Genome analysis

Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences

Timothy R. O’Connor¹, Curtis Dyreson² and John J. Wyrick¹,*

¹School of Molecular Biosciences and ²School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA

Received on July 28, 2005; revised on September 15, 2005; accepted on October 10, 2005
Advance Access publication October 13, 2005

ABSTRACT

Summary: To better understand the regulatory networks that control plant gene expression, tools are needed to systematically analyze and visualize promoter regulatory sequences in Arabidopsis thaliana. We have developed the Athena database, which contains 30,067 predicted Arabidopsis promoter sequences and consensus sequences for 105 previously characterized transcription factor (TF) binding sites. Athena provides four novel tools to facilitate the analysis of promoter sequences: a promoter visualization tool to enable the rapid inspection of key regulatory sequences in multiple promoters; a TF binding site enrichment tool to identify statistically over-represented TF sites occurring in a user-selected subset of promoters; a data-mining tool to rapidly select promoter sequences containing the specified combination of TF binding sites; and a tool to display the distribution of TF binding site positions in a selected set of promoter sequences.

Availability: http://www.bioinformatics2.wsu.edu/Athena

Contact: jwyrick@wsu.edu

INTRODUCTION

The plant Arabidopsis thaliana presents an excellent model system for the computational study of promoter sequences and transcription regulation. To aid in this study, a number of databases containing information about Arabidopsis promoter sequences and transcription factor (TF) binding sequences have been developed. Several databases, including PlantCARE (Lescot et al., 2002; Rombauts et al., 1999), PLACE (Higo et al., 1999) and ATTFDB (Davuluri et al., 2003), store consensus binding sequences for plant or Arabidopsis-specific TFs. ATCISDB, which is included in the Agris website along with ATTFDB, catalogs several putative binding sites on a genome-wide scale. This resource also provides visualization tools for individual promoter sequences (Davuluri et al., 2003). AthaMap integrates both Arabidopsis DNA sequence and predicted TF binding sites (Steffens et al., 2004). Through this resource, single segments of DNA are viewable in a textual format with TF binding sites indicated. PlantProm DB curates a collection of plant promoters with experimentally verified transcription start sites (Shahmuradov et al., 2003).

To supplement these transcription-related resources, there is a need for integrated tools that perform on-the-fly visualization of promoter sequences and streamlined data-mining applications of regulatory promoter sequences. Tools are also needed to assess the statistical significance of a putative enrichment of regulatory sequences in a selected subset of promoters. A further need is to integrate all of these tools and resources within one interface so that the output from one tool can be automatically piped into a related tool.

To meet these computational challenges, the Athena database employs a unified and automated interface to assist in the analysis of Arabidopsis promoter sequences. Athena integrates DNA sequence and Gene Ontology (GO) data to facilitate the visualization, statistical and positional analysis of promoter regulatory elements. These tools should advance the study of promoter sequences of individual genes and also provide a framework for the systematic analysis of transcriptional regulatory networks in A.thaliana.

SYSTEM DESCRIPTION

The Athena database and accompanying web interface were constructed using the MySQL DBMS, custom Perl scripts and dynamically generated HTML/JavaScript web pages. Athena contains up to 3 kb of promoter sequence for 30,067 predicted Arabidopsis genes and consensus sequences for 105 previously characterized TF binding sites. Athena also contains GO information for each predicted gene. Promoter sequence and GO data were obtained from TAIR (Rhee et al., 2003); TF binding site consensus sequences were imported from PLACE (Higo et al., 1999) and Agris (Davuluri et al., 2003). Using standard defaults for maximum promoter size (3 kb) and truncating promoter sequences that overlap with upstream genes, we find that the average Arabidopsis promoter has 29.9 predicted TF binding sites.

Promoter visualization

The promoter visualization tool graphically represents key regulatory elements in promoter sequences. In the visualization tool’s promoter selection page, the user can enter up to 100 gene accesses (e.g. At5g52310) and select various display options.

