A Two-Step Clustering for 3-D Gene Expression Data Reveals the Main Features of the Arabidopsis Stress Response.

Martin Strauch[1*], Jochen Supper[1], Christian Spieth[1], Dierk Wanke[2], Joachim Kilian[2], Klaus Harter[2] and Andreas Zell[1]

[1]Centre for Bioinformatics Tübingen (ZBIT)
Sand 1, 72076 Tübingen, Germany

[2]Centre for Plant Molecular Biology (ZMBP)
Auf der Morgenstelle 1, 72076 Tübingen, Germany

Summary

We developed an integrative approach for discovering gene modules, i.e. genes that are tightly correlated under several experimental conditions and applied it to a three-dimensional *Arabidopsis thaliana* microarray dataset. The dataset consists of approximately 23000 genes responding to 9 abiotic stress conditions at 6-9 different points in time. Our approach aims at finding relatively small and dense modules lending themselves to a specific biological interpretation. In order to detect gene modules within this dataset, we employ a two-step clustering process. In the first step, a *k*-means clustering on one condition is performed, which is subsequently used in the second step as a seed for the clustering of the remaining conditions. To validate the significance of the obtained modules, we performed a permutation analysis and determined a null hypothesis to compare the module scores against, providing a p-value for each module. Significant modules were mapped to the Gene Ontology (GO) in order to determine the participating biological processes.

As a result, we isolated modules showing high significance with respect to the p-values obtained by permutation analysis and GO mapping. In these modules we identified a number of genes that are either part of a general stress response with similar characteristics under different conditions (*coherent modules*), or part of a more specific stress response to a single stress condition (*single response modules*). We also found genes clustering within several conditions, which are, however, not part of a coherent module. These genes have a distinct temporal response under each condition. We call the modules they are contained in *individual response modules* (IR).

1 Introduction

Integrative analysis of coexpression across different microarray experiments allows for identifying functional relationships between genes. While previous work, e.g. by Lee *et al.* [7] and Elo *et al.* [3], has focussed on affirming the coexpression of genes in heterogeneous datasets, we extend this approach by systematically searching for genes contained in regulatory modules, which either show a common response in several experiments or exhibit a response specific for a certain experiment.

[*] correspondence should be sent to strauch@informatik.uni-tuebingen.de

The AtGenExpress project has provided a three-dimensional gene-sample-time *Arabidopsis thaliana* gene expression dataset, which is available at the TAIR database (*www.arabidopsis.org*). The dataset monitors the response to different abiotic stress conditions in the shoot and root tissues. All experiments were conducted in the same laboratory and were subject to the same normalisation protocol. Differences in gene expression between the experiments can thus be correlated to the stress conditions rather than the experimental setup. By applying our method to this homogeneous dataset we are able to distinguish general responses from those specific to a certain stress and, subsequently, to identify the main features of the *Arabidopsis* stress response.

The method we propose is a novel clustering approach for three-dimensional expression data. The majority of microarray experiments monitor the expression of genes over several time-points or conditions, providing a two-dimensional dataset. Such datasets can either be processed by full-space clustering or biclustering approaches. Datasets which, additionally, comprise a time dimension, are motivation for the development of clustering algorithms working on three-dimensional data.

Several clustering methods, such as k-means [14], have been applied to two-dimensional gene expression datasets. They are categorised as unsupervised (hypothesis-free) approaches and can thus be employed to reveal novel gene regulatory relations by detecting co-expression. Usually, either the gene or condition dimension is clustered.

Recently, several biclustering methods were proposed, which partition the condition as well as the gene dimension. Cheng and Church [2] introduced the biclustering approach in 2000, followed by further publications of Ihmels *et al.* [4], Tanay *et al.* [13] and Murali *et al.* [9]. A comparative study on biclustering algorithms has been performed by Prélic *et al.* [11]. These biclustering methods are all based on greedy search strategies and aim at finding correlating subsets of genes and conditions by clustering them simultaneously.

Working on three-dimensional data, Zhao *et al.* [16] recently developed the TRICLUSTER algorithm, which is an extension of the biclustering approach and Zhang *et al.* [5] introduced an extension of a full-space clustering approach. These publications are essentially the first to handle three-dimensional datasets. In this work we present a different extension of a full-space clustering approach.

Zhao *et al.* [16] applied their TRICLUSTER algorithm to an elutriation experiment from the Spellman *et al.* [12] cell cycle dataset. To obtain multiple entries in their sample-dimension, they interpreted several attributes of the raw data as different samples. Zhang *et al.* [5] introduced another algorithm to find clusters from multiple sclerosis (MS) microarray measurements. Their *sample* dimension consists of 13 MS patients, monitored over 7 time-points. The methods of Zhao *et al.* and Zhang *et al.* search for gene groups that have coherent patterns across both the samples and time-series.

