

**THE PREDICTION OF LOCAL MODULAR STRUCTURES
IN A CO-EXPRESSION NETWORK
BASED ON GENE EXPRESSION DATASETS**

YOSHIYUKI OGATA¹
yogata@kazusa.or.jp

NOZOMU SAKURAI¹
sakurai@kazusa.or.jp

HIDEYUKI SUZUKI¹
hsuzuki@kazusa.or.jp

KOH AOKI¹
kaoki@kazusa.or.jp

KAZUKI SAITO^{2,3}
ksaito@faculty.chiba-u.jp

DAISUKE SHIBATA¹
shibata@kazusa.or.jp

¹ *Department of Biotechnology Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan*

² *Graduate School of Pharmaceutical Science, Chiba University, Chiba 263-8522, Japan*

³ *Plant Science Center, RIKEN, Yokohama, Kanagawa 230-0045, Japan*

In scientific fields such as systems biology, evaluation of the relationship between network members (vertices) is approached using a network structure. In a co-expression network, comprising genes (vertices) and gene-to-gene links (edges) representing co-expression relationships, local modular structures with tight intra-modular connections include genes that are co-expressed with each other. For detecting such modules from among the whole network, an approach to evaluate network topology between modules as well as intra-modular network topology is useful. To detect such modules, we combined a novel inter-modular index with network density, the representative intra-modular index, instead of a single use of network density. We designed an algorithm to optimize the combinatory index for a module and applied it to Arabidopsis co-expression analysis. To verify the relation between modules obtained using our algorithm and biological knowledge, we compared it to the other tools for co-expression network analyses using the KEGG pathways, indicating that our algorithm detected network modules representing better associations with the pathways. It is also applicable to a large dataset of gene expression profiles, which is difficult to calculate in a mass.

Keywords: network topology, local modular structure, network density, co-expression analysis

1. Introduction

Complex systems are the focus of many studies in various fields, such as systems biology. To visualize and analyze such complex systems, network analysis is considered to be a powerful approach [1]. A network structure is composed of vertices that are linkable to multiple other vertices on the basis of a threshold for vertex-to-vertex connections; e.g., protein-to-protein in a protein interaction network [2-4], and gene-to-gene in a gene co-expression network [5-11], reviewed in [12-14]. Such a network includes ‘network modules’, in which the vertices are tightly interconnected [15,16]. The detection of network modules leads to the elucidation of novel vertex-to-vertex associations in various fields of networks.

In a co-expression network, pairs of genes are connected on the basis of similarity in their expression profiles, which are obtained from DNA microarray datasets, and are used for prediction their functional relatedness, such as the relationship of transcriptional

regulation [6], close relationships in a metabolic pathway [10,17] and subunits of a common protein complex [18]. To detect local network modules from among a co-expression network based on gene expression profiles, recent co-expression network approaches have optimized a cutoff value for gene-to-gene links in the network. Gupta et al. [19] proposed that the ‘average clustering coefficient’ index can be used to optimize a single cutoff value for their gene-to-gene correlation network. Margolin et al. [20] developed the ARACNE tool for detecting co-expressed genes, which exploits the ‘joint probability distribution’ index for evaluating gene-to-gene correlations on the basis of statistical physics and information theory for reducing the number of false-positive co-expression links. The DP-Clus tool uses the ‘cluster property’ index, which is calculated on the basis of network topology to improve the inference of protein-to-protein interactions by removing false-positive interactions [21]. To detect communities in a social network, in which a vertex represents a person and a pair of persons are connected on the basis of human relationship, Bagrow and Bollt [22] proposed the ‘l-shell’ algorithm, which can be executed without knowledge of the entire network. To use these algorithms, a cutoff value for the vertex-to-vertex association index must be selected.

To separate a local network module using an appropriate cutoff value, a combined evaluation of intra-modular and inter-modular connections is useful. A high cutoff value through the entire network may contribute to separating local network modules from each other and, conversely, may lose intra-modular connections. Although a low cutoff value may contribute to intra-modular connections, it may prevent the separation of local network modules. Network density is a representative index for intra-modular connections. As indices for inter-modular connections, the betweenness [23] and cohesiveness [24] indices are used. It is, however, difficult to combine these indices with network density due to the difference in object connections between them.

