The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis


The DIURNAL project (http://diurnal.cgrb.oregonstate.edu/) provides a graphical interface for mining and viewing diurnal and circadian microarray data for *Arabidopsis thaliana*, poplar, and rice. The database is searchable and provides access to several user-friendly Web-based data-mining tools with easy-to-understand output. The associated tools include HAYSTACK (http://haystack.cgrb.oregonstate.edu/) and ELEMENT (http://element.cgrb.oregonstate.edu/). HAYSTACK is a model-based pattern-matching algorithm for identifying genes that are coexpressed and potentially coregulated. HAYSTACK can be used to analyze virtually any large-scale microarray data set and provides an alternative method for clustering microarray data from any experimental system by grouping together genes whose expression patterns match the same or similar user-defined patterns. ELEMENT is a Web-based program for identifying potential cis-regulatory elements in the promoters of coregulated genes in *Arabidopsis*, poplar, and rice. Together, DIURNAL, HAYSTACK, and ELEMENT can be used to facilitate cross-species comparisons among the plant species supported and to accelerate functional genomics efforts in the laboratory.

INTRODUCTION

Various databases containing gene expression data and Web-based tools for analyzing microarray data have emerged as valuable resources for many aspects of plant research. These resources include AtGenExpress (Schmid et al. 2005; http://www.weigelworld.org/resources/microarray/AtGenExpress/), ArrayExpress (Brazma et al. 2003; http://www.ebi.ac.uk/arrayexpress/), Botany Array Resource (BAR; Toufighi et al. 2005; http://www.bar.utoronto.ca/), GENEVESTIGATOR (Zimmermann et al. 2004; https://www.genevestigator.ethz.ch/), GEO (Barrett et al. 2007; http://www.ncbi.nlm.nih.gov/geo/), NASC Arrays (Craigon et al. 2004; http://affymetrix.arabidopsis.info/), PlexDB/Barleybase (Shen et al. 2005; http://www.plexdb.org), TAIR (Garcia-Hernandez et al. 2002; Rhee et al. 2003; http://www.arabidopsis.org/), and VirtualPlant (http://virtualplant.org). Here, we introduce the DIURNAL project, consisting of a database containing diurnal and circadian expression data for approximately 22,800 *Arabidopsis* genes collected during 11 experiments, a Web interface for accessing this database, and complementary Web-based tools for analyzing microarray data and promoter sequences.

DIURNAL and its associated tools differ from other resources in several ways. DIURNAL is focused on plant-diurnal- and circadian-clock-regulated gene expression and currently supports *Arabidopsis*, rice, and poplar, three plant species with high-quality annotated genome sequences (*Arabidopsis* Genome Initiative 2000; International Rice Genome Sequencing Project 2005; Tuskan et al. 2006). A novel tool called HAYSTACK allows users to mine large microarray expression data sets by searching for specific user-defined patterns of expression. HAYSTACK is designed to find rare occurrences of very specific patterns in large data sets and provides an alternative method for clustering microarray data by grouping genes whose expression patterns match the same or similar HAYSTACK patterns. In addition to clustering genes based on the user-supplied model patterns, HAYSTACK can be used to identify genes that exhibit a pattern of expression similar to that of a particular gene of interest. Although we have used HAYSTACK in the DIURNAL project to identify diurnal and circadian regulated genes, this tool can be used to compare any large-scale data set representing at least three samples (e.g., treatments, genotypes, and time points) against a set of user-supplied model patterns. The resulting lists of coregulated genes in the HAYSTACK output can then be used with a third tool, ELEMENT, which is an enumerative promoter analysis program that analyzes the upstream regions of *Arabidopsis*, poplar, or rice genes to discover overrepresented elements that may represent novel transcription-factor-binding sites. Together, the DIURNAL suite of tools allows researchers to go from expression data to cis-regulatory elements and place this into the context of diurnal and circadian control, an important layer of regulation representing at least three samples (e.g., treatments, genotypes, and time points) against a set of user-supplied model patterns. The resulting lists of coregulated genes in the HAYSTACK output can then be used with a third tool, ELEMENT, which is an enumerative promoter analysis program that analyzes the upstream regions of *Arabidopsis*, poplar, or rice genes to discover overrepresented elements that may represent novel transcription-factor-binding sites. Together, the DIURNAL suite of tools allows researchers to go from expression data to cis-regulatory elements and place this into the context of diurnal and circadian control, an important layer of regulation in *Arabidopsis*, with 90% of genes exhibiting diurnal or circadian regulation under at least one growth condition (T. Michael et al., in prep.).
THE DIURNAL DATABASE

