– Review paper

Plant Tissue Cult. **14**(2) : 173-187, 2004 (December)

PTC

The *Arabidopsis* Information Resource : Impact for Identification of Important Genes in Less Known Plant Species

Greg Clark, Kanagasabapathi Sathasivan[1] and Ahmad S. Islam

Department of Molecular, Cell and Developmental Biology, University of Texas at Austin, Texas, 78712 USA

Abstract

This is the second article in the series on, "Valuable Internet Resources for Plant Molecular Biology Research" (Clark et al. 2003). In the last paper we mainly concentrated on the use of "The *Arabidopsis* Information Resource (TAIR)" for obtaining basic information about genes of interest, their location, structure and functions. Besides being a powerful research tool to *Arabidopsis* researchers, the goal of the TAIR website was to provide an easy-to-access database to the entire community of research scientists in biology helping them to solve research problems in their plants of interest including economically important plant species. We have shown in the current paper that the above objective, which prompted the launching of this website in 2000, has been realized admirably to the benefit of the entire biological research community. Giving examples, we have shown here how using the abundance of information available in this Super Resource Center, it is possible to retrieve all relevant information about important genes such as *4CL1* and *LOS1* and apply this knowledge in determining homologous regions controlling the same traits in the two jute species which yield fibers of commerce.

Introduction

In the past decade our understanding of plant biology has moved forward at an amazing speed because plant scientists chose *Arabidopsis thaliana* as a model system. This model plant is uniquely suited for an integrated approach to gain in-depth knowledge concerning genes and their functions both when they act individually or in a group as members of a gene family. The main reasons behind selecting this weedy plant as a model species is as follows: (a) from

[1]Author for correspondence.

seed-to-seed it is a two-month crop; (b) it can be grown easily under temperate conditions; (c) it produces a large number of seeds; (d) it has a small number of chromosomes (2n = 10) with a minimum number of repetitive DNA; (e) its genome is fully sequenced; (f) it has the most comprehensive reverse and forward genetic tools and resources that have enabled the *Arabidopsis* research community to study and begin to unravel functions of 60% of its 29,000 known genes; and finally (g) the research findings about the nature of individual genes and their functions can be extrapolated in solution of some of the most basic and applied problems in plant biology in the minimum possible time in a cost-effective manner.

In our previous review (Clark et al. 2003), we provided a general introduction and overview of TAIR (www.arabidopsis.org), one of the most important valuable Internet resources. Our goal in this review is to provide a more detailed description of some of the information sources and valuable tools for genetic analysis available to the worldwide community of biologists through the TAIR website.

An excellent starting point for those researchers not thoroughly familiar with the TAIR website is to choose the "help" option at the very top of the TAIR homepage and select the Quick-Start link (http://www.arabidopsis.org/help/quickstart.jsp) which will take the beginner to a general tutorial for using this website. There are extremely diverse types of information accessible through this resource from protocols and lab manuals to educational resources to maps of metabolic pathways.

It is important to realize that a scientist needs not work with *Arabidopsis* in the lab in order to benefit from the vast amount of genomic and proteomic information that are continuously generated by the *Arabidopsis* science community in various labs all over the world. The results generated by the *Arabidopsis* Functional Genomics Project have enhanced our insight in an unprecedented way. Central databases that are made publicly available, not only provide unparalleled insight into plant function by unraveling basic biological concepts, but also offer information about the genetic determinants that control important traits in this model plant and different crops in a similar fashion. Thus comparative studies on the physiology, biochemistry, and development of *Arabidopsis* with that of other plant taxa will reveal valuable information that may be fruitfully applied in the improvement of economically important agricultural crops. No wonder, why research on *Arabidopsis* is of paramount importance to our global society in general and developing countries in particular.

Funded by the National Science Foundation (NSF) and The Arabidopsis Functional Genomics Network (AFGN), the major aim of the Multinational Coordinated *Arabidopsis thaliana* Functional Genomics project, initiated by the Multinational *Arabidopsis* Steering Committee (MASC), has been the elucidation and understanding of the function of approximately 25,500 genes in *Arabidopsis* by 2010. In other words, the functional roles of more than 60% of all *Arabidopsis* genes are beginning to be defined. Such studies necessitated the identification of genes with T-DNA/Tn insertions and the establishment of a freely accessible transcriptome reference data set. There are 23,839 genes with T-DNA/Tn insertions (305 transposon/23,534 T-DNA; ca. 85%) available through TAIR as of the end of this year. In our first review we discussed the use of T-DNA insertional lines accessible through TAIR to study the function of an *Arabidopsis* gene of interest (Clark et al. 2003). Also available is almost 12,000 ORF clones useful for recombination cloning. The *Arabidopsis* Biological Resource Center (ABRC) is one among the sources that distribute the above material for recombination cloning. Databases with improved genome annotations have been expanded providing access to large functional data sets. Tools for global metabolic profiling reference data set, supported by a freely accessible transcriptome (all transcribed mRNAs within a cell) have been established so that researchers studying genomics can gain in-depth knowledge in functional biology.

