Bioinformatic Resources, Challenges, and Opportunities Using Arabidopsis As a Model Organism in a Post-Genomic Era[1]

Seung Yon Rhee*

Department of Plant Biology, 260 Panama Street, Carnegie Institution of Washington, Stanford, California 94305

Arabidopsis, a small annual plant belonging to the mustard family, is the subject of study for an estimated 7,000 researchers around the world. At the end of the year 2000, Arabidopsis will be the first higher plant genome completely sequenced by an international public effort, the Arabidopsis Genome Initiative (AGI; AGI et al., 2000). Similar collaborative efforts are under way to undertake a systematic approach to address biological problems in this organism. For the research community to benefit from the vast pool of information, there is a need for comprehensive databases, information retrieval and analysis systems that are easily accessible to researchers. Accessibility is not only the ability to search for information of interest in sophisticated ways, but also the ability to verify the quality and sources of data. In this paper, I will address some of the goals and challenges in the field of bioinformatics and describe one example of the currently available bioinformatic resources and opportunities for this model plant in detail, along with a few other resources.

OBJECTIVES AND CHALLENGES IN BIOINFORMATICS

An organism is a complex system. The reductionist approach taken in biology has driven us to a point where we can recreate one of the simplest life forms (virus) and decode the genetic makeup of one of the most complex life forms (human). Many believe that sequencing the genome of an organism is the first step in uncovering the rules that underlie development and behavior. Between the raw sequence data and the rules that govern a life form lies a gap not unlike that between a string of alphabetical characters and a great literary work like *War and Peace*. Many different approaches will be taken to characterize the structure and function of gene products, the relationships and interactions of gene products in a process, the relationships and interactions of processes, and the interactions between the organism and its abiotic and biotic environment. Some of these approaches are described in accompanying articles in this issue.

The objective in bioinformatics is to develop methods and tools to generate hypotheses from data obtained by a variety of approaches. There are many challenges that must be overcome to achieve this objective: (a) the increasing complexity of data and their associations to one another, (b) the diversity in the sources and formats in which data exist, (c) the variability of conditions and descriptions of experimentation, and (d) the quality of experimental evidence for data. Developing ways to meet these challenges must combine expertise not only in biology, but also in computer science, math, and engineering. This multidisciplinary approach is not limited to combining resources in different disciplines, but also includes training people who can combine studies in these disciplines to address complex problems in biology.

CURRENT BIOINFORMATICS RESOURCES FOR ARABIDOPSIS

There are many bioinformatics resources around the world that allow researchers to access and analyze the large amounts of genetic, genomic, and biological data through the World Wide Web. This paper is not an appropriate place to provide a broad survey of all these resources, and I will focus on describing the Arabidopsis Information Resource (TAIR). I will also briefly discuss other resources such as the Arabidopsis Genome Resource (AGR) and a few genome annotation databases. Table I provides addresses for these and other relevant web sites.

TAIR (WWW.ARABIDOPSIS.ORG)

The TAIR project was initiated last year and its goal is to provide comprehensive information about Arabidopsis in an industry standard relational database and provide many avenues of querying, browsing, and analyzing Arabidopsis data.

The basic structure of high-level data in TAIR's database centers around data objects, attribution, and

[1] S.Y.R. is supported by the National Science Foundation (grant no. DBI–9978564). This is Carnegie Institution of Washington Department of Plant Biology Publication 1,466.

* E-mail rhee@acoma.stanford.edu; fax 650–325–6857.

