Plant Cell Physiol. 48(3): 381–390 (2007)
doi:10.1093/pcp/pcm013, available online at www.pcp.oxfordjournals.org

Mini Review

Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology

Koh Aoki [1], Yoshiyuki Ogata [1] and Daisuke Shibata *

Kazusa DNA Research Institute, Kazusa-Kamatari 2-6-7, Kisarazu, 292-0818 Japan

Gene co-expression, in many cases, implies the presence of a functional linkage between genes. Co-expression analysis has uncovered gene regulatory mechanisms in model organisms such as *Escherichia coli* and yeast. Recently, accumulation of *Arabidopsis* microarray data has facilitated a genome-wide inspection of gene co-expression profiles in this model plant. An approach using network analysis has provided an intuitive way to represent complex co-expression patterns between many genes. Co-expression network analysis has enabled us to extract modules, or groups of tightly co-expressed genes, associated with biological processes. Furthermore, integrated analysis of gene expression and metabolite accumulation has allowed us to hypothesize the functions of genes associated with specific metabolic processes. Co-expression network analysis is a powerful approach for data-driven hypothesis construction and gene prioritization, and provides novel insights into the system-level understanding of plant cellular processes.

Keywords: Co-expression — Correlation — Metabolome — Network — Systems biology — Transcriptome.

Abbreviations: CESA, cellulose synthase; CTL, endochitinase-like; EST, expressed sequence tag; MEP, methylerythritol phosphate; MVA, mevalonate; PCC, Pearson correlation coefficient.

Introduction

Accumulation of genome-wide gene expression data has allowed biologists to investigate gene regulatory mechanisms using systems biological approaches. Recent developments in microarray technologies and bioinformatics have synergistically driven the progress of this field. With the accumulation of *Arabidopsis* gene expression data, the systems biological approach can now be applied to this model plant. The central concept of this approach is to depict organizational and functional relationships of the component molecules. Network analysis enables this as it conceptually represents the relationships between components by a network.

In many cases, a coordinated behavior of gene expression across a variety of experimental conditions indicates the presence of functional linkages between genes. For example, it has been demonstrated that the expression of genes associated with the same metabolic function is likely to show co-expression patterns (Ihmels et al. 2004, Kharchenko et al. 2005). An increasing number of studies have supported the versatility of co-expression analysis for inferring gene functions, although it has been recognized that co-expression does not necessarily mean a regulatory relationship (Stuart et al. 2003). In co-expression analysis, similarity of gene expression profiles is measured using correlation coefficients or any other distance measures. If the correlation between two genes is above a given threshold, the genes can be connected together to generate a network. A co-expression network thus illustrates correlation patterns between genes, and so represents the complexity of a cellular transcriptional network.

Analyses of cellular networks have revealed unforeseen similarities to non-biological complex network systems, including the Internet and society (Watts and Strogatz 1998, Barabasi and Albert 1999, Milo et al. 2002). One important finding is that a gene co-expression network has the universal topological features of complex network systems characterized by modularity, presence of hub and power-law degree distribution (Jeong et al. 2000, Jeong et al. 2001, Featherstone and Broadie 2002). However, although the recognition of such similarity to other complex systems has accelerated the understanding of the global topology of cellular networks, the global topology itself has revealed little about the regulatory mechanisms of specific biological processes. To obtain practical information for tackling biological problems associated with specific processes, it is necessary to focus more on the smaller architecture of the co-expression network.

[1]These authors contributed equally to this work.
*Corresponding author: E-mail, shibata@kazusa.or.jp; Fax, +81-438-52-3948.

In this mini review, we present an overview of recent efforts in the field of plant biology to elucidate gene regulatory mechanisms using co-expression analysis. First, we briefly introduce the cellular network architecture, network substructures, general strategies of co-expression analysis and public co-expression databases. Next, we summarize recent results of co-expression analysis and integrated co-expression analysis combining transcriptome and metabolome data. Finally, the limitations and future perspectives of co-expression network analysis are discussed. This mini review mainly focuses on co-expression networks. Readers should refer to other reviews for understanding general aspects of network biology (Barabasi and Oltvai 2004) and gene expression data analysis (Allison et al. 2006).

Terminology

The term 'co-expression' refers to a similarity of gene expression patterns across a variety of experimental conditions. The term 'network' is widely used in many fields of science. For example, in plant molecular genetics, a directed network is frequently used to illustrate positive and negative regulatory relationships between genes. Here, we refer to a network as an undirected graph composed of nodes and links representing genes and mutual co-expression relationships, respectively. The term 'topology' refers to patterns of node-to-node connectivity, or configuration of links.

Network architecture

The complexity of biological networks has a hierarchy (Oltvai and Barabasi 2002). Between the levels of genome-wide organization and individual molecular components (Fig. 1A), there are substructures such as modules (Rives and Galitski 2003), motifs (Milo et al. 2002) and pathways, characterized by topological properties such as degree distribution (Barabasi and Albert 1999), network density (Barnes 1969) and clustering coefficient (Watts and Strogatz 1998). Definitions of these concepts are summarized in Fig. 1B. There is particular interest in identifying topological modules. Given that most biological functions cannot be attributed to a single gene, a module is likely to represent a set of genes having a discrete function that arises from interactions among them (Hartwell et al. 1999). Conversely, the analysis of local modules may be more informative with respect to the regulatory mechanisms of the specific biological processes. Therefore, the identification of the modular structure is a primary goal in co-expression network analysis.

