The Arabidopsis Information Resource (TAIR)
Third Year Advisory Meeting
November 9, 2002 at Carnegie Institution, DPB

TABLE OF CONTENTS

II. Meeting Roster

TAIR Executive Board:		
Chris Somerville	Carnegie Institution, DPB	crs@andrew2.stanford.edu
Randy Scholl	ABRC	Scholl.1@asu.edu
Chris Town	TIGR	cdtown@tigr.org
David Meinke	Oklahoma State Univ.	meinke@okstate.edu
Bill Beavis	NCGR	wdb@ncgr.org
TAIR Advisory Board:		
Jeff Dangl	University of North Carolina	dangl@email.unc.edu
Dean Della Penna	University of Nevada, Reno	dellapen@msu.edu
Jeff Bennetzen	Purdue University	einstein@bilbo.bio.purdue.edu
Peter Karp	SRI	pkarp@ai.sri.com
Renate Schmidt	Max-Planck-Institut , Germany	rschmidt@mpiz-koeln.mpg.de
Peter McCourt	University of Toronto	mccourt@botany.utoronto.ca
Steve Tanksley	Cornell University	Sdt4@cornell.edu
Gavin Sherlock	Stanford Genome Resources	Sherlock@genome.stanford.edu
TAIR-NCGR		
Dan Weems	NCGR	dcw@ncgr.org
TAIR-Carnegie		
Sue Rhee	Carnegie Institution, DPB	rhee@acoma.stanford.edu
Eva Huala	Carnegie Institution, DPB	huala@acoma.stanford.edu
Marga Garcia-Hernandez	Carnegie Institution, DPB	Garcia@acoma.stanford.edu
Lukas Mueller	Carnegie Institution, DPB	Mueller@acoma.stanford.edu
Tanya Berardini	Carnegie Institution, DPB	tberardi@acoma.stanford.edu
Suparna Mundodi	Carnegie Institution, DPB	smundodi@acoma.stanford.edu
Leonore Reiser	Carnegie Institution, DPB	lreiser@acoma.stanford.edu
Peifen Zhang	Carnegie Institution, DPB	peifenz@acoma.stanford.edu
Nick Moseyko	Carnegie Institution, DPB	nickm@acoma.stanford.edu
Rashmi Nunn	Carnegie Institution, DPB	nunnr@acoma.stanford.edu
Guests		
Sylvia Spengler	NSF	sspengle@nsf.gov
Rebecca Joy	MASC Coordinator	rejoy@biotech.wisc.edu

III. Documentation

The purpose of this documentation is provide an overview of the TAIR project, current status, and future plan to bring everyone up to speed for a productive and in-depth discussion during the meeting.

TAIR Overall Progress and Current Status

We have recently completed the third year of the TAIR project and are embarking on the last two years of the grant. TAIR's main goal is to provide the Arabidopsis research community a gateway to the numerous and diverse information and tools about the research products, progress, and researchers engaged in understanding the underlying mechanisms for this model flowering plant. As the major research community information resource for the Arabidopsis community, we have the responsibility of providing the most comprehensive, up-to-date, and accurate information about the state of Arabidopsis research. The plans for achieving these goals are described in our original proposal to NSF, available online at: http://www.arabidopsis.org/about/proposal.html

The first year of the project was devoted to making a transition from the previous Arabidopsis Information Resource, AtDB, and rebuilding the infrastructure to accommodate the increasing diversity and quantity of Arabidopsis information, such as the complete sequence and annotation of the Arabidopsis genome. The second year was largely devoted to completing the new infrastructure and software architecture development, and engaged in bringing in large data sets such as the genome reannotation data from the Institute for Genome Resources (TIGR) and seed, DNA stocks, and community data from Arabidopsis Biological Resource Center (ABRC).

In the past year, we focused on populating the newly developed database (in addition to streamlining and enhancing database structure) and developing standard operation procedures (SOPs) for curating the massive information about the genome, transcriptome, and metabolome, in addition to the numerous in-depth characterization data contained in the literature. The SOPs range from specifying data formats for loading into the database, developing a series of sequence analysis protocols to provide accurate mapping and association of data objects to the annotated genome, and developing software and procedures for capturing experimental data from the literature.

Using our literature curation software, we have made approximately 6800 associations between ~4000 papers and ~2000 genes and have functionally annotated approximately 2800 genes from information contained in ca. 600 papers. Our literature curation database currently has about 5000 full-text articles (approximately 56% of the available Arabidopsis literature with abstracts, i.e. no supplements and conference proceedings included) and the entire literature curation process is managed and tracked by the software. To leverage the strength of the diversity in the biological knowledge of the curation team, we have been conducting bimonthly annotation meetings with two layers of project management system (curation/annotation managed by Drs Berardini and Mundodi and editing/quality control managed by Dr. Huala).

The scope of the TAIR project has grown significantly in response to the increase in the large-scale functional genomics efforts around the world, and we have successfully kept up with the increase in demand mostly by obtaining additional grants (including two supplements from this grant) to ramp up our personnel infrastructure (See Appendix A). The most critical aspect of this project (as with any project) is the hiring and retaining of trained individuals. NCGR unfortunately has experienced numerous occasions of personnel turnover, which has caused significant setbacks in the project throughout the last three years.

Some of the highlights of the past three years include:

1. One of the most comprehensive model organism database structures and schema documentation (http://www.arabidopsis.org/search/schemas.html)
2. Steady increase in the user access and registered users (approximately 500,000 page visits/month from 20,000 unique IP address; 11,000 and 4,000 registered individuals and laboratories, respectively)
3. Tremendous increase in the quantity and richness of the data available (online DB statistics at: http://www.arabidopsis.org/jsp/tairjsp/pubDbStats.jsp)
4. A comprehensive literature curation tool development and establishment of standard curation and data extraction procedures (See Appendix B)
5. Literature curation of genes: currently ~5400 functional annotations of ~2800 genes from ~600 articles) (See Appendix B)
6. Development of SeqViewer (http://www.arabidopsis.org/servlets/sv) and MapViewer (http://www.arabidopsis.org/servlets/mapper), two of the most frequently-used genome and map visualization tools
7. Development of a Sequence Annotation pipeline system, BACDB, to maintain the latest genome annotation and all mappings of sequenced entities (e.g. ESTs, full-length cDNAs, any sequenced polymorphisms including T-DNA insertions and SNPs) onto the genome
8. Assuming responsibility of ABRC's database, Arabidopsis Information Management System (AIMS): implemented functionalities for storing, ordering and curating public stock information in collaboration with ABRC (not originally planned in the proposal, and supported, in part, from a subcontract to ABRC's NSF grant)
9. Development of Arabidopsis metabolic pathway data and tools in collaboration with Peter Karp's group (two years ahead of schedule)
10. Establishment of collaboration with the Gene Ontology Consortium and Plant Ontology Consortium to develop the most comprehensive and widely-accepted controlled vocabularies for all biology (See Appendix C)
11. Establishment of collaborations among large databases, organizations, consortia, and individual researchers (See Appendix D)
12. Development and implementation of outreach and education component for TAIR users and Educators, supported by an NSF supplement to this grant (See Appendix E)

