W504–W509 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl204

Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis

Iain W. Manfield*, Chih-Hung Jen[1], John W. Pinney[1], Ioannis Michalopoulos[1], James R. Bradford[1], Philip M. Gilmartin and David R. Westhead[1]

Centre for Plant Sciences and [1]Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, West Yorkshire, LS2 9JT, UK

Received February 14, 2006; Revised and Accepted March 22, 2006

ABSTRACT

The *Arabidopsis* Co-expression Tool, ACT, ranks the genes across a large microarray dataset according to how closely their expression follows the expression of a query gene. A database stores pre-calculated co-expression results for ~21 800 genes based on data from over 300 arrays. These results can be corroborated by calculation of co-expression results for user-defined sub-sets of arrays or experiments from the NASC/GARNet array dataset. Clique Finder (CF) identifies groups of genes which are consistently co-expressed with each other across a user-defined co-expression list. The parameters can be altered easily to adjust cluster size and the output examined for optimal inclusion of genes with known biological roles. Alternatively, a Scatter Plot tool displays the correlation coefficients for all genes against two user-selected queries on a scatter plot which can be useful for visual identification of clusters of genes with similar *r*-values. User-input groups of genes can be highlighted on the scatter plots. Inclusion of genes with known biology in sets of genes identified using CF and Scatter Plot tools allows inferences to be made about the roles of the other genes in the set and both tools can therefore be used to generate short lists of genes for further characterization. ACT is freely available at www.Arabidopsis.leeds.ac.uk/ACT.

INTRODUCTION

Microarray data contain information on the relative expression levels in a tissue sample for the thousands of genes represented by probes on the array. Large collections of microarray data therefore contain information about concerted changes in transcript levels in these datasets beyond the original purpose of each experiment. The NASC/GARNet array data are one such data collection, containing results from many experiments analysing the responses in *Arabidopsis* to differing biotic and abiotic conditions and analysing mutants and a range of developmental stages (1).

A number of bioinformatics resources allow information to be recovered for individual genes from this and other microarray databases [e.g. The *Arabidopsis* Information Resource (TAIR) (2), NASCArrays tools (1), Stanford Microarray Database (3), Botany Array Resource (4) and Genevestigator (5)]. However, as the first microarray data became available, it was realized that this represented a mine of information for how genes were regulated and acted together (6) allowing predictions to be made about the co-regulation of genes from the correlation of their expression patterns. Indeed, in plant science, gene co-expression analysis has been used recently to predict biology and to inform experimental approaches, e.g. (7–9) and web-based tools reporting co-expression results based on *Arabidopsis* microarray data have become available [Botany Array Resource (4), Gene Recommender (10), CSB.DB (11) and *Arabidopsis* Co-expression Tool, ACT, (12)] making such tools available for all biologists. A range of different features are offered by these websites each with their own advantages.

ACT provides co-expression analysis for 21 891 genes, based on Affymetrix *Arabidopsis* Ath1 microarray data from the NASC/GARNet dataset. Our Clique Finder (CF) tool provides objective dissection of co-expression lists for genes consistently co-expressed with each other. The Scatter Plot tool allows visualization of the correlation values for all genes against two queries, with the facility to highlight sets of genes of interest, e.g. the members of a gene family. Identifying and visualizing marker genes with known biology (or 'guide genes'), (11) in co-expression datasets is a valuable

*To whom correspondence should be addressed. Tel: +44 113 343 2901; Fax: +44 113 343 3144; Email: i.manfield@leeds.ac.uk
Present addresses: Chih-Hung Jen, Genome Research Centre (VYMGC), National Yang-Ming University, 155 Li-Nong Street, Taipei, Taiwan
John W. Pinney, Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, M13 9PT, UK

approach to determining cut-off values giving sets of genes for further analysis. Here we illustrate the features of ACT using two transcription factors forming part of the circadian clock of *Arabidopsis*.

ACT, CLIQUE FINDER AND SCATTER PLOT SOFTWARE

Data handling and processing

ACT uses microarray data from the GARNet/NASCArrays (1) set, processed by the Affymetrix MAS5.0 analysis algorithms. Correlation values were based on the signal values output by this software. Probe sets showing no detection of expression in any experiment were deleted, and values of expression signals below a cut-off of 20 signal units in particular experiments were set to 20 to eliminate any chance correlations with these noisy low signal values.