We devised a novel index for evaluating inter-modular connections and combining with network density and developed an algorithm for detecting network modules using the combination. For combining these indices, we calculated their harmonic mean. To verify whether co-expression modules obtained from our algorithm are associated with biological knowledge, we applied it to Arabidopsis co-expression analysis and compared it to the publicly available algorithms ARACNE and DP-Clus for the similar purpose using the KEGG pathways. Through comparative analysis, we demonstrate that our algorithm detected co-expression modules with higher network indices and better assignment to the pathways than those obtained using the other tools.

2. Method and Results

2.1. Definitions

We define an index (NB) for evaluating connections between network modules and combining with network density (ND) as follows.

$$ND = \frac{\sum e(i)}{n \cdot (n-1)} \quad (1)$$

$$NB = \frac{\sum e(i)}{\sum d(i)} \quad (2)$$

In Eq. (1), n represents the number of vertices included in the module, i ranges according to $1 \leq i \leq n$, $e(i)$ represents the number of vertex-to-vertex links (edges) between the i th vertex in the module and the other module members over a threshold cutoff value for vertex-to-vertex association (TC ; e.g. the Pearson correlation coefficient in the application to Arabidopsis genes in the present research), and $d(i)$ represents the total number of vertex-to-vertex links between the i th vertex and the all possible vertices included in the whole network over a TC , irrespective of membership in the module (degree). Using these equations, the combined index (NC), the harmonic mean of ND and NB can be calculated in the following equation.

$$NC = \frac{2 \cdot \sum e(i)}{n \cdot (n-1) + \sum d(i)} \quad (3)$$

When NC of a network module is the maximal value of 1, the module is depicted as a complete graph with no vertex-to-vertex link to vertices outside of the module. In the present paper, we hypothesize that vertices in a network module with a higher NC value are more closely associated with each other.

We also define the combined index of a single vertex to a network module (VC) as follows.

$$VC(i) = \frac{2 \cdot e(i)}{n \cdot (n-1) + d(i)} \quad (4)$$

In Eq. (4), $e(i)$, $d(i)$, and n are similar to those in Eq. (3).

2.2. Microarray datasets

For a comparative analysis between our algorithm, ARACNE [20], and DP-Clus [21] using the KEGG PATHWAY dataset, including not only metabolic pathways, but protein complexes, we selected the AtGenExpress developmental dataset, composed of 237 DNA microarray data in various developmental stages [25], and 1752 Arabidopsis genes included in the 136 metabolic pathways and protein complexes for Arabidopsis [26]. Using the dataset, Pearson correlation coefficients for all pairs of the genes were calculated on the basis of similarity in their expression profiles. The size of this dataset does not exceed the limitation in ARACNE and DP-Clus implementation.

For a comprehensive co-expression analysis in Arabidopsis genes, we obtained the gene-to-gene correlation dataset from the ATTED-II database [27], comprising 22 263 genes and 1388 DNA microarray data, which are available at [28].

2.3. An algorithm to extract co-expression modules

Our algorithm starts with a single vertex, referred to as the seed vertex (SV), and then detects a network module including the SV in the following steps (Fig. 1).

