Comparative Analysis of Gene-Coexpression Networks Across Species

Shiquan Wu and Jing Li*

Electrical Engineering and Computer Science Department
Case Western Reserve University, Cleveland, OH 44106, USA
{shiquan.wu,jingli}@case.edu

Abstract. This paper presents a large scale analysis of gene-coexpression networks (GCNs) across four plant species, i.e. Arabidopsis, Barley, Soybean, and Wheat, over 1471 DNA microarrays. We first identify a set of 5164 metagenes that are highly conserved across all of them. For each of the four species, a GCN is constructed by linking reliable coexpressed metagene pairs based on their expression profiles within each species. Similarly, an overall GCN for the four species is constructed based on gene expression profiles across the four species. On average, more than 50K correlation links have been generated for each of the five networks. A number of recent studies have shown that topological structures of GCNs and some other biological networks have some common characteristics, and GCNs across species may reveals conserved genetic modules that contain functionally related genes. But no studies on GCNs across crop species have been reported. In this study, we focus on the comparative analysis of statistical properties on the topological structure of the above five networks across Arabidopsis and three crop species. We show that: (1) the five networks are scale-free and their degree distributions follow the power law; (2) these networks have the small-world property; (3) these networks share very similar values for a variety of network parameters such as degree distributions, network diameters, cluster coefficients, and frequency distributions of correlation patterns (sub-graphs); (4) these networks are non-random and are stable; (5) cliques and clique-like subgraphs are overly present in these networks. Further analysis can be carried out to investigate conserved functional modules and regulatory pathways across the four species based on these networks. A web-based computing tool, available at *http://cbc.case.edu/coexp.html*, has been designed to visualize expression profiles of metagenes across the four species.

1 Introduction

With the availability of huge amount of genomic data, gene functions are usually predicted by similarity-based sequence analysis [7,9]. A great challenge in the post-genomic era is to understand gene regulations, genetic pathways and functional relations/modules of biological organisms at a system level [13,17,18,19,22].

* Corresponding author.

I. Măndoiu and A. Zelikovsky (Eds.): ISBRA 2007, LNBI 4463, pp. 615–626, 2007.

For such a purpose, sequence-based analysis has its limitations because genes with even very similar sequences may not be functionally related to one another [11,19]. Therefore, it becomes essential to integrate both genomic information and microarray data (and some other data sources) in the discovery of gene regulatory and functional relations. However, it is still hard or even impossible to identify regulatory or functionally related genes if studies are limited to only a single species[19]. This motivates the investigation of gene regulatory and functional relations by integrating both genomic information and microarray data to study not only a single species, but across multiple species. Recently, much attention has been paid to the investigation of biological networks and/or conserved functional modules using multiple species, or multiple tissues. An earliest study has investigated gene-coexpression networks across humans, flies, worms, and yeast[19], and has discovered some global conserved genetic modules across these species. Since then, a number of studies[4,10,15] have been proposed to analyze complex gene-coexpression networks across species. Berg and Lässig[4] have proposed a Bayesian alignment method and have identified significant conservations of gene expression clusters and gene functions by analyzing GCNs between humans and mice. Lelandais et al. [15] have adopted the Multi-dimensional Scaling technique to compare GCNs from budding and fission yeasts and have extracted some common properties and difference between the two species. Guimera and Amaral[10] have proposed a method that can generate a 'cartographic representation' of biological networks which enables the identification of functional modules from those networks. They have applied the method on metabolic networks across twelve organisms and have discovered that nodes with different connectivity patterns are affected by different evolutionary constraints and pressures. Gene-coexpression networks across different tissues have also been studied by a number of groups [2,6,14] in order to identify conserved interactions among disease genes. Other researchers [3] have identified functionally related proteins by analyzing conserved protein-protein interactions across species.

These studies mainly focus on crossing humans, animals or diseases. None of them have investigated the properties of GCNs across plants. In this paper, we study the statistical properties [1] of gene coexpression networks across four plant species: Arabidopsis, Barley, Soybean, and Wheat using 1471 hybridizations, whose genomic and DNA microarray data are available at public webs. Arabidopsis is chosen here because it is a well-understood model organism and can be used to study the functionality of genes and/or functional modules in other three species, which are important crops in the world. To the authors' best knowledge, this is the first study on gene-coexpression networks of crop species together with a model organism Arabidopsis. It is of importance to understand the statistical properties of gene-coexpression networks in order to learn their functional relations, and to understand regulatory pathways among genes across these species. The current study on GCNs is an important step towards further understandings and studies of conserved functional modules from the four species.

