Journal of Plant Physiology 168 (2011) 611–618

Contents lists available at ScienceDirect

Journal of Plant Physiology

journal homepage: www.elsevier.de/jplph

Large cliques in *Arabidopsis* gene coexpression network and motif discovery

Xiaoqi Zheng[a,b], Taigang Liu[c], Zhongnan Yang[d], Jun Wang[a,b,*]

[a] *Department of Mathematics, Shanghai Normal University, Shanghai 200234, China*
[b] *Scientific Computing Key Laboratory of Shanghai Universities, Shanghai 200234, China*
[c] *College of Information Sciences and Engineering, Shandong Agricultural University, Taian 271018, China*
[d] *College of Life and Environmental Sciences, Shanghai Normal University, Shanghai 200234, China*

ARTICLE INFO

Article history:
Received 29 May 2010
Received in revised form 31 August 2010
Accepted 6 September 2010

Key words:
Arabidopsis
Gene coexpression network
cis-regulatory elements
Motif
Maximal cliques

ABSTRACT

Identification of *cis*-regulatory elements in *Arabidopsis* is a key step to understanding its transcriptional regulation scheme. In this study, the *Arabidopsis* gene coexpression network was constructed using the ATTED-II data, and thereafter a subgraph-induced approach and clique-finding algorithm were used to extract gene coexpression groups from the gene coexpression network. A total of 23 large coexpression gene groups were obtained, with each consisting of more than 100 highly correlated genes. Four classical tools were used to predict motifs in the promoter regions of coexpressed genes. Consequently, we detected a large number of candidate biologically relevant regulatory elements, and many of them are consistent with known *cis*-regulatory elements from AGRIS and AthaMap. Experiments on coexpressed groups, including E2Fa target genes, showed that our method had a high probability of returning the real binding motif. Our study provides the basis for future *cis*-regulatory module analysis and creates a starting point to unravel regulatory networks of *Arabidopsis thaliana*.

Introduction

One of the major challenges in current molecular biology is to understand the cellular systems that regulate gene expression. To achieve this aim, a key step is the identification of transcription factor binding sites (TFBSs), also called the *cis*-regulatory elements, in the regulatory regions of potentially coregulated genes. The binding site is a distinct nucleotide pattern of lengths from 5 to 15, which can be recognized and bound by a specific protein (transcription factor, TF) to determine the timing and location of transcriptional activity. A more complete understanding of transcription factors and their DNA binding activities will facilitate a more comprehensive and quantitative mapping of the regulatory pathways within cells, as well as a deeper understanding of the potential functions of individual genes regulated by newly identified DNA-binding sites (Stormo, 2000; Hu et al., 2005).

The determination of the binding site for a transcription factor can be done using different approaches. Experimentally, the most common way is the ChIP-chip (Lee et al., 2002), which combines the techniques of chromatin immunoprecipitation and microarray hybridization. A DNA segment that is bound specifically by a TF is purified and amplified, and then genomic target loci are identified by comparative hybridization of the immunoprecipitated and control DNA probes to a DNA microarray. However, ChIP-chip is currently not easily applicable in many higher eukaryotes (Sikder and Kodadek, 2005). Other established experimental methods such as electromobility shift assay (EMSA) or DNase I footprinting provide high-resolution views of single promoters, but are infeasible for large-scale analysis (Galas and Schmitz, 1978). With the availability of large scale genome sequencing and high-throughput gene expression analysis techniques, it is possible to predict TFBSs using computational tools. These methods are based mainly on the assumption that coexpression of genes arises from their transcriptional coregulation. So given a set of coexpressed genes, it is possible to retrieve their promoter sequences and then find the statistically overrepresented motifs. To date, more than a hundred predictive methods have been proposed, varying by the motif models, statistical measures and search strategies (Stormo, 2000; Wasserman and Krivan, 2003; Sandve and Drabløs, 2006; Das and Dai, 2007).