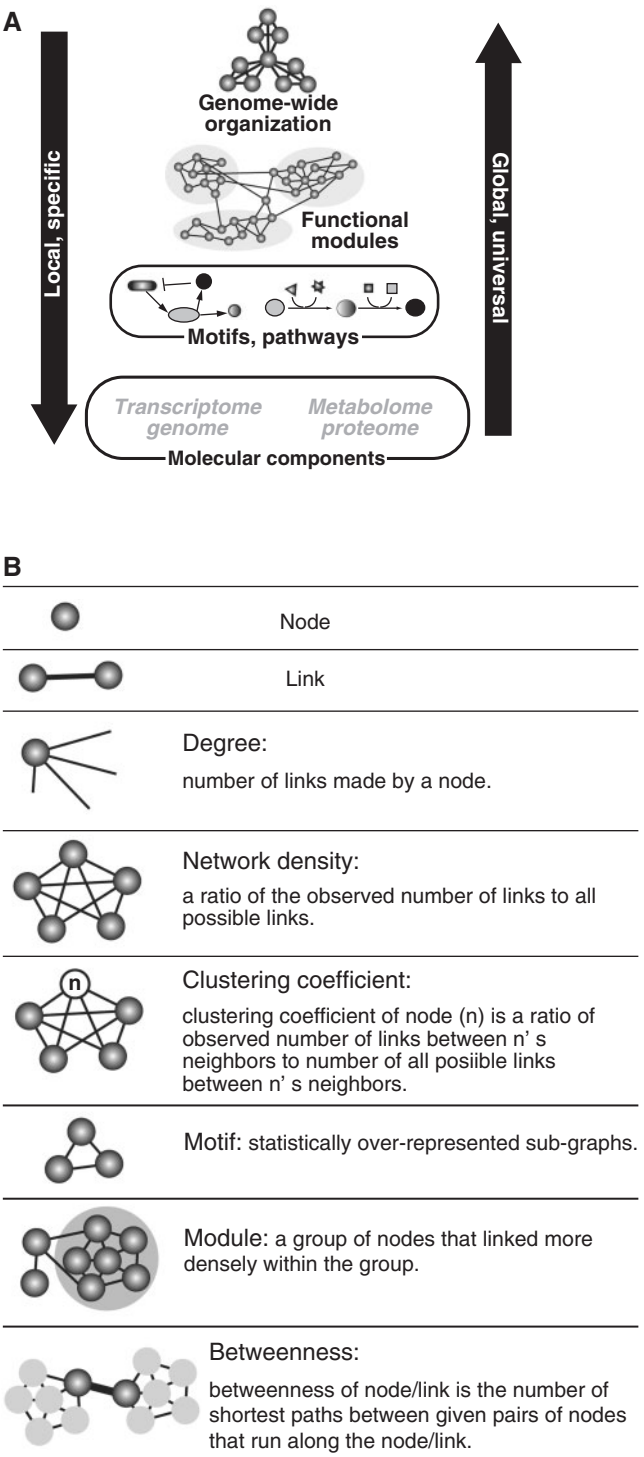

Fig. 1 Network architecture. (A) Hierarchical structure of cellular networks. From the top to the bottom, structures associate more closely with local and specific cellular processes. The concept of this illustration is adopted from Oltvai and Barabasi (2002) with permission. (B) Definitions of terms describing networks.

Protocol for co-expression network analysis

Fig. 2 illustrates practical protocols of co-expression network analysis, where strategies are conceptually classified into two categories. The first category uses a 'guide-gene' approach (Fig. 2, left). Prior to correlation coefficient analysis, an appropriate set of genes relating to the biological problem is selected based on experimental knowledge and literature information. The pre-selected set of genes are termed 'guide genes' (Lisso et al. 2005) or 'bait genes' (Wei et al. 2006). Here, we use the term 'guide genes', as the term 'bait gene' may cause confusion with the alternative meaning used in a molecular biological context referring to a yeast two-hybrid system. In the first round of co-expression analysis, correlation coefficients between the guide genes (guide genes 1) are retrieved from a correlation coefficient data set calculated from gene expression data (e.g. microarray). The visualization of co-expression using a network viewer such as that of Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/) (Batagelj and Mrvar, 2003) or BioLayout (http://www.biolayout.org/) (Enright and Ouzounis 2001) provides an intuitive grasp of co-expression modules. In this step, finding co-expression modules within the guide genes, as well as correlation between the guide genes and all other genes, is expected. A set of correlated genes found in the first round of analysis can be combined with another set of guide genes (guide genes 2), and the combination used as the guide genes in the second round of analysis. In summary, using a guide-gene approach, one can expect to find genes that directly or indirectly correlate with the genes of interest.

To demonstrate the guide-gene protocol, we show an example of our analysis on genes associated with the phenylpropanoid biosynthesis pathway. We selected 'flavonoid biosynthesis' genes and 'cinnamate–monolignol pathway/sinapoyl ester biosynthesis' genes from the pathway viewer KaPPA-View (http://kpv.kazusa.or.jp/kappa-view/, genes in pathway maps Ath00061 and Ath00064, respectively) (Tokimatsu et al. 2005) as guide genes 1 (Fig. 3A). To investigate co-expression patterns within guide genes 1, we searched the co-expression database ATTED-II (see next section) for pair-wise Pearson correlation coefficients (PCCs) using a PCC cut-off threshold of 0.6, and visualized the resulting co-expression network using Pajek. This first round of analysis revealed that the guide genes were classified into four distinct co-expression modules (Fig. 3B). Most of the genes in modules 1, 2 and 3 belonged to the flavonoid biosynthesis pathway (Ath00061). On the other hand, genes in module 4 belonged to the cinnamate–monolignol pathway (Ath00064). This result suggested that expression of flavonoid biosynthesis and cinnamate–monolignol pathway genes is differentially coordinated to produce pathway-specific metabolites. Next, to investigate the

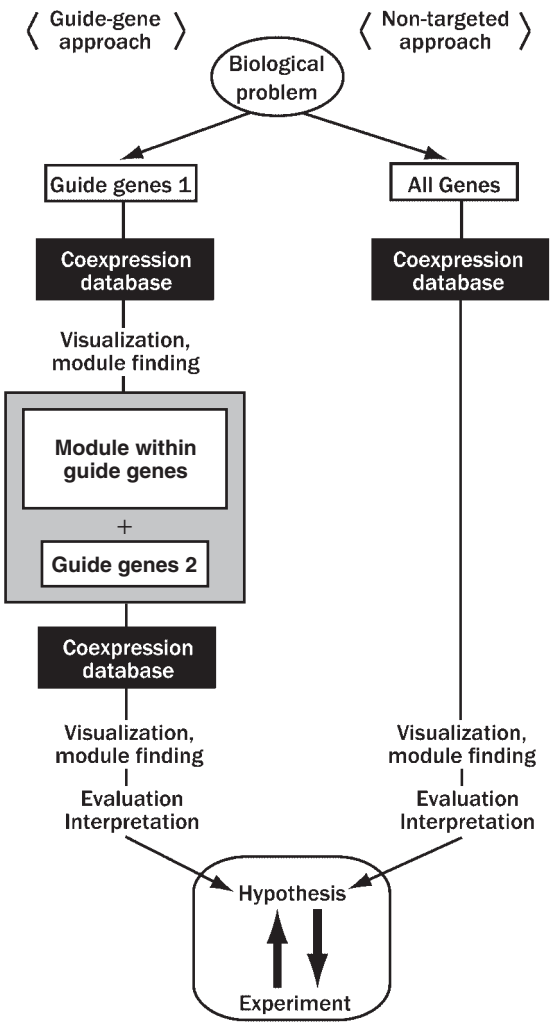

Fig. 2 Practical protocols of co-expression analysis. Left: guide-gene approach, in which co-expression profiles between and within selected guide genes are first investigated. Right: non-targeted approach, in which the modular structure is extracted from the entire network according to the topology of the links.