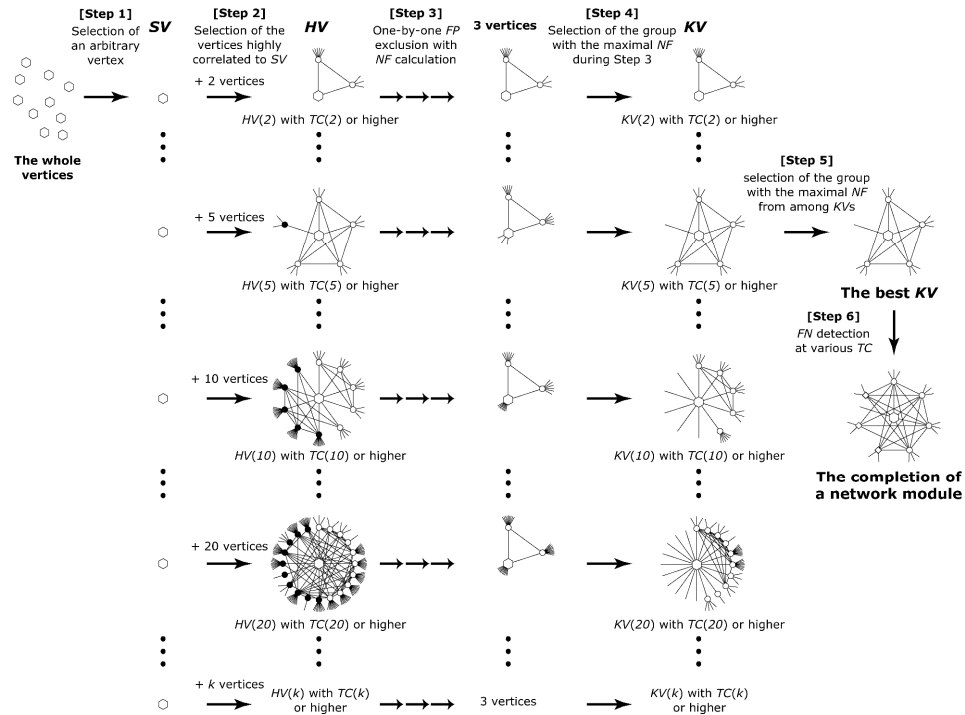


Figure 1: Schematic images of our algorithm. A hexagon and a circle represent a seed vertex (SV) and a vertex except for the SV . A vacant and a filled circle represent a vertex with a high and low VC value, respectively. A vertex-to-vertex link is depicted when a correlation coefficient for the relationship of the vertex pair are equal to or higher than a threshold cutoff value of the coefficient (TC). A link connected to a single vertex represents a connection between a group member and a non-member. The detailed description of this scheme is in the *An algorithm to extract co-expression modules* section.

Step 1. SV selection. An SV can be arbitrarily selected at the initiation of the algorithm. We selected all individual genes as SV s.

Step 2. Highly correlated vertex (HV) selection. For the SV , we set a series of vertex groups (HV s) composed of the SV (a hexagon in Fig. 1) and vertices (circles) with the highest correlation coefficients to the SV . For example, the $HV(i)$ group has $i + 1$ vertices including the SV and the vertices that show the first to i th highest correlation coefficients to the SV . The $HV(i)$ group members are interconnected on the basis of the cutoff value of $TC(i)$, which is the i th highest (i.e. the lowest within this group) correlation coefficient to the SV . In this step, namely, the SV is connected to all vertices.

Step 3. False-positive vertex exclusion and NC calculation. First, the NC value of each HV group is calculated. Next, the VC values of members of the group are calculated.

Then, the vertex with the lowest VC value (one of filled circles in Fig. 1) is excluded as a false-positive vertex from the group to create the new group. The NC calculation and the one-by-one exclusion are repeated until the group comprises three vertices. This step is performed for all HV groups selected in Step 2.

Step 4. Kernel vertex (KV) selection. Of all temporal groups during Step 3 for each HV group, the group with the highest NC value is selected as the KV group.

Step 5. The best KV selection. From among the KV groups, the group with the highest NC value is selected as the best KV group. When $KV(i)$ group, originating from the $HV(i)$ group, is selected as the best KV group, $TC(i)$ is used as the cutoff value of vertex-to-vertex correlation coefficient for the group.

Step 6. False-negative vertex detection. For the best KV group, vertices with high VC values to the group but non-members of the group (i.e. false-negative vertices for the group) are extracted from outside of the group. The VC value of each non-member is calculated at various TC (ranging from 0 to 1). If the highest value of a vertex is user-selected cutoff value (e.g. 0.5) or higher, the vertex is incorporated into the group. Finally, the group composed of the best KV group and the vertices extracted in this step are selected as members of a ‘network module’ originating from the SV .