However, as noted by many researchers, current prediction methods are successful for simple organisms like yeast, but perform significantly worse for higher multicellular organisms, such as humans, *Drosophila* and *Arabidopsis* (Sandve and Drabløs, 2006; Das and Dai, 2007). This is probably due to their larger genome sizes and more complex regulatory principles. Not all coregulated gene promoters share a common motif, because some of the identified genes in a given cluster might in fact be secondary response genes. On the other hand, because of the combinatorial nature of TFs, the

* Corresponding author at: Department of Mathematics, Shanghai Normal University, Shanghai 200234, China. Tel.: +86 021 64324284; fax: +86 021 64324284.
E-mail address: jwang@shnu.edu.cn (J. Wang).

0176-1617/$ – see front matter © 2010 Elsevier GmbH. All rights reserved.
doi:10.1016/j.jplph.2010.09.010

same motif can be found in the promoter regions of genes that are not coregulated. Thus, a coexpressed gene group, and especially a relatively small group, often does not possess enough information to enable an accurate binding motif. In recent years, some tools making use of the phylogenetic information have been proposed, e.g., phylogenetic footprint and phylogenetic shadow (Cliften et al., 2001, 2003; Berezikov et al., 2004). Unfortunately, for the model plant *Arabidopsis*, the lack of closely related genome sequences restricts the wide application of these methods.

Over recent years, research on biological networks such as coexpression, regulation, protein interaction and metabolic networks has become a central topic of bioinformatics (van Noort et al., 2004; Berg and Lässig, 2006; Jalan et al., 2010; Veiga et al., 2010). Due to the scale-free nature of biological networks, there exist few nodes (genes) with very high degrees (Albert and Barabási, 2002; Jalan et al., 2010). These nodes are responsible for holding the whole networks and are therefore essential to the network structure. For a regulatory network, these nodes are transcription factors that bind to many target genes. These TFs may be closely related to many important stages of organism development, such as germination, anther development, cell apoptosis, etc. Thus, further research on these TFs and their binding activity would inform our understanding of the transcriptional regulatory mechanisms of the corresponding organism. Target genes bound by the same transcription factor are coexpressed, and tend to form a cluster in the gene coexpression network. In theory, when using coexpression relationships to infer the common motifs, these large groups will yield more information for relatively accurate motif prediction.

In the present study, we focus on these large clusters in the *Arabidopsis* gene coexpression network and predict their corresponding motifs. After construction of the *Arabidopsis* gene coexpression network based on the ATTED-II database, we extracted a number of coexpressed gene groups (cliques in the coexpression network) using a subgraph-induced strategy and clique finder algorithm. Then promoter sequences of genes from each maximal clique were analyzed using four classical prediction tools for the prediction of TFBSs. A total of 4600 candidate TFBSs were detected, and many of are consistent with previously described *cis*-regulatory elements from AGRIS (Davuluri et al., 2003) and AthaMap (Steffens et al., 2004). This study provides the basis for future *cis*-regulatory module analysis and creates a starting point to unravel regulatory networks in *Arabidopsis thaliana*.

Methods

Construction of the Arabidopsis gene coexpression network

We used the expression data available from ATTED-II (Obayashi et al., 2007) to construct the gene coexpression network. ATTED-II provides gene-to-gene mutual ranks and correlation coefficients calculated from 58 publicly available experiments, 1388 GeneChips collected by AtGenExpress. There are a total of 20,906 files for 5 chromosomes in this dataset, with each file corresponding to an anchor gene. From each file, Pearson's correlation coefficients (PCCs) between the anchor gene and the rest genes were obtained.

To identify genes that are coexpressed, we computed the distribution of PCCs for all gene pairs in the 20,906 files and considered the 99% quantile of background distribution as significant (Fig. 1). As shown in Fig. 1, PCCs between pairwise genes followed a normal distribution with a peak value of 11,033,780 at PCC = −0.03, and the 99% quantile of the background distribution corresponded to PCC = 0.578. Then the gene coexpression network could be constructed through connecting the gene pairs with PCC ≥ 0.578 and disconnecting the rest. However, PCCs between the same pair of genes provided by different files may be inconsistent, so we removed those edges when constructing the network. Finally, we obtained a total of 1,087,660 valid coexpression relationships.

