The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant

Eva Huala[1,*], Allan W. Dickerman[2], Margarita Garcia-Hernandez[1], Danforth Weems[2], Leonore Reiser[1], Frank LaFond[2], David Hanley[2], Donald Kiphart[2], Mingzhe Zhuang[2], Wen Huang[1,2], Lukas A. Mueller[1], Debika Bhattacharyya[1], Devaki Bhaya[1], Bruno W. Sobral[2], William Beavis[2], David W. Meinke[3], Christopher D. Town[4], Chris Somerville[1] and Seung Yon Rhee[1]

[1]Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA, [2]National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA, [3]Department of Botany, Oklahoma State University, Stillwater, OK 74078, USA and [4]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received September 7, 2000; Revised and Accepted October 25, 2000

ABSTRACT

***Arabidopsis thaliana*, a small annual plant belonging to the mustard family, is the subject of study by an estimated 7000 researchers around the world. In addition to the large body of genetic, physiological and biochemical data gathered for this plant, it will be the first higher plant genome to be completely sequenced, with completion expected at the end of the year 2000. The sequencing effort has been coordinated by an international collaboration, the Arabidopsis Genome Initiative (AGI). The rationale for intensive investigation of *Arabidopsis* is that it is an excellent model for higher plants. In order to maximize use of the knowledge gained about this plant, there is a need for a comprehensive database and information retrieval and analysis system that will provide user-friendly access to *Arabidopsis* information. This paper describes the initial steps we have taken toward realizing these goals in a project called The Arabidopsis Information Resource (TAIR) (www.arabidopsis.org).**

INTRODUCTION

Several decades of research into the biology of *Arabidopsis thaliana* has yielded a wealth of genetic, physiological and biochemical information (1). With the completion and full annotation of the *Arabidopsis* genome sequence due by the end of the year 2000, the need for an excellent, comprehensive database for *Arabidopsis* information has become critical. The goal of TAIR is to provide a database that serves not only the needs of the *Arabidopsis* community but the biological research community as a whole, which requires easy access to *Arabidopsis* information to make maximum use of this model plant to solve research problems in other organisms, including economically important plant species.

The challenge for the TAIR database is to provide users with the means to efficiently and intuitively query, browse, graphically visualize and download a variety of complex data types including information about genes, clones, sequences, markers, mutants, seed stocks, members of the research community and research papers. In addition, the TAIR curators must be able to maintain data integrity by associating data with researchers, references and methods whenever possible, and continuously update existing data and adding new data types as they become available. Legacy data from the previous Arabidopsis database, AtDB (2), had to be accommodated and the transition from AtDB to TAIR made with no interruption of service.

WEB NAVIGATION STRUCTURE

All the tools developed through our project from the TAIR home page (http://www.arabidopsis.org/home.html). The web site that overlays the database was designed to be simple, portable and efficient. We implemented the web site mainly in HTML to ensure uniform functionality regardless of the hardware and software configurations of our users. The depth of the

*To whom correspondence should be addressed. Tel: +1 650 325 1521; Fax: +1 650 325 6857; Email: huala@acoma.stanford.edu
Present addresses:
Allan W. Dickerman, Virginia Bioinformatics Institute (0477), 1750 Kraft Drive, Corporate Research Center Building 10, Suite 1400, Blacksburg, VA 24061, USA
Donald Kiphart, Prediction Company, 236 Montezuma Avenue, Santa Fe, NM 87501, USA
Mingzhe Zhuang, Sugen, Inc., 230 East Grand Avenue, South San Francisco, CA 94080-4811, USA

web site was generally kept to three levels so that users would have to access no more than three pages to get to the information of interest. In addition, a navigation tool bar containing site search and help links and a footer containing a contact email address and information about the last update were included on all pages.

The website is divided into six major sections: TAIR DB (http://www.arabidopsis.org/search/), Tools (http://www.arabidopsis.org/tools/), Arabidopsis Information (http://www.arabidopsis.org/info/), News (http://www.arabidopsis.org/news/), External Links (http://www.arabidopsis.org/links/) and FTP directory (ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/). Documentation about our project and the organization of our web site can be found on About TAIR (http://www.arabidopsis.org/about/). All of these major sections are a part of the navigation tool bar.

MAP VIEWER

TAIR's comprehensive MapViewer (http://www.arabidopsis.org/servlets/mapper) is an integrated visualization tool for viewing genetic, physical and sequence maps for each *Arabidopsis* chromosome (Fig. 1). It allows users to search, browse, align, zoom, scroll and print maps and mapped objects in TAIR's database. Maps can be aligned by searching for a shared marker or clone, by entering the desired coordinates for each map or by scrolling. A control panel at the top allows all open maps to be scrolled, zoomed and searched together. Individual controls for each map on the left provide the same functions for a single map, and a clickable chromosome bar for each map shows the current location on the chromosome of the map view and allows easy access to other regions of the chromosome. Each entity on the maps is hyperlinked to an output page from the database, which displays all the information about this entity including associations to other data types, attribution, history and comments. There is an extensive help page on how to use the MapViewer, from interpretation of the data to navigation of the tool (http://arabidopsis.org/mapViewer/help/tairmapa.htm).