Correlation calculations

The WWW server is backed by a database containing all experimental data and annotations and GO terms. The database also contains pre-calculated correlation values over all experiments, allowing fast processing of these user queries. When the user defines a subset of the arrays, correlation calculations are carried out 'on the fly', since pre-storage of all possibilities is impractical and, in consequence, these user queries run more slowly.

The starting point for most users will be the Keyword Search tool which reports a list of genes likely to be of interest, with links to the pre-calculated co-expression data for each probe set. Alternatively, a tool is provided for conversion of AGI codes to Affymetrix probe IDs or, if known, the Correlation List can be recalled by entering a probe set ID of interest. A typical ACT output is shown in Figure 1. This tool returns a list of the array probe sets ranked by *r*-value for the correlation of their expression patterns. *P*-values and *E*-values are given as measures of the statistical significance of the observed correlation. By default, the 50 top-ranked probe sets are reported, but results for all 21 891 valid probe sets are available if requested. The full list is useful for examination of the anti-correlated genes. The AGI code represents a hyperlink to TAIR (www.arabidopsis.org) for more information and access to other external databases. Clicking on any Affymetrix probe ID in a list from the pre-calculated database returns the 50 best-correlated genes for this new query, allowing biologists to browse lists giving a subjective feeling for genes which may be consistently highly ranked on co-expression lists. For co-expression calculation using user-defined arrays, the output lists the arrays used for the calculation and then gives the full co-expression list from query gene itself through positively correlated genes to the most anti-correlated genes. A link for each experiment in the database opens a new window at NASC giving the information about each experiment. ACT lists can be saved as text files and opened in a spread-sheet program (in a tab-delimited format).

Themes to sets of co-expressed genes can be determined using our Word and GO counting tools to detect over-represented words or GO terms for the top-ranked probe sets. The significance of any over-representation is estimated by hypergeometric distribution analysis. These tools can be more informative than visual inspection of the annotations of the best-correlated genes as they provide a statistical basis to distinguish between common terms and genuinely over-represented ones.

Clique finder tool

The CF tool (illustrated in Figure 2) constructs clusters of genes with very similar expression patterns within the list of the top k genes correlated with a given query probe set. The algorithm allows overlap between clusters, so in contrast to traditional clustering methods for microarray data, each gene can potentially be shown to be involved in more than one type of biological response. The method used is based on the graph theoretical concept of a maximal clique.

Given a query probe set ID and a number of neighbours, k, as input, we first retrieve the top k probe sets from the database, ranked by Pearson correlation coefficient with respect to the query. Owing to the computational complexity of the clique finding algorithm, we currently support up to a maximum of $k = 100$ neighbours. A second database query then obtains all the correlation coefficients between all possible pairs of these genes. The genes are represented as vertices in a graph representation, and the links between each pair of genes are considered as weighted edges, where the edge weight is equal to the Pearson correlation coefficient between those two genes. We keep only the strongest $c\%$ of these edges according to a cut-off value set by the user, typically between 1 and 10%. This removes all anti-correlation edges and retains only those positive correlation edges with the strongest support. The graph representation is now an unweighted simple graph which is relatively sparse.

A standard algorithm (13) is now used to find all maximal cliques within the graph. A clique is a subset of vertices that are all connected to each other by edges, and in a maximal clique there are no more vertices that can be added to the clique such that this condition holds. A clique can reveal interesting biology because all its members are strongly correlated with each other. However, there is often significant overlap between cliques, and in this case it makes sense to combine them into clusters. Any clique sharing at least 50% of its genes with an overlapping clique is considered to be a 'neighbour' of that clique. A simple single-linkage clustering procedure joins all neighbouring cliques into clusters of probe sets. These clusters and the unclustered singletons are then output for inspection. Clicking on any probe set ID in the output in turn produces the CF result for that gene.

Scatter Plot tool

Another tool allows users to visualize the correlation of all genes against two probe sets simultaneously. Every probe set in the dataset is plotted on a scatter graph, where the two axes are the Pearson correlation coefficients against two different query probe sets (Figure 3). With two query probe sets involved in the same biological process, this tool gives the user an intuitive feel for the degree of correlation, and also makes it easy to identify groups of probe sets that are strongly correlated or anti-correlated with the query probe sets. Using an HTML image map, each probe set on the scatter plot has a link to its corresponding annotation information at TAIR.