The functional analysis of an *Arabidopsis* gene, once a researcher identi-fies a gene (e.g., by a particular expression pattern) may begin through the study of the corresponding T-DNA knock-out mutants available through TAIR (cf. previous PTC review article, Clark et al. 2003). Another great advantage enjoyed by current researchers is that they have free access to transcriptome data, initiated recently by Nottingham Arabidopsis Stock Center (NASC). Also available for the research community is arrays and the array facility and expression profile repository of the GARNET program in England. Another major development in this area is the recently released AtGenExpress reference transcriptome data set, compiled by a multinational initiative. It covers genome-wide transcript profiles of all major organs at various developmental stages, and in response to diverse environmental stimuli including abiotic and biotic stresses; and in response to phytohormones.

In our first PTC review article about the usefulness of Internet resources for gathering information about genes (Clark et al. 2003), we cited two examples, namely, *Phy B* and *Phy C.* By using TAIR, we were able to show that (a) these two genes are located on chromosome 2 and chromosome 5, respectively. Here we will use the same approach of using two more genes as examples in applying

TAIR analysis tools. First, we will describe an important gene, *4CL1* (4-coumarate:CoA ligase), which in conjunction with three other genes (cinnamoyl CoA reductase, caffeic acid O-methyltransferase and cinnamyl-alcohol dehydrogenase, synthesize lignin in woody and fibrous plants. Wen-Jing Hu and his colleagues at Plant Biotechnology Research Center, School of Forestry and Wood Products, Michigan Technological University have shown that by down-regulating *4CL1* gene that encodes 4-coumarate:Co A ligase (4CL) by antisense inhibition, it is possible to reduce levels of lignin by 45% percent concomitant with a 15% increase of cellulose level (Hu et al. 1999). They further showed that this change of lignin:cellulose ratio worked in a compensatory fashion without affecting the structural integrity of the poplar species. In fact, the growth of leaf, root, and stem increased significantly without affecting its structural integrity.

We got interested in the details of *4CL* gene because down-regulation of this gene is necessary in order to reduce levels of lignin for using cash crops such as jute for paper and pulp production. Except for a few (Sengupta and Palit 2004), hardly there is any information available on the molecular aspects of the jute genome, let alone the base sequence of *4CL* gene. We would like to show here how analysis tools available through TAIR can help us design experiments to clone the above gene from the two jute species, *Corchorus capsularis* and *C. olitorius* by giving researchers information on biochemical pathways for lignin synthesis (Cf. Fig. 1). It may be mentioned here that some significant progress by way of obtaining a good amount of mRNA in the two jute species has already been made in the lab of the second author (Kanagasabapathi Sathasivan) as a first step toward the construction of cDNA library (Farhan et al. 2004). Technical help from a USA-based company (Gene Arrays Inc.) has been sought to speed up the work of cDNA Library construction. Completion of this work is expected soon, paving the way for cloning the useful genes in the two jute species, taking powerful hints from TAIR and other important websites.

We will begin our search by typing the name of the gene, *4CL1* in the input box of the home page opposite TAIR "database" and click "Quick search". The query will retrieve two records, one protein and one locus. Click on the locus link which in turn displays a new page which summarizes the locus and gene details for *4CL1*. The locus identifier for *4CL1* is At1g51680 (At1 standing for *Arabidopsis thaliana* chromosome #1). When one clicks on "At1g51680", the locus detail page for *4CL1* is provided. This locus number is assigned by the AGI (Arabidopsis Genome Initiative) or TAIR curators and uniquely identifies a locus in the *Arabidopsis* genome. The locus detail page is a comprehensive

source of data for genome data in TAIR. In other words, the locus detail page is a 'container' for all of the information about the locus including Gene Ontology (GO) annotations, expression, and other gene and protein sequence and features. This locus detail page also includes additional information such as polymorphisms that are available, list of recent publications, gene structure, ESTs and cDNA sequences and functional information such as the role of the enzyme that this gene codes for in the last step of the general phenylpropanoid pathway. For example, from this locus detail page one can also click on the "View AraCyC information lignin biosynthesis" and a view of the general phenylpropanoid pathway which *4CL1* functions is provided (Fig. 1).