Despite the major achievements, we have experienced difficulties in achieving some of our projected timelines for the following areas:

1. Development of an industry-standard system infrastructure for providing high performance, secure (e.g. redundant), and 24-7 access to the public

2. Development of a systematic data exchange mechanism between TAIR production and test/development/curation environments that allow facile data integration into the Production system and preserve data integrity
3. Establishment of a mechanism to make TAIR software available to non-commercial and commercial users
4. Integration of genome-wide expression data from cDNA microarrays (e.g. from Arabidopsis Functional Genomic Consortium, AFGC) and high density oligo nucleotide arrays (e.g. Affymetrix)

Progress and Plan of Attack on the Problematic Areas

The following are more detailed information about the problematic areas outlined above and description on how we are planning on tackling these problems in the upcoming year. We will go over these issues during the meeting.

1. Development of an industry-standard system infrastructure for providing high performance, secure (e.g. redundant), and 24-7 access to the public

Issues involving system infrastructure with respect to performance, redundancy and public access can be broken down according to potential break points and then formulating a plan that minimizes down times while staying within budget. The mechanisms that we have put into effect are described below.

- Software Interrupts: Interruption in service can occur when defects in software result in either a server crash or server deadlock. A typical cause of a server crash would occur when an application uses up all available system memory on a server. An example where server deadlock can occur is when all the threads in a multi-tasking operating system are in use and not being released to other portions of the application. This condition typically happens because of synchronization deadlocks or infinite loops. To minimize the possibility of these types of interrupts, we have built and deployed a Test Server, an Inspection Server, and a Production Server in October, 2002. Beyond quality code generation practices, quality assurance is best maintained by a code inspection/test process. We have divided the testing of software into two phases:
  1. Development code is installed on the Test Web Server. Both NCGR and Carnegie have Internet access to the Test Web Server and testing is conducted by both parties.
  2. Once testing is completed on the Test Web Server, the code is loaded onto the Inspection Web Server. Before loading onto the Inspection Web Server, the code on the Inspection Web Server is made to be an exact duplicate of the Production Web Server. Once new code is loaded onto the Inspection Server, another round of testing is conducted by NCGR personnel. The testing phase is designed to ensure that code is not inadvertently migrated from the Test Web Server to the Production Web Server as multiple programs are under development and testing on the Test Web Server.

Quality assurance is not infallible. For every test regimen developed, there is the probability that some condition that can result in a failure gets through undetected. As a backup for this circumstance, the NCGR Systems Administrators have established a software program that attempts to access the Production Web Server on a continuous interval. If an interruption in service is detected, the SAs are automatically paged by the software, whereupon they perform tests to determine the cause of the interruption and restart it. This service, which was implemented in April 2002, is provided on a 24-7 basis.

- Web Server Hardware Interrupts: All computer hardware systems are going to fail at some point in their lifetime. The Production Web Server is a Sun Enterprise 450, consisting of four 400 MHz processors. The system is more than capable of handling the amount of Internet traffic imposed upon it from the TAIR user community. As such, if a hardware failure occurs on one of the processors, the three remaining are more than sufficient to handle the load. Previously, we ran the Test Web Server on the same hardware, but to insure that all processors are available to the Production Web Server, we have moved the Test Web Server to its own computer system in October, 2002. In the event that a critical hardware failure occurs, NCGR has a 24-7 maintenance agreement with Sun Microsystems to repair any hardware failures within 4 hours of the fault. In the event that Sun cannot repair the Production Web Server within their guaranteed time period, we have the ability of using the Inspection Web Server as a temporary backup system until Sun has remedied the hardware fault. We admit that this Inspection/Backup Web Server will be slower than the Production Web Server since it has only a single 500 MHz processor, but the probability of this type of failure is low and we don't foresee the downtime to be extensive. Meanwhile, we will be able to provide the Arabidopsis community uninterrupted access to TAIR.
- Database Server Hardware Interrupts: The Production Database Server is a Sun Enterprise 450 much like the Production Web Server. The major difference between the two is that the disk devices are setup for automatic mirroring of the database to separate disks. In the unfortunate circumstance of a disk failure on this system, Sybase will automatically transfer data access to the mirrored disks. Other, more drastic hardware failures can also be handled. NCGR has a Test Database Server. This system is currently a Sun Ultra II but is in the process of being upgraded to a Sun Enterprise 4000 and is expected to be available to TAIR by the end of November 2002. NCGR generates database backup to tape on a daily basis. In the event of a catastrophic hardware failure, we can dump the latest backup to the Test Database and reconfigure the Production Web Server to access data from the Test Database Server.
- Analysis Server Hardware Interrupts: For the past 3 years we have employed two TimeLogic Decypher computer systems for the purpose of running NCBI BLAST. Due to the prohibitive cost of maintaining and upgrading these computers, NCGR has opted to build an Analysis Cluster comprising of 3 Linux 2 processor 933 MHz computers that have been designed for optimization by employing a load-sharing algorithm. This system has proven to be superior to the outdated TimeLogic machines with performance times up to twice as fast. The system has also been designed so that additional nodes can be easily added to it in the event that demand increases to a point where performance is degraded. The load-sharing concept also means that if one of the systems fails, the user community should not perceive any interruption in service.
- Network Interrupts: To date, the vast majority of downtime periods for the TAIR Web Server have been due to forces outside of NCGR. One means of reducing the possibility of network interrupts is via a redundant Internet connection. We have established redundant Internet service by paying our Internet Provider for two T1 Internet lines. Even with two Internet connections, we are not guaranteed of uninterrupted service. The TAIR Web site has experienced 8 occasions where service was interrupted by our Internet Provider or in one case where a backhoe cut through both of the Internet cables. The only solution to this type

of Internet interruption is to buy Internet service from multiple Internet providers that do not route through the same trunk lines. We have researched this possibility and regret that the cost is prohibitive at the current time. We currently spend about $12,000/month on the current configuration but the estimates for a completely redundant system have been quoted around three times that of our current rate. It should also be noted that the software noted above that the System Administrators utilize to test for TAIR Internet server failures is also implemented at their homes so that Network interrupts can also be detected. In such an event, the SAs work with the Internet Provider to determine the cause of the interruption and bring about a resumption of services in as timely a manner as possible.