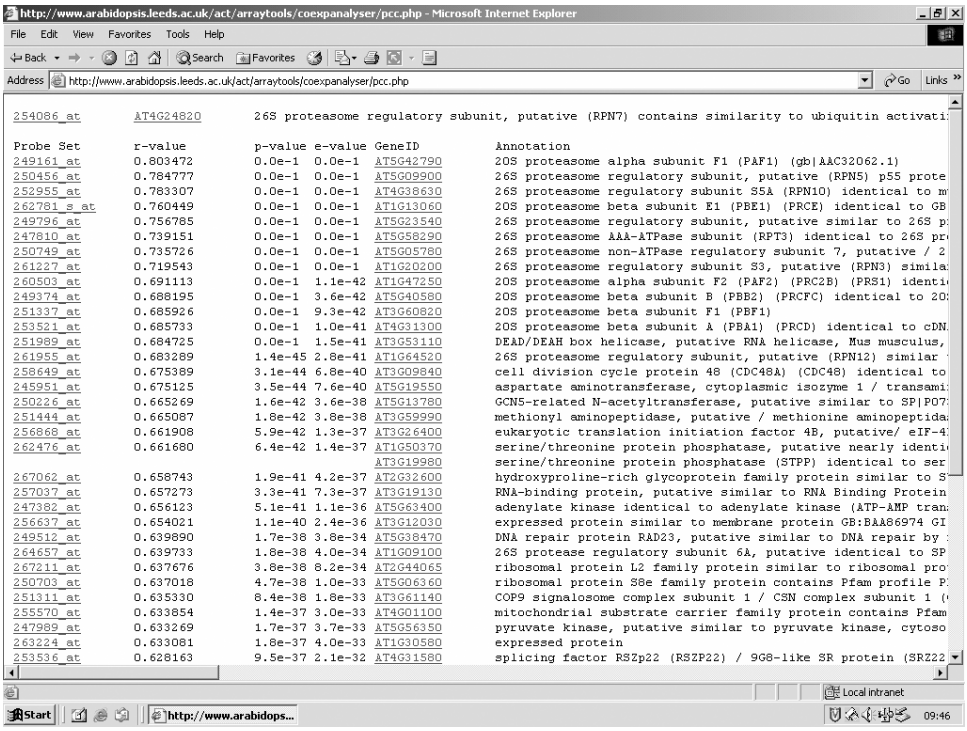

Figure 1. Screen shot of a typical ACT output showing co-expression of genes encoding sub-units of the proteasome. Gene identifiers, correlation *r*-value, measures of statistical significance and annotation are shown.

Implementation

The microarray data were stored in a MySQL database. The correlation calculations were implemented in C and the WWW interface (including Correlation List and Scatter Plot tools) was implemented using the Apache WWW server and Perl/PHP. The Clique Finder algorithm was implemented in Java.

USING ACT, CLIQUE FINDER AND SCATTER PLOT TOOLS

Co-expression output

The circadian clock in plants regulates many aspects of plant growth and development including changes in gene expression that are central to many core functions. In *Arabidopsis*, some of the genes which constitute the clock have been identified but many of the signalling inputs and outputs are still to be characterized. Two components of the 'central oscillator' are myb transcription factor genes, *cca1* and *lhy*. The pre-calculated co-expression list for *lhy* is shown in Table 1 (for space reasons, most of the information reported on the web pages has been removed) revealing co-expression of *lhy* and *cca1* with each other, with another myb gene (At3g09600) and with a CONSTANS-like transcription factor. One approach to corroborating such correlations is to calculate the co-expression values using a different set of data and comparing the two lists. The result of such a calculation, from a small number of arrays (42 arrays from three experiments selected based on biological knowledge of *lhy* and *cca1* expression patterns) is also given in Table 1. The high r-values for genes at the top of each list indicate strong co-expression of these genes with the query. The different *r*-values reflect the use of different datasets for the calculations; the datasets for the user-defined calculation were derived from experiments using similar tissues thus producing higher *r*-values compared with the pre-calculated database which is based on a wide range of tissues. Genes common to these lists of the top-ranked 15 genes are indicated in bold type with genes of one list also present in the top 100 (i.e. top ½%) of the other list indicated in italics. Clearly there are many genes common to both lists, supporting the suggestion that these are a set of genes which are co-expressed and therefore whose expression is indeed likely to be regulated in a similar manner. This represents a valuable prediction, especially for the unannotated genes in these lists.