Here, we relax this constraint such that the genes are only required to cluster within each condition. The biological observation behind this approach is that certain gene groups are expressed in response to various stress conditions, however the temporal control may differ among the conditions. Such a behaviour is particularly suitable for the investigation of co-regulation and can not be detected by the previously published approaches. In order to avoid confusion of terms, we will refer to our findings as modules and not as (bi/tri)clusters. Modules can be coherent, such as the clusters found e.g. by the TRICLUSTER algorithm. Moreover, we detect genes which cluster under several conditions but show a different time trajectory under each condition (Fig. 3). We refer to the respective modules by the term ”individual response” (IR) modules.

We employ an efficient two-step clustering utilising the properties of the dataset at hand. The first property follows from the assumption that genes which form modules under various stress conditions should also cluster within a single stress condition. To find these modules, we apply a standard clustering technique (*k*-means) to an initial single stress condition and use the results as seeds. The second property concerns the time-dimension, which can not be partitioned like the other dimensions, thus reducing the search space. Apart from the run time improvement, the two-step approach enables us to control and analyse each step individually, making the overall module formation more intuitive.

To interpret the reliability and biological meaning of the obtained modules, an assessment of statistical significance is necessary. Kerr *et al.* [6] proposed bootstrapping to assess the stability of clusters, and Xie *et al.* [15] proposed a permutation based significance test to generate a null hypothesis. In this work, we conduct a permutation analysis and fit a null hypothesis distribution to the obtained scores. Finally, to associate the resulting modules with their respective function, a mapping to the Gene Ontology (GO) [1] is performed.

Due to the nature of the approach outlined above, which does not partition the time dimension, we do not consider it to be a triclustering method. For the analysis of stress responses, however, regarding gene-time-trajectories under different stress conditions appears to be a promising concept. Our method encompasses a broader definition of clusters, which we refer to as modules, and, moreover, we show it to be more robust in the face of noisy signals than the TRICLUSTER algorithm.

2 Materials and Methods

2.1 The Clustering Approach

2.1.1 Notation

The gene-condition-time expression matrix is a matrix $E = G \times C \times T$ of real numbers, where $G = \{g_1, g_2, ..., g_{g_N}\}$ is a set of genes, $T = \{t_1, t_2, ..., t_{t_N}\}$ is a set of time-points and $C = \{c_1, c_2, ..., c_{c_N}\}$ a set of experimental conditions, respectively. Each entry e_{ijk} of E specifies the expression level of gene g_i under condition c_j at time-point t_k.

Our aim in introducing this method is to find clusters within one condition $c_{initial} \in C$ and to observe how the genes contained therein behave under all the other conditions. To mine the complete dataset, for every initial condition each significant partitioning is considered as a seed. As we do not subdivide the time dimension of the three-dimensional dataset we operate on time trajectories $\mathbf{g}_c = \{g_{ct_1}, \cdots, g_{c|T|}\}$ of fixed length $|T|$, where g_{ct} denotes the expression value for gene g under condition c at time-point t.

Formally, our interest is in mining gene modules, i.e. 2-tuples $M = (G', C')$ consisting of a set of genes $G' \subseteq G$ and a set of conditions $C' \subseteq C$. We consider a module M to be significant if the average Pearson distance of the genes $g \in G'$ under the conditions $c \in C'$ is below the threshold τ. The choice of τ is motivated in section 2.2. The Pearson distance of two genes g^1 and g^2 is calculated by comparing the time trajectories $\mathbf{g}_c^1 = \{g_{ct_1}^1, \cdots, g_{c|T|}^1\}$ and $\mathbf{g}_c^2 = \{g_{ct_1}^2, \cdots, g_{c|T|}^2\}$ of fixed length $|T|$.

Let $g_{(c)}^{centroid}$ denote the average time trajectory of the genes $g \in G'$ under condition $c \in C'$. Then, for a significant module (G', c) the following holds:

$$\mu_{Pearson}(\mathbf{g}_c^{G'}) < \tau \text{ with } \mu_{Pearson}(\mathbf{g}_c^{G'}) = \frac{1}{|G'|} \sum_{g' \in G'} \rho(\mathbf{g}_c^{g'}, g_{(c)}^{centroid}) \tag{1}$$

Here, ρ computes the Pearson distance, i.e. $1-$ Pearson coefficient. Using the above definition of a module we allow three kinds of modules in our results:

1. *Single response modules*, where $|C'| = 1$ and $\mu_{Pearson}(\mathbf{g}_{C'}^{G'}) < \tau$, i.e. genes that cluster together under exactly one condition, but not under any other condition.
2. *Coherent modules*, where $|C' > 1|$ and $\mu_{Pearson}(\mathbf{g}_{C'}^{G'}) < \tau$. Here, the time-trajectories of all genes follow roughly the same shape (depending on the threshold τ) under all conditions $c \in C'$.
3. *Independent Response (IR) Modules*, where $|C' > 1|$ and $\mu_{Pearson}(\mathbf{g}_c^{G'}) < \tau \;\; \forall\, c \in C'$. Here, all genes cluster together under each condition $c \in C'$, but with a different trajectory shape for each condition.