Organisms experience daily environmental changes in light (photocycles) and temperature (thermocycles) that vary by season and latitude. Consequently, organisms have evolved an endogenous circadian clock with a period of about 24 hours, which ensures that internal biological processes are appropriately synchronized with the daily changes in the environment (Michael et al. 2003; Woelfle et al. 2004; Dodd et al. 2005). Transcript abundance can be found peaking at almost every hour during the day/night cycle, and this regulation forms the foundation for time-of-day-specific biological activities.

The Arabidopsis microarray data available in the DIURNAL database was collected using the Affymetrix ATH1 GeneChip microarray platform, which represents about 22,800 genes as annotated by TAIR (ftp://ftp.arabidopsis.org/home/tair/Microarrays/Affymetrix). To identify cycling genes, we used these 336 model patterns with HAYSTACK to interrogate Affymetrix ATH1 GeneChip data sets for 11 time courses with 12 time points each. These data are presented and analyzed in a manuscript that has been submitted for publication (T. Michael et al., in prep.). Plant material and growth conditions for the Arabidopsis time courses are described at http://diurnal.cgrb.oregonstate.edu/diurnal_details.html.

Interface Design, Features, and Navigation

Layouts, fonts, color schemes, and navigation were designed to make the interface user-friendly. The main page of DIURNAL contains a link to an “About Diurnal” page that summarizes the project. This page has a menu bar with links to the other tools of the project: ELEMENT and ORTHOMAP. An additional tool, ORTHOMAP, provides predicted orthologs, homologs, or simple best BLAST matches for Arabidopsis, rice, and poplar and links to the respective external annotation resources at TAIR, TIGR, or JGI. This tool allows a user to obtain this information for a gene or group of genes of interest, with or without accessing different databases or conducting sequence comparisons. The menu bar and links appear on every page to enable efficient browsing between tools. In addition, the tools are connected in such a way that, for example, a list of genes identified by a query of DIURNAL or a list of predicted orthologs identified by a query of ORTHOMAP can be sent to ELEMENT for promoter analysis. The databases underlying DIURNAL, ELEMENT, and ORTHOMAP are based on MySQL, and the Web interfaces have been implemented using the Apache HTML delivery engine (http://www.apache.org) Perl and Mason (http://www.masonhq.com).

DIURNAL: A Diurnal/Circadian Gene Expression Data-mining Tool

DIURNAL is useful for a researcher who is interested in the diurnal or circadian expression profiles of a particular set of genes, either for diurnal/circadian experiments or for planning other types of experiments. The interface allows users to input a list of genes and select among the experiments in the database. The data resulting from the query is displayed graphically and available for download as a text file. In addition, we enable direct exporting of gene lists resulting from queries into ELEMENT for promoter analysis.

The entry page of DIURNAL includes a text box for entering/CDS identifiers and check boxes for selecting experimental conditions/array data sets. This page represents the “Basic Search,” which has options for displaying the best-fitting HAYSTACK model and the expression profiles for multiple genes on a single graph, and for normalizing the expression levels among multiple genes being graphed. The normalization option makes it easier to compare the phases and waveforms of expression for sets of genes with large differences in expression levels. The basic search returns a graph or graphs of the diurnal or circadian expression profiles for the gene(s) submitted by the user (Fig. 1A). Below the graph is a table that summarizes the condition, phase, best-fitting model, and correlation value representing the quality of the match between the model and the experimental data. Above each graph are links to the raw text data and to the annotation page for the particular gene. By clicking on the graph, the user is presented with a larger version of the graph that can be downloaded in PNG image file format.