Clique finder

However, visual examination of two lists is very slow and there is subjectivity as to how far down two lists a user would be prepared to look for genes in common. Beyond the visual examination of two co-expression lists for genes in common, the CF tool uses a more complex algorithm for the prediction of biological relevance, searching a co-expression list (corresponding to a single query gene) for other genes that are consistently co-expressed with each other (Figure 2). The CF output for the *cca1* myb gene is presented in Table 2 (copied from the Web page and edited slightly). At the top of the page are the identifiers and annotations for the query gene and below this are the parameters used in the CF search. The 'more edges' and 'fewer edges' buttons on the Web page

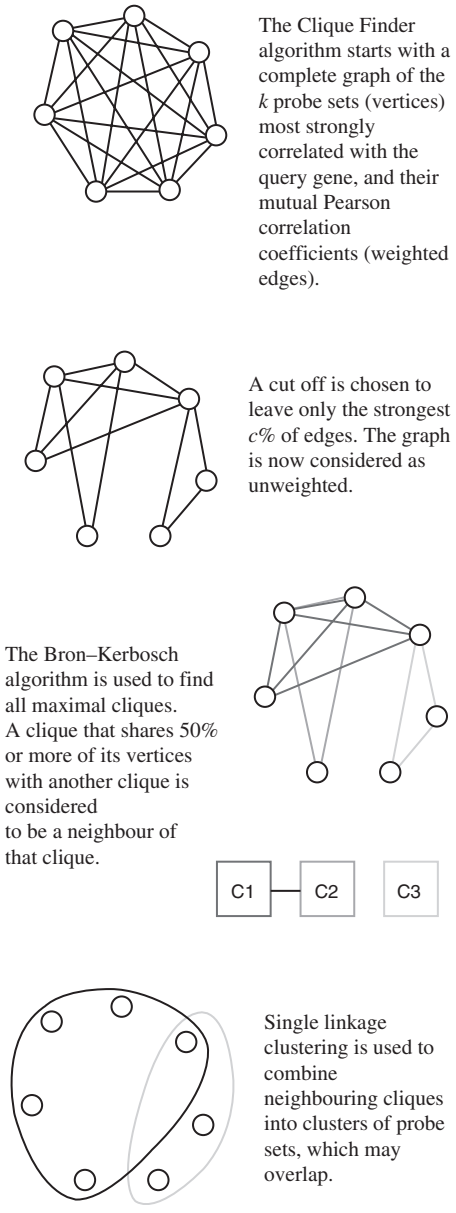

Figure 2. The Clique Finder algorithm for identification of groups of consistently co-expressed genes.

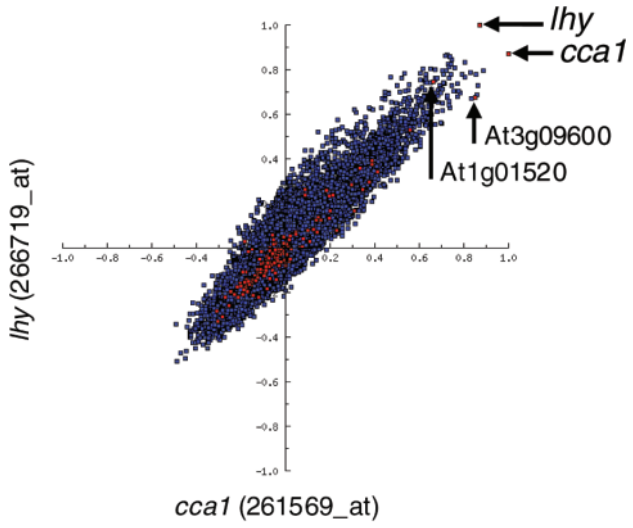

Figure 3. Co-correlation Scatter Plot. The *r*-values for all genes against two queries, *cca1* (266719_at) and *lhy* (261569_at), are displayed on a scatter plot. The values for all of the myb genes represented on the array are highlighted in red. The labelled genes are discussed in the text.

allow the biologist to explore how these parameters affect cluster size and representation of genes with known biology in each cluster.