The biological motivation behind each definition is different. Single response modules allow us to correlate genes to specific conditions, while coherent modules are similar to biclusters and may reveal co-regulation under multiple conditions. These genes are potentially controlled by the same transcription factors and display a general stress response.

Finally, IR modules capture a more complex type of co-regulation, i.e. they hint at the existence of stress regulation specific to every condition alongside with a common transcriptional control. Examples of the three module types mentioned are given in the results section: Fig. 3 (IR module), Fig. 4 (coherent module) and Fig. 5 (single response module).

2.1.2 Algorithm

For the initial clustering step one stress condition $c_{initial}$ has to be selected, which is then clustered by *k*-means based on Pearson distances. The genes $G' \subseteq G$ considered in this step are the genes which meet the fold change criterion under condition $c_{initial}$. In order to obtain dense clusters, we allow only such genes in G' to be clustered, that have a Pearson distance below a threshold of $\delta = 0.05$ to at least one other gene. We thus achieve an incomplete clustering, which does not assign every gene to a cluster. This provides better gene sets than a complete clustering which is forced to assign a gene to a cluster even if it does not fit particularly well into a specific cluster. Lowering or raising the distance threshold δ results in tighter or looser clusters, respectively. To determine the number of clusters k for *k*-means, an additional PCA analysis is applied, such that k is set to the number of principal components that explain $(1 - \alpha)\,\%$ of the variation, where a default parameter of $\alpha = 0.05$ yields the best results.

The initial clustering as described above results in an incomplete partitioning of the genes $G' \subseteq G$ under the initial condition $c_{initial}$ into l_n clusters of genes: $L_{c_{initial}} = \{l_1, l_2, ..., l_n\}$. For each of these clusters, the average Pearson distance (Eq.1) is computed and compared to the threshold τ (Sec. 2.2). Clusters whose average Pearson distance falls below this threshold are regarded as significant and used as seeds.

After the initial clustering step, the significant clusters l_i can be regarded as (preliminary) modules $M' = (G', C')$ with $C' = c_{initial}$ and $G' = l_i$. At this point we know that the genes in G' cluster together under $c_{initial}$, i.e. using the trajectory notation $\mu_{Pearson}(\mathbf{g}_{C'}^{G'}) < \tau$. In order to check whether these genes cluster together only under this single condition or if they are also coexpressed under other conditions, we examine the genes under all the other conditions $C \setminus c_{initial}$: We check whether $\mu_{Pearson}(\mathbf{g}_{c'}^{G'}) < \tau$ for one condition $c' \in C \setminus c_{initial}$ at a time and append c' to the set of conditions C' if the average Pearson distance is below the threshold. Finally, we have constructed a significant module M = (G',C') from a significant cluster l_i, which has been found under condition $c_{initial}$. This process is repeated for all significant clusters found under all possible start conditions $c_{initial}$, e.g. salt/shoot, salt/root, cold/shoot, etc. until the whole dataset is covered.

2.2 Evaluation of Statistical Significance

In order to measure the reliability of cluster findings and to determine a reasonable value for the threshold τ, we calculate *p*-values based on null hypotheses. Because the null distribution of the test statistic is unknown, a permutation analysis was employed to estimate this distribution. Therefore, 50 000 initial clusterings were performed on randomly shuffled datasets. [1] For each cluster size, a probability distribution was fitted to the histogram of the obtained cluster densities. We considered the normal, log-normal, beta and gamma distributions. This approximation was then validated by a chi-square test (χ^2). The best distribution, with respect to the chi-square test, was defined as a null distribution.

In order to determine the threshold value, we chose to set τ to the cluster density corresponding to a p-value of 0.05. For big cluster sizes, the 50,000 permutations run did not yield sufficient data for a smooth histogram. For runtime reasons, we then set τ to the lowest cluster density found for a cluster of the respective size in randomly shuffled data. As a consequence, for these cases binary p-values result: $p = 0$ for significant and $p = 1$ for insignificant.

As a post-processing step to evaluate our findings we also perform a significance analysis on the modules detected by our algorithm. Gene groups that are significantly coexpressed under several conditions or genes specific for a certain stress condition should be associated with their respective function. Therefore, using NetAffx™ [8], we perform a mapping to the gene ontology (GO) to find the active biological processes along with *p*-values. Note, that the NetAffx™ p-values refer only to the biological function enrichment, whereas the p-values described above refer to the significance of finding a cluster of the given size in randomly shuffled data.

[1] Genes, conditions and time-points were permuted, leaving only the expression values intact but no further information.