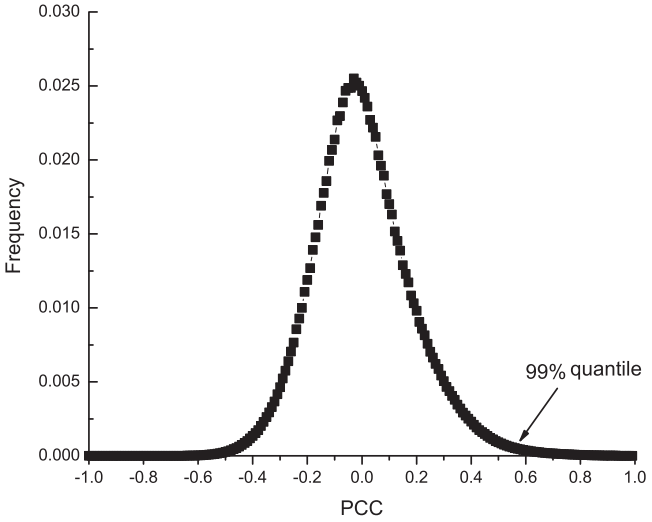

Fig. 1. Distribution of the Pearson's correlation coefficients (PCCs) for all gene pairs in *Arabidopsis thaliana*. As shown, PCCs between pairwise genes followed a normal distribution with a peak value of 11,033,780 at PCC = −0.03. We considered the 99% quantile of the background distribution as significant.

Extraction of coexpression groups

To detect motifs, it is first necessary to obtain reliable coexpressed gene groups (CEGs) in the *Arabidopsis* gene coexpression network. For each file downloaded from ATTED-II, which denotes an anchor gene and its PCCs with other genes, all genes coexpressed with it can be considered as a coexpression group (Haberer et al., 2006). However, according to this definition, the coexpression relationships are not transitive. For example, the PCCs between two genes g_1 and g_2, g_1 and g_3 both exceed 0.578, but it does not guarantee that g_2 and g_3 are coexpressed, which contradicts the meaning of coexpression. To overcome this problem, we applied a graph-clustering method based on the maximal clique algorithm as follows.

Clique and the maximal clique are useful tools in graph theory (Bondy and Murty, 1976; Papadimitriou and Yannakakis, 1981). For an undirected graph G, a *clique* is defined to be a complete subgraph of G, i.e., a collection of vertices that are all connected with each other. A clique is called *maximal* if there are no more vertices that can be added to the clique. A *maximum clique* of a graph is a maximal clique with maximum number of vertices. A clique in the *Arabidopsis* gene coexpression network has the property that any two genes in this clique have very similar expression patterns. So in the present study, we defined the CoExpression Groups (CEG) of *Arabidopsis* genes to be the maximal cliques in the coexpression network.

In order to obtain maximal cliques in the *Arabidopsis* gene coexpression network, it is intuitive to deal with the entire network using some clique-finding algorithm. However, finding cliques in a large graph is computationally intensive (Balas and Yu, 1986; Babel, 1991; Pardalos and Xue, 1994). In the present study, we explored a subgraph-induced scheme as follows. For a maximal clique C: $\{g_1, g_2, g_3, \ldots, g_N\}$, we considered the induced sub-graph N_{g_1} which consists of all genes that connected with g_1 in the *Arabidopsis* gene coexpression network (g_1 included). Since $g_2, g_3, \ldots, g_N$ are all connected with g_1, they are included in the sub-graph N_{g_1}, as does the clique C. In addition, C is maximal in the whole coexpression network, and N_{g_1} is a subgraph, so C is also a maximal clique in N_{g_1}. From this observation, we can calculate cliques in all sub-networks