Cluster 2 contains the three myb genes seen in the co-expression lists, namely *cca1*, *lhy* and the uncharacterized myb gene At3g09600. Another transcription factor (CONSTANS-LIKE 2) and some unannotated genes are present in cluster 2 supporting the suggestion that they are indeed regulated in a similar manner. Overall, CF cluster 2 is very similar to the genes common to the two co-expression lists presented in Table 1, suggesting that, of the 100 top-ranked genes, ~9 of them are indeed predicted to be co-regulated. The remaining genes of this list of 100 may be regulated in a different manner or have other signalling inputs thus changing their behaviour. Support for the validity of Cluster 2 as a co-regulated set of genes comes from published work analysing effects of red light on gene expression in *Arabidopsis* (14); expression of the genes highlighted in bold type in Table 2 responds to illumination with red light. This observed enrichment suggests that the other genes of the set may also be red light-responsive, matching the behaviour of the 'guide genes', and further suggests that ACT and CF can be used to suggest roles for poorly-characterized genes.

The choice of parameters for the Clique Finder algorithm will determine how many genes are included in each cluster or are unclustered and therefore how many genes are included in short lists for further analysis. More aggressive parameters, giving smaller cluster sizes, would be appropriate for low-throughput follow-up analyses (such as characterisation of mutant plants), whereas less stringent criteria will give larger clusters more appropriate for high-throughput analyses such as printing of custom microarrays or bioinformatic analyses. While larger clusters may include more 'false positives' (genes incorrectly included in the cluster), they might also include more 'true positives' and this would offer the opportunity to identify a biological role for a larger set of genes if an appropriate screen is available. Conversely, characterization of small clusters, perhaps excluding well-characterized 'guide genes' giving a 'false negative' result, represents a lost opportunity to identify functions for the uncharacterized genes also incorrectly excluded from the cluster.

Co-correlation scatter plots

Co-expression lists are not necessarily a good format for looking at many more than a few top-ranked genes. Therefore, we developed a Scatter Plot tool for visualization of correlation results for all 21 891 genes in our database with two query genes. The output from this tool can reveal groups of genes better correlated with one query than the other, or well separated from the bulk of the genes, which may empirically

Table 1. Comparison of ACT output from the pre-calculated database with co-expression results based on a user-selected set of arrays for myb transcription factor gene, *lhy*

Co-expression result from a pre-calculated database			Co-expression result from a user-selected set of arrays		
r-Value	GeneID	Annotation	*r*-Value	GeneID	Annotation
(1.0)	**At1g01060**	**LHY myb transcription factor**	**(1.0)**	**At1g01060**	**LHY myb transcription factor**
0.88	*At1g64500*	*glutaredoxin family protein*	0.97	**At2g46830**	**CCA1 myb transcription factor**
0.87	**At2g46830**	**CCA1 myb transcription factor**	0.95	**At3G09600**	**myb transcription factor**
0.86	**At3g02380**	**zinc finger protein (COL2)**	0.94	At3g47420	glycerol-3-phosphate transporter,
0.85	**At3g09600**	**myb transcription factor**	0.93	**At3g09600**	**myb transcription factor**
0.85	**At3g54500**	**expressed protein**	0.92	At2g47490	mitochondrial substrate carrier
0.84	**At3g09600**	**myb transcription factor**	0.90	*At5g14760*	*L- aspartate oxidase*
0.83	*At1g55960*	*expressed protein*	0.89	*At5g15850*	*zinc finger protein (COL1)*
0.83	*At2g15020*	*expressed protein*	0.89	*At1g62180*	*5'-adenylylsulfate reductase*
0.80	*At1g65870*	*disease resistance protein*	0.89	**At3g02380**	**zinc finger protein (COL2)**
0.80	*At5g64940*	*ABC1 family protein*	0.89	**At3g54500**	**expressed protein**
0.79	*At2g24540*	*kelch repeat F-box protein*	0.88	**At2g19650**	**DC1 domain protein**
0.78	*At4g24700*	*expressed protein*	0.88	At5g18670	beta-amylase
0.77	**At2g19650**	**DC1 domain protein**	0.87	At1g14280	phytochrome kinase,
0.77	*At2g26080*	*glycine dehydrogenase*	0.87	*At5g22390*	*expressed protein*

The query gene is perfectly correlated with itself and therefore this *r*-value is given in brackets. Only the top-ranked 15 probes (one gene is represented by two probes) are presented from each list. Genes present on both lists are highlighted in boldface and genes ranked in the top 100 of the other list are indicated in italics.