The rest of this paper is organized as follows. In Section 2, we first introduce the data sources that include genomic sequences and 1471 DNA microarray expression profiles of the four plant species. A set of 5164 metagenes is then obtained by comparing the genes across the four species using BLAST. A web-based computing tool is designed to view the expressions of metagenes across various species, experiments, and hybridizations. Five gene-coexpression networks are then constructed based on the Pearson's correlation coefficient from the DNA microarray expression profiles of metagenes. We also propose a simple algorithm to calculate the frequency distribution of correlation patterns. In Section 3, a comparative analysis is conducted on the five gene-coexpression networks. We have obtained the following statistical properties for these networks: (1) the degree distributions of the five coexpression networks follow the power-law, *i.e.*, $P(k) \sim k^{-\gamma}$, which means the probability of a node with a degree of k ($P(k)$) is proportional to $k^{-\gamma}$, where $\gamma \geq 0$ is the exponent of the power-law [1]; (2) the five gene-coexpression networks have the small-world property [20], *i.e.*, the network diameters are small and the cluster coefficients are great; (3) the five gene-coexpression networks share very similar values across a variety of network parameters such as degree distributions, network diameters, cluster coefficients, and the frequency distributions of interaction/correlation patterns; (4) these networks are non-random and their properties are stable under randomly introduced noise, even with as large as 20% changes of edges; (5) cliques and clique-like subgraphs are overly present in these gene-coexpression networks. In Section 4, we conclude the paper by discussing some potential work in predicting functional modules and regulatory pathways using these coexpression networks from multiple species.

2 Materials and Methods

2.1 Sequence and Expression Data

The materials used in this study include the genomic sequences and a large set of DNA microarray expression profiles of the four plant species, which were collected from several public sources on the Internet: *http://affymetrix.com, http://arabidopsis.org, http://tigr.org, http://ausubellab.mgh.harvard.edu/imds, http://psi081.ba.ars.usda.gov/SGMD, http://soybeangenome.org, http://harvest-web.org, http://plexdb. org, http://www.ncbi.nlm.nih.gov/projects/geo, http:// smd.stanford.edu, etc.* For Arabidopsis, we select 617 DNA microarray expression profiles. For Barley, Soybean, and Wheat, we respectively have 671, 53, and 130 microarrays. These 1471 DNA microarray expression profiles contain diverse conditions of microarray experiments (*e.g.*, various experimental organisms, different experimental types, a wide range of experimental factors, *etc.*)

The aim of this study is to investigate the common orthologous genes of the four species and the statistical properties of the gene-coexpressions across the species, experiments, and hybridizations.

There are three major steps in this study: (1) identifying metagenes across the four species; (2) constructing five gene-coexpression networks based on microarray expression profiles of the metagenes: one for each of the four individual species, and one for the overall gene-coexpressions across the four species; (3) investigating the statistical properties of the five gene-coexpression networks by a comparative analysis.

2.2 Identifying Metagenes

By applying "all-against-all" BLAST[19] to all genes of each pair of species, 5146 metagenes are obtained and shown in Table 1 (a complete list of the 5164 metagenes is available on our website). These 5164 metagenes are only a small fraction of all the genes with expression data from each species. Each metagene is defined as a set of four genes, one from each of the four species. Any two genes in a metagene are each other the best hit by BLAST using their protein sequences. For example, Metagene 1 consists of four genes: "244901_at" in Arabidopsis, "Barley1_53087" in Barley, "GmaAffx.25198.1.S1_at" in Soybean, and "Ta.22468.1.S1_at" in Wheat (see the first row in Table 1). These four genes are the best hits each other by BLAST using their protein sequences. Metagenes setup a mapping between the genes of one species and those of another species. By this mapping, it is possible to analyze gene expressions across various species, experiments, and hybridizations. The expressions of metagenes across species, experiments, and hybridizations can be viewed by our web-based computing tool at *http://cbc.case.edu/coexp.html*.