relationship between the phenylpropanoid pathway and metabolically upstream pathways, the genes in the four modules were combined with guide genes 2 containing 'aromatic amino acid biosynthesis' genes, 'Calvin cycle' genes and 'pentose phosphate pathway' genes taken from KaPPA-View (genes in pathway maps Ath00017, Ath00112 and 0001, respectively). We again searched ATTED-II for pair-wise PCCs using a cut-off threshold of 0.6. This second round of analysis revealed that two genes in the aromatic amino acid biosynthesis pathway (Ath00017), the 5-enolpyruvylshikimate-3-phosphate synthase gene (*EPSP synthase*) and the 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase gene (*DAHP synthase*), were

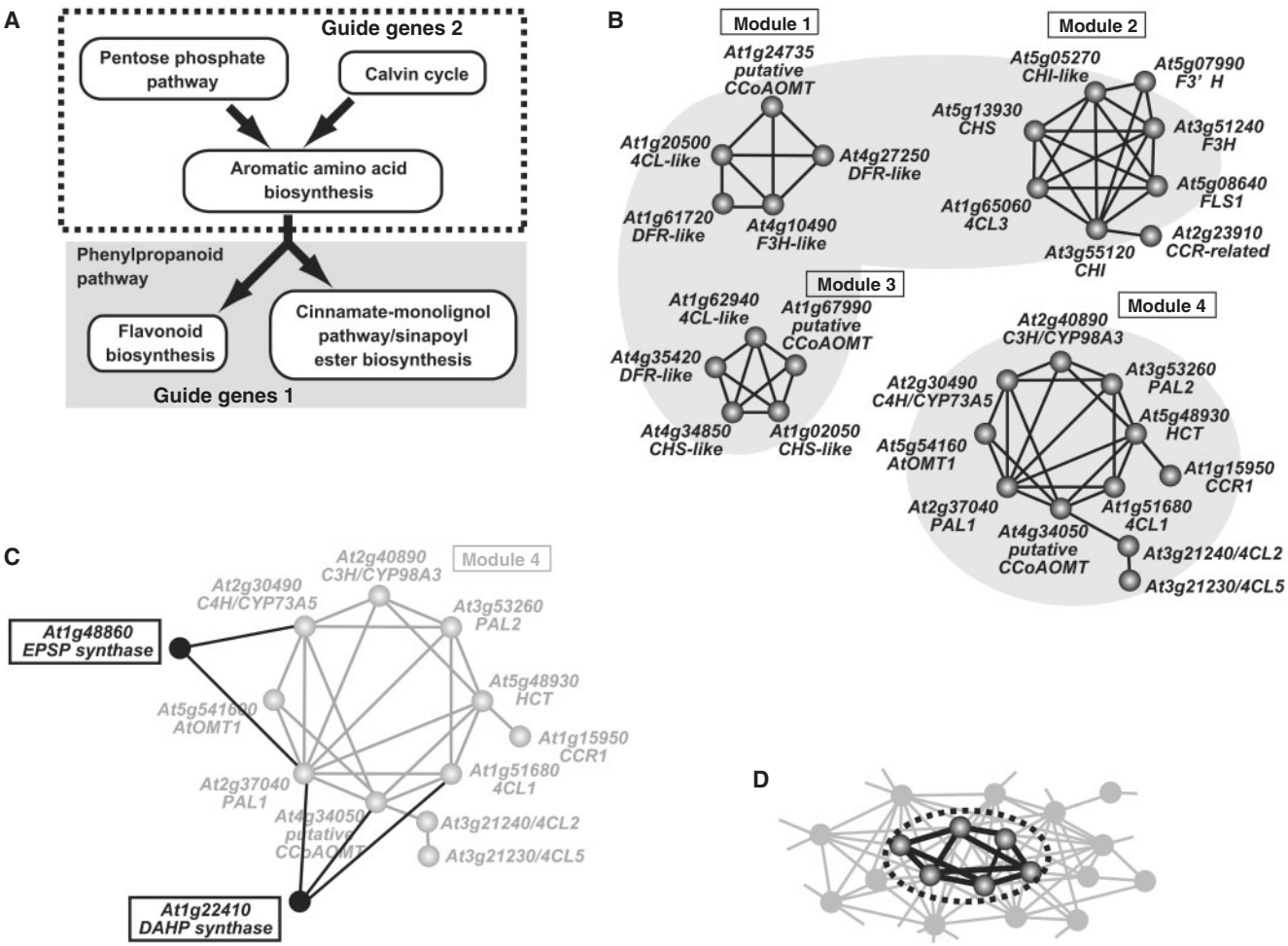

Fig. 3 Example of co-expression analysis using the guide-gene approach. (A) To analyze phenylpropanoid pathway genes, genes involved in flavonoid biosynthesis (KaPPA-View map number Ath00061) and cinnamate–monolignol pathway/sinapoyl ester biosynthesis (KaPPA-View map number Ath00064) were selected as guide genes 1. To investigate the relationship between the phenylpropanoid pathway and upstream pathways, genes involved in aromatic amino acid biosynthesis (KaPPA-View map number Ath00017), the Calvin cycle (KaPPA-View map number Ath00112) and the pentose phosphate pathway (KaPPA-View map number 0001) were selected as guide genes 2. (B) Co-expression network of guide genes 1 (phenylpropanoid pathway genes). Most of the genes in modules 1, 2 and 3 were involved in flavonoid biosynthesis (Ath00061). All genes in module 4 were involved in the cinnamate–monolignol pathway (Ath00064). *CCoAOMT*, cafferoyl-CoA 3-*O*-methyltransferase; *4CL*, 4-coumarate-CoA ligase; *DFR*, dihydroflavonol 4-reductase; *F3H*, flavanone 3-hydroxylase; *CHI*, chalcone isomerase; *CHS*, chalcone synthase; *F3′H*, flavonoid 3′-hydroxylase; *FLS*, flavonol synthase; *CCR*, cinnamoyl-CoA reductase; *C3H*, coumarate 3-hydroxylase; *C4H*, cinnamate 4-hydroxylase; *AtOMT1*, *Arabidopsis thaliana O*-methyltransferase 1; *PAL*, phenylalanine ammonia-lyase; *HCT*, hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase. (C) Upstream pathway genes co-expressed with cinnamate–monolignol pathway genes. Co-expression of two genes involved in aromatic amino acid biosynthesis (Ath00017) with cinnamate–monolignol pathway (Ath00064) genes in module 4 was found using guide genes 2. *EPSP synthase*, 5-enolpyruvylshikimate-3-phosphate synthase; *DAHP synthase*, 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase. (D) A drawback of the guide-gene approach. A module found by this approach (circles and lines in the black broken circle) may be a part of a larger and more densely connected module (gray circles and gray lines).