The left side of all pages within DIURNAL contains a menu box with a link to “Advanced Search.” On the Advanced Search page, a user can query the database to return all genes matching a set of user-defined criteria. The user can define the phase, model, and correlation cutoff values and use a pulldown menu to select an experimental condition to query. Results are represented as a table listing genes matching the query criteria (Fig. 1B). Each row of the table represents a single gene and the fields of the table contain the Affymetrix probe-set identifier, the locus identifier, correlation value, phase of expression based on the best-fitting HAYSTACK model, and the name of the best-fitting model. The rows are linked to a detailed graphical display of the time course expression profile for the particular gene. As in the basic search, the graph image and raw data are downloadable, and a link is provided to the respective annotation database. An advanced search allows the user to cluster cycling probe sets/genomes by their phase or waveform (best model) and thus can generate gene lists suitable for input to ELEMENT or another promoter analysis algorithm. To facilitate such analysis, we have provided an option to send the gene list resulting from a DIURNAL advanced search directly to ELEMENT for prediction of potential cis-regulatory elements.

When a user enters a list of genes and chooses the desired options, the microarray data are selected for display by looking up the appropriate Affymetrix probe-set identifier in a table within our database. There is not always a unique relationship between the Affymetrix probe-set identifier and a locus identifier, so more than one row of the table may correspond to the same gene, but representing data from different probe sets. The data generated by the query is displayed on the appropriate results pages, depending on the options chosen and whether a basic or advanced search was performed.

Another output option of DIURNAL is a tab-delimited
Figure 1. DIURNAL. (A) Example of the output of the DIURNAL Basic Search showing a graph of a gene expression profile for AT1G01060 under the short-day condition with a summary of model match statistics. (B) Example of the HTML table output from the DIURNAL Advanced Search showing genes and model match statistics fitting the search criteria. Clicking on a row of the table displays a graph for the gene and its best-fitting HAYSTACK model.
plain text file that contains the expression profile data for queried probe sets/genes. The data presented in this file are the same data as those used to generate the graphs, with each row of data representing a single gene or probe set. The data columns list the probe-set identifier, locus identifier, best model, the experimental condition/data set, and correlation, whether the row is data (DATA) or fitted model (MODEL), followed by the series of expression values for the time course.

DIURNAL may be used in several different ways by a biologist. First, the tool can be used to characterize the diurnal or circadian temporal expression pattern of a gene of interest, especially by querying the diurnal conditions that approximate standard laboratory conditions such as short days or long days. Second, it can be used to identify genes coexpressed and potentially coregulated with a gene of interest. This type of analysis can suggest potential protein interactions as well because it has been demonstrated that gene expression among interacting proteins can co-evolve (Fraser et al. 2004). In vivo experiments altering gene expression or using the two-hybrid system can further elucidate potential interactions among the genes identified. Finally, DIURNAL provides a method of clustering coexpressed genes whose promoters may be analyzed with either ELEMENT or other algorithms.

Because the DIURNAL interface can accept lists of multiple genes for comparison, it can also be used to predict functional redundancies among members of a gene family. For example, a user could submit a list of gene identifiers for a family of genes and use the DIURNAL output to classify them according to their temporal expression patterns. Example of such analyses are shown in Figure 2. Members of the C2C2-YABBY transcription factor family function to specify abaxial cell fate in Arabidopsis (Siegfried et al. 1999). A query of DIURNAL using the gene identifiers for the six members of the C2C2-YABBY TF family reveals diurnal expression profiles consistent with functional redundancy among some family members (Fig. 2A). Five out of six family members are represented on the Affymetrix ATH1 microarrays, and three out of these five YAB genes share similar expression profiles, with peak expression occurring just before dawn, at phases 23 and 22, respectively. The similar phasing of expression of these two family members could indicate functional redundancy at this time of day. In contrast, YABBY5 (At2g26580) cycles with a phase of 16 hours after dawn, whereas CRC (AT1G69180) and YAB4 (AT1G23420) do not cycle. Another example involves an 11-gene family of trehalose-6-phosphate synthases involved in biosynthesis of trehalose, a sugar whose accumulation is implicated in drought stress tolerance (Leyman et al. 2001; Karim et al. 2007). All 11 family members are represented on the ATH1 microarrays. In short days, 7 out of 11 cycle and 6 of these are phased to 18–19 hours after dawn (Fig. 2B). In thermocycles (LLHC), 8 out of 11 cycle and 4 of these are phased to 18–19 hours after dawn (Fig. 2B). In thermocycles (LLHC), 8 out of 11 cycle and 4 of these are phased to 18–19 hours after dawn (Fig. 2B). In thermocycles (LLHC), 8 out of 11 cycle and 4 of these are phased to 18–19 hours after dawn (Fig. 2B). In thermocycles (LLHC), 8 out of 11 cycle and 4 of these are phased to 18–19 hours after dawn (Fig. 2B). In thermocycles (LLHC), 8 out of 11 cycle and 4 of these are phased to 18–19 hours after dawn (Fig. 2B). In thermocycles (LLHC), 8 out of 11 cycle and 4 of these are phased to 18–19 hours after dawn (Fig. 2B).