DATABASE QUERY/BROWSE INTERFACE

The TAIR Database Search page is the entry point for searching the major classes of data housed in TAIR. The current version allows searching for clone, marker and gene information with implementation of community, reference and sequence searching planned for the coming year. Currently the database houses information on over 25 000 genes, 20 600 clones, 2144 markers, 7000 researchers and 10 000 references. The search page provides two main search options for the user: a general search which queries many different data types and a specific search which searches only a single data type but allows the user to customize the search. Options for customization

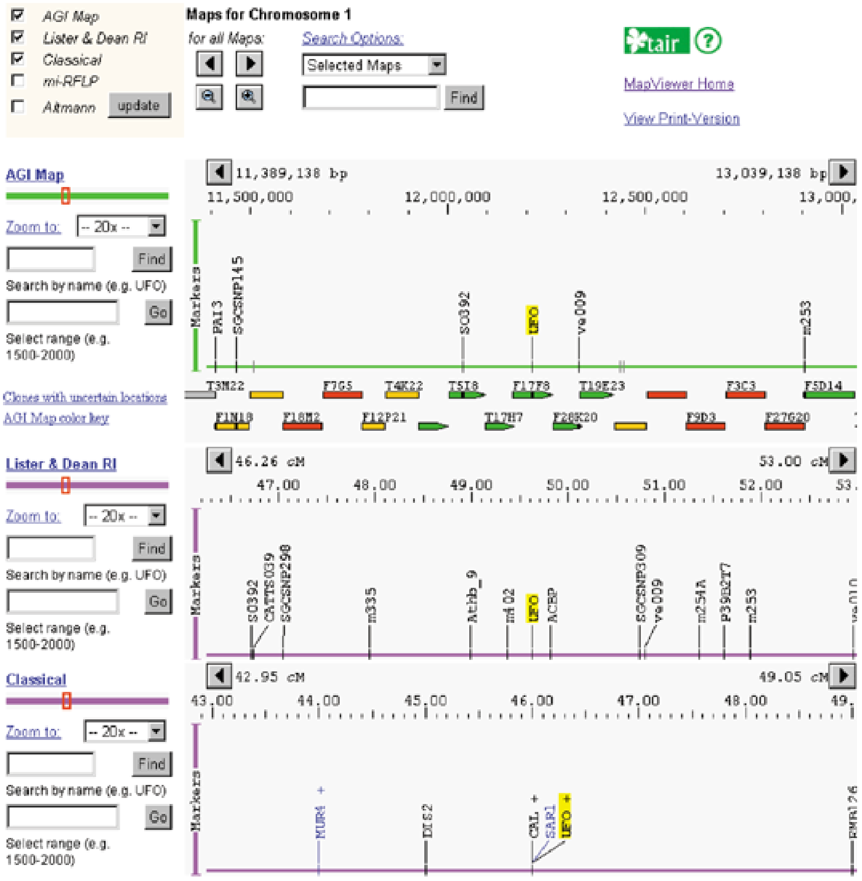

Figure 1. MapViewer showing AGI, RI and Classical Genetic maps for chromosome 1 following a search for the UFO marker (highlighted in yellow).

by feature include restriction of a clone search to only those clones with a certain vector type, or clones which are cDNAs, have end sequences, are fully sequenced, or have been used to make a genetic marker. Marker searches may be restricted to a certain class of markers, such as CAPS markers or all PCR-based markers, or limited to those which show a polymorphism between a chosen pair of ecotypes. Gene searches can be limited to those genes which have a predicted structure, have been cloned or sequenced, or can be found on a map. In addition to restricting searches by feature, all three advanced search pages provide the option of restricting the search by map, chromosome and location, or specifying a range of locations.

DATA DETAIL PAGES

Search results are presented on a summary page, which lists all results of the search and can be used to access a data detail page for each object, download data, or to view the object's map position using the TAIR MapViewer. The detail page presents a comprehensive summary of all data associated to the chosen object in the TAIR database, in addition to links to associated objects. For clones the detail page includes information on clone-ends, vector type, and associated accession numbers hot-linked to the sequence record. For markers, details shown include aliases, type, length, associated phenotype or digest pattern, special conditions, primer sequences and map positions. Gene information includes ORF name, product name and description, associated clones and sequences, and other data. All detail pages include aliases, associated sequence information and attribution of the information to a community member.

DATABASE STRUCTURE, DESIGN AND IMPLEMENTATION

The TAIR database is intended to store all types of biological data for *Arabidopsis* plus the metadata needed to attribute the data to the individual scientists and publications. The data model is built around a variety of data types, including clones, genes, sequences, genetic markers, polymorphisms and transcripts that inherit attributes from a fundamental TairObject class. The basic structure of the database, shown in Figure 2, links the TairObject class to annotation (function, map position, expression, etc.), and attribution (source of data, update history and references). Diagrams that illustrate this structure can be found at (http://www.arabidopsis.org/search/schemas.html).