co-expressed with cinnamate–monolignol pathway genes in module 4 (Fig. 3C). These genes did not have co-expression links with modules containing flavonoid biosynthesis genes. The result suggested the hypothesis that aromatic amino acid biosynthesis is more tightly coordinated with the cinnamate–monolignol pathway than with flavonoid biosynthesis. In addition, it suggested that *EPSP synthase* and *DAHP synthase* could be regulatory points to control metabolic flow from sugar phosphate to monolignol.

The other category uses a 'non-targeted' approach (Fig. 2, right). In this approach, a knowledge-independent module search of the entire network is performed based on the topology of the network. 'Module' can be defined as a group of densely connected nodes that have a sparsely

Table 1 Public databases of *Arabidopsis* gene co-expression

Website	Reference	Description[a]
Arabidopsis Coexpression Data Mining Tool (http://www.arabidopsis.leeds.ac.uk/act/)	Jen et al. (2006)	443 arrays, single-gene correlation coefficient query, expression pattern displayer, *cis*-element analyzer, ID/function linker tool for Affymetrix array, etc
AthCoR@CSB.DB (http://csbdb.mpimp-golm.mpg.de)	Steinhauser et al. (2004)	123 arrays, multiple-gene correlation coefficient query, Spearman's Rho rank, Kendall's coefficient of rank, Pearson's linear product–moment; multiple output mode, etc.
ATTED-II (http://www.atted.bio.titech.ac.jp)	Obayashi et al. (2007)	1,388 arrays, multiple-gene correlation coefficient query, *cis*-element prediction, expression data graph, gene correlation table is available, etc.
Genevestigator (http://www.genevestigator.ethz.ch)	Zimmermann et al. (2004)	2,620 arrays, single-gene correlation coefficient query, digital Northern, response viewer, gene chronologer (growth stage), gene atlas (organ/tissue), meta-analyzer, mutant surveyor, etc.
The Botany Array Resource (http://bbc.botany.utoronto.ca/)	Toufighi et al. (2005)	1,430 arrays, single-gene correlation coefficient query, electronic Northern, illustration of gene expression map, promoter analysis, users upload their expression data and compute correlation coefficients, gene correlation data are available, etc.

[a] Number of array experiments used for correlation coefficient calculation (in December 2006), and other search and analytical functions.

connected periphery. Several algorithms have been proposed to extract such groups computationally, which will be reviewed in a later section.

Because the guide-gene approach allows moderately sized analyses compared with the computationally-intensive non-target approach, it is appropriate for gene prioritization in a single-investigator study. However, we note a pitfall of this approach. As the guide-gene approach uses a pre-selected gene set, it cannot exclude the possibility that an expression module found by this approach may be part of a larger and more densely connected module (Fig. 3D). It is important to test whether connectivity within the module is higher than that to the outside. Practically, and most simply, this test can be done by searching the co-expression links of the module-member genes against all genes.

After finding the co-expression module(s) using either approach, the validity of the modules is evaluated by statistical tests based on random resampling (Lisso et al. 2005, Wei et al. 2006) or permutation (shuffling of the data to create pseudo data sets) (Stuart et al. 2003), and the results are interpreted using biological knowledge such as annotation and pathway information. Finally, generation of an appropriate hypothesis is expected to result from the co-expression network analysis. Here, we emphasize that the use of co-expression network analysis is an efficient way to develop a hypothesis, but not to prove that hypothesis. Co-expression does not necessarily indicate a direct regulatory relationship. Therefore, the hypothesis derived from the analysis needs experimental verification to ensure that the observed co-expression is biologically relevant.

Arabidopsis co-expression databases

Researchers can calculate gene-to-gene correlation coefficients using their own expression data sets. Alternatively, researchers can retrieve correlation coefficient data from public databases (Table 1). These databases provide results of large-scale correlation coefficient calculations using expression data from various experimental conditions (ranging from 123 to 2,620 arrays) deposited by international microarray consortiums such as AtGenExpress (Mirror site in Japan, http://pfg.psc.riken.jp/AtGenExpress/links.html), NASCArray (http://affymetrix.arabidopsis.info/) (Craigon et al. 2004), Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (Edgar et al. 2002) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) (Brazma et al. 2003). At present (December 2006), all the databases in Table 1 accept single-gene queries for a correlation coefficient search. AthCoR@CSB.DB (http://csbdb.mpimp-golm.mpg.de) (Steinhauser et al. 2004) and ATTED-II

(http://www.atted.bio.titech.ac.jp) (Obayashi et al. 2007) accept multigene queries as well. *Arabidopsis* Co-expression Data Mining Tool (ACT) (http://www.arabidopsis.leeds.ac.uk/act/) (Jen et al. 2006), ATTED-II, Genevestigator (http://www.genevestigator.ethz.ch) (Zimmermann et al. 2004) and The Botany Array Resource (BAR) (http://bbc.botany.utoronto.ca/) (Toufighi et al. 2005) are implemented with gene expression visualizing tools. The expression visualizing tools of ACT, Genevestigator and BAR accept multigene queries. Additionally, lists of correlation coefficients are downloadable in ATTED-II and BAR.

To obtain an overview of the topological features of the entire *Arabidopsis* co-expression network, we conducted a survey on the link number, node number and network density with respect to the nodes that have at least one link at certain PCC cut-offs based on correlation coefficients provided by ATTED-II (Fig. 4A, B). The number of links and nodes decreases with increasing PCC cut-off threshold (Fig. 4A). Network density, however, displays a minimal value at a PCC ranging from 0.55 to 0.66, and shows a slight increase at a PCC cut-off greater than this range (Fig. 4B). Below the PCC cut-off 0.55–0.66, many low PCC links connect nodes together. Naturally, the high network densities in the low PCC range do not necessarily mean significant correlations. In contrast, above the PCC cut-off 0.55–0.66, an increase in network densities is attributed to the presence of high PCC links densely connecting a decreasing number of nodes. This implies that biologically significant modules are expected to be found above the PCC cut-off where the network density displays a minimal value.