2. Development of a systematic data exchange mechanism between TAIR production and test/development/curation environments that allow facile data integration into the Production system and preserve data integrity

Currently we have established 3 TAIR databases. One is the Production Database, another is the Test Database and a third is installed at Carnegie. The Test and Carnegie Databases are for test purposes. Recently we have been encountering situations where testing fails due to these two databases being out of sync with the production database. We have been researching how to get the test and Carnegie databases synchronized with the production database while maintaining the functionality intended for each. The problem inherent with these databases is the simple fact that the database schema is continually being modified as the scope of the TAIR project grows. Replication software from the database vender can be utilized but replication software is intended for use where the database schema is unchanging. Another option that is available is to use a Sybase program called BCP (Bulk Copy) to update the Test and Carnegie databases but BCP has the same problem as replication software in that it assumes identical database schemas, including implementation order, between the source and destination databases. In addition, the speed of BCP is horrendously slow when used on a database as large as the TAIR database. A third possibility to solve this problem is to perform a binary dump of the production database and rebuild the Test and Carnegie databases from this binary dump. Though this approach could work for the Carnegie database, this approach would be problematical for the Test Database as we perform tests on database schema changes and software updates on that system before performing the same operations on the production database. Overwriting that work would be counterproductive.

We believe we have a solution to the above problems that will become available after November 2002. During that month the Test Database Server will be upgraded from a Sun Ultra II to a Sun Enterprise 4000. Not only is the Sun E4000 a faster and more reliable computer, but it also has considerably more disk space. We propose to create another copy of the TAIR database on the Test Server, which we will call the Inspection Database. Our Database Administrators assure us that they can write software that will acquire a copy of the Production Database dump and use it to automatically build the Inspection and Carnegie Databases at some agreed-upon interval. In addition, the NCGR Test Database will remain as a database that will be updated only manually at wider intervals to allow continued schema and data loading tests. The Inspection Database can then be used by software that tests automatic database-to-database loading that is currently under development and where the synchronization problem arose. We estimate this system can be in place about a month after the Carnegie Database Server has been upgraded to Sybase 12.5.

3. Establishment of TAIR software availability mechanism using a non-commercial and commercial license agreements

The TAIR Executive Committee has been evaluating two source code licenses that NCGR developed for making TAIR source code available to commercial and non-commercial applicants. The current wording of the two source code licenses indicate that the source code is available for an unlimited term, that the use must be for internal use only (i.e. services not sold to a third party) and that TAIR will not provide any support to the applicants. One difference between the two licenses is that grant back is expected from non-commercial users but not from commercial users. Grant back means that TAIR has the right of royalty-free license to copies of the software if the user performs code modifications that improve it. Another difference between the two licenses is that the source code would be provided to non-commercial applicants at no charge but commercial applicants would pay a one-time fee that would be equally divided between the TAIR development teams at Carnegie and NCGR. The latest version of the non-commercial and commercial license agreements, which have been essentially approved by the Executive Committee. We would like to ask the Advisory board to examine these agreements and provide feedback at the meeting. Our hope is to finalize the agreements at the end of the advisory board meeting.

To implement the delivery of the source code, we would require the applicant to sign the appropriate source code license and deliver it to TAIR. Upon receipt of the signed document, the user would be given access to an FTP account where a copy of the source code will reside.

4. Integration of genome-wide expression data from cDNA microarrays (e.g. AFGC) and high density oligo nucleotide arrays

The goal is to provide access to all publicly available Arabidopsis expression data in a searchable, browseable, and downloadable format. Sources of data include individual submissions, the literature, or other databases such as ArrayExpress or GEO. We are currently focusing on the integration of genome-wide expression data from AFGC and Affymetrix gene chip microarray data from individual researchers, and individual gene expression information (e.g. Northern, *in situ* hybridization) from the literature. In the future, we plan on integrating other types of expression technologies (e.g. SAGE, LYNX, etc.)

Our plan was to make the AFGC expression data available starting January of 2002 (as the AFGC project ended at the end of 2001) and discussion about the transition of data and tools from AFGC to TAIR began at the end of 2000. There have been discussions of transferring two major types of resources from AFGC: software (expression viewer, array element search, motif finder) and data (experiment abstracts, RNA information, raw data, processed data currently existing as flat files or in the Stanford Microarray Database, SMD). Although we have incorporated as much as possible from the SMD database design, we have decided against importing the database structure and associated software because of the significant differences in the use case and emphasis of the database functionality between SMD and TAIR.

The following are some of the transitions and information made available to date:

1. Array Elements Annotation
    - available as bulk, downloadable file (released 7/11/01 and subsequent updates released)
    - searchable and downloadable with links to TAIR data pages, AFGC's expression viewer, and SMD's spot history (released 8/27/2002)
2. Software
    - Expression Viewer (released 8/27/2002 and updated with all AFGC data on 10/30/2002)
    - Array Element Search (released 8/27/2002)
    - Motif Finder (to be released 11/4/2002)
3. Experiment and expression data
    - Database Implementation (completed 9/30/2002)
    - AFGC data received (initial set on 11/16/2001, complete set on
    - 9/26/2002)
    - Data reformatting and curating (ongoing, mostly completed 10/20/2002)
    - Data loading (ongoing, expected to be completed in November 2002)
    - Search and Download Software development (ongoing)
    - Microarray data submission process design (completed 11/1/02) http://acoma.stanford.edu/~tacklind/array/microarray_submission_new.html
    - Microarray data submission software (ongoing)
    - Expression search and download (design initiated)

The most important factor that played into the delay of the microarray data integration was the limited personnel. We had only a 0.75% FTE curator (Marga) who has been communicating with AFGC, curating the data, designing the data model and user functions and interfaces, and attending the MGED meetings. The second factor has been the complexity of the information and the lack of in-house and collaborative expertise in this area. We have tackled these issues in the following ways:

- hired two additional curators (supported by other grants) who have experience in performing and analyzing Arabidopsis microarray data: Dr. Suparna Mundodi (started 3/1/02) who worked for the AFGC and Dr. Nick Moseyko (started 10/1/02) who has experience performing and analyzing Affy microarray experiments using Arabidopsis
- hired a half-time database developer, David Dixon, from an NSF supplement to the grant.
- hired a postdoctoral fellow (from another grant) who will work on developing methods for analyzing and integrating microarray data with other information in TAIR
- initiated discussions with ArrayExpress at EBI to collaborate on software development and data exchange (initial discussion on 9/02, follow-up meeting planned for 11/21/02)

We spent a lot of effort in formalizing the user function specifications and the database design. Since the information management of microarray data is a relatively new area and consensus is evolving among the community members, this is an area we needed to put much research into establishing a good structure that is consistent with the most current accepted model (e.g. MIAME compliant) and is amendable for change.