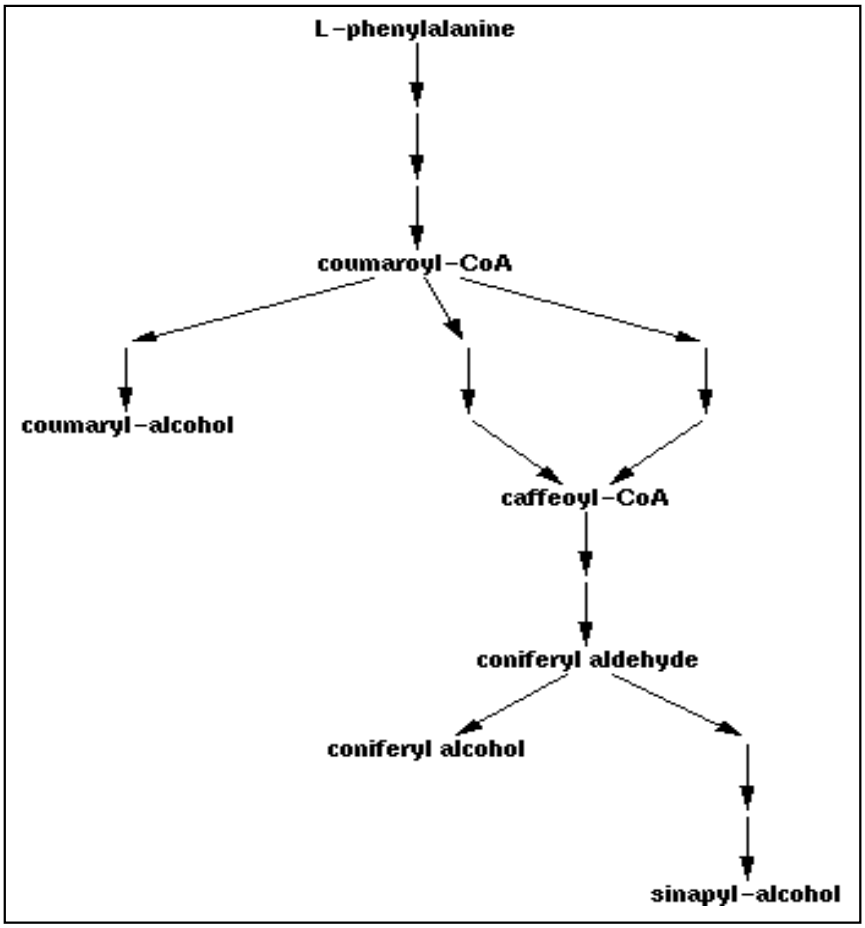

Fig 1. The second chemical in the flow chart is 4-coumaroyl CoA, the substrate for lignin and flavonoid biosynthesis. This is produced by 4-coumarate:CoA ligase *(4CL)*, a key regulatory enzyme in lignin biosynthesis.

AraCyc : *Arabidopsis thaliana* Biochemical Pathways

Another way to view this pathway is by directly using the AraCyC analysis tool which allows researchers to visualize many different biochemical pathways characterized for *Arabidopsis thaliana*. Starting at the TAIR home page click on "AraCyC pathways" under the heading "Analysis Tools" which will bring one to the AraCyC introduction page, (http://www.arabidopsis.org/tools/aracyc/). After arriving at this introduction page one can click on the option "Main Query Page". From this page one can enter the term lignin biosynthesis or the locus identifier for *4CL1* into the blank space highlighted in orange next to the word "Query" and the phenylpropanoid pathway as shown in Figure 1 will be provided. This main query page allows one to search for specific pathways, genes, proteins, or compounds. From the AraCyC introduction page there are two other options: "Metabolic Map" and "Omics Viewer". The Metabolic Map tool provides a "bird's eye" view of *Arabidopsis* metabolism. Before using this tool it is recommended to click on the "here" option for a Quick time demo that provides a basic tutorial for using this Metabolic Map. The Omics Viewer combines experimental data such as microarray expression, proteomics, with the metabolic profiling on the Metabolic Map. There is also a Quick time demo available for the Omics Viewer.

Using SeqViewer to Obtain Sequence Information

TAIR's SeqViewer allows users to view and search the whole *Arabidopsis* genome down to the nucleotide level and displays the exact location of genes, markers, transcripts, polymorphisms, T-DNA/Transposon transgene insertion flanking sequences, and clone ends. Starting at TAIR's home page, go to the second column captioned, "Analysis Tools" and click the first subheading, SeqViewer. In the search box that will appear underneath the caption Whole Genome View Options enter the name of the gene of interest, *4CL1*, by choosing, "Gene" in the drop down menu and click on the button, "Submit". The result is displayed on the whole chromosome view as a red tick mark on chromosome 1. Click on the red mark after entering the name of the gene or the locus identifier (At1G51680) on the top of the left hand input box, marked "Zoom to" and hit "Find". This opens the 200 Kb close up view centered on the gene highlighted in yellow. If you place your mouse on the gene highlighted in yellow, a pop up window will appear with a number of options. Click the last option, "nucleotide sequence view" and it will display the annotated nucleotide sequence including all gene details such as start and stop codons, introns, exons, untranslated regions. Forward and complementary strands can be seen by switching the option in the drop down menu on the right hand side of the page.