Are the numbers of microarray experiment data used to calculate correlation coefficients large enough to generate robust co-expression networks? It has been reported that accuracy in identification of co-regulated genes from co-expression analysis plateaus at 50–100 experiments in the case of yeast (Yeung et al. 2004). To test this in *Arabidopsis*, we randomly selected microarray experiments from ATTED-II, and calculated the network density of the entire network generated from random sets of these experiments (Fig. 4C). The result demonstrated that network density essentially reached equilibrium once >100 arrays were used. This result suggests that the sizes of the microarray data sets used for the calculation of correlation coefficients in the public co-expression databases are reasonably large enough to generate condition-independent co-expression networks.

Co-expression analysis to identify new genes and functional modules

Co-expression analysis has already been applied to some plant biological problems, and successfully generated

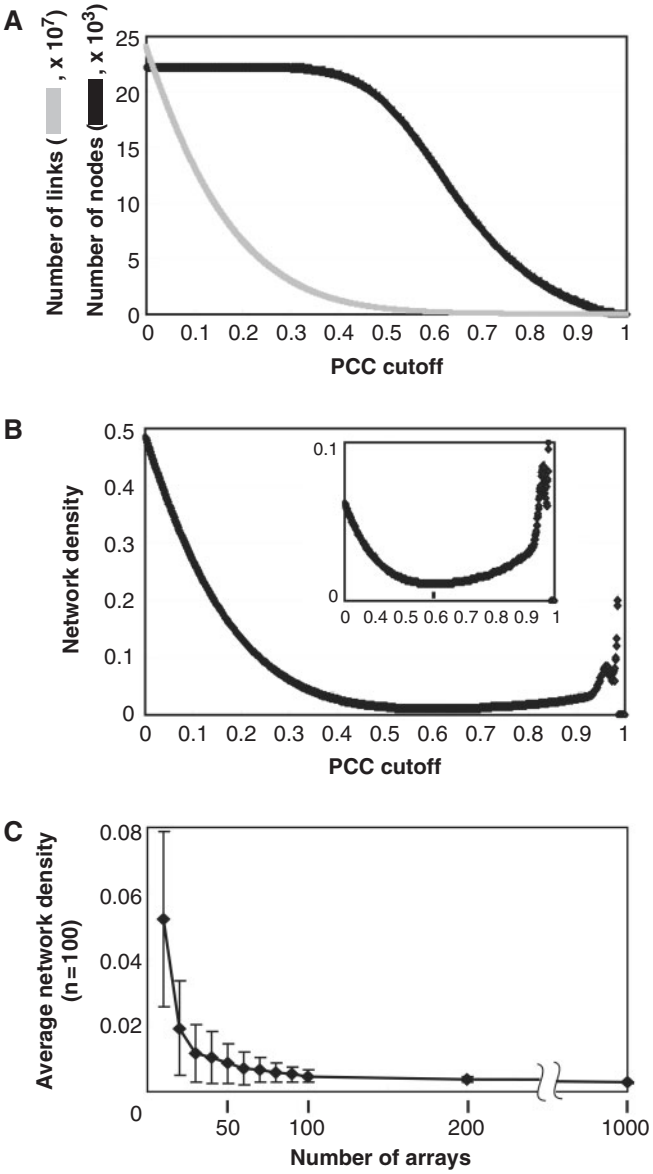

Fig. 4 Overview of global features of the *Arabidopsis* co-expression network. (A) The number of links (gray) and number of nodes (black) in the entire network at positive PCC cut-off values. (B) Network density of the entire network at positive PCC cut-off values. The magnified curve (inset) demonstrates that the network density shows its minimum at a PCC cut-off of 0.55–0.66 (network density is 0.011), and then increases above a PCC cut-off of 0.66. Correlation coefficients were retrieved from ATTED-II. (C) Correlation between numbers of array data used for PCC calculation and global feature of the network. The indicated numbers (10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200 and 1,000 array data) of array data were randomly selected from ATTED-II, and network densities of the entire network generated from the selected array data were calculated at a PCC cut-off of 0.7. Random sampling of each indicated number of array data was performed for 100 repeats. Network density essentially equilibrated once more than 100 array data were used. Similar results were obtained for PCC cut-offs of 0.5, 0.6, 0.8 and 0.9 (data not shown).

biologically relevant hypotheses with respect to functional relationships between genes. In this section, we summarize the results from recent reports.

Guide-gene approach: cell wall formation

Cellulose is synthesized by plasma membrane-localized complexes containing cellulose synthase (CESA) subunits. Persson and co-workers analyzed genes co-expressed with *CESA* genes using a linear regression method based on expression data from the NASCArray (Persson et al. 2005). They identified distinct *CESA* gene clusters associated with primary (*CESA1*, *3* and *6*) and secondary (*CESA4*, *7* and *8*) cell wall formation. Several genes known to be involved in cellulose synthesis, such as *COBRA*, endochitinase-like gene 1 (*CTL1*) and their paralogs, were co-expressed with *CESA* clusters. These findings suggested that the two groups of *CESA* genes act as central parts of functional modules for primary and secondary cell wall synthesis, respectively. Furthermore, the results suggested the hypothesis that *COBRA* and *CTL* are components or regulators of *CESA* functional modules.

To test the biological significance of the apparent co-expression, the authors selected genes (At5g54690, At3g16920, At5g03170 and At4g27435) that are highly co-expressed with secondary cell wall-related *CESA* genes and examined the phenotypes of the T-DNA insertion mutants. All insertion lines appeared to have alterations in cell wall composition based on their Fourier transform infrared (FTIR) spectra. These results supported the hypothesis that these genes are functionally associated with cell wall formation.

The authors also estimated degrees of co-expression of pathways to *CESA* modules by calculating pathway scores (Persson et al. 2005). Both *CESA* modules were coordinated with several common pathways, including the homogalacturonan degradation pathway. On the other hand, the brassinosteroid pathway had higher co-expression for *CESA1*, *3* and *6*. Lignin and dTDP-rhamnose biosynthesis pathways showed higher co-expression for *CESA4*, *7* and *8*. These co-expression patterns suggested that cell wall synthesis is functionally coordinated with other metabolic pathways.