2.4. Testing

To verify the applicability of our algorithm to associating a network module detected from our algorithm with a biological network, we performed a comparative analysis using Arabidopsis genes to the ARACNE tool [20] and the DP-Clus tool [21], which can be used to detect co-expression modules from gene expression datasets and are appropriate for a fair comparison to our algorithm, although they are designed for other purposes. In the analysis, the indices in their network topology and in their associations to 136 pathways and protein complexes in the KEGG PATHWAY dataset were compared. As a gene expression dataset for this comparative analysis, we selected the AtGenExpress developmental dataset (see Microarray datasets), because ARACNE and DP-Clus can accept the sizes of the dataset, while our algorithm is feasible, irrespective of data size. The ARACNE and DP-Clus tools require setting threshold cutoff values for the gene-to-gene association index. For fair comparison between our algorithm, ARACNE, and DP-Clus in extracting co-expression modules, we used several cutoff values for their execution; i.e. e-20, e-30, e-40, e-50 for ARACNE and a range of gene-to-gene correlation coefficient of 0.45 to 0.90 at intervals of 0.05 for DP-Clus. To perform the comparative analysis, we selected their adequate cutoff values with which the average size of the co-expression modules from each approach is similar to that of the KEGG PATHWAY dataset, resulting in values of e-40 for ARACNE and 0.75 for DP-Clus. For the co-expression modules obtained from the three network approaches, we

calculated the precision values of the modules to individual pathways in the following equation and compared those between approaches.

$$(\text{precision}) = (\text{module members assigned to a pathway}) / (\text{all of the module members}) \quad (5)$$

The average precision values in a total of 1752 modules are 0.31 for our algorithm, 0.19 for ARACNE, and 0.18 for DP-Clus. Figure 2 shows the relationship of the precision values to the ratio of genes included in co-expression modules with the precision value or higher to 1752 genes. In Fig. 2, the ratios in any precision values are highest in using our algorithm, especially the number of genes in co-expression modules with 0.9 or higher precision values (228 genes) are five-fold (45 genes) and ten-fold (22 genes) as much as ARACNE and DP-Clus.

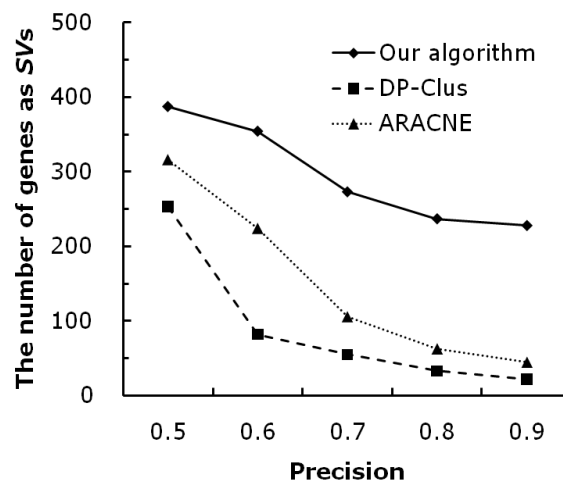


Figure 2: The number of genes included in co-expression modules well-assigned to the KEGG pathways. The precision value represents the ratio of genes included in a specific KEGG pathway among genes included in a co-expression module. For the precision value of 0.9, our algorithm detects 228 genes; five-fold and ten-fold as much as those in ARACNE (45 genes) and DP-Clus (22 genes), respectively.

We performed comparative ROC curve analysis, a representative approach to evaluate the ability to detect, in the assignment of co-expression modules detected by our algorithm and the other tools to a specific KEGG pathway. We selected the largest ‘Ribosome’ pathway, to which 190 genes are assigned, for the analysis. The precision values of all co-expression modules in the three approaches to the pathway were calculated. On the basis of the assignment of SVs included in the modules to the pathway, the false positive rates (*FPRs*) and the true positive rates (*TPRs*) were calculated to depict ROC curves of the approaches (Fig. 3). In the ROC curves, our algorithm shows the better result than those of ARACNE and DP-Clus, especially in the region of low *FPR*. These findings indicate that co-expression modules obtained using our algorithm are assigned to the KEGG pathways better than the other tools. The information of co-expression modules with high precision values (≥ 0.5) is listed in Table 1.