2.3 Arabidopsis Expression Data and Preprocessing

The AtGenExpress project is a multinational effort to uncover the transcriptome of *Arabidopsis thaliana*. Within this project, numerous time-series measurements based on the Affymetrix chip ATH1-121501 are provided. Among these, we extracted data from the abiotic stress treatment experiments (cold, osmotic, salt, drought, genotoxic, UV-B, wound, heat) for both tissues, root and shoot, as well as the corresponding control measurements. Each of these time series contains 6 to 9 measurements after exposure to the stress condition, each time-point having two biological replicates. To obtain the fold-change, the ATH1 chips were normalised with GCRMA [10], the biological replicates were averaged and finally the $\log_2$ was taken. After normalisation, the expression values are in a range between 1 and 16. The standard deviation of the biological replicates is 0.26.

Prior to the clustering, a filter on the fold change was employed: In preliminary experiments a 2-fold expression change was found to be significant. Thus any gene with a fold change smaller than 2 in every time-point, with respect to the control measurement, is left out. Hence, for each stress dataset, this filtering results in different gene groups.

2.4 Artificial Dataset and Scoring

An artificial dataset was constructed for evaluation purposes, which consists of 100 genes monitored under 30 conditions over 9 different points in time. The three-dimensional data matrix was filled with random gene-time trajectories drawn from a normal distribution. Then, 10 non-overlapping, perfectly co-regulated coherent modules were planted, each comprising 10 genes and covering 3 conditions. The time-dimension was not partitioned. The modules were constructed using an additive model, which randomly chooses a gene-time trajectory as a centroid, around which the remaining 9 genes are grouped. This dataset represents noise level $\sigma = 0$, as the modules are perfectly co-regulated. Further noise levels are simulated by adding different random values to each point of the gene-time trajectories. The values are drawn from normal distributions centered around the mean of the data matrix and with different standard deviations $\sigma = (0.1, 0.3, 0.5, 0.7, 0.9)$, resulting in the different noise levels. These noise leves are a measure comparable to the standard deviation of the biological replicates (0.26 for the Arabidopsis dataset). Note, however, that these measures are not equal.

A score based on the genes which are part of the modules was used to evaluate the results. The score, as well as the test framework is comparable to the methods employed in the biclustering study by Prélic *et. al* [11]. Two modules M_1 and M_2, e.g. the planted module and a module detected by a clustering technique are compared based on the corresponding sets of genes G_1 and G_2. In the case of an ideal recovery, i.e. when $M_1 = M_2$, the score is 1.

$$S(G_1, G_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|} \tag{2}$$

Using the above formula, two modules can be compared. In order to score the performance in recovering all modules, the following formula is employed. Here, two sets of result modules

$R_1 = (M_1, ..., M_{|R_1|})$ and $R_2 = (M_1, ..., M_{|R_2|})$ are compared based on the genes G_n and conditions C_n which are part of the respective modules.

$$S^*(R_1, R_2) = \frac{\sum_{(G_1,C_1)\in R_1} \; max_{(G_2,C_2)\in R_2} \; S(G_1, G_2)}{|R_1|} \tag{3}$$

2.5 Complexity

Given a condition-specific subset of genes $G' \in G$ with size $|G'|$ and the number of conditions $|C|$. The overall complexity of the two-step clustering algorithm is as follows:

$$O\left(\; (|G'|^2 + |G'| + |C| \cdot |G'|) \cdot |C| \; \right) \tag{4}$$

First, we need a Pearson correlation table containing pairwise correlations of the genes in G', which takes $O(|G'|^2)$ to compute. Then, a linear time k-means clustering is applied to G', which adds $O(G')$ to the overall complexity. Following the resulting initial clusters through the other conditions takes $|C| \cdot |G'|$, as we compute densities and look up p-values for all clusters (each of them contains $|G'|$ genes) under all conditions C. The above procedure is repeated for each of the $|C|$ conditions.

As for typical microarray datasets $C \ll G$, usually no runtime problems arise through multiplication with the factor $|C|$. Furthermore, the condition-specific gene subsets G' result from a fold-change filtering, which on the Arabidopsis thaliana dataset leaves subsets G' of no more than 500-600 genes. This is much smaller than the size of the dataset, which comprises ca. 23000 genes.

Computing time was below 2 minutes on a desktop machine (AMD Athlon XP ™, 1150 Mhz) for each of the datasets.

3 Results

3.1 Artificial Data

Both, the two-step clustering and the TRICLUSTER algorithm were run on the artificial datasets described in 2.4. For the two-step clustering, default parameters were employed. To ensure comparability, no fold-change filtering was applied prior to executing the algorithm. TRICLUSTER was run with minimum cluster size constraints [T,C,G] = [2,3,10], as the clusters expected to be found comprise 10 genes, which are co-regulated under 3 conditions. The range window size was set to values between 0.01 and 0.05, depending on the noise level of the current dataset. For details on the TRICLUSTER algorithm see Zhao *et. al* [16].