Table I. *Uniform resource links for bioinformatic resources in Arabidopsis*

Resource	Address
TAIR	www.arabidopsis.org
AGR	ukcrop.net/agr/
Munich Information Center for Protein Sequences Arabidopsis Database	www.mips.biochem.mpg.de/proj/thal/
The Institute for Genome Research Arabidopsis Annotation Database	www.tigr.org/tdb/ath1/htmls/ath1.html
Kazusa Arabidopsis Opening Site	www.kazusa.or.jp/kaos/
Database of Arabidopsis Annotation	luggagefast.stanford.edu/group/arabprotein/
Arabidopsis Repeated Sequence Database	nucleus.cshl.org/protarab/AtRepBase.htm

annotation (www.arabidopsis.org/search/schemas.html). Data objects (TAIR_Objects) currently include map elements (clones, genes, sequences, genetic markers, polymorphims, transcripts, etc.). A map element is any object that can have a position assigned on another object. Annotation refers to descriptions of data objects, such as position, function, or expression. Attribution contains information about the source (people, organizations, and references) of the data and any changes to the data (history) as well as comments made by TAIR curators and personal communications from individual researchers. References include journal articles, web sites, databases, and software analysis

The TAIR database search page (www.arabidopsis.org/search/) provides an entry point for searching the major classes of data housed in TAIR. The current version permits searching clone, genetic marker, and gene information. Users can perform a general search by name that queries many different data types, or a specific search that searches only a single data type but allows the user to refine the search. Options for customized searching for clones include limiting by vector types clones that are cDNAs, have end sequences, are fully sequenced, and/or have been used to make a genetic marker. Markers can be searched by limiting to a certain class of markers, such as cleaved-amplified polymorpic sequences markers or

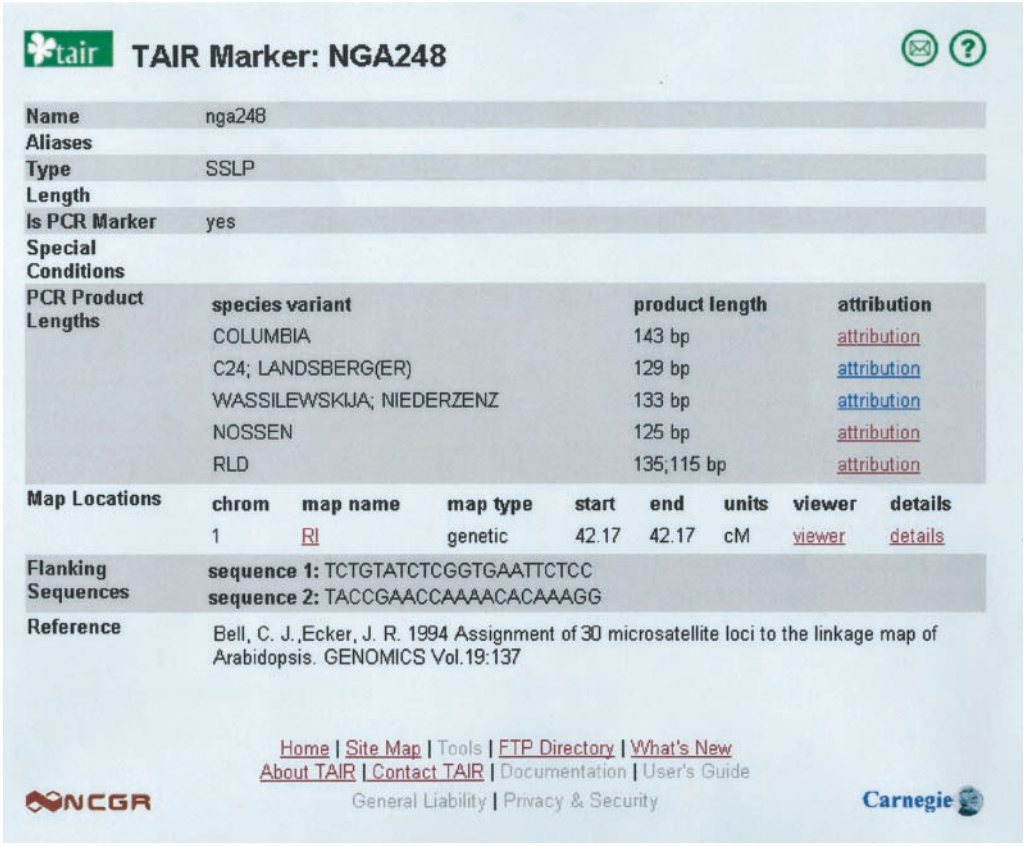

Figure 1. A detail page of genetic marker nga248. The detail page for markers shows all the information about a marker in TAIR, including name, aliases, the ecotypes that give rise to polymorphism, the type and nature of polymorphism (fragment lengths in this case), primer sequences, references, and attributions. (http://arabidopsis.org/search/).