Alternately, click the highlighted locus identifier number and it will pull up the page entitled, "Gene Model", displaying several relevant subsections concerning the gene of interest, such as T-DNA flanks, DNA base sequence constituting both genomic (2801 N-bases) and cDNAs (1686 N-bases; the gene without the introns; Fig. 2), number of introns (non-coding regions; indicated by blank space), exons (coding regions). In addition, are shown ESTs (expressed sequence tags), or short cDNA sequences obtained from cDNA libraries representing a small segment of an entire gene.

```
   1  ATGGCGCCAC AAGAACAAGC AGTTTCTCAG GTGATGGAGA AACAGAGCAA
  51  CAACAACAAC AGTGACGTCA TTTTCCGATC AAAGTTACCG GATATTTACA
 101  TCCCGAACCA CCTATCTCTC CACGACTACA TCTTCCAAAA CATCTCCGAA
 151  TTCGCCACTA AGCCTTGCCT AATCAACGGA CCAACCGGCC ACGTGTACAC
 201  TTACTCCGAC GTCCACGTCA TCTCCCGCCA AATCGCCGCC AATTTTCACA
 251  AACTCGGCGT TAACCAAAAC GACGTCGTCA TGCTCCTCCT CCCAAACTGT
 301  CCCGAATTCG TCCTCTCTTT CCTCGCCGCC TCCTTCCGCG GCGCAACCGC
 351  CACCGCCGCA AACCCTTTCT TCACTCCGGC GGAGATAGCT AAACAAGCCA
 401  AAGCCTCCAA CACCAAACTC ATAATCACCG AAGCTCGTTA CGTCGACAAA
 451  ATCAAACCAC TTCAAAACGA CGACGGAGTA GTCATCGTCT GCATCGACGA
 501  CAACGAATCC GTGCCAATCC CTGAAGGCTG CCTCCGCTTC ACCGAGTTGA
 551  CTCAGTCGAC AACCGAGGCA TCAGAAGTCA TCGACTCGGT GGAGATTTCA
 601  CCGGACGACG TGGTGGCACT ACCTTACTCC TCTGGCACGA CGGGATTACC
 651  AAAAGGAGTG ATGCTGACTC ACAAGGGACT AGTCACGAGC GTTGCTCAGC
 701  AAGTCGACGG CGAGAACCCG AATCTTTATT TCCACAGCGA TGACGTCATA
 751  CTCTGTGTTT TGCCCATGTT TCATATCTAC GCTTTGAACT CGATCATGTT
 801  GTGTGGTCTT AGAGTTGGTG CGGCGATTCT GATAATGCCG AAGTTTGAGA
 851  TCAATCTGCT ATTGGAGCTG ATCCAGAGGT GTAAAGTGAC GGTGGCTCCG
 901  ATGGTTCCGC CGATTGTGTT GGCCATTGCG AAGTCTTCGG AGACGGAGAA
 951  GTATGATTTG AGCTCGATAA GAGTGGTGAA ATCTGGTGCT GCTCCTCTTG
1001  GTAAAGAACT TGAAGATGCC GTTAATGCCA AGTTTCCTAA TGCCAAACTC
1051  GGTCAGGGAT ACGGAATGAC GGAAGCAGGT CCAGTGCTAG CAATGTCGTT
1101  AGGTTTTGCA AAGGAACCTT TTCCGGTTAA GTCAGGAGCT TGTGGTACTG
1151  TTGTAAGAAA TGCTGAGATG AAAATAGTTG ATCCAGACAC CGGAGATTCT
1201  CTTTCGAGGA ATCAACCCGG TGAGATTTGT ATTCGTGGTC ACCAGATCAT
1251  GAAAGGTTAC CTCAACAATC CGGCAGCTAC AGCAGAGACC ATTGATAAAG
1301  ACGGTTGGCT TCATACTGGA GATATTGGAT TGATCGATGA CGATGACGAG
1351  CTTTTCATCG TTGATCGATT GAAAGAACTT ATCAAGTATA AAGGTTTTCA
1401  GGTAGCTCCG GCTGAGCTAG AGGCTTTGCT CATCGGTCAT CCTGACATTA
1451  CTGATGTTGC TGTTGTCGCA ATGAAAGAAG AAGCAGCTGG TGAAGTTCCT
1501  GTTGCATTTG TGGTGAAATC GAAGGATTCG GAGTTATCAG AAGATGATGT
1551  GAAGCAATTC GTGTCGAAAC AGGTTGTGTT TTACAAGAGA ATCAACAAAG
1601  TGTTCTTCAC TGAATCCATT CCTAAAGCTC CATCAGGGAA GATATTGAGG
1651  AAAGATCTGA GGGCAAAACT AGCAAATGGA TTGTGA
```