Table 2. CF output for a myb transcription factor showing only one of the three clusters produced

Query probe: 266719_at AT2G46830 **myb transcription factor (CCA1)**
Neighbour list size: 100; edge limit: 4.0%; number of clusters found: 3
MORE edges (5.0%) => larger clusters
FEWER edges (3.0%) => smaller clusters
Cluster 2 (9 probes)
Mean *r* to query = 0.756271
Mean *r* within cluster = 0.791100

261569_at	0.870453	AT1G01060	**myb transcription factor LHY**
261958_at	0.796923	AT1G64500	**glutaredoxin family protein**
265892_at	0.787191	AT2G15020	expressed protein
263796_at	0.778055	AT2G24540	**kelch repeat-containing F-box protein**
251869_at	0.777933	AT3G54500	expressed protein
258497_at	0.728698	AT3G02380	**zinc finger protein CO-LIKE 2**
265939_at	0.704785	AT2G19650	DC1 domain-containing protein
258724_at	0.687695	AT3G09600	**myb family transcription factor**
258723_at	0.674703	AT3G09600	**myb family transcription factor**

This output is edited from the format produced by the website. The expression of genes highlighted in boldface has been shown to be red-light responsive (see text for details).

suggest r-value cut-offs. Additionally, this tool can be used to highlight a group of genes, e.g. all members of a gene family. In *Arabidopsis* there has been expansion of gene families producing, e.g. 190 myb genes (15) with probes for 177 of these genes on the Ath1 Affymetrix array. There are three myb genes co-expressed in a cluster identified by CF, but other myb genes were also highly ranked in the co-expression list.

Lhy and *cca1* show similar expression patterns and are therefore ideal query genes for the Scatter Plot tool giving a positive correlation against the other genes in the database (Figure 3). The two query genes have correlation values of 1.0 with themselves, are strongly correlated with each other and are therefore located at the top right of the figure. This visual presentation allows an empirical identification of a set of genes more strongly expressed with the query genes than the bulk of the genes.

Highlighting all myb genes on the Ath1 array on the Scatter Plot (Figure 3) reveals that expression of most myb genes is poorly correlated with *cca1* and *lhy*, but At3g09600 and an additional gene (At1g01520) are strongly co-expressed with the two query genes and may merit further analysis. Indeed, both At3g09600 and At1g01520 have been suggested as genes which may play roles in the circadian clock in addition to *cca1* and *lhy* (16). The Scatter Plot visual analysis indicates that expression of other myb genes, with similar sequences to *lhy* and *cca1*, is not correlated with the genes analysed here and therefore they are likely to play different roles.

DISCUSSION

There are many possible statistical approaches to measure correlation, including the Pearson correlation coefficient, the Spearman rank and others (17). Each has theoretical advantages and disadvantages, but it is as yet uncertain which gives the best results on microarray data. ACT uses the simplest of these measures, the Pearson correlation (*r*). We have found this to be effective (12) and that similar results are produced by other approaches. It has the advantage that calculation of the statistical significance of the observed correlation (*P*-value) is straightforward. It is clear that no single correlation value (*r*) or *P*-value cut-off could be used routinely for selecting a set of genes showing strong

co-expression as these values will be affected by factors including the datasets used and the biological processes involved. Rather, interpretation of co-expression patterns is facilitated by biological knowledge of the relevant system and therefore our tools encourage an exploration of the data and allow visualization of results to help users identify short lists of genes for further analysis.

Many users of ACT will choose to analyse output lists for single genes, examining the annotations for over-represented themes and keywords of interest. Our Word and GO counting tools provide a statistical basis for interpreting such themes. Sets of co-expressed genes will be useful inputs into tools such as Genevestigator (5) providing additional types of information. In addition, sets of co-expressed genes may be mapped onto databases of Gene Ontology and metabolic pathway information (18,19) to help identify the biological processes in operation. Similarly, analysis of the promoters of a set of co-expressed genes for over-represented motifs [e.g. (20)] may give confidence in transcription factor-binding site prediction which would not be possible by comparison of a single promoter against a database of known motifs.

The results for an individual microarray experiment are likely to be the sum of a number of (potentially interacting) processes. From amongst a set of genes identified by microarray analysis with significant fold changes in their expression levels, ACT and CF may be useful to identify the different sets of genes which are co-expressed with each other but where each set of genes is responding to a different stimulus. Indeed, the inclusion of genes with small expression level fold changes in such sets may be supported by ACT if they are strongly co-expressed with other genes with larger expression level fold changes.