Database Conceptual Design and Implementation

In the creation of the database model several of the existing database structures were considered (ArrayExpress, SMD, GeneX, GEO, GATC, and ArrayDB). Also, the availability of large samples data from AFGC and a few based on Affymetrix chips was very helpful in the modeling. The current data model follows the specifications of the MAGE group, which deals on the standardization of microarray data representation and exchange. Before arriving at the model we have decided to implement in TAIR, at least 34 other models were drawn on the way. We believe this "final" model will allow us to address the uses of the database that are outlined in a document, available at: http://acoma.stanford.edu/~curator/use_case.html under "USE CASES FOR MICROARRAY EXPRESSION ". The database diagram and specifications of the final version can be found at: http://arabidopsis.org/search/ERwin/Tair.htm (Click on Expression/Display 1).

Data Collection, Curation & Processing

The data currently in our hands include the complete set of the public data from AFGC and two Affy data sets from Thomashow's group. These data include not only the results, but also the information regarding array elements, array design, and experimental design (when available).

AFGC staff extracted the Arabidopsis data from SMD in the format we requested at two different times. The first set comprises the data for 232 hybridizations available as of Nov. 2001. It consisted of 38 files, of which 25 required minimum or zero curation/processing (22 are the different array layouts, and the other 3 contained the numerical results data). The rest of files required curation/processing, which resulted into 14 data files (they need to be re-formatted again to accommodate the newest database structure). The second batch of AFGC data was received in September 2002 and is currently in house as flat files and needs to be curated and processed as the first set.

At the meeting, we will discuss the design of the data model, and the different ways we intend to make the information available for searching, downloading, visualizing, and analyzing the data.

Overall plan for Year 4

There are five major goals planned for year 4:

1. Overcome the problems outlined in the previous section
2. Keep up-to-date the Arabidopsis genome annotation
3. Establish specific goals and plans for establishing TAIR as the main infrastructure for disseminating the functional genomics efforts (e.g. 2010 projects)
4. Develop a multi-faceted Quality Control system for maintaining the integrity, quality, and accuracy of the data within and incoming from the research community
5. Integrate phenotype, allele, images, and proteome information from the literature and through establishment of collaboration with data providers

Genome Annotation Maintenance

In the past year, we have been involved in several aspects of maintaining the genome annotation. Some of the areas are briefly described below. We will discuss these areas in more detail at the meeting.

Although the genome sequence of Arabidopsis was 'completed' at the end of 2000, there are a few areas of the genome that were never completely sequenced, and other areas for which the completed sequence was not deposited in GenBank. We have been working with the advisory board of the Arabidopsis Genome Initiative (AGI) to monitor the post-announcement progress toward 'completion' of the genome. Some of the information we have been gathering (in collaboration with TIGR) is available online at: http://arabidopsis.org/info/agicomplete.html In addition to checking GenBank for the latest status of the remaining clones still being sequenced or in need of an updated description in GenBank, we have been monitoring and analyzing any Arabidopsis sequences that do not match the genome (e.g. dbEST, PLN, GSS databases in GenBank as well as direct user communications) and communicating these to TIGR.

In addition, we have been working with TIGR and MIPS to generate, maintain and track the history of the AGI locus code designations. Both MIPS and TIGR have been involved in 'reannotation' of the genome since the announcement of the complete genome sequence, and this has included obsoleting, merging, splitting, and adding new genes to the sequence. In order to keep track of these changing genome annotations and minimize conflicts in the usage of the AGI locus codes, we have been analyzing the different versions of the TIGR and MIPS genome releases and tracking the history of the locus code usage, which is downloadable and searchable online at: http://arabidopsis.org/tools/bulk/locushistory/index.html.

TIGR has been the main group to reannotate the genome in the last two years (2000-2002) and we have been planning to take over this responsibility after TIGR's funding ends. Thus, we have interacted closely with TIGR on issues of data exchange (e.g. XML format, functional annotation standards and exchange). We have met on a bimonthly basis over the phone for approximately 6 months specifically to discuss genome annotation data exchange and are currently using a common web page to record each group's annotation status to minimize duplication of effort. TIGR has also been responsible for releasing annotation data to GenBank and we have been in communication with NCBI recently to develop systematic methods to submit the genome annotation data to GenBank after TIGR's funding ends. Toward this goal, we have been learning the Sequin system, established at GenBank for the last ten years, to update the structural annotation of genes.

Establishing TAIR as the main infrastructure for disseminating the functional genomics data

With the completion of the genome sequence has come a new emphasis on functional genomics within the *Arabidopsis* community.

Large-scale funding initiatives have been established in several countries and are providing the opportunity and means for laboratories to undertake high-throughput functional examination of gene families. TAIR has a central role to play in the world of *Arabidopsis* functional genomics research, both as a supplier of information and biological reagents to those undertaking this research and as an archive and center for distribution of the data generated through it.

For most labs undertaking functional genomics research, the ultimate goal of the investigation is to determine the developmental, physiological and/ or cellular function of each member of a family of genes related by sequence homology or similar expression profiles, and to determine where these functions fit into the biology of the plant as a whole.

One important aspect of the functional genomics initiative is the amount of work that needs to be done in a relatively short amount of time. The international community has set forth a goal of determining the function of every one of the over 25,000 *Arabidopsis* genes by the year 2010. A key to the success will be the expedient availability of biological reagents, data, and data mining tools.