Table 1. Metagenes across Arabdopsis, Barley, Soybean, and Wheat

No.	**Arabidopsis**	**Barley**	**Soybean**	**Wheat**
1	244901_at	Barley1_53087	GmaAffx.25198.1.S1_at	Ta.22468.1.S1_at
2	244936_at	Barley1_53095	GmaAffx.17247.1.S1_at	TaAffx.124056.1.S1_at
···	······	······	······	······
5164	267646_at	Barley1_07944	Gma.3262.1.S1_at	Ta.10084.1.S1_at

2.3 Constructing Gene-Coexpression Networks

Gene-coexpression networks are constructed based upon metagenes' DNA microarray expression profiles. Relabel the 1471 hybridizations and denote H_1, H_2, $\cdots$, and H_{617} the 617 hybridizations of Arabidopsis; H_{618}, H_{619}, $\cdots$, and H_{1288} the 671 hybridizations of Barley; H_{1289}, H_{1290}, $\cdots$, and H_{1418} the 130 hybridizations of Soybean; and H_{1419}, H_{1420}, $\cdots$, H_{1471} the 53 hybridizations of Wheat. These 1471 DNA microarrays define the following 5164 × 1471 matrix of intensities, where the k_{th} row of the matrix represents the expression intensities of the k_{th} metagene across the 1471 hybridizations. Whereas each column represents the expression intensities of all metagenes under a hybridization. The hybridizations within each experiment have been normalized when we downloaded the data. However, when the hybridizations from various experiments and different species are put together, a new normalization is needed. The expression intensities are normalized across species and experiments using the Quantile normalization method[5]. The purpose is to adjust the effects arising from variation of different experiments rather than from biological differences [5,21].

To construct a gene-coexpression network $G = (V, E)$, we take each metagene as a node in V. For each pair of metagenes: k and j, the Pearson's correlation coefficient $(r(k, j))$ based on their expression profiles can be calculated as $r(k, j) = \frac{\Sigma KL - \frac{\Sigma K}{\Sigma L}}{\sqrt{(\Sigma K^2 - \frac{(\Sigma K)^2}{N})(\Sigma L^2 - \frac{(\Sigma L)^2}{N})}}$, where K and L represent the k_{th} and l_{th} row vectors of the intensity matrix, and N is the number of hybridizations. If $|r|$ is greater than a predefined cutoff value (the choice of the exact value of the threshold will be discussed below), the expressions of metagenes k and j are highly correlated and an edge is added between the pair. When constructing a gene-coexpression network, a proper cutoff value is necessary so that only significant correlations are included in a coexpression network. In this study, a similar approach as in[14] has been used in determining threshold values for different datasets. Basically, under the null hypothesis of no correlation, the Pearson correlation coefficient corresponds to a t-distribution with degrees of $N - 2$. An overall error rate of 0.05 is chosen after Bonferroni correction of multiple testing. In addition, only top and bottom 0.5% of correlations will be included for further study[14]. The combination of criteria corresponds to cutoff values from 0.8 to 0.9 in this study.

There are five gene-coexpression networks being constructed as follows. For Arabidopsis, the gene-coexpression network $G_{AT} = (V, E_{AT})$ is constructed based on the microarray data of Arabidopsis: H_1, H_2, $\cdots$, and H_{617}, with a cutoff value 0.8. Similarly, for Barley, Soybean, and Wheat, their gene coexpression networks $G_{BB} = (V,\ E_{BB})$, $G_{GM} = (V, E_{GM})$, $G_{TA} = (V, E_{TA})$ are respectively constructed by using H_{618}, H_{619}, $\cdots$,H_{1288} with cutoff value 0.8 for G_{BB}; H_{1289}, H_{1290}, $\cdots$, H_{1418} with cutoff value 0.85 for G_{GM}; H_{1419}, H_{1420}, $\cdots$, $H_{1471}$with cutoff value 0.9 for G_{TA}. Finally, $G = (V, E)$ is constructed as an overall gene coexpression network across four species by using all 1471 DNA microarrays with cutoff value 0.8. Therefore, G_{AT}, G_{BB}, G_{GM} and G_{TA} respectively represent metagene pairs that are coexpressed with significant correlations within the experiments of each individual species. Whereas, G represents metagene pairs coexpressed with significant correlations in all experiments across the four species. Each of the networks has 5164 nodes and the number of edges is in the range of 50k-70k (see Table 2). A possible explanation that different networks have different cutoff values is that the numbers of hybridizations for different networks vary dramatically, from around 600 for G_{AT} and G_{BB} to 100 for G_{GM} and 50 for G_{TA}. Greater cutoff values for smaller data sets are chosen to ensure only significant correlations being included.