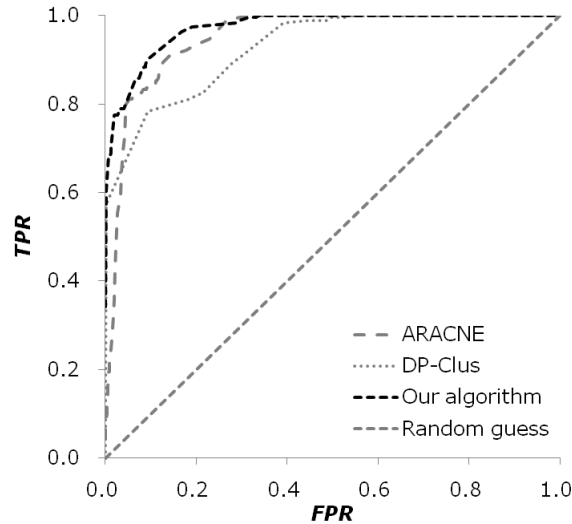


Figure 3: ROC curves of the three approaches in the assignment of co-expression modules to ‘ribosome’ pathway. The *FPR* and *TPR* indices represent false positive rate and true positive rate, respectively. The *FPR* index is calculated by the ratio of genes incorrectly assigned to the pathway by an approach among genes that are not included in the pathway. The *TPR* index is calculated by the ratio of genes correctly assigned to the pathway by an approach among genes included in the pathway. The ROC curve of our algorithm shows the better result than those of the other two algorithms.

2.5. Implementation

To verify the applicability of our algorithm to a large dataset, we applied it to a large expression dataset of Arabidopsis genes (see Microarray datasets). In this application, many co-expression modules with high *NC* values were detected. Of 22 263 genes, 8446 genes (approximately 38 %) are included in co-expression modules with 0.5 or higher *NC* values. The *TC* values of co-expression modules, in which the *NC* values are optimized, vary in the both organisms as shown in Fig. 4; i.e. 0.59 ± 0.14 (mean \pm standard deviation). It indicates that the *TC* values should be selected in individual co-expression modules in terms of tight intra-modular connections. The *TC* peak in Fig. 4 (i.e. 0.59) are similar to a single *TC* value in the network that was selected in previous reports. Saito *et al.* [13] used 0.6 as a single threshold cutoff for the coefficient. Aoki *et al.* [12] mentioned that, for *TC* ranging from 0.55 to 0.66, the density of the entire network displays a minimal value, i.e. that the number of false-positive gene-to-gene links is minimized within the range. To minimize such false-positive links, Gupta *et al.* [19] optimized the average clustering coefficient value in a well-controlled dataset and determined a single *TC* value of 0.9 (the square of Pearson correlation coefficient). The broadened distribution of *TC* values in optimized co-expression modules (Fig. 4) suggests that *TC* values should be selected for each module when using the vast number of microarray datasets, including various experimental designs.

Table 1: Co-expression modules with high precision values to the KEGG pathways.

Genes in a co-expression module (A)	Genes in a specific KEGG pathway (B)	Precision (B / A)	KEGG pathway name	Pathway ID	One of SVs
115	112	0.97	Ribosome	ath03010	At2g17360
22	21	0.95	Proteasome	ath03050	At3g22630
25	19	0.76	Oxidative phosphorylation	ath00190	At1g01050
16	11	0.69	DNA replication	ath03030	At4g02060
15	8	0.53	Phenylpropanoid biosynthesis	ath00940	At5g42590
10	7	0.70	Fatty acid biosynthesis	ath00061	At1g62640
10	7	0.70	Photosynthesis	ath00195	At4g12800
5	4	0.80	Alpha-linolenic acid metabolism	ath00592	At1g17420
7	4	0.57	Starch and sucrose metabolism	ath00500	At2g36390
3	3	1.00	Valine, leucine and isoleucine biosynthesis	ath00290	At1g31180
3	3	1.00	Flavonoid biosynthesis	ath00941	At3g51240
3	3	1.00	Phosphatidylinositol signalling system	ath04070	At2g41210
3	3	1.00	SNARE interactions in vesicular transport	ath04130	At5g58060
4	3	0.75	Valine, leucine and isoleucine degradation	ath00280	At1g50110
4	3	0.75	Phenylalanine, tyrosine and tryptophan biosynthesis	ath00400	At1g48860
4	3	0.75	Methane metabolism	ath00680	At1g22440
4	3	0.75	Mismatch repair	ath03430	At4g02460
5	3	0.60	Glycolysis / Gluconeogenesis	ath00010	At3g52930
5	3	0.60	Carbon fixation in photosynthetic organisms	ath00710	At1g63290

Co-expression modules that show precision values of 0.5 or higher and comprise 3 or more genes are listed.