HAYSTACK (http://haystack.cgrb.oregonstate.edu/) is a program for identifying genes whose expression levels behave similarly across all samples in a microarray data set. Multiple approaches have been developed for organizing and inferring patterns emerging from microarray data (for review, see Belacel et al. 2006). Conventional microarray-clustering approaches are based on identifying distinct or separable groups of genes based on a distance metric (Hierarchical, K-Means, Self-organizing Maps, Support Vector Machines), or principal component analysis. In contrast, our approach is hypothesis-driven and depends on predefined models to identify statistically similar groups of coexpressed genes. By identifying rare occurrences of specific biologically relevant expression patterns in the experimental data, we are able to dramatically reduce the search space for subsequent analyses, including promoter analysis to identify important cis-regulatory elements.

HAYSTACK uses a pattern-matching algorithm to identify genes whose expression patterns fit a user-defined model. HAYSTACK is designed to find rare occurrences of very specific patterns in a large data set and provides an alternative method for clustering microarray data, by grouping genes whose expression patterns match the same or similar HAYSTACK patterns. The algorithm is based on determining the Pearson correlation coefficient between gene expression profiles and user-defined models. HAYSTACK determines the correlation of an experimental data series with each supplied model pattern and applies a series of statistical tests and ad hoc filters to identify genes of interest and their corresponding best-fitting model. We have used HAYSTACK to compare microarray time course data against a collection of diurnal/circadian models to identify cycling genes. The Web version of HAYSTACK (http://haystack.cgrb.oregonstate.edu/) can be used to compare a large-scale data set against a set of user-supplied model patterns to search for biologically relevant patterns in the data.

We are interested in time-of-day specific and circadian transcriptional networks; therefore, we have focused on time course data to highlight the simplicity and power of HAYSTACK. We developed multiple cycling patterns based on diurnal and circadian time courses available in the literature: asymmetric, rigid, spike, cosine, sine, and/or box-like patterns (Harmer et al. 2000; Smith et al. 2004; Blasing et al. 2005; Edwards et al. 2006). To capture both cycling and phase information in the time course data, HAYSTACK patterns were used to mine data from 11 Arabidopsis time courses. The resulting analyses are available through the DIURNAL interface.

To use HAYSTACK, a user starts by uploading a file containing the model patterns (which can be easily constructed using a text editor or Microsoft Excel) and a file containing the microarray data of interest, arranged as a data series in the same format as the model patterns. The user then selects optional statistical criteria and ad hoc filters, and HAYSTACK calculates the correlation coefficient between the expression values across the microarray.
Figure 2. (A) Graph and summary table of diurnal expression profiles of the C2C2-YABBY transcription factor family in long days. (B) Graph and summary table of diurnal expression profiles of the trehalose-6-phosphate synthase family in short days. For clarity, expression profiles of genes that do not cycle in this condition have been omitted.
The data set for that gene and each of the user-supplied models. HAYSTACK is not limited to any particular microarray platform, genome annotation, or even expression values. It can be used to compare any large-scale data set representing at least three samples (e.g., treatments, genotypes, and time points) against the user-supplied model patterns. A minimum of three data points are necessary for the program to perform valid correlation calculations between the model and experimental data series. The program returns to the user (via e-mail) a link to a results file containing those genes satisfying the user-specified criteria (Fig. 3A). The results from HAYSTACK may be viewed in additional ways. First, the results can be passed directly to a plotting program on the HAYSTACK Web site for easy visual inspection of the data series and corresponding best-match model (Fig. 3B). Alternatively, the results are available in a text file format that may be downloaded and viewed separately.