Some TAIR features important to those beginning an investigation into a gene family

- Correct and make searchable gene annotations
- Sequence and data manipulation software, or links to open source software
- Microarray /transcript profiling data for mining
- Insertion line flanking sequences (FSTs) with facile searching for gene insertion knock-out /mutant lines, with links to providers
- Additional information about microarray/ transcript profiling and other service providers

A major goal for TAIR as these projects get underway will be developing mechanisms for handling and displaying the large quantities of data that will be generated by functional genomics projects. Types of data will include, but not be limited to:

- Gene product identity and structure, including data about functional motifs found within RNA or protein
- In some cases, multiple products from individual genes; includes post-translational processing and active vs. inactive forms of proteins
- RNA and protein expression profiles varying in developmental time, tissue, and condition; including photographic evidence
- Network/ pathway/ process with which the gene product is involved; including models
- Proteomic data
- Metabolomic data

- Data concerning regulatory network that controls expression/ modification of gene and protein
- Mutant/ genetic variant phenotype data, including transcript, proteomic and metabolomic profiles

Many of the issues surrounding establishing TAIR as the main infrastructure for dissemination of functional genomics data are the same issues that have been discussed for management of data emerging from other types of research. Some of these issues include:

- Getting the data into the database, especially a) IP issues that come up with funding from private industry, and b) getting researchers to submit data to TAIR.
- Many functional genomics projects are currently setting up their own, separate databases, data from which should be eventually folded into TAIR.
- Controlled vocabulary issues, including protocol descriptions, phenotype descriptions
- Efficient data mining methods

Developing a Multi-faceted Quality Control System

We have spent a lot of effort in increasing the content of the database in the past year. Data in TAIR can be categorized into two classes, static (e.g. physical mapping data, genetic marker information) and dynamic (e.g. structure and function of genes). In addition to the data that are provided from the researchers, we curate and add many different data types and associations (e.g. genome position using sequence analysis, functional description of genes from the literature, and associations of polymorphisms and clones to genes using sequence analysis). In order to provide the most accurate and up-to-date information, we have implemented a number of standard operation procedures for annotating the genome, associating data to the genome, and extracting information from the literature. Some of the curation pipelines are quite complex and undergo a few dependent processes. In order to bring these curated information to the production environment, which is also undergoing updates from direct user interactions, we need to develop a more robust mechanism of exchanging data between the test/curation systems and the production system at TAIR. There are four main areas of data curation and exchange QC we envisage:

A. Establishing a robust in-house data analysis and QC mechanisms
B. Developing a robust testing environment for bulk, curated data entry
C. Establishing strict minimal data requirements and formats for incoming data
D. Establishing facile user communication and feedback on the quality of the data

During the meeting, we will discuss some of the ways we have been tackling this issue and additional areas/methods we should consider incorporating to ensure the highest quality of information.

Integrating phenotype, allele, images, and proteome information from the literature and through direct interaction with data providers

We have been focusing our literature curation efforts on describing the function and role of individual genes in the past year. Our next target for literature curation is allele/phenotype information and protein information (e.g. post-translational modification, subunit/complex information, tryptic digest patterns, three-dimensional structure information). In addition, we plan on developing methods to curate gene interactions information (e.g. genetic pathways, transcriptional regulation information). The use of image data for describing allele information (e.g. mutant phenotype) will be critical and we are currently developing methods for storing and displaying image data. We will discuss how we will be curating these complex data types during the meeting.

Areas of Advice Wanted:

1. What should be the priorities in the outlined future and present goals?
2. Should we be more proactive on outreach and education?
3. Any issues neglected?
4. Any future goals neglected?
5. What types of additional data should we capture for metabolic pathways?
6. What would be the ideal level of information we should try to capture from the literature?
7. What additional data types should we be capturing from the literature?
8. Are there any suggestions for reusable software that may be useful to integrate into TAIR?
9. What are some of the ways we can reach out to plant researchers not working on Arabidopsis?
10. What are some of the ways to increase researcher feedback in ensuring data quality?
11. Are there any meetings and venues we should consider attending and conducting workshops to inform researchers of our activities?

Appendix A. Project Personnel History

Carnegie Institution, DPB

PI/Co-I

First Name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Chris	Somerville	1.00	0	9/1/99	Present
Sue	Rhee	1.00	72	9/1/99	Present

Scientific Curators

First Name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Eva	Huala	0.75	80	9/1/99	Present
Margarita	Garcia-Hernandez	0.80	90	9/1/99	Present
Tanya	Berardini	1.00	50	1/15/02	Present
Nick	Moseyko	1.00	0	9/1/02	Present
Lukas	Mueller	1.00	76	6/19/00	Present
Leonore	Reiser	1.00	100	12/01/99	Present
Suparna	Mundodi	1.00	0	3/1/02	Present
Peifen	Zhang	1.00	0	4/15/02	Present
Aisling	Doyle	1.00	100	11/1/00	10/30/02
Gabriel	Lander	1.00	0	8/26/02	Present
Jungwon	Yoon	0.75	20	9/1/00	Present

Programmers

First Name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Bryan	Murtha	1.00	100	2/1/01	8/3/01
Chunxia	Xu	1.00	100	10/10/01	Present
Danny	Yoo	1.00	0	6/1/01	Present

Postdoctoral Fellows and Associates

First Name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Alan	Chou	1.00	0	3/18/02	7/9/02
Mark	Lambrecht	1.00	0	4/1/01	2/28/02
Yigong	Lou	1.00	0	9/1/02	Present
Shijun	Li	1.00	0	10/21/02	Present

Visiting Graduate Students

First name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Anell	Bengt	1.00	0	9/15/00	12/31/01
Smita	Mitra	1.00	0	8/28/00	10/5/00

Interns

First name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Debika	Bhattacharyya	1.00	100	7/1/00	8/31/00
Jill	Larimore	hourly	100	1/3/01	4/30/02
Holly	Nottage	hourly	100	2/1/01	7/15/01
Julie	Tacklind	1.00	70	12/5/01	Present

National Center for Genome Resources --BILL/DAN PLEASE FILL IN THE BLANKS

PI/Co-I

First Name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Bruno	Sobral				gone 2001
Allan	Dickerman				gone 2001
Bill	Beavis				Present
Dan	Weems				Present

Programmers

First Name	Last Name	FTE	% Support from TAIR	Start Date	End Date
Don	Kiphart				gone 2000
Frank	LaFond				gone 2000
David	Hanley				gone 2001
JJ	Zhuang				gone 2000
Wen	Huang				gone 2000
Yihe	Wu				gone 2001
Neil	Miller				Present
Guanghong	Chen				Present
Mary	Montoya				Present
David	Dixon				gone 2002

Appendix B. Progress on the Literature Curation Tool Development and Implementation

We have developed a literature curation tool called PubSearch, which stores literature, gene, functional annotation, and keyword data in a stand-alone database and allows curators to establish associations between these data types using a web browser. In PubSearch, first-pass associations between terms (gene names and keywords) and articles are made automatically by a string matching program that indexes terms to articles. Commonly occurring words such as AND, THE, IF (stop words) are filtered out to minimize meaningless associations from being stored. For terms with a higher signal-to-noise ratio, curators verify the matches via the web browser user interface.