A summary of the runs on artificial data is given in Figure 1. Apparently, the TRICLUSTER algorithm performs well at noise level $\sigma = 0$ and is even slightly better than the two-step clustering. However, with increasing noise level, the quality of the TRICLUSTER results clearly falls below the score attained by the two-step clustering. Naturally, it is hard to compare parameter-intensive clustering methods. We have confirmed the ability of the TRICLUSTER

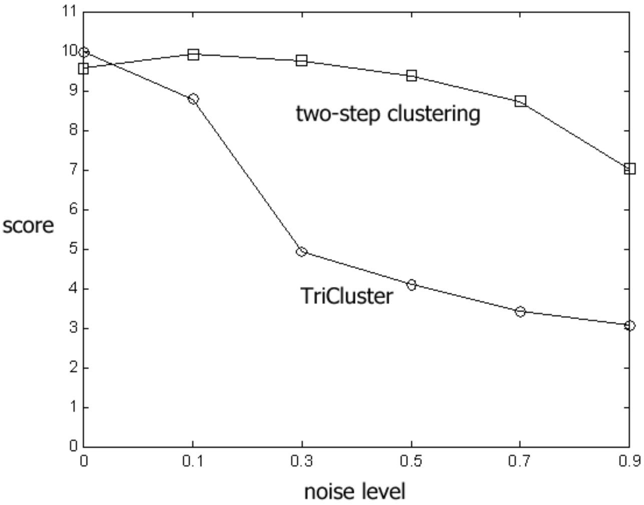

Figure 1: Results from the evaluation on artificial data. The plot shows the scores obtained at different noise levels ranging from $\sigma = 0$ (clear signal) to $\sigma = 0.9$ (distorted signal). In case all 10 modules are found without additional or missing genes, the optimal score of S=10 can be obtained.

algorithm to detect clusters in three-dimensional data. However, runs on biological data show its weakness in the face of noisy signals (see Section 3.2) and make the two-step clustering appear to be the method of choice for our purpose.

Surprisingly, the two-step clustering attains its best result at noise level $\sigma = 0.1$. The slightly weaker result under $\sigma = 0$, i.e. with no noise added to distort the signal, is always due to some modules being found in two parts. Thus, the optimal score for a module with no genes missing can not be obtained, even though all target genes have been identified.

A possible explanation for this behaviour might be the fact that under noise level $\sigma = 0$ the condition-specific pre-filtering already separates the extremely homogeneous signal from the background even before the k-means clustering is applied, which then attempts to find more subpartitions than are actually present in the data. Such a constellation, however, will most likely never occur on biological datasets, which are affected by signal noise.

3.2 Arabidopsis Gene Expression Data

We found the two-step clustering to be more useful for application to the *Arabidopsis* stress response dataset. Using the TRICLUSTER algorithm we obtained numerous dense clusters, which were, however, stationary, i.e. clusters with flat gene-time trajectories under all conditions. In order to uncover the stress response of *Arabidopsis* we are looking for changes in the trajectory shapes under different stress conditions.

Several significant modules could be obtained by employing the two-step clustering. An overview of the modules found in the root and shoot tissues is given in Figure 2. It could be shown that a diverse stress response occurs in the shoot tissue, where the response to wounding and UV-B stress is especially prominent. These two responses are, understandably, absent in the root. Less modules could be identified in the root, which reacts only to cold, salt, osmotic and heat stress. Obviously, the root is often confronted with osmotic and salt stress, as well as it is affected by temperature changes.