HAYSTACK can be applied intuitively in several different ways by a biologist. For example, to generate leads for further characterization in the laboratory, a researcher may want to identify all genes in a microarray data set that exhibit an expression profile similar to a particular pattern. Alternatively, a researcher may want to identify genes coexpressed with a particular gene of interest. In this case, the supplied model pattern could be the expression profile of the gene of interest. A third use for HAYSTACK is to seed promoter analysis using ELEMENT or other promoter element discovery programs such as Promomer (Toufighi et al. 2005), SIFT (Hudson and Quail 2003), or MotiSampler (Thijs et al. 2001, 2002). An interesting potential use of HAYSTACK
involves using it to search nonmicroarray data such as the significance statistics for potential cis elements generated by enumerative promoter-searching tools such as ELEMENT. We used this approach to search the serialized Z-scores for overrepresented words identified in our diurnal/circadian studies and thus identified co-occurring elements that form the basis of diurnal/circadian transcriptional network modules (see Fig. 5 below) (T. Michael et al., in prep.).

Other Web-based bioinformatics tools such as the “Expression Angler” or “Sample Angler” at BAR (Toufighi et al. 2005) can identify genes that respond similarly across samples (i.e., genes whose expression profiles are highly correlated), whereas HAYSTACK provides a model-based method for clustering microarray data—genes that are returned with best matches to the same model are potentially coregulated. Some qualifications must be made regarding the results from HAYSTACK, and some of the user options are intended to overcome these shortcomings. For example, spurious model/data matches involving data series in which all of the points are below the background level (noise) are possible. To address this possibility, we implemented a “Background Cutoff,” which is the minimum acceptable value for the highest value in the data series. Another possible basis for spurious matches involves situations in which a data series is highly correlated with a model, but the difference between the maximum and minimum values in the data series is insignificant. This problem arises because the Pearson correlation is amplitude-independent. We address this issue by providing a “Fold Cutoff” option that allows a user to determine the minimum acceptable fold difference (i.e., max/min) between the maximum and minimum values in the data series. These options provide flexibility for users to decide what parameters make sense in the context of their experiments.

ELEMENT: A TOOL FOR IDENTIFYING POTENTIAL CIS-REGULATORY ELEMENTS IN PLANTS

The regulation of gene expression in eukaryotes is largely mediated by transcription factors (TFs) that bind within regulatory regions (promoters) upstream of the coding sequence. Transcription factors recognize specific DNA motifs, bind, and in turn interact with each other and the basal transcriptional machinery to regulate the expression of adjacent genes. With the recent availability of high-quality sequenced and annotated genomes, large public microarray databases, and easy access to microarray technology for individual laboratories, there is a need for bioinformatics tools to predict components of transcriptional networks, including transcription-factor-binding sites. Groups of coexpressed genes identified using microarrays may be coregulated and thus can form the foundation for analyses of promoter sequences to identify important cis-regulatory elements.

A number of algorithms have been developed to identify known and putative regulatory elements in the promoter sequences of coregulated genes (for review, see Rombauts et al. 2003; Tompa et al. 2005). The fundamental assumption underlying all of these computational approaches is that coregulated genes should contain similar regulatory motifs in their promoters, and these motifs should be significantly overrepresented in the set of coregulated promoters. There are two general computational approaches for identifying potential cis-regulatory elements. One approach is an enumerative method, and the other is an alignment method. The alignment methods are exemplified by programs that use a Gibbs sampling method (Thijs et al. 2002).

The enumerative methods estimate the probability of occurrence of short DNA sequences, or “words,” by comparing the count in a set of coregulated sequences to an expected count based on random sampling or statistical modeling of a background distribution (van Helden et al. 1998; Hudson and Quail 2003; Kreps et al. 2003; Marino-Ramirez et al. 2004; Nemhauser et al. 2004; Koussevitzky et al. 2007). Therefore, each algorithm requires some background model to calculate an expected frequency for each word. The composition of the sequences underlying the background model is critical because the various features (e.g., exons, introns, and intergenic regions) within a genome exhibit different oligomer compositions. Both enumerative and alignment approaches have been applied to promoter analysis in plants, and the putative coregulated sequences chosen for analysis were typically derived from hierarchical clustering or other analyses of microarray data (Harmer et al. 2000; Chen et al. 2002; Hudson and Quail 2003; Hulzink et al. 2003; Nemhauser et al. 2004; Koussevitzky et al. 2007).