2.4 Statistical Analysis of Network Parameters

The aim of this study is to investigate the statistical properties of network parameters that include degree distributions, network diameters, and clustering coefficients from the five gene-coexpression networks. These parameters have been widely discussed for large-scale complex networks and the calculation can be performed based on their definitions [1].

In addition to expression links, some subgraphs in gene-coexpression networks may represent important functional modules, and it is of great interest to understand or identify special patterns of subgraphs that are overly represented from GCNs across species. As the first step, we have studied the frequency distribution of correlation patterns/subgraphs in this paper that have similarly been taken into considerations by other researchers [19]. In total, 29 correlation patterns/subgraphs each with 3 to 5 nodes have been included (see Fig. 2). For each of the patterns, its frequency in a network is defined as the number of its occurrences in the network. If a substructure has been counted as one pattern (say Pattern 17 in Fig. 2), none of its subgraphs with the same number of nodes (say Pattern 12 and 16) will be counted as occurrences of smaller patterns. A simple computer program has been implemented to count the frequencies of the 29 correlation patterns in a network based on the following algorithm. The running time of the algorithm depends on the degree distribution of a network. But it is much faster than the naive method that examines every subset with 5 nodes.

Algorithm: Counting Pattern Frequency

```
 Input  Network $G = (V, E)$ (denote $V$ by integers: 1,2,...,5164).
Output  Frequency $f_k$ (denote $P_k$ the $k_{th}$ pattern, $1 \leq k \leq 29$).

Step 1  For each node $i$, find its neighbors:
        $N_i = \{j|(j, i) \in E, j > i\}, 1 \leq i \leq 5164$.

Step 2  For each $N_i$, check every subset $U$ of $N_i$ with $|U| \leq 4$,
        if $\{i\} \cup U$ is a pattern, say $P_k = G_{\{i\} \cup U}$, and $G_{\{i\} \cup U}$
        is not contained in any other patterns with $|U| + 1$ nodes,
        count the occurrence into the frequency of $P_k$, i.e. $f_k$++.
```

3 Results

In this section, we present the statistical properties of network parameters obtained by the comparative analysis of the five gene-coexpression networks. First, a brief summary on the network parameters of the five networks is given in Table 2. It can be seen that all the parameters are very similar across the five gene coexpression networks. This is probably because the four plant species have relatively small evolutionary distances from each other.

As observed in many gene expression networks by other researchers, the GCNs obtained here also have the small-world property, *i.e.*, they all have small network diameters (either 4 or 5) and great cluster coefficients (at least 0.6). The degrees of the five GCNs follow the power-law distributions with very similar power-law exponents γ (Table 2). The values of γ (1.13-1.28) are consistent with many other gene-coexpression networks obtained by other researchers(see [2] and references therein). The degree distributions of the five GCN are displayed

Table 2. Summary of network parameters

Network	Node	Edge	Power-law exponent	Network diameter	Cluster coefficient
G_{AT}	5164	56392	1.1520	5	0.7042
G_{BB}	5164	52714	1.2147	4	0.8264
G_{GM}	5164	72968	1.1298	4	0.7827
G_{TA}	5164	63382	1.1569	4	0.6338
G	5164	51905	1.2777	4	0.6444