To date, the best-elaborated data types describe the structural genomic components such as chromosomes, clones, sequences, markers and genes. These data types are broadly unified as being features of a chromosome, sharing properties such as length, location (both absolute and relative to other elements in the same linear space), and in many cases a nucleotide sequence. These properties are manifested by all objects that inherit attributes from the MapElement class, a subclass of TairObject that includes markers, clones and other discrete biological entities. We have adapted the model of genomic maps from the Object Management Group (http://cgi.omg.org/cgi-bin/doc?dtc/99-12-01) to represent the relationship of one MapElement being located on an encompassing, larger MapElement, which includes the possibility of nested maps. In our model any map element (e.g., clone, sequence, gene) can potentially be a map as well as being positioned on a higher-level map.

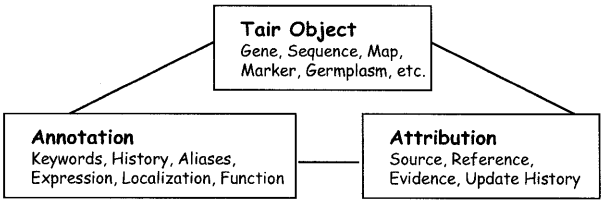
Tair Object
Gene, Sequence, Map, Marker, Germplasm, etc.
Annotation
Keywords, History, Aliases, Expression, Localization, Function
Attribution
Source, Reference, Evidence, Update History

Figure 2. The high-level organization of data in the TAIR database.

We made an early design decision to use an object-oriented (OO) approach to data representation. The OO approach, with subclasses inheriting data fields of the parent classes, is implemented in a relational database (Sybase) using a series of parent and child tables. A parent table, for example TairObject, contains or is linked to generic information, and includes a type field that indicates the subtype that the particular record belongs to. This subtype is taken to indicate which subtable, out of a defined set of options, contains the lower-level details of the object plus a primary-key index into the parent table. Thus, a particular clone in TAIR will have its information distributed in one row in each of the Clone, MapElement and TairObject tables, plus rows of additional tables that link to each of these main tables. One benefit of this design is the standardization of data relationships. The superclass TairObject provides a reliable foundation that can be counted on to provide the links to contributing scientists, literature references, etc. regardless of what kind of data one is manipulating. Both TAIR personnel and TAIR users benefit from the constancy of these generic features that exist across many data types. This also simplifies code development by reusing generic methods to retrieve, store, modify and display these aspects of the base classes. The OO design also allows elaboration of the database schema by extending the existing TairObject base class and allowing it to inherit generic associations to attribution and annotation, avoiding the need to re-implement them.

SOFTWARE DESIGN AND IMPLEMENTATION

The object-oriented design of the database integrates very naturally with Java, the object-oriented programming language used for the MapViewer and the TairObject report generator. Both these programs run as servlets, which are Java programs running on the server. The Apache web server forwards HTTP requests to these servlets, which process the requests and send appropriate HTML and graphics back out over the Internet. The Java servlet program runs continuously rather than restarting for each HTTP request as a typical CGI program would. The map viewer exploits this by doing the time-consuming process of reading map data into memory at start-up and by preserving a user's state across HTTP requests. The TairObject report generator is designed to take a request that specifies a particular TairObject, such as a clone specified by name or numerical identifier, reading a limited network of data surrounding that object from the database, and formatting it as HTML with hyperlinks to connected data. The TairObject servlet is invoked upon clicking on a data element in the map viewer and as the second step of the database query interface. Perl is used for many CGI programs at TAIR such as the initial stages of the database queries and the BLAST and FASTA searches.

FUTURE PLANS

In the coming year, we will reiterate the process of database structure and user interface development to enhance the data content and functionality. The major data content enhancement will come from elaboration of the genome annotation and incorporation of genetic mapping data, stock (germplasm and DNA) data from the Arabidopsis Biological Resource Center (ABRC), and gene expression data from microarray and gene chip experiments. We are also collaborating with the Gene Ontology Consortium (http://www.geneontology.org) and other groups to develop controlled vocabularies for annotating plant genes using a consistent set of terms to facilitate cross-species comparisons.

ACKNOWLEDGEMENTS

TAIR is supported by NSF Grant DBI-9978564. This is Carnegie Institution of Washington Department of Plant Biology publication 1460.

REFERENCES

1. Meinke,D.W., Cherry,J.M., Dean,C., Rounsley,S.D. and Koornneef,M. (1998) *Arabidopsis thaliana*: a model plant for genome analysis. *Science*, **282**, 679–682.
2. Flanders,D.J., Weng,S., Petel,F.X. and Cherry,J.M. (1998) AtDB, the Arabidopsis thaliana database, and graphical-web-display of progress by the Arabidopsis Genome Initiative. *Nucleic Acids Res.*, **26**, 80–84.