3. Discussion

Our algorithm assembled network modules that are well-assigned to biological knowledge such as metabolic pathways. Especially, co-expression modules involved in ‘ribosome’, ‘oxidative phosphorylation’, ‘proteasome’, and ‘phenylpropanoid biosynthesis’ showed high (> 0.9) precision values of the module members to the pathways or complexes (Table 1), suggesting that such pathways and functional groups are well co-expressed. Li [17] and Wei *et al.* [10] reported associations between metabolic pathways and co-expression. However, many co-expression modules show low precision values to the pathways because of the following possible causations. First, all of metabolic pathways are not necessarily co-expressed, possibly due to regulatory mechanisms other than transcriptional regulation such as post-transcriptional regulation, as mentioned by Saito *et al.* [13]. Second, the paucity of information on the assignment of enzyme genes to metabolic pathways may cause a low rate of assignment of enzyme

genes included in co-expression modules. Third, co-expression relationships are not necessarily equivalent to direct regulatory relationships. Co-expression network analysis is just pre-screening for reconstructing regulatory networks. Furthermore information on the assignment to biological pathways may lead to a reduction of co-expression modules with low assignment to the pathways.

Co-expression modules with optimized NC values show broadened distribution of TC values throughout all modules. Such distribution may arise from biases included in combined datasets, e.g. a dataset including microarray data of different experimental designs for different purposes. In the analyzed expression dataset of Arabidopsis genes, which is composed of 1388 microarray data, includes 275 assay data (approximately 20 %) from leaves and 9 assay data (< 1 %) from pollen, indicating that expression similarity in leaves between genes may have a strong influence on their gene-to-gene correlation coefficients and vice versa in pollen. Under such influence, a single threshold cutoff of the coefficient may be adequate for limited co-expression modules. Despite of the heterogeneity of a gene expression dataset, Obayashi *et al.* [27] reported that the applicability of such dataset in co-expression analysis.

Our algorithm can be applied to any sizes of datasets comprising vertex-to-vertex correlation data. The algorithm starts with the association of a single vertex, such as hierarchical clustering, and shows constant and processible computational loads through the implementation. In contrast, many publicly available tools for clustering approach such as ARACNE [20] and DP-Clus [21] start with the matrix operation of the whole dataset and thus depend upon the size of a dataset for their steady implementation.

Our algorithm is an approximate approach for optimizing an NC value for a network module. In its application to Arabidopsis gene expression datasets, it occasionally causes different memberships in co-expression modules involved in the common biological event, indicating the insufficient maximization of NC values. The real maximization for a gene cluster including an SV , from among 22 263 genes, requires verification of all possible combinations of genes including the SV . The number of such combination, however, reaches astronomical levels in the following equation.

$$\sum_{i=3}^N {}_N C_i \quad (6)$$

In Eq. (6), N represents the number of genome-wide genes. Using our algorithm, the number of such combinations is reduced into N^2 -order as follows.

$$\frac{(N-2) \cdot (N-1)}{2} \quad (7)$$

Although our algorithm may provide some modules with insufficient maximization of NC , it provided the NC values (0.76 on average) better than those in ARACNE (0.33) and DP-Clus (0.48).

Additionally, a gene may be included in multiple co-expression modules through our algorithm, indicating the possibility that a gene is associated with multiple biological processes.

4. Conclusions

Our algorithm can be applied to vertex-to-vertex correlation datasets, based on which a network including modular structure is constructed, such as gene expression datasets, irrespective the size of the dataset.

Acknowledgments

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO) program, which is part of the “Development of Fundamental Technologies for Controlling the Material Production Process of Plants” project.