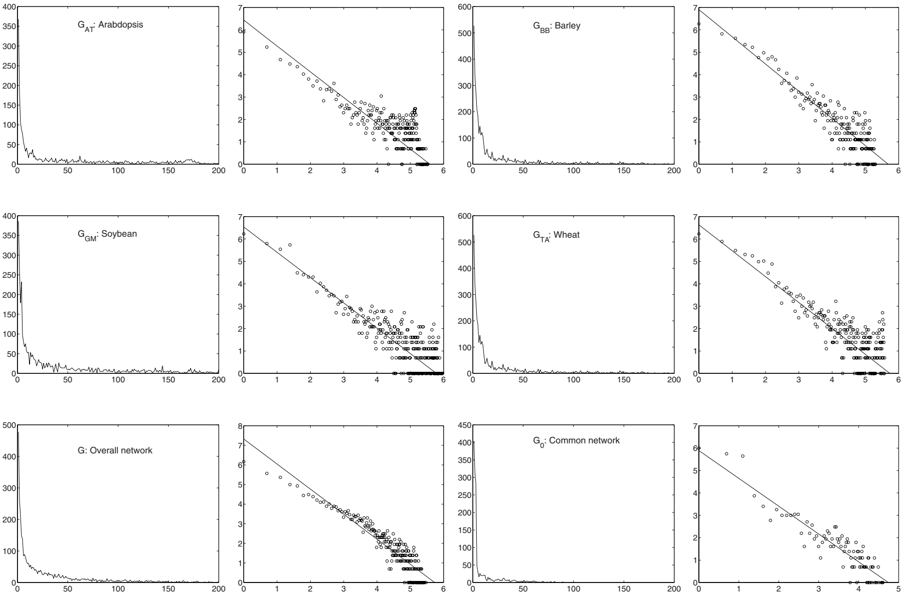

Fig. 1. The first and third columns are the degree distributions of G_{AT}, G_{BB}, G_{GM}, G_{TA}, G and G_0, respectively, where $x-$axis represents the number of degrees and $y-$axis represents the number of nodes. The second and forth columns are the log-log plots of the degree distributions of $G_{AT}, G_{BB}, G_{GM}, G_{TA}, G, G_0$.

in Fig. 1, together with their log-log plots. In addition, we have defined a new network $G_0 = (V, E_0)$ (called the common network) by taking all the common edges from the four gene-coexpression networks for individual species, i.e. $E_0 = E_{AT} \cap E_{BB} \cap E_{GM} \cap E_{TA}$. The degrees of G_0 also follows a power-law distribution, shown in Fig. 1, but the total number of edges in G_0 is much smaller compared with other five networks.

Distributions of correlation patterns/subgraphs. Genes and proteins always interact with each other in groups to perform certain biological functions. It is important to understand their interaction/correlation patterns. As the first step, we evaluate the frequency distributions of 29 correlation patterns (also see [16], each of which has 3 to 5 nodes) in the gene-coexpresion networks obtained in this study. Each pattern may represent a particular type of the interactions/correlations of the genes/metagenes. For comparison purpose, we further introduce two new networks. The perturbed network under noise G_S is defined by introducing as much as 20% noise on edges, *i.e.*, deleting 10% of the existing edges and adding 10% of new edges in the network G. And a random network G_R is generated, which consists of the same number of nodes (5164). An edge will be added to each pair of nodes in G_R with a small probability so that the total number of edges in G_R will be similar as the number of edges in other networks. The frequency distributions of the 29 patterns in all eight networks, *i.e.*, $G_{AT}, G_{BB}, G_{GM}, G_{TA}, G, G_S$, G_0 and G_R, are counted using the algorithm described in subsection 2.4 and shown in Fig. 2.

