source biological web services proposal. Brief Bioinform **3:** 331–341

Wortman JR, Haas BJ, Hannick LI, Smith RK Jr, Maiti R, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, et al (2003) Annotation of the Arabidopsis genome. Plant Physiol **132:** 461–468

Yazaki J, Kojima K, Suzuki K, Kishimoto N, Kikuchi S (2004) The Rice PIPELINE: a unification tool for plant functional genomics. Nucleic Acids Res **32:** 383–387