all PCR-based markers, and/or to those that show a polymorphism between a chosen pair of ecotypes. Genes can be searched by open reading frame name, gene symbol, full name, or product name. In addition, gene searching can be limited to those genes whose structures have been experimentally determined, cloned, and/or sequenced. In addition to restricting searches by these features, all three customizable search pages provide the option of restricting the search by map and chromosome, or specifying a range of locations on a chromosome. In the future, more sophisticated queries across different data types will be available to allow more complex and flexible queries across the entire database.

Search results are summarized on a middle page, which is a one-line description of the matching results. Users can access a detail page for each entry by clicking on the name, or view the entry's map location using the TAIR MapViewer. The detail page presents a comprehensive summary of all data associated to the chosen object in the TAIR database. For clones, this includes clone-ends, vector type, and associated sequence accession numbers hot-linked to the sequence record. For markers, details shown include type, length, associated phenotype or digest pattern, special conditions, primer sequences, and map positions (Fig. 1A). Gene information includes open reading frame 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, which includes association to the community member(s), references, update history, and comments. A long-term goal is to have more customizable result pages to allow the browsing and downloading of specific data of interest.

In addition to the search tools, several sequence analysis tools (www.arabidopsis.org/tools) are available at TAIR; BLAST, FASTA, and PatMatch allow users to analyze a sequence against a number of Arabidopsis sequence data sets via the Web. The data sets currently include all Arabidopsis proteins, all Arabidopsis DNA sequences, bacterial artificial chromosome end and expressed sequence tag sequences only, and others. In the future, more specific data sets such as genes, markers, transcripts, as well as nonplant sequences, will be available. The BLAST and FASTA input forms include a variety of parameters for pair-wise sequence alignment algorithms (Altschul et al., 1997; Pearson, 2000). For BLAST, users can submit up to 20,000 nucleotides in multi-FASTA format. PatMatch (developed at Arabidopsis Database; Rhee et al., 1999) allows users to find motifs by entering a regular expression pattern or a simple string of less than 20 characters. It provides an alternative to BLAST and FASTA, which are not suited for short-string searching.

Another software tool recently developed at TAIR is a map viewer (www.arabidopsis.org/servlets/mapper), which integrates the visualization and analysis of different maps on each chromosome (Fig. 2). It allows searching, browsing, and aligning of maps and map elements on each chromosome from TAIR's database and displays the information graphically. This tool was developed to facilitate forward and reverse genetics, where researchers can start with a mutant phenotype and get to the gene of interest, or start with a gene or gene family of interest and find out what the roles of these genes might be by looking for mutations in the gene. Each entity on the maps is hyperlinked to a page with detailed information from the database. In addition, there is extensive help on how to use the map viewer, from interpretation of the data to navigation of the tool. The help document is hyperlinked from many places on the map viewer (http://arabidopsis.org/mapViewer/help/tairmapa.htm).

In the upcoming year, TAIR will be reiterating 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 annotation of the genome, incorporation of genetic mapping data, stock (germ plasm and DNA) data from the Arabidopsis Biological Resource Center, and gene expression data from microarray and gene chip experiments.

Gene annotation will be elaborated to include data from cDNA sequencing projects, gene family analysis, expression, and other experimental data associated with the genes. Sources of annotation will include molecular databases such as GenBank, SwissProt, functional genomics groups, bioinformatics groups, literature, in-house analysis, and individual researchers. Each annotation will be tagged with attribution as well as the source (type of experiment or analysis) for the annotation. TAIR will put efforts into identifying and associating all aliases of gene names from these sources.