Guide-gene approach: isoprenoid biosynthesis

The isoprenoid biosynthesis pathway consists of two spatially separated pathways. One is the mevalonate (MVA) pathway in the cytosol, and the other is the methylerythritol phosphate (MEP) pathway in the plastid. Wille and co-workers analyzed a network of 40 isoprenoid-related genes using PCC calculation and a modified Gaussian graphical model (Wille et al. 2004). Their results identified the MVA and MEP pathways as two separate modules. Acetoacetyl-CoA thiolase 1 (*AACT1*) and hydroxymethylglutaryl-CoA reductase 1 (*HMGR1*), genes for key enzyme of the MVA pathway, were negatively correlated with the expression of MEP pathway genes. Isopentenyl diphosphate isomerase 1 (*IPPI1*) in the plastidic pathway was positively correlated with MVA pathway genes. They thus hypothesized that these genes may be candidates for the regulator of cross-talk between the two pathways. In this report, the correlation between the isoprenoid pathway and downstream metabolic pathways was inferred by incorporating an additional 795 genes in the network analysis. The result demonstrated that MVA pathway genes were closely correlated with plastoquinone and phytosterol pathway genes, while MEP pathway genes were correlated with carotenoid and chlorophyll pathway genes.

Guide-gene approach: whole metabolic network

A comprehensive co-expression analysis of metabolic pathway genes has been reported recently (Wei et al. 2006). In this report, the authors retrieved 1,330 genes associated with metabolism from the AraCyc database (Mueller et al. 2003; http://arabidopsis.org/tools/aracyc/). Using these metabolism-related genes as guide genes, they analyzed the gene co-expression profile by linear regression, and confirmed that a few previously reported opinions for other cases were also true in the case of the *Arabidopsis* co-expression network. First, genes belonging to the same metabolic pathway are likely to be co-expressed, as has been demonstrated for metabolic pathway genes of yeast (Ihmels et al. 2004, Kharchenko et al. 2005). Secondly, the degree of distribution of the node was skewed, with the majority of metabolic genes having a small number of links while a few had a large number of links, as also seen in yeast (Magwene and Kim 2004). Additionally, they developed a pathway-level co-expression analysis. In this analysis, guide genes from the metabolic pathway of interest are selected first (which they called 'bait genes'), and then genes that are co-expressed with more than one member of the guide genes were searched for in the whole genome. Pathway-level co-expression analysis allows selection of candidate genes that may be responsible for regulation or coordination of the expression of metabolic genes.

In addition, Wei et al. (2006) observed that single-copy metabolic genes tend to have multiple links, while genes with multiple paralogs had fewer links. Interestingly, the study of gene co-expression using various *Arabidopsis* tissues has demonstrated that large gene families had highly correlated expression patterns within the families (Schmid et al. 2005). These results cannot be compared directly because they focused on different aspects of co-expression; Wei and co-workers focused on co-expression with all other genes, while Schmid and co-workers focused on within-family co-expression.

However, these results raise a general question in terms of correlation between the size of a gene family and the degree of co-expression. In yeast, it has been reported that there is a higher probability of compensation for duplicate genes (Gu et al. 2003). In addition, it has been reported that the disruption of well-connected genes is likely to result in the exhibition of a severe phenotype (Jeong et al. 2001). These results suggest that copy numbers and link numbers of genes correlate with the robustness of the cellular system. Further investigation is required to clarify the dynamics between acquisition of co-expression linkage and gene duplication.

Non-targeted approach: top-down and bottom-up module detection

Module detection followed by inspection of member gene annotations is one of the key steps in network analysis necessary to infer gene functions. For example, if an unknown gene was found in a densely connected module in which other member genes were known to be involved in a certain cellular process, it would be hypothesized that the unknown gene had functional relationships with that process. A non-targeted approach aims to detect local modular structures from the entire co-expression network, according to the topology of the links. In comparison with the guide-gene approach that depends on knowledge of biological processes, a non-targeted approach facilitates knowledge-independent detection of modules. Thus, discovery of novel modules that may not be obtained using the guide-gene approach is expected.

The top-down strategy of the non-targeted approach is summarized as the process of finding densely connected regions separated by sparsely connected regions. This approach has been employed to detect modules, or 'community structure', in social and biological networks (Girvan and Newman 2002). The algorithm is based on the iterative removal of links with high 'betweenness' (Fig. 1B), i.e. removal of links along which many of the shortest paths between pairs of nodes run. Links with high 'betweenness' are likely to represent the periphery of modules (Fig. 5A). By removing these links, the authors separated modules from one another, and successfully demonstrated the underlying structures of the network.

In contrast to this top-down module extraction approach, a bottom-up approach has been used to detect modules, or 'clusters', in protein–protein interaction networks of *E. coli* and yeast (Altaf-Ul-Amin et al. 2006). The authors grew seed clusters by adding neighboring nodes if the addition of the node did not decrease the network density or cluster property (Fig. 5B; for definition of 'cluster property', see Altaf-Ul-Amin et al. 2006). In many clusters detected using this algorithm, proteins associated with similar functional classes were densely linked, suggesting that the algorithm predicts biologically relevant protein complexes.

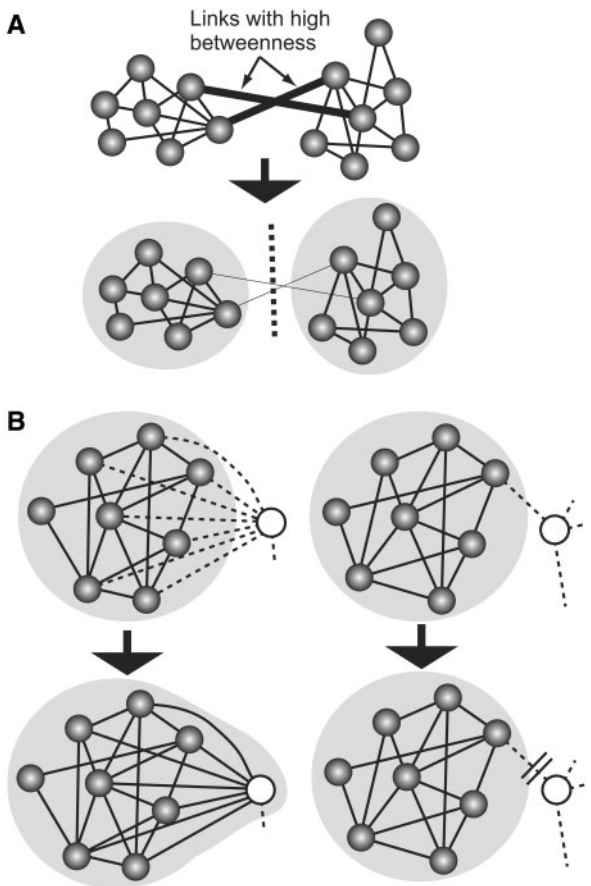

Fig. 5 Non-targeted strategies of module detection. (A) Top-down module extraction approach. Links with high 'betweenness' are removed iteratively to separate the network into modules. (B) Bottom-up module detection approach. If a neighboring node (white) does not decrease the network density of the seed cluster (gray), it is included in the cluster (left). If a neighboring gene decreases the network density, it is not included in the cluster (right).