Fig. 2. Clicking highlighted *4CL1* in the "Gene Model" page retrieves the next page called "Locus: At1g51680." Somewhere in the middle of the page you will see two subsections: full length CDS, full length genomic under Nucleotide Sequence. Hit *4CL1* to see1686 N-bases comprising the cDNA of this gene.

Using GenBank Database to Search for *4C1* Homologues

In order to compare the *4CL1* gene of *Arabidopsis* with its counterparts in a few woody and fibrous species, the next approach will be to consult GenBank data for homology of this gene. Surprisingly, the gene under the same name *4CL* is present in a number of plants belonging to diverse groups such as rice, pine, aspen, tobacco and certain forage grasses (Spangenberg 2000) and legumes (Guo et al. 2001).

Next, using Clustal W, (see detailed instructions later in this article) an important molecular tool, multiple alignments between the *4CL* genes derived from the above species are done in order to determine the extent of homology between these species. The purpose of this exercise has been to identify conserved regions involved in lignin biosynthesis to design primers which could be used for amplifying cDNA clone by RT-PCR. This approach could be used to clone specific and related genes of interest. The cDNA sequence of *Arabidopsis* and aspen will help us locate the homologous regions involved in lignin biosynthesis in the two species of jute, once their cDNA data are in our hands.

Using TAIR to Obtain Information for the *LOS1* gene

As a second example gene, we will retrieve relevant information about, *LOS1* (low response to osmotic stress), an important gene that confers cold tolerance to *Arabidopsis*. For jute breeders this piece of information is very important because of the need to produce cold tolerant jute varieties suitable for planting in the northern part of Bangladesh, where summer season starts later than the rest of the country. Some accessions in the Gene Bank of Bangladesh Jute Research Institute (BJRI) have proved to be cold tolerant (Hossain et al. 2003). Using *LOS1* gene as a probe, it may be possible to identify the homologous region in the cold tolerant accessions of BJRI and clone it with the aim of producing transgenic lines of jute capable of growing where summer sets in late.

As in the previous example, type the gene symbol, *LOS1* in the input box of the home page opposite to "TAIR database" and hit the box "Quick Search". Hit the highlighted word, "locus". The locus identifier for *LOS1* (AT1G56070) along with other information such as description of the gene in terms of its function will be displayed in the next page. The numerical 1 after the first two letters AT indicates that this gene is also resident on chromosome 1. Next click on the locus identifier, AT1G56070, and as in the first example the locus detail page will appear and provide access to a more complete set of information including the full length genomic (3813 bps) and cDNA (2531 bps) sequences, and 363 ESTs. The locus detail page also provides a small description of function of

this gene in that it encodes a translation elongation factor 2-like protein that is involved in cold-induced translation. The page ends with a list of papers published about this gene.

In order to use SeqViewer to gather more detailed sequence information for the *LOS1* gene, repeat the process as described earlier for *4CL1*. This time also one will see a red tick mark on chromosome 1. Click on the red mark after entering the name of the gene *LOS1* or the locus identifier (At1G56070) on the top of the left hand input box, marked "Zoom to" and hit "Find". This opens the 200 Kb close up view centered on the gene highlighted in yellow. By placing your mouse on the gene highlighted in yellow, a pop up window will appear with a number of options, the last one being, "nucleotide sequence view". It will show a brief summary of the gene and a link to the Nucleotide Sequence View which displays the annotated nucleotide sequence. Then click on the locus name in the SeqViewer, this will take one back to the original locus page. Giving a summary the pop up ensures that you are looking at the same gene with which you had started your investigation. These are only couple of examples of how the molecular information available, can be utilized for further research. Following are some detailed instructions in searching the molecular data bases.

Using the Molecular Data Bases for Teaching and Research Purposes

Extensive amount of information is freely available in the Internet to search for genes and proteins of specific interest. It is essential that students and researchers learn about the utilization of these resources and understand what has been already done and we can utilize such knowledge for further research. Simple and detailed instructions are given here to help search for DNA and protein sequences and utilize them for multiple alignment and presentations. This information can be used to clone similar or same genes from crop plants and to study their expressions and potential applications in agriculture.