To accommodate consistent and facile access of gene annotation, gene expression, and germplasm data, a structured ontology system must be established to describe the function, cellular localization, process, and anatomy/development. In addition, a list of controlled vocabulary for environmental conditions/treatment will be developed. TAIR is collaborating with the Gene Ontology Consortium (www.geneontology.org) to develop the ontology structure for gene annotation, MaizeDB to develop the anatomy/development ontology for plants, and the Arabidopsis Functional Genomics Center to develop the environmental conditions/treatment controlled vocabulary.

AGR (HTTP://UKCROP.NET/AGR/)

AGR is focused on integrating AGI sequence data with the physical and genetic maps of Arabidopsis to provide the necessary components for the study of

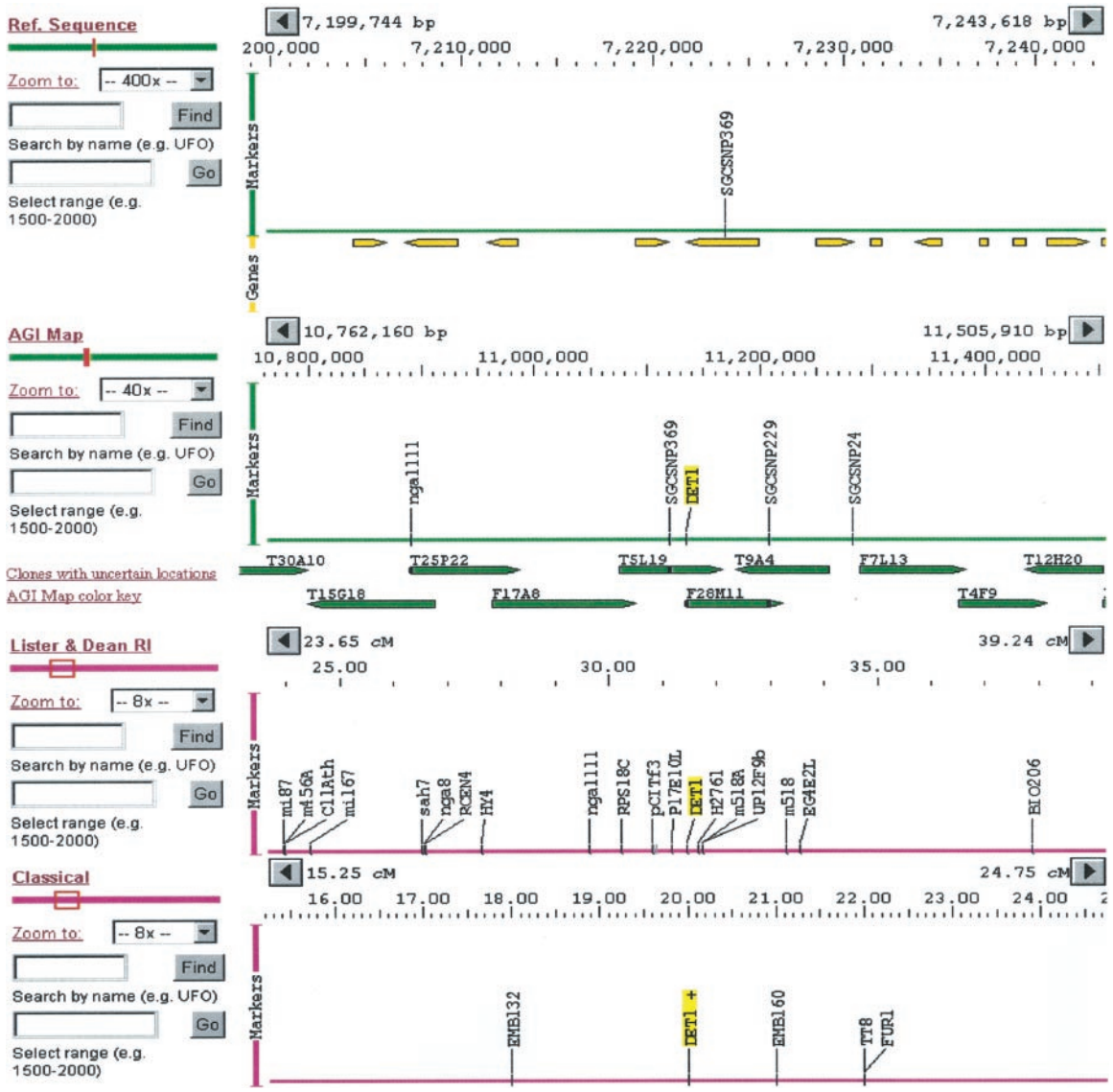

Figure 2. A close-up view of TAIR's Map Viewer. The AGI map, RI map, and classical genetic maps are aligned around the entity DET1. The Reference Sequence Map was searched for a polymorphism near DET1, SGCSNP369. Each map has its own search and zoom functions and the zooming can be customized. Each entity on these maps is linked to a detail page from the database (http://www.arabidopsis.org/servlets/mapper).

gene function and the identification of crop plant orthologs of Arabidopsis genes. AGR maintains the public recombinant inbred (genetic) maps for Arabidopsis and integrates this information with physical map data. AGR's primary purpose is to focus on comparative genomics between Arabidopsis and crop plant species. Sequence homology information is maintained with respect to the public sequence databases (SWISSPROT, TREMBL, dbEST, and EMBL) by searches using BLAST. New Java applets have been developed to provide interactive displays of map and sequence data. AGR is currently the only public database containing insertional flank sequences searchable by BLAST and linked to germ plasm requests.

ARABIDOPSIS GENOME ANNOTATION DATABASES