PubSearch uses a simple database schema in a MySQL database management system (DBMS) (version 3.21), which can be queried and updated using a password-protected login mechanism via the internet using a web-browser. The middleware is written in Java (version 1.3) and uses Java Servlet and Java Server Page (JSP) technology. The system is currently running on a Linux RedHat7.2 system with Tomcat (version 4.0) as the servlet engine. A demo of the current version of this tool and its documentation can be accessed from:

http://tesuque.stanford.edu:9999/pubtest/index.jsp
Username: demo Password: demo

The tool has been used and refined for the past 6 months by 7 curators at TAIR and 5 Arabidopsis curators at the Institute for Genome Resources (TIGR) to curate over 12,000 articles. The tool is much more convenient and user-friendly than our old system involving flat files and our curation work has become much more efficient as a result.

The system currently allows curators to update gene information (e.g. aliases, description, summary, sequences) and annotate the roles of the genes in many different aspects (e.g. subcellular location, expression patterns, molecular function, biological process, subunit/protein complex, gene family, regulation, etc.). Each functional annotation uses a controlled vocabulary (we currently have five working categories of controlled vocabulary), evidence information (currently seven categories each with several subtypes of evidence source), and the literature or other reference that the annotation is derived from. In addition, the software allows full-text searching of ~5000 articles. Because all the literature information and the curated data are stored and tracked in the database, we can apply many different strategies for deciding how to prioritize the annotation efforts. We have successfully used two strategies (e.g. completing the subcellular location annotations and focusing on literature with matches to plant-specific controlled vocabularies). We have curated approximately 600 papers and 2800 genes using the software to date and are currently curating approximately 100 genes per week as a group. In addition, we can estimate how much more is needed to 'finish' the literature in our system currently before needing to utilize additional methods for targeting papers to curate.

Use of this comprehensive software, which is increasing in its complexity and functionality, we are curating information from the literature as a group. By having six post Ph.D. level curators and 2 B.A./M.A. level curator assistants use the same software and track the procedures, questions, and justifications for making the annotations, we are progressing towards the establishment of a extensive curation guide and SOPs. In addition, we are implementing an 'editorial' procedure where four senior curators randomly check annotations to reduce any inconsistencies in the underlying assumptions of the individual curators. All curators meet bi-monthly to discuss any issues on annotations. We are currently annotating approximately 50 genes per week, which includes functional annotation using the controlled vocabularies and merging/inserting/obsoleting gene information.

We have recently acquired a two-year NIH grant in collaboration with Rat Genome Database to further develop this stand-alone tool for use by any model organism database.

Appendix C. Progress on Controlled Vocabulary Development and Functional Annotation

Progress from TAIR focused on the following areas; 1) hiring personnel, 2) updating the ontologies to accommodate annotation of plant genes, 3) manually annotating Arabidopsis genes to GO, and 4) assisting other plant groups to annotate genes using GO. These activities were made possible, in part, by a subcontract to Jackson Lab on the GO NIH grant.

The two personnel are both stellar biologists: Dr. Tanya Berardini (starting January 15, 2002) who is experienced in plant development and phylogeny studies, and Dr. Suparna Mundodi (staring March 1, 2002) who has expertise in plant defense and microarray analysis. Tanya and Suparna were trained by Dr. Leonore Reiser (who participated in GO since year 2000 from TAIR's own funding and is now a full-time education outreach coordinator at TAIR) in the first half of this year. Tanya and Suparna are now active, trained GO curators, who oversee the annotation and GO development at TAIR (developing and maintaining curation guide, organizing and moderating bi-weekly annotation meetings) and coordinate collaborative efforts with other plant groups (e.g. TIGR, Gramene, MaizeDB).

The first and foremost task of TAIR, the only plant consortium member with CVS write access to the ontologies, has been to ensure that the GO structures are suitable for annotation of and comparison among plant genes. Thus, we are devoting a large portion of our effort in updating the GO structures. So far, we have added 457 terms and updated approximately 300 terms. Several sections of the biological process ontology were examined in great detail and modified. The major revisions were done in collaboration with Dr. David Hill at the Mouse Genome Database. Some revisions (e.g. embryogenesis and morphogenesis, histogenesis and organogenesis, and pattern formation) involved the addition of many higher order terms that were necessary to group the more species-specific terms that were already present in the ontology. Some sections were added specifically to annotate plant-specific processes (e.g. gametogenesis (sensu Magnoliophyta), double fertilization, and germination). Some sections needed major elaboration via the addition of children and reorganization of higher-level structures (e.g. photosynthesis, secondary metabolism, sugar and signal transduction terms, and response to external stimuli).

This year also marked the beginning of serious GO annotation efforts at TAIR, spear-headed by Suparna and Tanya. Since we have started doing GO annotations, our manually curated annotations have increased from about 50 (February 2002 GO meeting) to 2563 (September 2002 GO meeting). There are currently 4105 manually curated GO annotations to 2371 genes at TAIR. As a group, we have been meeting regularly and discussing a variety of strategies for literature curation of genes using GO. These strategies are implemented in our literature curation software, PubSearch, and a couple of strategies have been carried out successfully. For example, we have essentially completed the GO component annotation based on experimental data (IDA) in the Arabidopsis literature (496 genes and 569 annotations). Another strategy recently implemented focuses on papers containing the GO terms that TAIR added AND describe genes that do not yet have manual GO annotations. We identified approximately 1400 of these genes in our PubSearch database. Through this process, we are checking the integrity of the terms we added to GO (i.e. checking the True Path Rule and whether these terms reflect valid processes, functions and components). Upon completion of this task, we will continue the strategy by annotating genes described in papers containing non-TAIR GO terms. There are approximately 3200 such genes in our database.

In addition to the above two main efforts, Suparna developed a general GO-slim terminology, in collaboration with Amelia Ireland from EBI. Suparna recently completed a second version of a plant GO-slim file, which will shortly be ready to be hosted on the GO site. Likewise, Tanya has been involved in several phone conversations and extended email discussions with curators from other plant databases (Pankaj Jaiswal at Gramene, Leszek Vincent at MaizeDB, and Linda Hannick at TIGR) regarding adding and modifying plant-related terms in GO and on various annotation methods.