Identifying DNA or protein sequences:

1) Go to the NCBI website at http://www.ncbi.nlm.nih.gov/
2) Enter the sequence you are looking for in the search box and make sure the drop box next to the search box is set to "**nucleotide**" or **"Protein"** (the drop box is located immediately to the left of the search box) then click "Go". For example, if you are looking for the sequence of the UBQ10 gene in *Arabidopsis* you will enter "UBQ10 Arabidopsis" in the search box and click "Go".
3) This will bring up several selections. Click on the blue lettering of the selection you want. For example, after entering "UBQ10 Arabidopsis" you

can click on the **accession number** "NM_178970" (make sure that you click on one that has "**mRNA, complete cds**" in the title, meaning it is the complete mRNA sequence without introns).

4) The blue lettering is known as the "Accession number" and should be written down for future reference (rather than running a search for the desired sequence you can just enter the Accession number in the search box from "step 2").
5) Scroll down to the bottom of the screen and copy the sequence directly under "Origin" (including the **"//"**). Don't worry about the numbers in the sequence, they won't interfere with anything.
6) Repeat steps 2-5 until you have several similar sequences.

Creating a Multiple Alignment:

1) Go to the **ClustalW** site at http://clustalw.genome.jp/
2) Output format should be set to: **CLUSTAL**
3) Pairwise Alignment should be set to: **SLOW/ACCURATE**
4) The **DNA "bubble"** should be selected and not the PROTEIN bubble.
5) Title your sequences by first entering a **">"** followed by the desired title. For example, an acceptable title would be: **>ARAB UBQ10**
6) Hit enter so that the cursor is now directly below your title and paste your sequence making sure that at the end of the sequence there is a **"//"** (this can either be directly after the last amino acid/DNA sequence or on a new line. Numbers in the sequence is OK).
7) Repeat steps 5-6 until you have entered several different sequences you wish to compare.
8) Click "**Execute** Multiple Alignment".
9) SAVE THE **CLUSTALW.ALN** FILE TO THE DESKTOP
10) If you want phylogenetic tree, click on either the blue-colored "NJ-tree" or "Rooted Dendrogram" at the bottom of the clustallW report in step 8.
11) At the top of your CLUSTALW report, you will find percentages of amino acids in the sequence alignment that are identical when comparing only two sequences at a time. For example, if your report says "Sequences (1:2) Aligned. Score: 87.2", this means that when the first two sequences saved on your floppy were aligned, 87.2% of the amino acids were identical in both sequences.

Making the Multiple Alignment Easy to Read:

1) Go to the ESPript 2.2 site at http://prodes.toulouse.inra.fr/ ESPript/cgi-bin/ESPript.cgi
2) Click on the "Browse" button in the "Main alignment file" box and then select the **.aln** file you saved to the desktop.
3) Check the "number sequences" box.
4) Check the "display consensus" box under the "Similarity calculations" title (it's the 1st "required parameters" box).
5) Check the "Flashy" bubble under the "Output layout" title (it's the 2nd "required parameters" box).
6) Under the "Output file or device" title check the box next to the format you would like to view the alignment in (most common/preferable is the "**pdf file**").
7) Click on "SUBMIT" button (bright yellow button located at the top left of the page).
8) Click on "RES" located just to the right of the "SUBMIT" button (in between "DOC" and "IN").
9) Click on the blue lettering next to the output format you would like. Select **"pdf" format.**
10) Save this file to a desired location with a title you will recognize for future reference.
11) You can use this PDF file for presentations, papers or designing primers etc.

Using *Arabidopsis* Microarray Data Available at TAIR

As discussed earlier, one area of great advance in our understanding of the general roles of *Arabidopsis* genes comes from transcriptome data. Knowledge of the expression profile of a gene of interest provides important clues to the possible functions for that gene product. Additionally, knowledge of the expression profiles of many genes during development or in response to environmental cues greatly contributes to our understanding of the networks of proteins functioning in these processes. Microarray analysis is a method of analyzing changes in transcript abundance between two conditions for many genes simultaneously. This kind of expression profiling data for *Arabidopsis* is now available at the finger tips of researchers world-wide.

