BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 19 2008, pages 2263–2264
doi:10.1093/bioinformatics/btn417

Gene expression

Arabidopsis thaliana regulatory element analyzer

Ananyo Choudhury* and Ansuman Lahiri

Department of Biophysics Molecular Biology and Genetics, University of Calcutta, 92 APC Road, Kolkata 700009, India

Received on March 8, 2008; revised on August 5, 2008; accepted on August 6, 2008

Advance Access publication August 11, 2008

Associate Editor: David Rocke

ABSTRACT

Summary: In the *Arabidopsis thaliana* regulatory element analyzer (AtREA) server, we have integrated sequence data, genome-wide expression data and functional annotation data in three application modules which will be useful to identify major regulatory targets of a user-provided *cis*-regulatory element (CRE), study different features of CRE distribution and evaluate the role of a set of CREs in the regulation of gene expression—independently as well as in combination with other user-provided CREs.

Availability: AtREA is freely available at http://www.bioinformatics.org/grn/atrea.html.

Contact: ananyo.c@gmail.com

1 INTRODUCTION

A *cis*-regulatory element (CRE) is a short stretch of DNA sequence which is recognized by its cognate transcription factor (TF) and thereby enables the expression of a gene located adjacent to it to be regulated by it. The analysis of over-representation of known and novel CRE(s) in upstream sequences of a set of co-expressed genes is a widely used method to associate the CRE(s) with regulation of gene expression under the given set of conditions. Many web servers are available for such analysis in *Arabidopsis* (Fu *et al*., 2004; Galuschka *et al*., 2007; Kankainen *et al*., 2006; O'Connor *et al*., 2005; Obayashi *et al.*, 2007; Ruan and Zhang, 2005). Although such web applications are very useful, the study of the distribution of a CRE in ontology groups or multiple microarray datasets is not straightforward. Moreover, they do not offer the facility to evaluate different CRE features or to investigate the role of co-occurrence of different CREs in a promoter. To address these problems, we have constructed a web server for a more comprehensive analysis of known or predicted CREs in *Arabidopsis*.

2 IMPLEMENTATION

AtREA has been linked to an integrated database that contains 1 kb upstream sequence, gene ontology (GO) (Ashburner *et al*., 2000), MIPS FUNCAT (Ruepp *et al*., 2004) and ARACYC Pathways (Zhang *et al*., 2005a) annotation and microarray expression data for >22 000 *Arabidopsis* genes (corresponding to affymetrix 22k probe set). Upstream sequence [1 kb upstream of transcription start site (TSS) for genes with annotated TSSs and upstream of the translational start for the remainder, based on TAIR 7], GO and pathway annotations of these genes were downloaded from TAIR database (Swarbreck *et al*., 2008). The FUNCAT annotations were retrieved from the MIPS database (Mewes *et al*., 2008).The TAIR expression dataset consisting of 1388 normalized, affymetrix 22k slides was retrieved from http://www.atted.bio.titech.ac.jp. Replicate slides from this set were averaged and slides labeled as no-treatment, control or 0 h were removed and the resulting slides were used to construct the expression dataset of AtREA. The scripts used in AtREA are written in CGI-PERL and runs on an apache server. The interface of AtREA consists of three different sections that host three different modules.

Distribution of CRE: genome-wide analysis of distribution of a CRE has also been shown to be useful in the identification of major functional groups regulated via the CRE (Long *et al.*, 2004). The 'CRE distribution analysis module' in AtREA has been designed to study the distribution of a CRE in ontology classes in *Arabidopsis*. The ontology groups that can be analyzed using this module include GO, MIPS FUNCAT and metabolic pathways (ARACYC) ontologies. This module can also be used to study the distribution of a CRE in expression classes, which were derived from and summarize the existing microarray data in *Arabidopsis*. A significant feature of this module is that, in addition to simple CRE consensus sequences (e.g. ACGTGTC), this module can also receive CRE position frequency matrices, CRE pairs (along with minimum and maximum distances between the component CRE sequences) and combinations of CREs (containing a maximum of four different CRE units) as input. The purpose of this module is to provide a quick and comprehensive understanding of functional targets of a CRE and identification of conditions in which the given CRE may be involved in transcription regulation.