Appendix D. Collaborators of TAIR

TAIR relies heavily on its collaborations with other groups, large and small, in sharing data, software, and expertise. We have made much progress in collaborating with bioinformatics, genomics, and biology-based consortia, databases, and laboratories around the world. Some of the major collaborative groups include:

- The Institute for Genome Resources (TIGR, Chris Town, Owen White, Brian Haas, Linda Hannick)
- Arabidopsis Biological Resource Center (ABRC, Randy Scholl, Debbie Crist, Emma Knee)
- Gene Ontology Consortium (GO, Flybase, MGI, SGD, Wormbase, TIGR, Swiss-Prot, Dictybase, and numerous other databases)
- SALK Institute (Joe Ecker, Paul Shinn)
- Arabidopsis Tilling Project (ATP, Steve Henikoff, Elizabeth Greene)
- Stanford Genome Center (SGC, Mark Piercy)
- Plant Ontology Consortium (POC, MaizeDB, Gramene, IRRI)
- Arabidopsis Functional Genomics Consortium (AFGC, Shauna Somerville, Pam Green)
- National Center for Biotechnology Information (NCBI, Jim Ostell, Tatiana Tatusov)
- European Bioinformatics Institute (EBI, Helen Parkinson, Alvis Brazma)
- Plant Systems Biology, Ghent University (Martin Kuiper)
- Cold Tolerance Project, Michigan State University (Mike Thomashow)
- Large-scale Protein Localization Project, SUNY Stonybrook (Vitaly Citovsky)
- Rat Genome Database (Simon Twigger)
- SRI International (Peter Karp)
- Cereon Genomics, Inc. (Steve Rounsley)

Appendix E. Education/Outreach Activities:

We received a supplement from the TAIR NSF grant to initiate education and outreach to the TAIR research community and teachers. The supplement has been essential in creating and developing education and outreach components in the following eight ways at TAIR:

1. Outreach and education to the research community

Goals: To educate both novice and ‘expert’ users about extent and utilization of data and tools available at TAIR. To obtain user feedback (both positive and negative) and use cases that informs the design and redesign of TAIR database and tools.

Personnel: All curators
We have conducted the following four workshops at:

- ASPB meeting, Denver, CO (2 sessions)
- Arabidopsis International Congress, Seville, Spain (2 sessions)
- University of California at Berkeley, CA (1 session)
- Stanford University, Stanford, CA (2 sessions)

Approximately 250-300 researchers attended the workshops in total and the response was overwhelmingly positive. We plan on continuing and possibly extending the number of workshops in the upcoming year. During these workshops, a survey was distributed and the questions and answers to the survey is available online at: http://tesuque.stanford.edu/~iris/cgi-bin/analysisSurvey.pl

2. Education pages

Goals: To create web pages and resources that will 1) alert the TAIR community to outreach/education opportunities and, 2) provide educators and students with a forum for contact and collaboration between educators and scientists.

Personnel: Leonore Reiser, Jungwon Yoon

Progress/Status:
Created new section for education (under About Arabidopsis) including:

1. Web page for scientists highlighting the new NSF proposal guidelines for incorporating outreach into research proposals. This page also links to other relevant content such as a list of outreach organizations.
2. Web pages with descriptions and contact information for various outreach programs and websites for plant biology, biotechnology, genomics and genetics education. Subsequently moved all the programs into TAIR Community where it is now searchable (see Community Registration Updates).
3. Created an email forum for teachers /students and educators use to communicate with each other and potentially provide a forum for collaboration. This was created in July 2002 and there has been zero traffic on this list.

Timeline: Structure is essentially completed, now in updates phase. Updates may include replacement of Programs and Websites pages with browseable list generated from TAIR DB.

3. Community/Updates and Registration

Goals: Facilitate inclusion of more community into TAIR-specifically pre-college teachers, students and outreach personnel/programs. Allow more specific searching by including additional parameters in the search.

Personnel: Leonore Reiser, Neil Miller, Iris Xu, Dan Weems

Progress/Status:

1.Updated Community/Organization/Person and Affiliation tables to accommodate new community types and relationships (added 'contact person' and the ability to add contact person during registration). Other ancillary updates to the registration process and search interface including searching by job title and organization type.

2.Updated Community detail pages to display associated References.

Remaining Tasks:

1.Develop an SOP for curation of community defined (user defined) keywords. This requires obtaining user defined keywords from TAIR, curation in Pub and return to TAIR along with updated data associations.

2.Develop and deploy JSPs to allow users to associate themselves to publications (and also to obsolete incorrect associations).

4. Protocols

Goals: To make protocols in TAIR more accessible by making them searchable from TAIR database. To allow users to add protocols to the database.

Personnel: Leonore Reiser, Julie Tacklind, Iris Xu, Dan Weems, Neil Miller, Marga Garcia Hernandez.

Progress/Status:

1.Modification of database structure. A new table (Protocols) has been added to the UML and specifications sent to NCGR for implementation in the new Sybase 12.5 database.

2.Protocols already on TAIR's website have been consolidated on one page (Protocols and Lab Manuals) that is linked from the home page.

3.Added new Protocols/Lab Manuals (such as growth protocols from NASC/ABRC/TILLing and PREP Handbook)

4.Converted (where necessary) protocols to PDF documents and deposited these into the TAIR FTP site.

5.Manually curated protocols from TAIR and created flat-files for bulk loading into TAIR.

6. Created mockups for protocol searching, submission, and detail pages. Additional updates to these forms are being made.

7. In the course of curation it was determined that an ontology of methods was required for annotation. Method 'keywords' have been obtained from the following sources: Microarray data, Protocols and GO evidence code descriptions and imported into DAGEditor. The ontology is being structured using this tool.

Remaining tasks:

1. Data loading into TAIR.
2. Development and deployment of search, display and submission of protocols JSPs.
3. Finalize a base-line version of the experimental methods ontology.
4. Develop an SOP for curation of community defined (user defined) keywords. (see Remaining Tasks under Community Registration Updates).
5. Curation of protocols from Weeds World/AIS.

5. Integrated Help/Users Guide

Goals: To assist the user community (esp. researchers, teachers and students) in learning about and how to use TAIR resources (tools and data).

Personnel: Leonore Reiser, Jungwon Yoon, Lukas Mueller, Guanghong Chen, Sue Rhee, Tanya Berardini, Gabe Lander

Progress/Status:

<u>Phase 1 Goals: (to be completed by November 15)</u>

1. Create a standardized format for help pages that makes it easy to create and modify help pages.
    - Converted the existing header and footer into java script
    - Implemented use of style sheets for consistent format

- Updated existing help pages to new format (with the exception of Tool help documents)

2. Improve navigation of help documents
   - Move all help documents to /help subdirectory in TAIR. This has been implemented for the bulk tools and search help and will be implemented for analysis tools.
   - All pages that have been relocated still exist and have been set to redirect after 3 seconds to the updated-corresponding page in the new location.
   - Revision of help to include 3 primary domains: Database Search Help, Tools Help and Community Registering/Stock Ordering Help.
3. Create help documents for all current searches/tools (where appropriate).
   - This is essentially complete with the exception of :Polymorphism Detail page help, Sequence bulk download help, GO Annotation bulk download help.
4. Make help pages explicitly searchable.
   - Implement Google search to search only documents within /help and subdirectories. The help pages could eventually reside in TAIR DB and be searched via a new interface.