In this review we are going to highlight one of the tools available through TAIR to obtain gene expression data for *Arabidopsis*. The Microarray Element

Search tool allows scientists to visualize and mine cluster data from a small subset of microarray experiments (http://www.arabidopsis.org/tools/bulk/microarray/index.jsp). Although the Microarray Element Search can be used for finding out what arrays include which genes of interest and can serve as an entry point for getting this data the expression viewer is limited to only 579 out of the 3000 slides in TAIR. The Microarray Experiment Search tool can be used for finding raw and analyzed data for all microarray experiments stored in TAIR and essentially allows a researcher to find out what microarray data is available to determine gene expression under a given set of conditions (http://www.arabidopsis.org/servlets/Search?type=expr&search_action=new_search). Another search tool available is the Microarray Expression Search tool which can be used for finding the expression levels of a gene or set of genes of interest in all or a subset of hybridizations stored in TAIR (http://www.arabidopsis.org/servlets/Search?action=new_search&type=expression).

To use the Microarray Element Search tool, start at the home page for TAIR and then simply click on "Microrray Analysis" under the heading "Analysis Tools" and you will arrive at the introduction page which provides basic information. Using the *4CL1* gene as an example, one can further examine the expression profile for this gene by clicking on the "Microarray Elements Search tool" option at the bottom of the page. Then by simply entering the accession number for the *4CL1* gene on this page the researcher has accessed microarray datasets for this gene. Another option for obtaining these same datasets is starting from the main locus link page for *4CL1.* Under the heading "RNA data" one can click on "Viewer" under the subheading expression viewer. This will bring you to a page titled "Expression Analysis Across All the Experiments" from which one can pick from different datasets with clustered data. In this example, *4CL1*, data is clustered according to chemical treatment, ecotype comparison, or hormone treatment. Simply click on one of the datasets and hit the submit button and one can now view the microarray data for that particular treatment. At the top of the page there will be a color scale indicating fold change between two conditions with different shades of green representing a certain fold level of repression and shades of red representing induction level. Also, near the top of the page is a "Help" button which provides needed introductory information before trying to interpret the microarray data provided below.

In order to examine expression levels of the *4CL1* gene between different *Arabidopsis* ecotypes click on ecotype expression and submit, then scroll down the new page titled "Expression of G9F10T7 in response to various Ecotype Comparisons". On the left side one will see the accession number for the *4CL1*

gene and toward the right one sees black or gray as the color scale indicating that there were no differences in transcript levels for that ecotype, or no data was available for that ecotype, respectively. So, interestingly, it appears that the expression of this enzyme is unchanged when compared among different ecotypes. Next we will examine expression of *4CL1* after plants are exposed to different hormone treatments by clicking on hormone treatments on the previous page. Notice that when the dataset for hormone treatments are retrieved at the bottom of the page there is expression data for a total of 20 genes with the *4CL1* gene locus identifier at the top of the list of locus identifier numbers. The other 19 loci identifiers represent genes which share the most similar expression profile with the *4CL1* gene within that treatment group. Knowing the genes that share a similar expression profile can sometimes be helpful in understanding gene functions. For example, it would be very interesting if one of the other locus identifiers showing a similar expression profile to the *4CL1* gene for the hormone treatment group matched one of the other three enzymes discussed earlier that function together to synthesize lignin. In examining the microarray data clustered for hormone treatments there does not appear to be a consistent effect on *4CL1* expression by any of the specific hormone treatments.

By clicking on "chemical treatment" we can view more microarray data for the *4CL1* gene from the G9F107 array element. In this case there is only data for the *4CL1* gene but as one scans across the data it appears that the 4CL1 gene is induced slightly more than 2-fold after 30 minutes and 1 hour in a chitin elicitation timecourse study. This apparent induction of *4CL1* expression by chitin treatment is worth taking a closer look at because there is consistency in the results as reported by 3 separate hybridizations within the timecourse. Thus this finding warrants obtaining further information and by clicking on the hybridization name of the experiment "CB1-Chitin Timecourse 30 min A" one is provided with a hybridization collection detail page which allows one to see the corresponding experimental details. This page initially provides information about the slides and datasets, for example, it appears that the slides represent biological replicates but that there are no reverse replicates available and that the experimental variables are water treatment versus chitin treatment. By clicking on "Experiment Summary" the name of the authors, a description of the experiment and relevant publications are provided.

There are several important things to consider when using these TAIR microarray datasets as a resource. There are two types of replicates used in performing microarray analysis, biological replicates and technical replicates. Biological replicates are experimental repeats that are derived from separate

biological samples, for example, the initial source for RNA that is being used for comparison has come from different isolations from different plants. On the other hand, technical replicates are experimental repeats that use probes made from RNA of the same isolation but different reverse transcription reactions or one reverse transcription reaction used for several different probes. It is important to have data for which there are dye swaps between Cy3 and Cy5 labels, called reverse replicates, because the two dyes have different degradation rates. Ideally, data is more trustworthy when one can access multiple biological and technical replicates in addition to the reverse replicate data. Expression data mined from microarrays can be made stronger by confirming the expression pattern using an alternative method such as Northern blot analysis or quantitative RT-PCR. With an awareness of these complexities the expression data available through TAIR can be used as a strong indicator of a gene's expression pattern which can then be used as a guide in experimental design.