Evaluation of CRE features: different features like numbers of occurrence of a CRE, distances of a CRE from TSS, and variations in the CRE consensus sequence have been shown to affect the regulatory significance of *Arabidopsis* CREs (Geisler *et al*., 2006; Suzuki *et al.*, 2005; Zhang *et al*., 2005b). Such features of any CRE in induced/repressed genes from an expression experiment or a functional class or a user-specified gene set can be analyzed by the 'CRE features analysis module'—the second module of AtREA. This module currently accepts CREs in simple consensus sequence format as input. The 'position' option in this module evaluates the positional trends in the distribution of a CRE in a set of genes in comparison to the background, which contains all of the 22 000 1 kb upstream sequences of *Arabidopsis*. The program divides the 1 kb upstream sequences into five 200 bp position windows and compares the distribution of the CRE among these

*To whom correspondence should be addressed.

five windows. The 'strand' option in this module similarly compares the distribution of CRE(s) in DNA strands, both in the supplied gene set and in the background. The 'frequency' option in this module has been incorporated to identify the impact of multiple occurrences of a CRE in the expression of genes. Most CRE(s) in plants are reported in consensus or regular expression formats. As a result, it is often important to study the role of variations in the consensus sequences in gene expression. The 'variant analysis' option in this module generates all possible single nucleotide variants of the input CRE consensus and compares the distribution of these variants with the original user-defined CRE both in the background and in the supplied gene set. The CRE features module can, therefore, be useful in detecting preferences in localization of a CRE, finding the position window and the strand in which a CRE is most efficient in expression regulation, identifying a minimum number of instances of the CRE in a promoter that is required for its function and in isolating variant sequences of a CRE that are also relevant in gene expression regulation.

Promoter state analysis: transcription regulation under many conditions, like abscisic acid (ABA) signaling in *Arabidopsis*, involves multiple TFs and CREs (Shinozaki and Yamaguchi-Shinozaki, 2007). To understand regulation under such conditions we have included a 'promoter state analysis module' in AtREA. This module of AtREA takes as input a set of genes (from expression dataset or from a user-defined gene list) and a set of CRE(s) that may be involved in expression regulation under the given condition (which can be derived from literature/experimental data and/or computational analysis). The module first segregates the entire upstream sequence set (background) into promoter states based on the presence/absence (binary mode) or the frequency of occurrence (CRE instances mode) of each of the CRE from the user-defined set of CRE(s). The module then compares the expected frequency of occurrence of each of the states (calculated from their respective genomic occurrences) to their actual frequency of occurrence in the selected gene. This module can therefore contribute in identification of a CRE or CRE combinations that show significant over-representation in induced/repressed genes in an experimental condition and in detection of CRE combinations that show significant difference in expression compared to expression of the component CRE(s).

3 CONCLUSIONS

Analysis of CRE distribution is a widely used technique in uncovering transcription regulation in *Arabidopsis*. Due to the variety of analysis options, integrated data and simple interface, we expect AtREA to be useful in the characterization of known and novel CREs in *Arabidopsis*, identification of their functional targets and evaluation of their role in expression, individually and in combination with other CREs.

Funding: Council for Scientific and Industrial Research (9/28 607)/2003 -EMR 1).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Fu,Y. *et al.* (2004) MotifViz: an analysis and visualization tool for motif discovery. *Nucleic Acids Res*., **32**, W420–W423.

Galuschka,C. *et al.* (2007) AthaMap web tools for the analysis and identification of co-regulated genes. *Nucleic Acids Res.*, **35,** D857–D862.

Geisler,M. *et al*. (2006) A universal algorithm for genome-wide in silicio identification of biologically significant gene promoter putative cis-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in Arabidopsis. *Plant J.*, **45**, 384–398.

Kankainen,M. *et al*. (2006) POXO: a web-enabled tool series to discover transcription factor binding sites. *Nucleic Acids Res*., **34**, W534–W540.

Long,F. *et al.* (2004) Genome-wide prediction and analysis of function-specific transcription factor binding sites. *In Silico Biol*., **4**, 395–341.

Mewes,H.W. *et al*. (2008) MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res.*, **36**, D196–D201.

O'Connor,T.R. *et al.* (2005) Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics*, **21**, 4411–4413.

Obayashi,T. *et al.* (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res.*, **35**, D863–D869.

Ruan,J. and Zhang,W. (2005) CAGER: classification analysis of gene expression regulation using multiple information sources. *BMC Bioinformatics*, **6**, 114.

Ruepp,A. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*., **32**, 5539–5545.

Shinozaki,K. and Yamaguchi-Shinozaki,K. (2007) Gene networks involved in drought stress response and tolerance. *J. Exp. Bot.*, **58**, 221–227.

Suzuki,M. *et al.* (2005) Quantitative statistical analysis of cis-regulatory sequences in ABA/VP1- and CBF/DREB1-regulated genes of Arabidopsis. *Plant Physiol.*, **139**, 437–447.

Swarbreck,D. *et al*. (2008) The Arabidopsis information resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*., **36**, D1009–D1014.

Zhang,P. *et al*. (2005a) MetaCyc and AraCyc: metabolic pathway databases for plant research. *Plant Physiol*., **138**, 27–37.

Zhang,W. *et al*. (2005b) Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana. *Bioinformatics*, **21**, 3074–3081.