The goal of the ELEMENT program is to provide a user-friendly Web-based tool that uses the enumerative method to identify statistically overrepresented 3–8 mer DNA words in a group of coexpressed genes in Arabidopsis, rice, or poplar. An earlier version of ELEMENT (Nemhauser et al. 2004) calculated Z-scores for each DNA word based on a comparison of the number of occurrences of that word in the upstream sequences of the coexpressed genes against a background distribution derived by random sampling of the upstream sequences of all genes represented on the microarray. The current Web-based version of ELEMENT has been improved considerably. It supports Arabidopsis, poplar, and rice and allows a user to choose various promoter lengths for analysis and to apply a false discovery rate (FDR) filter as desired (Benjamini and Hochberg 1995; Storey and Tibshirani 2003). There are currently a few other Web-based enumerative promoter analysis tools for Arabidopsis, such as the BAR Promoter (Toufighi et al. 2005), SIFT (Hudson and Quail 2003), and TAIR’s motiffinder (http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp). Besides ELEMENT, we are aware of no other Web-based bioinformatics tools for analysis of rice and poplar promoters.

The ELEMENT platform consists of databases of putative Arabidopsis, rice, and poplar regulatory DNAs, word statistics for all 3–8 mer DNA words occurring in these promoter sequences, software implemented in Perl to analyze promoters and apply statistical screening criteria, a series of accessory scripts to summarize the results of these analyses, and a Web interface implemented in Mason and HTML. ELEMENT uses a database of precalculated statistics for 3–8 mer words in the promoters of all genes repre-
SENT on the *Arabidopsis*, rice, and poplar Affymetrix arrays to estimate the Z-score, a measure of the distance in standard deviations of a sample from the mean, for each word. The observed frequency of a particular 3–8 mer word in a group of promoters is compared with the expected frequency of that word derived from the promoters of randomly sampled genes represented on the Affymetrix microarrays for the species, thus providing a background model based on sequences of an appropriate composition for the species and promoter length under consideration. For example, the background model statistics for *Oryza sativa* ssp. *Japonica* are derived from the frequencies of all 3–8 mer words in the upstream sequences of 34,967 nontransposable-element-related *Japonica* rice genes represented on the Affymetrix rice microarrays.

ELEMENT takes as an input a list of standard locus identifiers for the respective species. We have also provided direct links from DIURNAL and ORTHOMAP so that lists of coexpressed genes or groups of predicted orthologs/homologs may be sent directly to ELEMENT for analysis. The user selects the species and promoter length to be analyzed and adjusts the false-discovery rate and minimum number of occurrences for overrepresented elements as desired. ELEMENT uses, as its reference sets, the 500 bp, 1 kbp, 2 kbp, and 3 kbp upstream of annotated gene models of *Arabidopsis*, rice, and poplar. Because ELEMENT relies on genome annotations, and the transcription start sites are not always well annotated, some gene models lack annotated 5'-untranslated regions (5'UTRs), and thus, in some cases the 3' end of the predicted promoter region is defined as the beginning of the open reading frame.

ELEMENT returns to the user (via e-mail) a link to results files containing (1) a table of overrepresented elements and their corresponding statistics, with any user-selected filtering criteria applied (Fig. 4A), and (2) a visual alignment of the significant elements aligned to the promoter sequences (Fig. 4B). The latter analysis enables a user to easily identify clusters of nested or partially overlapping overrepresented words or groups of significant words and their relative positions within the promoters. A user may also decide to apply no filtering, in which case, the program returns the statistics for all 3–8 mers.

Results from ELEMENT may be viewed in several ways. First, the results can be viewed on the ELEMENT Web site in an HTML table format (Fig. 4A) that can be sorted online by word, number of occurrences, or significance statistics (Z-score, p-value, or corrected p-value). Alternatively, the results are available as a text file that may be downloaded for import into a spreadsheet or other program. The results tables also flag overrepresented elements that match known *cis* elements in the PLACE (Higo et al. 1998) or PlantCARE (Rombauts et al. 1999) databases. Therefore, in addition to predicting novel promoter elements, ELEMENT also finds known *cis* elements.