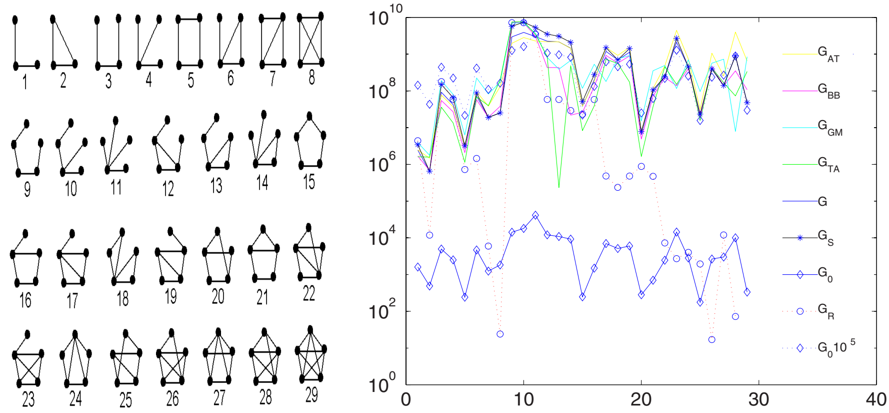

Fig. 2. Left: the 29 correlation patterns (also see [16]). Right: the frequency distributions of the 29 patterns; Top part: $G_{AT}, G_{BB}, G_{GM}, G_{TA}, G, G_S$ (presented by "$*$"), $G_0 \cdot 10^5$(presented by "$\diamond$"); Bottom part: G_0 (presented by "$\diamond$"); G_R is indicated by "$\circ$": up and down between top and bottom. $x-$axis: 29 patterns. $y-$axis: frequency.

The five gene-coexpression networks $(G_{AT}, G_{BB}, G_{GM}, G_{TA}, G)$, as well as the perturbed network G_S, have very similar frequency distributions over all the 29 patterns. The frequencies in graph G_0 are much lower because G_0 consists of a much smaller number of edges. But the distribution pattern is the same as those of other coexpression networks. To make it clear, we multiple the frequency distribution in G_0 by 10^5 and denote the new scaled distribution by $G_0 \cdot 10^5$. Fig. 2 shows that all the 7 gene coexpression networks have very similar distributions across all the patterns, but the random network G_R has a very different distribution. Further examinations reveal that a subset patterns (such as patterns 2, 7-8, 22-29) in the 7 coexpression networks have very different frequencies from the random network G_R. Those patterns are either cliques (patterns 2, 8 and 29) or some condensed patterns that are very similar to cliques (patter 7 and patterns 22-28). The over presence of cliques or clique-like subgraphs in GCNs may reflect the facts that those genes in a clique may encode proteins that form a protein complex, or they may be regulated by a common transcription factor. More investigations are needed on these overly presented patterns.

In order to quantitatively measure the overall differences of frequency distributions of the 29 patterns among G_{AT},G_{BB},G_{GM},G_{TA}, G, G_S, and G_R, a distance measure is defined as $d(G_i, G_j) = \sum_{k=1}^{29} w_k |log(n_{ik}) - log(n_{jk})|$, where n_{ik} and n_{jk} are the frequencies of Pattern k of G_i and G_j, respectively, and w_k is the weight of Pattern k ($w_k \geq 0$ and $\sum_{k=1}^{29} w_k = 1$). In this study, we simply take an equal weight for each pattern, *i.e.*, $w_k = 1/29$ for any k. The pairwise distances among all the networks are given in Table 3. The distances show that G_{AT},G_{BB},G_{GM},G_{TA}, G, and G_S (also $G_0 \cdot 10^5$) are very close each other (all pairwise distances of the 6 networks are less than 1.66). In contrast, the random network G_R is quite different from them (the distance between G_R and any other one is around 6, see the last row).

Table 3. Distances on frequency distributions

Distance	G_{AT}	G_{BB}	G_{GM}	G_{TA}	G	G_0	$G_0 \cdot 10^5$	G_S	G_R
G_{AT}	0	0.99	1.07	1.30	0.49	11.39	1.19	0.61	6.81
G_{BB}	0.99	0	1.11	0.93	0.63	10.57	1.33	0.76	5.91
G_{GM}	1.07	1.11	0	1.27	1.21	11.29	1.35	1.14	6.41
G_{TA}	1.30	0.93	1.27	0	1.26	10.32	1.66	1.38	6.18
G_G	0.49	0.63	1.21	1.26	0	11.00	1.06	0.22	6.33
$G_0 \cdot 10^5$	1.19	1.33	1.35	1.66	1.06	11.40	0	1.14	6.73
G_S	0.61	0.76	1.14	1.38	0.22	11.22	1.14	0	6.37
G_R	6.81	5.91	6.41	6.18	6.33	19.13	6.73	6.37	0

Non-randomness and stability of coexpression networks. We believe that the correlation links and the overly presented patterns in the networks are statistically significant and they might be biologically meaningful. First of all, the analysis was performed on a large data set that consists of 1471 hybridizations. It is unlikely to obtain significant correlations and expression patterns by chance over such a large data set. Furthermore, we have constructed a random network G_R and a perturbed network G_S. The network parameters of the two networks are shown in Table 4 and their degree distributions are shown in Fig. 3. It is obvious that the gene-coexpression networks are quite different from the random network in terms of degree distributions, cluster coefficients, and pattern frequency distributions. On the other hand, all the parameters from the perturbed network are very similar with those from all other gene-coexpression networks. This indicates that the results obtained from this study are quite robust and can not be generated by chance.