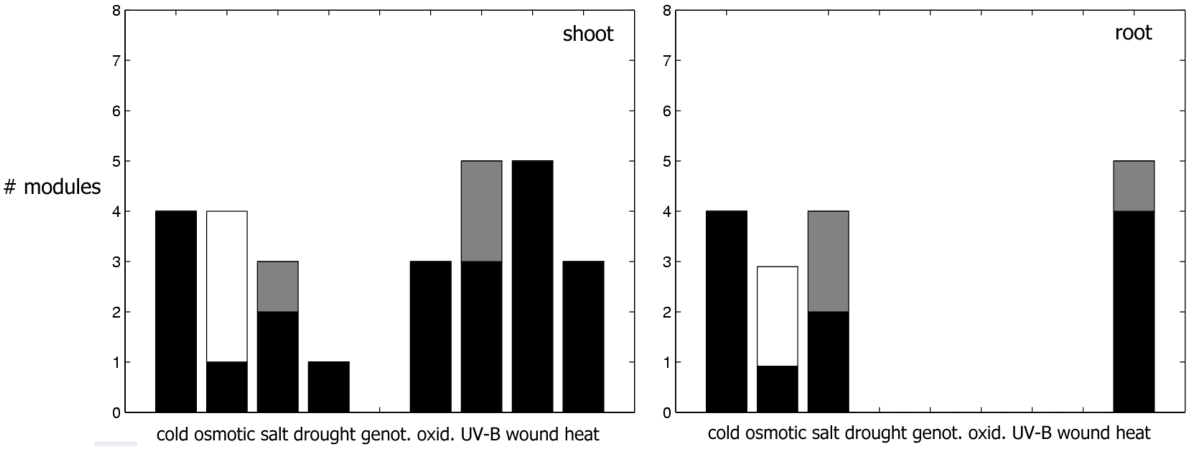

Figure 2: Number of significant modules detected under the respective initial conditions in the shoot and root tissues. Single response modules are represented by black, IR modules by grey and coherent modules by white bars.

In the following, three prominent signals, which are moreover representative of the three module types, are discussed in further detail: By clustering the heat dataset from the root tissue we obtain a module consisting of 6 genes (Fig. 3), 5 of which map to the GO term ”response to heat” (*p*-value: $9.83 \cdot 10^{-233}$). Furthermore, all genes map to the GO term ”response to unfolded protein” (*p*-value: $3.69 \cdot 10^{-201}$), as well as to ”protein folding” (*p*-value: $1.67 \cdot 10^{-55}$). All of these genes code for heat shock proteins, whose chaperone activity is especially needed under heat stress but also of general use. Interestingly, under every included condition the clusters have small *p*-values, however these clusters have quite different trajectories with respect to each other. This module is an example of an IR module. In fact, more modules detected in the *Arabidopsis* dataset belong to the IR than to the coherent type, suggesting the existence of a common regulatory component plus condition-specific influences in these cases. The most abundant module type, however, is the single response module, which is due to the approach taken by our algorithm: We search for single response modules in the first place and, if possible, extend these into IR or coherent modules.

In Figure 4, a coherent module for the conditions *salt* and *osmotic* is presented. Obviously, there is a close relationship between the two conditions involved. The most significant GO mappings are ”response to water”(*p*-value: $3.25 \cdot 10^{-14}$), ”chitin catabolism” (*p*-value: $2.63 \cdot 10^{-16}$) and ”response to abscisic acid stimulus” (*p*-value: $6.94 \cdot 10^{-30}$), the latter being responsible for closing the plant’s stomata to prevent loss of water. Apparently, the GO mapping confirms the clustering result. Numerous modules were found which cover the response to osmotic stresses, some of which are, unlike the example shown, specific for salt stress and not active under general osmotic stress.

In contrast to the modules just mentioned, Figure 5 shows a module which is specific just for exposition to UV-B radiation. This module is enriched with GO annotations like ”ripening” (*p*-value: $7.56 \cdot 10^{-13}$) , ”respiratory gaseous exchange” (*p*-value: $2.84 \cdot 10^{-36}$) or ”ethylene biosynthesis” (*p*-value: $3.17 \cdot 10^{-12}$). This coincides with ethylene being a plant hormone which stimulates the ripening of fruit, the opening of flowers and is involved in biotic and abiotic stress signalling responses.