Conclusion

Thanks to the enormous amount of efforts by TAIR. It has assembled and catalogued a multitude of valuable information generated as a result of the successful completion of the first phase of the Multinational Coordinated *Arabidopsis thaliana* Genome Research Project. The wealth of information generated by TAIR in a user friendly manner has become extremely useful to the worldwide research community of biologists, particularly those who are working in various fields of molecular biology. The data provided by TAIR have made it abundantly clear the immense complexity of the plant biological system and the difficulty to unravel the function of every gene in the system.

In spite of this complexity, *Arabidopsis* research has provided valuable information about the structure of important genes and their respective function. Such cutting edge resources and analytical tools developed through collaborative worldwide efforts have now made it possible to work with other plant species for identifying the genes of agronomic importance. In this article we have tried to demonstrate that once the recently initiated construction of cDNA library of the two jute species is completed, TAIR resources will enable the workers in this field to identify two among many important gene systems, namely *4CL1* and *LOS1* involved in the last step of lignin biosynthesis and in imparting cold tolerance, respectively. It will open up a new vista for molecular breeders to produce commercial jute varieties with lower levels of lignin and varieties with cold tolerance and other traits of agronomic importance.

Acknowledgment

The authors are grateful to Dr. Seung Yon (Sue) Rhee, Ph.D., TAIR Director, and Dr. Leonore Reiser, Ph.D., TAIR Curator, for critically reading the article and offering many helpful suggestions. They would also like to thank Farhan Khan at The University of Texas for his technical contributions and suggestions.

References

Boerjan W, Ralph J and **Baucher M** (2003) Lignin biosynthesis. Annu Rev Plant Biol. **54**: 519-46.

Clark G, Islam AS and **Sathasivan K** (2003) Valuable Internet Resources for Plant Molecular Biology Research, Plant Tissue Cult. **13**: 85-97.

Gene Arrays Inc. 201 Perry Parkway, MD 20877, USA. GA offers high-quality customized cDNA and Genomic libraries for specific needs of fellow researchers.

Guo D, Chen F, Blount J, Inoue K and **Dixon RA** (2001). Down-regulation of caffeic acid 3-*O*-methyltransferase and caffeoyl CoA 3-*O*-methyltransferase in transgenic alfalfa (*Medicago sativa* L.). Effects on lignin structure and implications for the biosynthesis of guaiacyl and syringyl lignin. Plant Cell **13**: 73-88.

Guo Yan, Liming Xiong, Manabu Ishitani and **Jian-Kang Zhu** (2002) An *Arabidopsis* mutation in translation elongation factor 2 causes superinduction of CBF/DREB1 transcription factor genes but blocks the induction of their downstream targets under low temperatures. Proc Natl Acad Sci U S A **99**(11): 7786–7791.

Hossain M Belayat, Aleya Awal, Mohammad Aminur Rahman, Samiul Haque and **Haseena Khan** (2003) Distinction between cold sensitive and cold tolerant jute by a combined RAPD and AFLP study. J. Biochem. and Mol. Biol. **36** (5)427-432; (BSRK and Springer-Verlag).

Hu WJ, Harding SA, Lung J, Popko JL, Ralph J, Stokke DD, Tsai CJ, and **Chiang VL** (1999) Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nature Biotechnology **17**: 808-812.

Huala Eva et al (2001) The *Arabidopsis* Information Resource (TAIR): A comprehensive database and web-based information and retrieval system for a model plant. Nucleic Acid Res. **29**: 102-105

Khan Farhan, Ahmad Islam and **Kanagasabapathi Sathasivan** (2004) A Rapid Method for High Quality RNA Isolation from Jute: *Corchorus capsularis* L. and *Corchorus capsularis* L. Plant Tissue Cult. **14**(1): 63-68.

Sederoff R (1999) Building better trees with antisense. Nature Biotechnology **17**: 750-751.

G. Sengupta and **P. Palit** (2004) Characterization of a lignified secondary phloem fibre-deficient mutant of jute *(Corchorus capsularis).* Ann. Bot. **93**(2): 211 - 220.

Spangenberg German, Karen Baera, Amanda Bartkowski, Robyn Heath, Helen Huxley, Angela Lidgett, Damian Lynch, Russell McInnes and **Jutta Nagel** (2000) Isolation and characterization of lignin biosynthesis from forage grasses. International Plant and Animal Genome (VIII) Conference (abstract).