Biological meaning of the gene coexpression networks. The networks we construct represent significant correlations among metagenes across the species over a large set of microarray data. Various types of subgraphs in these coexpression networks may imply biological meaningful properties or functional relations of genes. We first take the top three hub nodes from the Arabidopsis network

Table 4. Network parameters of G_R and G_S

Network	Node	Edge	Power-law exponent	Network diameter	Cluster coefficient
G_R	5164	53461	Non Power-law	4	0.008
G_S	5164	73119	1.2672	6	0.604

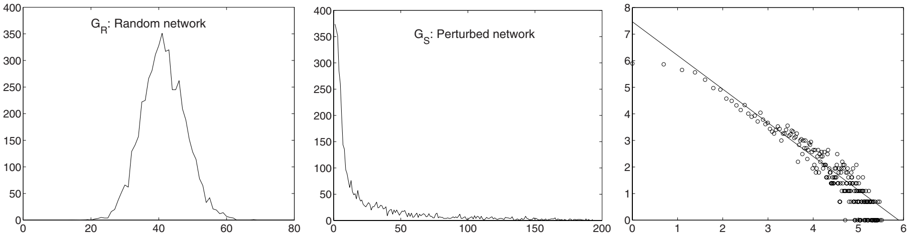

Fig. 3. Left: the non-power-law degree distribution of G_R. $x-$axis: number of degrees; $y-$axis: number of nodes. Middle: the degree distribution of G_S. Right: the log-log plot of the degree distribution of G_S.

and check their GO (Gene Ontology) annotations(*http://www.arabidopsis.org*). We find that the genes represented by the hub nodes are involved in protein expressing, folding, and binding, which are essential in protein synthesis and protein-protein interactions. The corresponding metagenes also have large number of links in the coexpression networks of other three species (due to the page limitation, details of the results in this subsection can be found from our website). The above observations suggest that our coexpression networks may reveal certain important biological properties. Gene functions from Barley, Soybean, and Wheat, which are mostly unknown, can be predicted through our coexpression networks and functional annotations of Arabidopsis.

To further explore the networks, we choose those highly significant links by a cut-off value of 0.99. The node with the largest number of links in Arabidopsis (gene 246075_at) has 47 neighbors. This gene has its GO annotation as transferase activity, transferring glycosyl groups, and UDP-galactosyltransferase activity. Among all the 48 genes, more than 70% of all the gene pairs (48 choose 2) are linked. Most genes in this subgraph share similar GO annotations such as catalytic activity, cellulose synthase activity, transferase activity, and kinase activity. Therefore this subgraph may present one or several groups of functionally related genes. A few genes with unknown biological functions in this subgraph may be predicted based on annotations of other genes within the same group.

4 Discussions

In this study, we first obtained metagenes that are orthologous genes in common to the four plant species by sequence analysis. Those metagenes might have been conserved from their common ancestor through evolution and might play an important role in biological functions and regulations. Gene-coexpression networks were then constructed based on their expression profiles. We have investigated the statistical properties of those gene-coexpression networks. The degrees of all gene-coexpression networks follow power-law distributions and they have the small-world property with small network diameters and great cluster coefficients. The values of those parameters from all the expression networks are very similar, probably because the four species are very close in evolution. The properties are quite different from those of a random network and are robust under perturbation. We have also investigated the frequency distributions of 29 correlation patterns and have found that cliques and clique-like patterns are overly present in these networks but not in a random network with similar size. This result implies that some of those patterns may represent certain important functional modules. Further studies are needed to explore the biological meanings of those patterns.

This study has two significant features. First, it is the first study on the gene-coexpression networks across crop species, whereas previous studies mainly focus on the gene-coexpression networks of single crop species[8], or across humans, animals and diseases[2,3,4,6,10,14,15,19]. Secondly, this study first investigates the statistical properties of the frequency distributions of correlation patterns in gene-coexpression networks and has identified that cliques and clique-like patterns are overly present in these networks. Previous studies mainly discuss network parameters such as degree distributions, diameters, cluster coefficients.