To demonstrate the utility of ELEMENT, we have used it to identify both known and novel regulatory motifs that function in the diurnal/circadian regulation of *Arabidopsis* gene expression. We analyzed 11 diurnal and circadian time courses in the reference plant *Arabidopsis*. By using HAYSTACK, we were able to identify putative coexpressed/coregulated genes for each phase of the day. The list of genes in each phase served as the input for ELEMENT, which identified overrepresented 3–8 mer DNA words in 500-bp upstream regions.

Our ELEMENT analysis revealed multiple variants of the Morning Element (ME) (Fig. 5A) (Harmer and Kay 2005), Evening Element (EE), and GATA (Fig. 5B) (Schindler and Cashmore 1990; Anderson and Kay 1995; Harmer et al. 2000), and G-box (Fig. 5C) (Giuliano et al. 1988; Michael and McClung 2002, 2003; Hudson and Quail 2003) that match known *cis* elements in the PLACE (Higo et al. 1998) or PlantCARE (Rombauts et al. 1999) databases.

Our ELEMENT analysis predicted previously unknown *cis*-regulatory elements not identified by other methods (T. Michael et al., in prep.). We have validated a group of these predicted elements using in vivo luciferase imaging, specifying their activity to a phase of day during which no diurnal- or circadian-associated *cis* elements had previously been identified. These predictions of novel diurnal/circadian-associated elements were robust enough to be well-conserved in rice and poplar. Therefore, using ELEMENT, we were able to predict both known and novel *cis*-regulatory elements and define specific aspects of their activity that were previously unknown.

**CONCLUSIONS AND FUTURE DIRECTIONS**

In conclusion, DIURNAL, HAYSTACK, and ELEMENT are a multifunctional, user-friendly, and complementary collection of Web-based tools. Our Affymetrix microarray data for plant diurnal and circadian time courses are accessible through the DIURNAL interface, which provides a powerful means to query and visualize the data. Two additional tools, HAYSTACK and ELEMENT, provide useful methods for querying microarray data. HAYSTACK analysis constitutes an alternative...
method for clustering microarray data by grouping together genes whose expression patterns match the same or similar user-defined HAYSTACK patterns. HAYSTACK can be used to compare virtually any large-scale data set against a set of user-supplied model patterns. ELEMENT identifies predicted cis-regulatory elements within the promoters of groups of coexpressed genes from Arabidopsis, rice, or poplar and can be used to facilitate cross-species comparisons of promoter architecture.

Data from our poplar and rice microarray time courses will be added to the DIURNAL database over the next year. We plan to add diurnal and circadian time course data for additional species when such data become available. We also plan to expand ELEMENT to support promoter analysis for additional plant species with sequenced genomes or genomes that will be sequenced in the near future, including sorghum, medicago, maize, Arabidopsis lyrata, papaya, and Brachypodium distachyon. We hope that these resources will help to accelerate functional genomics efforts directed at the mechanisms underlying diurnal and circadian biology in plants. For example, in Arabidopsis, the ability to order knockout or knockdown T-DNA insertion lines directly from stock centers makes it possible to rapidly perform in vivo tests of in silico predictions derived from HAYSTACK analysis of microarray data. Promoter element predictions from ELEMENT can in turn be tested experimentally to identify functional cis-regulatory elements.

ACKNOWLEDGMENTS

This work was supported by a National Science Foundation Plant Genome grant (DBI 0605240) (T.C.M, J.C., and S.A.K), the Howard Hughes Medical Institute, and National Institutes of Health (NIH) grants GM56006 and GM67837 (S.A.K.) and GM52413 and GM62932 (J.C.). T.P.M and T.C.M were supported by Ruth L. Kirschstein NIH Postdoctoral Fellowships.
REFERENCES


Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in Arabidopsis leaves. Plant Physiol. 136: 2687.


The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis


Access the most recent version at doi:10.1101/sqb.2007.72.006

References

This article cites 46 articles, 21 of which can be accessed free at:
http://symposium.cshlp.org/content/72/353.full.html#ref-list-1

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to Cold Spring Harbor Symposia on Quantitative Biology go to:
http://symposium.cshlp.org/subscriptions

© Cold Spring Harbor Laboratory Press