References

- [1] Clauset, A., Moore, C., Newman, M.E.J., Hierarchical structure and the prediction of missing links in networks, *Nature*, 453:98-101, 2008.
- [2] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D., A combined algorithm for genome-wide prediction of protein function, *Nature*, 402:83-86, 1999.
- [3] Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carroll, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J., Rothberg, J.M., A protein interaction map of *Drosophila melanogaster*, *Science*, 302:1727-1736, 2003.
- [4] Parrish, J.R., Yu, J., Liu, G., Hines, J.A., Chan, J.E., Mangiola, B.A., Zhang, H., Pacifico, S., Fotouhi, F., DiRita, V.J., Ideker, T., Andrews, P., Finley, R.L., A proteome-wide protein interaction map for *Campylobacter jejuni*, *Genome Biol.*, 8:R130, 2007.
- [5] Ben-Dor, A., Yakhini, Z., Clustering gene expression patterns, *J. Comput. Biol.*, 6:281-297, 1999.
- [6] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298:799-804, 2002.
- [7] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., Network motifs: simple building blocks of complex networks, *Science*, 298:824-827, 2002.
- [8] Katagiri, F., Glazerbrook, J., Local context finder (LCF) reveals multidimensional relationships among mRNA expression profiles of *Arabidopsis* responding to pathogen infection, *Proc. Natl. Acad. Sci. USA*, 100:10842-10847, 2003.

- [9] Stuart, J.M., Segal, E., Koller, D., Kim, S.K., A gene-coexpression network for global discovery of conserved genetic modules, *Science*, 302:249-255, 2003.
- [10] Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G.P., Somerville, C., Loraine, A., Transcriptional coordination of the metabolic network in Arabidopsis, *Plant Physiol.*, 142:762-774, 2006.
- [11] Ma, S., Gong, Q., Bohnert, H.J., An Arabidopsis gene network based on the graphical Gaussian model, *Genome Res.*, 17:1614-1625, 2007.
- [12] Aoki, K., Ogata, Y., Shibata, D., Approaches for extracting practical information from gene co-expression networks in plant biology, *Plant Cell Physiol.*, 48:381-390, 2007.
- [13] Saito, K., Hirai, M.Y., Yonekura-Sakakibara, K., Decoding genes with coexpression networks and metabolomics – ‘majority report by precogs’, *Trends Plant Sci.*, 13:36-43, 2008.
- [14] Ogata, Y., Shibata, D., Practical network approaches and biologic interpretations of co-expression analyses in plants, *Plant Biotechnol.*, 26:3-7, 2009
- [15] Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W., From molecular to modular cell biology, *Nature*, 402:c47-c52, 1999.
- [16] Rives, A.W., Galitski, T., Modular organization of cellular networks, *Proc. Natl. Acad. Sci. USA*, 100:1128-1133, 2003.
- [17] Li, K., Genome-wide coexpression dynamics: Theory and application, *Proc. Natl. Acad. Sci. USA*, 99:16875-16880, 2002.
- [18] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863-14868, 1998.
- [19] Gupta, A., Maranas, C.D., Albert, R., Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites, *Bioinformatics*, 22:209-214, 2006.
- [20] Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., Califano, A., ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics*, 7:S7, 2006.
- [21] Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S., Development and implementation of an algorithm for detection of protein complexes in large interaction networks, *BMC Bioinformatics*, 7:207, 2006.
- [22] Bagrow, J.P., Bollt, E.M., A local method for detecting communities, *Phys. Rev. E*, 72:046108, 2005.
- [23] Girvan, M., Newman, M.E.J., Community structure in social and biological networks, *Proc. Natl. Acad. Sci. USA*, 99:7821-7826, 2002.
- [24] Bock, R.D., Husain, S.Z., An adaptation of Holzinger’s B-coefnlncients for the analysis of sociometric data, *Sociometry*, 13:146-153, 1950.
- [25] Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., Lohmann, J.U., A gene expression map of Arabidopsis thaliana development, *Nat. Genet.*, 37:501-506, 2005.
- [26] Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., Kanehisa, M., KEGG Atlas mapping for global analysis of metabolic pathways, *Nucl. Acids Res.*, 36:W423-W426, 2008.
- [27] Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., Kinoshita, K., ATTED-II provides coexpressed gene networks for Arabidopsis, *Nucl. Acids Res.*, 37:D987-D991, 2009.
- [28] <http://www.weigelworld.org/resources/microarray/AtGenExpress/>