As these module detection strategies depend solely on the topological property of the network, we expect the strategies also to be applied to co-expression networks.

Integrated analysis of gene co-expression combined with other omics data

Gene expression controls accumulation of metabolites, which in turn regulate the gene expression. Thus, a combined analysis of metabolite accumulation and gene co-expression provides new insights into regulatory processes of metabolite production. Hirai and co-workers combined metabolome and transcriptome data to identify control mechanisms regulating responses to sulfur and nitrogen deficiency (Hirai et al. 2004). Expression profiles of 13,000 expressed sequence tags (ESTs) and metabolic

profiles of 3,000 mass peaks were obtained by cDNA macroarray and Fourier transform mass spectrometry, respectively. Fold changes in expression intensities of ESTs and mass peaks were combined into a single matrix, and then the expression patterns were classified according to similarity by a batch-learning self-organizing map method. Consequently, the authors found regulatory linkage among nutrient deficiency, primary metabolism and glucosinolate metabolism (Hirai et al. 2004). Moreover, they showed the possibility that co-expression analysis of transcripts and metabolites could identify regulatory metabolites and genes of metabolic pathways (Hirai et al. 2005). For example, *O*-acetylserine was clustered together with genes induced by sulfur deficiency, suggesting that the genes were coordinately regulated by *O*-acetylserine under sulfur deficiency. Several putative transcription factor genes were clustered with glucosinolate biosynthesis genes, suggesting that the transcription factors were candidate genes for controlling glucosinolate metabolism.

Similar gene–metabolite co-expression analysis successfully identified terpene synthase genes involved in volatile compound formation in cucumber (Mercke et al. 2004). Combined gene–metabolite co-expression analysis using a plant overexpressing *PAP1* transcription factor also identified novel glucosyltransferase genes involved in anthocyanin biosynthesis (Tohge et al. 2005). In these reports, the gene functions predicted by co-expression analyses were confirmed experimentally using bacterial expression system or T-DNA insertion lines.

Nikiforova and co-workers used network analysis to investigate gene co-response to sulfur deficiency (Nikiforova et al. 2005). They rearranged the original co-expression network consisting of 6,454 genes and 81 metabolites into a cause-to-effect network starting from sulfur. This approach predicted that the sulfur deficiency caused an enhanced lateral root formation via auxin- and calcium-related signaling pathways.

The integrated analysis of metabolite accumulation and gene expression has recently been applied to non-model plants. Rischer and co-workers analyzed a gene–metabolite co-expression network of the medicinal plant *Catharanthus roseus* (Rischer et al. 2006). To overcome the lack of a gene expression profiling method such as microarray, the authors used cDNA AFLP (amplified fragment length polymorphism) technology to acquire quantitative gene expression profiles of *C. roseus*. Gene-to-gene and gene-to-metabolite correlation networks allowed them to hypothesize novel cytochrome P450 genes involved in terpenoid indole alkaloid biosynthesis and novel AP2-domain transcription factor genes possibly regulating terpenoid indole alkaloid biosynthesis. This report demonstrated a potential use of integrated co-expression analysis to examine the metabolic regulation of non-model plants.

For simultaneous visualization of transcription and metabolite networks, pathway viewer tools, which overlay gene expression data onto metabolic pathway maps, provide a bird's eye view of experimentally observed changes. This type of pathway viewer includes The Pathway Tools Omics Viewer (http://www.arabidopsis.org:1555/expression.html) (Paley and Karp 2006), MapMan (http://gabi.rzpd.de/projects/MapMan/) (Thimm et al. 2004) and KaPPA-View (http://kpv.kazusa.or.jp/kappa-view/) (Tokimatsu et al. 2005).

Conclusions and perspectives

Accumulation of *Arabidopsis* transcriptome data has facilitated the genome-wide analysis of gene co-expression profiles. Several co-expression databases provide condition-independent correlation coefficients computed from large sets of microarray data. These databases have allowed the search of co-expressed genes with genes of interest. Co-expression networks constructed from pair-wise correlation coefficients have provided an efficient way to identify functional transcription modules associated with specific biological processes. The biologically relevant hypotheses developed using co-expression analysis have assisted in the design of hypothesis-driven experiments and gene prioritization for those experiments. In summary, co-expression analysis, using microarray data accumulated so far, is now within reach of many researchers, even if they do not compute the correlation coefficients themselves.

Correlation coefficients provided in the databases are a convenient measure for estimating gene-to-gene co-expression. However, we emphasize that it is crucial to review original expression data. Genes naturally exhibit high correlation if entire expression patterns across diverse conditions are similar. On the other hand, genes also exhibit high correlation if they are expressed together under a few conditions and are otherwise silent. Thus, reviewing original expression data provides insights into the reason why genes of interest show high correlation. Some of the co-expression databases implement a browser of original expression data, which helps in the discrimination of meaningful co-expression profiles from less meaningful ones.

Co-expression analysis has laid the foundation for the system-level understanding of physiological processes. The next steps include development of methodologies to integrate multiple omics data sets, as has been proposed for human and zebrafish (Aerts et al. 2006, Butte and Kohane 2006). Associations between genome, transcriptome, proteome, metabolome and phenome will be considered together to uncover regulatory relationships that cannot be extracted from a single omics data set. This line of study may reveal the function of genes that do not show an apparent co-expression with any other genes. In addition,

the next steps include the analysis of time series expression data. The extent of time displacement existing between gene expression and its end-points (e.g. metabolite accumulation, phenotype change) needs to be gauged when relating gene expression to other omics data using classical correlation methods. A time scale of response and re-equilibration of gene expression may include information such as the nature of interaction within the cellular system (Nicholson et al. 2004). Finally, with the development of the methodology, correlation-based analysis will shed new light not only on the static but also on the dynamic aspects of behavior of plant cellular systems.