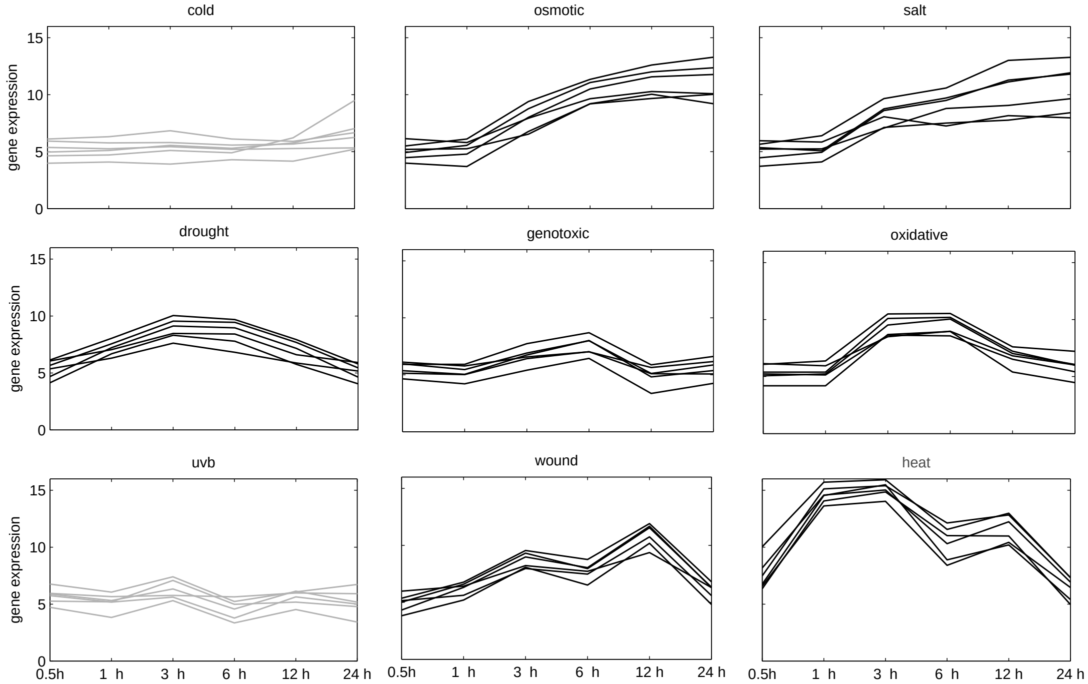

Figure 3: Module generated with heat/root as initial condition (*p*-value: $1.87 \cdot 10^{-5}$). The shaded conditions are not included into the cluster, whereas all conditions with a *p*-value below 5 % are. This leads to an IR module containing all conditions except UV-B and cold.

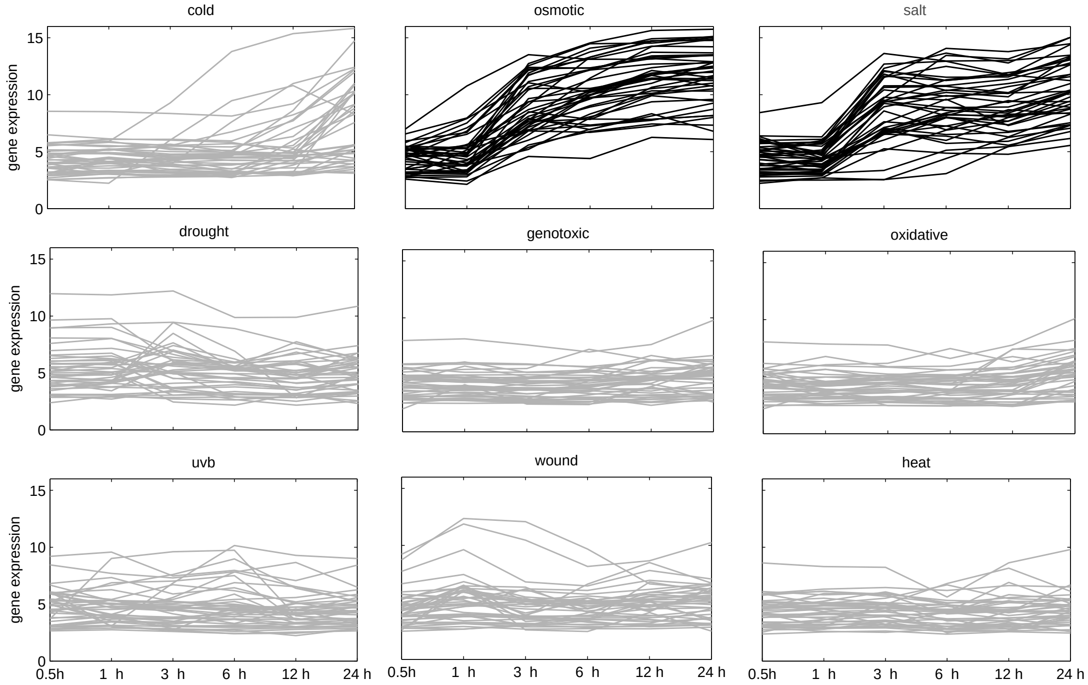

Figure 4: Coherent module found for salt/shoot as initial condition (*p*-value: $7.12 \cdot 10^{-7}$). All other conditions do not receive a p-value below 5 % and are therefore not included into the module.