Modelling gene networks will involve 'the collection, description and systematization of network elements' (21) requiring information with a high level of coverage of the possible elements of a system. ACT provides co-expression results for more probe sets than other similar tools, including genes likely to be expressed at a low level and in a small proportion of the experiments. Comparison of our co-expression predictions for a group of myb transcription factors with independent results from the literature supports our approaches. ACT therefore provides tools to allow inclusion of many genes in co-regulated sets (or exclusion from those sets) allowing predictions to be made about signalling networks which can then be tested experimentally.

ACKNOWLEDGEMENTS

We are grateful to our sponsors for financial support. I.W.M. and I.M. acknowledge support from the Biotechnology and Biological Sciences Research Council (UK) (Grant number: 24/G18363). C.-H.J. acknowledges support from UK Overseas Research Scholarships (ORS) and the University of Leeds. We are very grateful for the public availability of the NASC/GARNet microarray database and acknowledge all those who contributed to this publicly accessible data. J.W.P. acknowledges support from the Medical Research Council (UK). Funding to pay the Open Access publication charges for this article was provided by the BBSRC.

Conflict of interest statement. None declared.

REFERENCES

1. Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res*., **32**, D575–D577.
2. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G.H., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al*. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.
3. Ball,C.A., Awad,I.A.B., Demeter,J., Gollub,J., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Matese,J.C., Nitzberg,M., Wymore,F. *et al*. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*., **33**, D580–D582.
4. Toufighi,K., Brady,S.M., Austin,R., Ly,E. and Provart,N.J. (2005) The Botany Array Resource: e-Northerns, Expression Angling, and Promoter analyses. *Plant J*., **43**, 153–163.
5. Zimmermann,P., Hennig,L. and Gruissem,W. (2005) Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci*., **10**, 407–409.
6. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
7. Rautengarten,C., Steinhauser,D., Bussis,D., Stintzi,A., Schaller,A., Kopka,J. and Altmann,T. (2005) Inferring hypotheses on functional relationships of genes: analysis of the *Arabidopsis thaliana* subtilase gene family. *PLoS Comput. Biol*., **1**, 297–312.
8. Lisso,J., Steinhauser,D., Altmann,T., Kopka,J. and Mussig,C. (2005) Identification of brassinosteroid-related genes by means of transcript co-response analyses. *Nucleic Acids Res*., **33**, 2685–2696.
9. Persson,S., Wei,H.R., Milne,J., Page,G.P. and Somerville,C.R. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl Acad. Sci. USA*, **102**, 8633–8638.
10. Owen,A.B., Stuart,J., Mach,K., Villeneuve,A.M. and Kim,S. (2003) A gene recommender algorithm to identify coexpressed genes in *C.elegans*. *Genome Res*., **13**, 1828–1837.
11. Steinhauser,D., Usadel,B., Luedemann,A., Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics*, **20**, 3647–3651.
12. Jen,C.H., Manfield,I.W., Michalopoulos,I., Pinney,J.W., Willats,W.G.T., Gilmartin,P.M. and Westhead,D.R. (2006) ACT: a WWW-based *Arabidopsis* Co-expression Tool and database for microarray based gene expression analysis. *Plant J*., **46**, 336–348.
13. Bron,C. and Kerbosch,J. (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.
14. Monte,E., Tepperman,J.M., Al-Sady,B., Kaczorowski,K.A., Alonso,J.M., Ecker,J.R., Li,X., Zhang,Y.L. and Quail,P.H. (2004) The phytochrome-interacting transcription factor, PIF3, acts early, selectively, and positively in light-induced chloroplast development. *Proc. Natl Acad. Sci. USA*, **101**, 16091–16098.
15. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C.Z., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al*. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
16. Carré,I.A. and Kim,J.Y. (2002) MYB transcription factors in the *Arabidopsis* circadian clock. *J. Exp. Bot*., **53**, 1551–1557.
17. D'haeseleer,P. (2005) How does gene expression clustering work? *Nat. Biotechnol*., **23**, 1499–1501.
18. Zeeberg,B.R., Feng,W.M., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*., **4**, R28.
19. Zhang,P.F., Foerster,H., Tissier,C.P., Mueller,L., Paley,S., Karp,P.D. and Rhee,S.Y. (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol*., **138**, 27–37.
20. Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol*., **9**, 447–464.
21. Schlitt,T. and Brazma,A. (2005) Modelling gene networks at different organisational levels. *FEBS Lett*., **579**, 1859–1866.