Athena was used to visualize 2000 bp of the promoter sequence of the gene erd10 (At1g20450), which is involved in dehydration response in plants. A portion of the resulting web page containing the compact promoter visualization output is shown in Figure 1A. The web page also shows a TF table beneath the promoter image, which provides a key to the color-coded TF binding sites, lists the number of instances of each TF binding site in the promoter and gives the significance of enrichment (P-value) for each TF site in the

*To whom correspondence should be addressed.
promoter sequence (calculated using a hypergeometric probability distribution). An example of a TF table is shown in Figure 1B. The alternative form of promoter visualization is the cartoon display (data not shown), which is better at illustrating overlapping TF binding sites than the compact display. The visualization display is interactive; by checking the boxes adjacent to each TF site in the table, the user can select which TF binding sites to display in the promoter image.

Data mining

The data-mining tool enables the user to identify promoters based on the presence of selected TF binding sites in the promoter sequence, or the GO classification of the corresponding gene. Multiple TF sites can be selected using AND Boolean logic; promoters containing all of the selected TF sites will be identified. Multiple GO terms can be selected using OR Boolean logic.

The data-mining tool was used to select promoters whose corresponding genes were classified under the following three GO terms: response to water deprivation, salinity response and cold acclimation. The mining tool identified 55 genes classified under at least one of these terms (data not shown). The data-mining results include a table of TF binding sites that are overrepresented (enriched) in the promoter sequences of dehydration/cold response genes (Fig. 1B). Five of the enriched TF binding sites are classified under the ABRE/ABF or DRE binding site families, which are known to have a significant role in regulating gene expression in response to the three selected stresses (Yamaguchi-Shinozaki and Shinozaki, 1994). The mining results page also includes the list of 55 genes selected, a table of TF sites that are present but not enriched and a table of enriched GO terms (data not shown).

More advanced data-mining queries can be performed using the analysis tool selection page. The analysis selection tool includes optional promoter position constraints, which allow the user to search for TF sites that occur in specific positions in the promoter sequences (e.g. -200 to -100 bp upstream). Athena also includes a custom motif search tool, which allows the user to search for promoters containing a user-specified consensus sequence.

Figure 1. (A) Compact visualization of the erd10 (At1g20450) promoter. Each TF binding site is indicated by a color-coded hash mark matching those of the TF table. The predicted gene sequence is represented as a gray rectangle; an arrow indicates the start of transcription. Aqua rectangles indicate predicted CpG islands. (B) Enriched TF binding sites present in the promoters of dehydration/cold response genes. The number of promoters containing at least one instance of the TF binding site and the total number of TF binding sites in the selected set of sequences are given in the ‘P’ and ‘S’ columns of the TF table, respectively. (C) Histogram of ABRE-like binding site distribution in dehydration/cold response promoter sequences. The gray distribution backdrop is the random (expected) distribution of these TF binding sites. (D) Histogram of DRE-core motif position distribution. The position of the DRE-core motif was plotted across all dehydration/cold acclimation promoters with the same parameters as in C.
Analysis of TF site position bias

The Athena analysis tools can be used to examine the distribution of TF binding sites in promoter sequences. The analysis page contains a histogram tool that graphically displays the positions of a selected TF binding site among the specified promoter sequences. If multiple TF sites are selected, then the histogram tool will display the aggregate positions of all selected TF sites in a single plot.

The histogram tool was used to display the distribution of ABRE-like TF binding sites in the promoter sequences of the previously selected dehydration/cold response genes. Inspection of Figure 1C indicates that the ABRE-like TF sites are enriched between positions –40 and –160 in the promoter sequences of dehydration/cold response genes. The observed position bias is not simply due to the base composition of the ABRE-like consensus sequence; for random permutations of this consensus sequence do not show a similar promoter position bias (data not shown). In contrast, the DRE core site does not show a strong positional bias in the promoters of dehydration/cold response genes (Fig. 1D). These results suggest the hypothesis that the positioning of the ABRE-like TF site in promoter sequences of dehydration/cold response genes may be important for its transcriptional regulatory function.

ACKNOWLEDGEMENTS

We are grateful to John Browse, Mark Lange, Nicholas Provart, Heidi Strom and Bryan Thines for helpful comments, suggestions and discussions on the Athena database and this manuscript. We thank Monique Kohagura for assistance in developing some of the Athena tools and web pages. We thank Justin Fischer and Jason Sikes for web and server support. This work was supported, in part, by American Cancer Society grant RSG-03-181-01-GMC.

Conflict of Interest: none declared.

REFERENCES