There are several other databases that allow users to access the Arabidopsis genome sequence and annotation data from the AGI. They each provide searching and browsing tools for the sequence of the Arabidopsis genome as well as annotation of the genome analyzed by the individual groups' suite of programs. Researchers can also download these data in bulk from their public file transfer protocol directories. The following are some of the major databases in this category: Munich Information Center for Protein Sequences Arabidopsis Database, The Institute for Genome Research Arabidopsis Annotation Data-

base, Kazusa Arabidopsis Opening Site, and Database of Arabidopsis Annotation.

CONCLUSIONS

In this paper, I have outlined the ultimate goals desired and the initial steps taken in the areas of genomics and bioinformatics in the Arabidopsis community. The next step is to add value to the genome by verifying the structure and function of the genes uncovered by the genome sequencing and annotation efforts. In addition, the large number of genes whose functions are unknown will need to be studied in a number of ways. This road to discovery will require the expertise of researchers who are focused on specific areas of biology and who are applying more systematic ways of analyzing data in a holistic way. Tools and resources that can serve as the hub of the data flow in the community will be central in moving the level of knowledge forward. For this, we need a database structure that can handle complex data types and be expandable, tools for querying, visualizing, and analyzing the data, and more standardized ways of not only designing and performing experiments, but also describing and analyzing the data.

ACKNOWLEDGMENTS

I thank Eva Huala, Marga Garcia-Hernándes, Leonore Reiser, Lukas Mueller, and Chris Somerville for their helpful comments on the manuscript.

Received September 7, 2000; accepted September 19, 2000.

LITERATURE CITED

AGI et al. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature (in press)

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

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol **132:** 185–219

Rhee SY, Weng S, Bongard-Pierce DK, Garcia-Hernandez M, Malekian A, Flanders DJ, Cherry JM (1999) Unified display of *Arabidopsis thaliana* physical maps from AtDB, the *A. thaliana* database. Nucleic Acids Res **27:** 79–84