Pathways and gene regulatory networks are usually predicted by comparative genomics using sequence information, which can also be applied on crops such as Barley, Soybean, and Wheat. However, metagenes and coexpression networks can be used as a new method to predicting functionally related genes, functional modules and regulatory pathways. By across-species inference, the known functionally related genes, functional modules and regulatory pathways of one species can be used to predicting those of other species. Arabidopsis is a well-studied species. Many functionally related genes, functional modules and regulatory pathways have been identified in Arabidopsis. Whereas, little has been known on the pathways, regulations, functions, and modules about Barley, Soybean, and Wheat. Gene-coexpression networks can make it possible to predict functionally related genes, functional modules and regulatory pathways in the three species by those in Arabidopsis. If a group of coexpressed metagenes are functionally related in Arabidopsis, by comparing gene-coexpression networks, it is possible to predict that those metagenes may also be functionally related in Barley, Soybean, and Wheat. We will address this issue in our future studies.

Acknowledgements. This research was supported in part by NIH/NLM grant LM008991 and a start-up fund from Case Western Reserve University to JL.

References

1. Albert, B. and Barabási, A.-L.: Statistical mechanics of complex networks, *Review of Modern Physics*, 74(2002)47-97.
2. Aggarwal, A., Guo, D. L., etc.: Topological and Functional Discovery in a Gene Coexpression Meta-Network of Gastric Cancer, *Cancer Research*, 16(2006)232-241.
3. Bandyopadhyay, S., Sharan, R. and Ideker, T.: Systematic identification of functional orthologs based on protein network comparison, *Genome Research*, 16(2006)428-435.
4. Berg, J. and Lässig, M.: Cross-species analysis of biological networks by Bayesian alignment, *PNAS*, 103(2006)10967-10972.
5. Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P.: A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance, *Bioinformatics*, 19(2003)185-193.
6. Choi, J. K., Yu, U., Yoo, O. J. and Kim, S.: Differential coexpression analysis using microarray data and its application to human cancer, *Bioinformatics*, 21(2005)4348-4355.
7. Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G.: *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge Univ. Press (1998).
8. Faccioli, P., Provero, P., etc: From single genes to co-expression networks: extracting knowledge from barley functional genomics, *Plant Molecular Biology*, 58(2005)739-750.
9. Gusfield,D.: *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press (1997).
10. Guimera, R. and Amaral, L. A. N.: Functional cartography of complex metabolic networks, *Nature*, 443(2005)895-900.
11. Gerlt, J. A. and Babbitt, P. C.: Can sequence determine function?, *Genome Biology*, 1(2000):REVIEWS0005.
12. Hanfrey, C., Sommer, S., etc.: Arabidopsis polyamine biosynthesis: absence of ornithine decarboxylase and the mechanism of arginine decarboxylase activity, *Plant Journal*, 27(2001)551-60.
13. Kitano, H.: Systems Biology: A Brief Overview, *Science* 295(2002)1662-1664.
14. Lee, H. K., Hsu, A.K., etc.: Coexpression analysis of human genes across many microarray data sets, *Genome Research*, 14(2004)1085-1094.
15. Lelandais, G., Vincens, etc.: Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms, *Bioinformatics*, 22(2006)1359-1366.
16. Pržulj, N., Corneil, D. G. and Jurisica, I.: Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(2004)3508-3515.
17. Sali, S.: Functional links between proteins, *Nature*, 402(1999)23-26.
18. Strogatz, S. H.: Exploring complex networks, *Nature*, 410(2001)268-276.
19. Stuart, J. M., Segal, E., Koller, D., and Kim, S. K.: A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules, *Science*, 302(2003)249-255.
20. Watts, D. J. and Strogatz, H.: Collective dynamics of 'small-world' networks, *Nature*, 393(1998)440-442.
21. Yang, Y. H., Dudoit, S., etc.: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, 30N4(2002)e15.
22. Zhou, X. J. and Gibson, G.: Cross-species comparison of genome-wide expression patterns, *Genome Biology*, 5(2004)232-233.