Acknowledgments

We thank Dr. Takeshi Obayashi (The Institute of Medical Science, The University of Tokyo) for providing us with robust multiarray analysis (RMA)-normalized gene expression data used to calculate the correlation coefficients in ATTED-II. We also thank Drs. T. Ohno, R. Sano, K. Sugiyama, T. Dansako, M. Takeda, N. Sakurai, H. Suzuki and Y. Iijima (Kazusa DNA Research Institute) for helpful discussions. This work was performed as part of the technology development projects of the 'Green Biotechnology Program', which was supported by New Energy and Industrial Technology Development (NEDO).

References

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al. (2006) *Nat. Biotechnol.* 24: 537–544.

Allison, D.B., Cui, X., Page, G.P. and Sabripour, M. (2006) *Nat. Rev. Genet.* 7: 55–65.

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K. and Kanaya, S. (2006) *BMC Bioinformatics* 7: 207.

Barabasi, A.L. and Albert, R. (1999) *Science* 286: 509–512.

Barabasi, A.L. and Oltvai, Z.N. (2004) *Nat. Rev. Genet.* 5: 101–113.

Barnes, J.A. (1969) *In* Social Networks in Urban Situations. Edited by Mitchell, J.C. pp. 51–76. Manchester University Press, Manchester.

Batagelj, V. and Mrvar, A. (2003) *In* Graph Drawing Software. Edited by Jünger, M. and Mutzel, P. pp. 77–103. Springer, Berlin.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., et al. (2003) *Nucleic Acids Res.* 31: 68–71.

Butte, A.J. and Kohane, I.S. (2006) *Nat. Biotechnol.* 24: 55–62.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) *Nucleic Acids Res.* 32: D575–577.

Edgar, R., Domrachev, M. and Lash, A.E. (2002) *Nucleic Acids Res.* 30: 207–210.

Enright, A.J. and Ouzounis, C.A. (2001) *Bioinformatics* 17: 853–854.

Featherstone, D.E. and Broadie, K. (2002) *Bioessays* 24: 267–274.

Girvan, M. and Newman, M.E. (2002) *Proc. Natl Acad. Sci. USA* 99: 7821–7826.

Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W. and Li, W.H. (2003) *Nature* 421: 63–66.

Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) *Nature* 402: C47–C52.

Hirai, M.Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D.B., et al. (2005) *J. Biol. Chem.* 280: 25590–25595.

Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T. and Saito, K. (2004) *Proc. Natl Acad. Sci. USA* 101: 10205–10210.

Ihmels, J., Levy, R. and Barkai, N. (2004) *Nat. Biotechnol.* 22: 86–92.

Jen, C.H., Manfield, I.W., Michalopoulos, I., Pinney, J.W., Willats, W.G., Gilmartin, P.M. and Westhead, D.R. (2006) *Plant J.* 46: 336–348.

Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) *Nature* 411: 41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) *Nature* 407: 651–654.

Kharchenko, P., Church, G.M. and Vitkup, D. (2005) *Mol. Syst. Biol.* 1: E1-E6.

Lisso, J., Steinhauser, D., Altmann, T., Kopka, J. and Mussig, C. (2005) *Nucleic Acids Res.* 33: 2685–2696.

Magwene, P.M. and Kim, J. (2004) *Genome Biol.* 5: R100.

Mercke, P., Kappers, I.F., Verstappen, F.W., Vorst, O., Dicke, M. and Bouwmeester, H.J. (2004) *Plant Physiol* 135: 2012–2024.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) *Science* 298: 824–827.

Mueller, L.A., Zhang, P. and Rhee, S.Y. (2003) *Plant Physiol.* 132: 453–460.

Nicholson, J.K., Holmes, E., Lindon, J.C. and Wilson, I.D. (2004) *Nat. Biotechnol.* 22: 1268–1274.

Nikiforova, V.J., Daub, C.O., Hesse, H., Willmitzer, L. and Hoefgen, R. (2005) *J. Exp. Bot.* 56: 1887–1896.

Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K. and Ohta, H. (2007) *Nucleic Acids Res.* 35 (Database issue): D863-D869.

Oltvai, Z.N. and Barabasi, A.L. (2002) *Science* 298: 763–764.

Paley, S.M. and Karp, P.D. (2006) *Nucleic Acids Res.* 34: 3771–3778.

Persson, S., Wei, H., Milne, J., Page, G.P. and Somerville, C.R. (2005) *Proc. Natl Acad. Sci. USA* 102: 8633–8638.

Rischer, H., Oresic, M., Seppanen-Laakso, T., Katajamaa, M., Lammertyn, F., Ardiles-Diaz, W., Van Montagu, M.C., Inze, D., Oksman-Caldentey, K.M. and Goossens, A. (2006) *Proc. Natl Acad. Sci. USA* 103: 5614–5619.

Rives, A.W. and Galitski, T. (2003) *Proc. Natl Acad. Sci. USA* 100: 1128–1133.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) *Nat. Genet.* 37: 501–506.

Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O. and Kopka, J. (2004) *Bioinformatics* 20: 3647–3651.

Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) *Science* 302: 249–255.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. and Stitt, M. (2004) *Plant J.* 37: 914–939.

Tohge, T., Nishiyama, Y., Hirai, M.Y., Yano, M., Nakajima, J., et al. (2005) *Plant J.* 42: 218–235.

Tokimatsu, T., Sakurai, N., Suzuki, H., Ohta, H., Nishitani, K., Koyama, T., Umezawa, T., Misawa, N., Saito, K. and Shibata, D. (2005) *Plant Physiol.* 138: 1289–1300.

Toufighi, K., Brady, S.M., Austin, R., Ly, E. and Provart, N.J. (2005) *Plant J.* 43: 153–163.

Watts, D.J. and Strogatz, S.H. (1998) *Nature* 393: 440–442.

Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G.P., Somerville, C. and Loraine, A. (2006) *Plant Physiol.* 142: 762–774.

Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., et al. (2004) *Genome Biol.* 5: R92.

Yeung, K.Y., Medvedovic, M. and Bumgarner, R.E. (2004) *Genome Biol.* 5: R48.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004) *Plant Physiol.* 136: 2621–2632.

(Received October 13, 2006; Accepted January 19, 2007)