<u>Phase 2 Goals:</u>

1. Web-based users guide based upon a problem solving approach. The most frequently asked questions/common tasks will be targeted from user queries (searchable via Jitterbug curator question tracking software and questions stored as 'FAQs'). The users guide will be integrated with the help documentation and eventually tutorials. For example, a section of the users guide that illustrates how to use TAIR to obtain information about and mutant strains for members of a gene family would also include links to help pages for the SeqViewer,Polymorphism Search and Detail pages, TAIR registration and stock ordering.
2. Automatically generated table of contents using Perl script. Once all of the help documents are in a standard format, the major topics can be easily pulled out at organized (based upon previous and next links in the header) into a table of contents. This will make it easier to ensure the contents always are up to date.
3. Tutorials-Development of interactive tutorials that are integrated with TAIR help, users guide and glossary. A simple prototype tutorial was developed in response to a request from a user (Dr. Susan Blauth –University of Redlands, who attended a TAIR workshop at the ASPB). Dr. Blauth has provided feedback as to the format, content and flow and will provide additional feedback after her class has utilized the tutorial. To view the tutorial with frames (http://tair.stanford.edu/~lreiser/help/tutorial/gene_tutorial.html) and without frames (http://tair.stanford.edu/~lreiser/help/tutorial/genetuttext2.html). Additional tutorials will likely use FLASH which will allow for more informative demonstration , i.e. it will be easier to clearly show how to navigate though and use TAIR effectively.

6. Glossary

Objectives: Create a searchable glossary of terms both for specific data types, attributes used by TAIR as well as other commonly used terms/concepts. Create hyperlinks between these terms (when used in TAIR) and their descriptions. For some terms, find or create visual explanations.

Personnel: Leonore Reiser, Iris Xu, Neil Miller, Dan Weems, Tanya Berardini, Sue Rhee

Progress/Status:

1. Updating table definitions (defining TAIR DB terms). A new table to hold TAIR table definitions and interface for updating this table was generated using PERL and updated to a JSP.

2.Glossary: Minimal effort has been put into this part as many of the potential terms and definitions will come from existing controlled vocabularies (e.g. anatomy, developmental stages, methods).

Remaining Tasks:

1.Modification of definitions updater to allow updating of additional fields and insertion of new fields.

2.Insertion of definitions using this tool.

7. Collaborations/Cooperation

Dr. Erin Dolan Outreach Coordinator-Fralin Biotechnology Center: NIH SEPA proposal, integration of PREP handbook into TAIR.

Ellyn Daugherty and TEAM GUS- San Mateo High School: Assisting a high school teacher and 4 students in a project to analyze Arabidopsis plants transformed with a 35S::GUS fusion. Seeds were obtained from Drs Yukiko Mizukami and John Harada. Students are growing the plants, PCR amplifying the GUS gene and sequencing the product. Their queries are informative with respect to what information students and teachers need and find useful. This has already resulted in some changes/enhancements to the About Arabidopsis web page and protocols section. Ms. Daugherty plans to include this project in her book on Biotechnology for high school students but also would permit inclusion into TAIR.

Barbara Soots Outreach Coordinator-Center for Engineering Plants Resistance Against Pathogens (CEPRAP). : Assist in development of plant biology projects using Arabidopsis for high school teachers and students. (see Meetings/Workshops attended).

Dolan DNA Learning Center: Provided a letter of support for Dolan DNA Learning Center NSF CCLI proposal for Arabidopsis project. Agreed to participate in first teacher workshop.

8. Meetings/Workshops

Leonore Reiser participated in the following meetings related to science/plant biology education with the objectives of 1) identifying potential collaborators and 2) identifying ways in which TAIR can support science education and 3) problems with TAIR that inhibit student/teacher use.

1. CEPRAP Biotechnology Outreach. August 17, 2002: Invited by Barbara Soots to attended a meeting of ca.16 teachers who participate in CEPRAP’s biotechnology program for high schools. Presented a brief introduction to Arabidopsis and TAIR (http://acoma.stanford.edu/~lreiser/Davis.html).

2.SCATS (Schools and Colleges Advancing the Teaching of Science)workshop. September 24,2002.Co-presented a workshop with Barbara Soots for high school teachers looking to incorporate plant biology into their curricula. Presented a test project using Arabidopsis MADS-box mutants (seeds supplied by the ABRC and Dr. Martin Yanofsky). The rationale for using these particular mutants was that by studying these plants certain state standards could be addressed (e.g. evolution, plant reproduction, structure-function, genetics). The teachers are currently growing the plants and analyzing the phenotypes. Three follow-up workshops are planned for 2003.(http://acoma.stanford.edu/~lreiser/SCATS.html).

3.BEDROCK workshop. September 17-20, 2002. Participated in a BioQuest workshop for community college teachers who want to incorporate bioinformatics into their curriculum. Assisted Foothill College professor Kathleen Duncan in developing a module for her introductory biology class including a phylogenetic analysis of Arabidopsis and other cyptochromes.(http://bioquest.org/bedrock/sunnyvale_workshop/project5.htm).

4. A meeting has been proposed: Plant Biology Education MiniMeeting (projected date – location: Dec. 14, 2002- Berkeley, CA). The agenda includes: Presentation of existing projects, assessment of teacher and student needs, identifying tie ins to state and national science education standards, dissemination of projects/materials on a statewide or national scale, brainstorming new , identifying ways to collaborate and minimize overlap- thus maximizing limited resources.

Organizers:
Erin Dolan, Peggy Lemaux, Barbara Soots, and **Leonore Reiser**

Invited participants:
Paul Williams-Science House
Martin Chrispeels-UC San Diego
Sue Rhee-TAIR
Fred Hempel-Mendel Biotechnology
Ellyn Daugherty-San Mateo High School
Jennifer Aizenman-Dolan DNA Learning Center
Pat Seawell- BABEC
Madison West High Teachers
Gary Graper, Betsy Barnard, and Sara Patterson-UW Madison Dept of Horticulture
Dan McDonnell-LAUSD science coordinator
Rebecca Smith, UCSF's Science and Health Education Program
Ilse Ortabasi -Kindermagic
Julia Bailey-Serres-UC Riverside
Sherry Seethaler,-UC Berkeley Education Department
Dave Gilchrist-UC Davis
Doug Cook-UC Davis