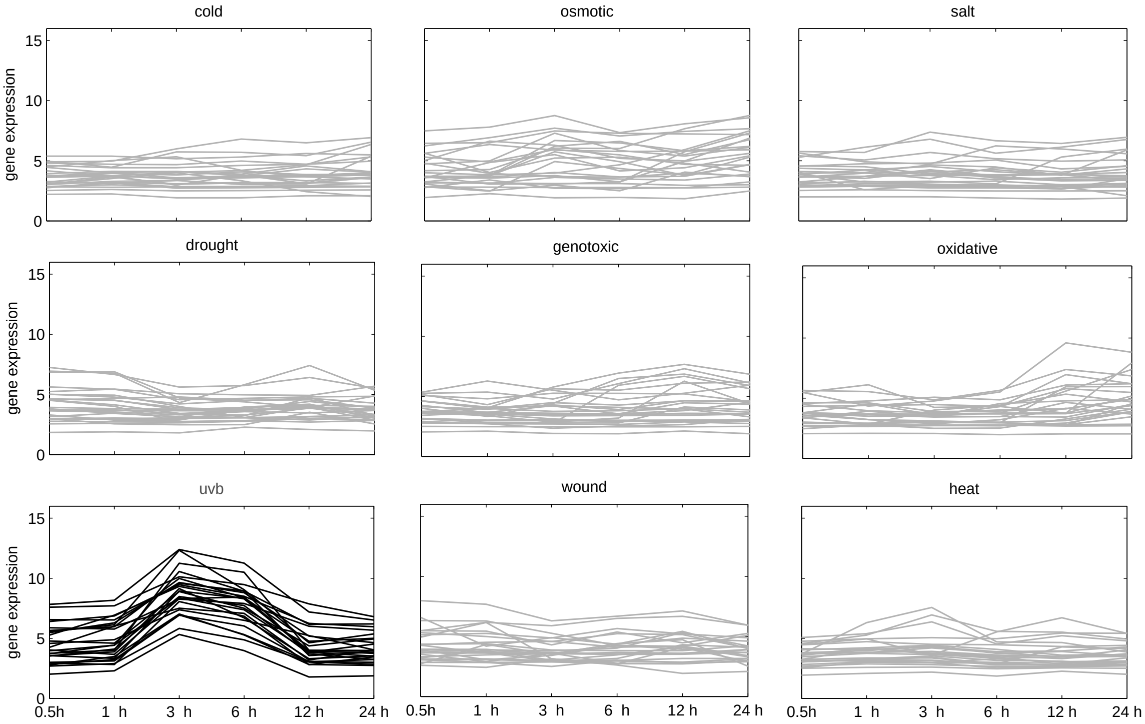

Figure 5: A module obtained by starting with UV-B/shoot as initial condition (*p*-value: $9.08 \cdot 10^{-7}$). Obviously, the genes participating in the initial cluster do not cluster under the other conditions. This is an example for a single response module, which is specific for the highlighted condition UV-B.

4 Conclusion

We have applied a two-step clustering approach to a three-dimensional *Arabidopsis thaliana* dataset. Such a three-dimensional dataset with a sufficient sampling rate in each dimension is very rare and makes it particularly suitable to apply extended clustering methods. However, because of the three-dimensional nature of the dataset, conventional clustering and biclustering algorithms will be of limited use. Our clustering approach is able to discover statistically significant single and coherent as well as IR modules. Several modules with meaningful biological processes could be revealed and several stress conditions could be related to specific stress response processes. For instance, the heat stress module shows a highly significant enrichment for ”response to heat” and ”response to unfolded protein” under almost all conditions.

The simple incomplete clustering applied separately to each condition considerably improves the capability even of standard clustering techniques to identify dense clusters in front of a noisy background. This condition-specific pre-filtering results in a greater resistance to noise than global approaches such as the TRICLUSTER algorithm, which could most likely be improved by adopting this strategy.

Overall, the introduced method is suitable for finding genes that act as part of a stress response, either specifically to a single condition or to a number of different conditions. Our method, in contrast to previously published work, is capable of discovering IR modules. These findings of stringent co-regulated gene modules open up the opportunity for direct experimental investigations in the laboratory, as well as for interesting computational applications.

Software

A Matlab implementation of the proposed algorithm including a GUI and the Arabidopsis dataset can be downloaded at `http://www-ra.informatik.uni-tuebingen.de/software/IAGEN/index.html`.

Acknowledgements

We thank Lizhuang Zhao for providing us with an implementation of his TRICLUSTER algorithm. This work was supported by the National Genome Research Network (NGFN II) of the Federal Ministry for Education and Research in Germany under contract number 0313323 and by the Arabidopsis Functional Genomics Network (AFGN) of the DFG (HA2146/5-2).

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[2] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press, 2000.

[3] L.L. Elo, R. Lahesmaa, and T. Aittokallio. Inference of gene coexpression networks by integrative analysis across microarray experiments. *Journal of Integrative Bioinformatics*, 3(2):33, 2006.

[4] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31(4):370–7, Aug 2002.

[5] D. Jiang, J. Pei, M. Ramanathan, C. Tang, and A. Zhang. Mining coherent gene clusters from gene-sample-time microarray data. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 430 – 439, 2004.

[6] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci U S A*, 98(16):8961–8965, Jul 2001.

[7] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14(6):1085–1094, Jun 2004.

[8] G. Liu, A. E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, and M. A. Siani-Rose. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*, 31(1):82–86, Jan 2003.

[9] T. M. Murali and S. Kasif. Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput*, pages 77–88, 2003.

[10] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys*, E68:011906, Jul 2003.

[11] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, May 2006.

[12] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, Dec 1998.

[13] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18 Suppl 1:136–144, 2002.

[14] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22(3):281–5, Jul 1999.

[15] Y. Xie, W. Pan, and A. B. Khodursky. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21(23):4280–4288, Dec 2005.

[16] L. Zhao and M. J. Zaki. Tricluster: an effective algorithm for mining coherent clusters in 3D microarray data. In *SIGMOD ’05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 694–705, 2005.